Computational Prediction of Host-Pathogen Protein-Protein Interactions

Matthew D. Dyer\textsuperscript{1,2} , T. M. Murali\textsuperscript{3} , and Bruno W. Sobral\textsuperscript{2}

\textsuperscript{1}Genetics, Bioinformatics, and Computational Biology Program,
\textsuperscript{2}Virginia Bioinformatics Institute, and
\textsuperscript{3}Department of Computer Science.
Virginia Polytechnic Institute and State University.
Motivation

• Introduction to the problem
• Application to Host-Pathogen Systems (*H. sapiens* - *P. falciparum*)
Outline

- Motivation
  - Introduction to the problem
  - Application to Host-Pathogen Systems (*H. sapiens* - *P. falciparum*)

- Literature Review
  - Methods
  - Data sources
Motivation
  • Introduction to the problem
  • Application to Host-Pathogen Systems (H. sapiens - P. falciparum)

Literature Review
  • Methods
  • Data sources

Methods
  • Definitions
  • Complications when we look at host-pathogen systems
  • Protein-domain profile model for prediction
  • Evaluation
    • Triplet proximity
    • Triplet coexpression
    • Weighted functional enrichment
Motivation
- Introduction to the problem
- Application to Host-Pathogen Systems (H. sapiens - P. falciparum)

Literature Review
- Methods
- Data sources

Methods
- Definitions
- Complications when we look at host-pathogen systems
- Protein-domain profile model for prediction
- Evaluation
  - Triplet proximity
  - Triplet coexpression
  - Weighted functional enrichment

Results, conclusions, and future work
Infectious diseases result in millions of deaths each year
Infectious diseases result in millions of deaths each year.

Millions of dollars are spent annually to better understand how pathogens infect their hosts.
Infectious diseases result in millions of deaths each year

Millions of dollars are spent annually to better understand how pathogens infect their hosts

*Plasmodium falciparum* (Malaria)
- 300 to 500 million clinical cases / year
- 1.5 to 2.7 million deaths / year
- Currently no effective vaccine
- Parasite resistance to current drugs
- Preventative drugs too expensive for people living in infected areas
High Risk Areas
Limited Risk Areas
Computational Prediction of Host-Pathogen Interactions

Malaria Life-cycle

Mosquito Stages
- Development of sporozoites in mid gut
- Mosquito infection by ingestion of gametocytes
- Fertilization of gametocytes
- Microgametocyte
- Macrogametocyte

Invasion of salivary glands
- Invasion of hepatic cells by sporozoites

Liver Stages
- Formation of exo-erythrocytic merozoites in hepatic schizont
- Invasion of hepatic cells by sporozoites
- Release of exo-erythrocytic merozoites

Erythrocytic Stages
- Release of erythrocytic merozoites
- Asexual Erythrocyte Stages
- Système
- Trophozoite
- Schizont

Sexual Stages
Computational Prediction of Host-Pathogen Interactions

Malaria Life-cycle

Mosquito Stages
- Ookinete penetration of midgut wall
- Oocyst development
- Development of sporozoites in mid gut
- Invasion of salivary glands
- Salivary gland
- Zygote formation
- Mosquito infection by ingestion of gametocytes

Liver Stages
- Formation of exo-erythrocytic merozoites in hepatic schizont
- Invasion of hepatic cells by sporozoites
- Release of exo-erythrocytic merozoites

Erythrocytic Stages
- Merozoites
- Release of erythrocytic merozoites
- Asexual Erythrocyte Stages
- Sexual Stages
- Microgametocyte
- Macrogametocyte

Human infection by sporozoites in salivary secretion
An important aspect of any host-pathogen system is the mechanism by which a pathogen infects its host

- Surface proteins and molecules form the foundation of communication between a host and pathogen
- Protein-protein interactions play a vital role in initiating infection
An important aspect of any host-pathogen system is the mechanism by which a pathogen infects its host. 
- Surface proteins and molecules form the foundation of communication between a host and pathogen. 
- Protein-protein interactions play a vital role in initiating infection.

*P. falciparum* (MSP1s) - Merozoite surface proteins allow the parasite to invade a red blood cell.
An important aspect of any host-pathogen system is the mechanism by which a pathogen infects its host:
- Surface proteins and molecules form the foundation of communication between a host and pathogen.
- Protein-protein interactions play a vital role in initiating infection.

*P. falciparum* (MSP1s) - Merozoite surface proteins allow the parasite to invade a red blood cell.

Identifying which interactions enable a pathogen to infect its host provides us with potential targets for therapeutics.
Method for predicting PPIs between host proteins and pathogen proteins

Predicted *H. sapiens* (○) - *P. falciparum* (♦) network
Method for predicting PPIs between host proteins and pathogen proteins

Methods for evaluating host-pathogen PPI predictions

Predicted *H. sapiens* (○) - *P. falciparum* (◇) network
Organism X

Predicting PPIs
Predicting PPIs

Organism X
Computational Prediction of Host-Pathogen Interactions

Predicting PPIs

Host

Pathogen
Predicting PPIs
Computational Prediction of Host-Pathogen Interactions

Dyer, Murali, and Sobral

2001

2007

Literature Review
Domain Profiles

Domain Profiles

**Domain Profiles**

**Domain Profiles**

**Domain Profiles**


**Sequence Homology**

Huang *et al.* BMC Bioinformatics (2007) 8(1):152
Computational Prediction of Host-Pathogen Interactions

**2001**

- **Domain Profiles**

- **Sequence Homology**

---

**Literature Review**

**Source Organisms**

**Identify Orthologs**

**And Map Interactions**

**Target Organism**

VirginiaTech
Computational Prediction of Host-Pathogen Interactions

2001

- Domain Profiles

- Sequence Homology
  Huang et al. BMC Bioinformatics (2007) 8(1):152

2007

Source Organisms

Identify Orthologs
And Map Interactions

Target Organism

Literature Review
**Domain Profiles**


**Sequence Homology**

Huang *et al.* BMC Bioinformatics (2007) 8(1):152
2001

- **Domain Profiles**

- **Sequence Homology**

- **Machine Learning**

2007


### Genomic Data

<table>
<thead>
<tr>
<th>Feature</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Domains</td>
<td>Sprinzak et al. (2001), Ng et al. (2002), Deng et al. (2002), Kim et al. (2002)</td>
</tr>
<tr>
<td>Homology</td>
<td>Sprinzak et al. (2003), Haiyuan et al. (2004)</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>Deng et al. (2002), Bar-Joseph et al. (2003), Jensen et al. (2003), Zhang et al. (2004)</td>
</tr>
<tr>
<td>Protein Expression</td>
<td>Ghaemmaghami et al. (2003)</td>
</tr>
<tr>
<td>Yeast Two-Hybrid</td>
<td>Utz et al. (2000), Ito et al. (2001), Zhang et al. (2004)</td>
</tr>
<tr>
<td>Synthetic Lethal</td>
<td>von Mering et al. (2002), Tong et al. (2004), Wong et al. (2004)</td>
</tr>
<tr>
<td>Tandem Affinity Purification</td>
<td>Gavin et al. (2002), Badger et al. (2003), Zhang et al. (2004)</td>
</tr>
<tr>
<td>Transcription Factor</td>
<td>Zhang et al. (2004)</td>
</tr>
<tr>
<td>Knockout Phenotype</td>
<td>Zhang et al. (2004)</td>
</tr>
<tr>
<td>Colocalization</td>
<td>Kumar et al. (2002), von Mering et al. (2002), Jensen et al. (2003), Zhang et al. (2004)</td>
</tr>
</tbody>
</table>
Literature Review

**Domain Profiles**

**Sequence Homology**

**Machine Learning**
Computational Prediction of Host-Pathogen Interactions

2001

- **Domain Profiles**

- **Sequence Homology**

- **Machine Learning**

- **Graph Structure**
  - Goldberg and Roth. PNAS. (2003). 100(8):4372 - 4376

2007
Computational Prediction of Host-Pathogen Interactions

Predicting PPIs

Host

Pathogen
Computational Prediction of Host-Pathogen Interactions

Predicting PPIs

Host

Pathogen
Experimental studies test only a small number of interactions at a time.
Experimental studies test only a small number of interactions at a time.

- A number of data types used to train the previously-mentioned methods such as gene expression and knockout phenotype are not available for host-pathogen systems.
  - Simultaneous gene expression of both host and pathogen.
We integrate a number of public intraspecies PPI datasets with protein domain-profiles to develop a novel framework for predicting and studying host-pathogen PPI networks.

- Compute statistics of how often proteins containing specific domain pairs interact and use these statistics to make predictions.
- Evaluate predictions using three computational tests.
Graph $G(V, E)$ - a set of vertices and a set of edges which connect the vertices (intraspecies PPI network)
**Definitions**

- **Graph** $G(V, E)$ - a set of vertices and a set of edges which connect the vertices (intraspecies PPI network)

- **Bipartite graph** $BG(V_1, V_2, E)$ - two sets of vertices that are disjoint and a set of edges that connect a vertex from set one with a vertex of set two. (interspecies PPI network)
Definitions

Graph $G(V, E)$ - a set of vertices and a set of edges which connect the vertices (intraspecies PPI network)

Bipartite graph $BG(V_1, V_2, E)$ - two sets of vertices that are disjoint and a set of edges that connect a vertex from set one with a vertex of set two. (interspecies PPI network)

Triplet - two proteins in either the host or pathogen that are predicted to interact with the same protein in the other system (H-H-P, H-P-P)
$D(g,d)$ - the event that protein $g$ contains domain $d$
\( D(g,d) \) - the event that protein \( g \) contains domain \( d \)

\( I(g,h) \) - the event that protein \( g \) and protein \( h \) interact
$P$ - set of proteins with at least one domain and one interaction
$P$ - set of proteins with at least one domain and one interaction

$P_d$ - set of proteins that in $P$ that contain domain $d$ (□)
$S$ - set of interactions between pairs of proteins in $P$
S - set of interactions between pairs of proteins in $P$

$S_{d,e}$ - set of interactions where one protein contains $d$ (■) and the other $e$ (◇)
S - set of interactions between pairs of proteins in $P$

$S_{d,e}$ - set of interactions where one protein contains $d$ (□) and the other $e$ (◇)

Compute $Pr\{I(g,h) \mid D(g,d) D(h,e)\}$ using Bayes rule
For every pair of domains $d$ and $e$ estimate each of the probabilities on the right hand side of the equation from the known data.

\[
\Pr\{I(g, h) | D(g, d)D(h, e)\} = \frac{\Pr\{D(g, d)D(h, e)|I(g, h)\} \Pr\{I(g, h)\}}{\Pr\{D(g, d), D(h, e)\}}
\]
For every pair of domains \( d \) and \( e \) estimate each of the probabilities on the right hand side of the equation from the known data

\[
\Pr\{I(g, h) | D(g, d)D(h, e)\} = \frac{\Pr\{D(g, d)D(h, e) | I(g, h)\} \Pr\{I(g, h)\}}{\Pr\{D(g, d), D(h, e)\}}
\]

\( \Pr\{D(g, d)D(h, e) | I(g, h)\} \) is the fraction of interactions where one protein contains domain \( d \) (■) and the other domain \( e \) (◇) given we known they interact

\[
\Pr\{D(g, d)D(h, e) | I(g, h)\} = \frac{|S_{d,e}|}{|S|}
\]
Pr\{l(g,h)\} represents the probability that a pair of proteins interact.
\[ \Pr\{I(g, h)\} \text{ represents the probability that a pair of proteins interact} \]

Use \( \binom{|P|}{2} \) instead of \( P^2 \) to avoid counting self-interacting proteins.
Protein-domain Profiles

\[ \Pr\{D(g,d) \cap D(h,e)\} \] is the probability that if we randomly choose two proteins, one with have domain \( d \) (□) and the other domain \( e \) (◇)

- Introduce a small correction factor to account for the situation when the same protein contains both domains

\[
\frac{|P_d||P_e| - |P_d \cap P_e|}{|P|^2}
\]
Pr\{D(g,d) \ D(h,e)\} is the probability that if we randomly choose two proteins, one with have domain $d$ (■) and the other domain $e$ (◇)

- Introduce a small correction factor to account for the situation when the same protein contains both domains

$$\frac{|P_d||P_e| - |P_d \cap P_e|}{\binom{|P|}{2}}$$

$P_d = \{g, h, i\}$

$P_e = \{g, j\}$

Possible = ...

VirginiaTech
Probability that two proteins \((g\text{ and }h)\) interact given they contain domains \(d\) and \(e\)

\[
Pr\{I(g, h) | D(g, d) D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}
\]
Probability that two proteins (g and h) interact given they contain domains d and e

\[ \Pr\{ I(g, h) | D(g, d) D(h, e) \} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|} \]

Multiple pairs of domains may predict that the same pair of proteins interact

- Notation: \( M_g \) - set of all domains contained in protein g

\[ \Pr\{ I(g, h) \} = 1 - \prod_{d \in M_g} \prod_{e \in M_h} (1 - \Pr\{ I(g, h) | D(g, d) D(h, e) \}) \]
Probability that two proteins \((g\) and \(h)\) interact given they contain domains \(d\) and \(e\)

\[
\Pr\{I(g, h) | D(g, d) D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}
\]

Multiple pairs of domains may predict that the same pair of proteins interact

- Notation: \(M_g\) - set of all domains contained in protein \(g\)

\[
Pr\{I(g, h)\} = 1 - \bigcap_{d \in M_g} \bigcap_{e \in M_h} (1 - \Pr\{I(g, h)|D(g, d)D(h, e)\})
\]

Do not correct for possible correlation between domain occurrences

- Current available data is too sparse to be useful
Computational Prediction of Host-Pathogen Interactions

Analysis Flow

Decompose Domain Pairs And Increment Counts

D1  D2

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8 4 0 5 1
0 6 1 7
0 0 3
2 0
2
Analysis Flow

Decompose Domain Pairs And Increment Counts

P1

P2

Host

Pathogen

D1

D2

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Computational Prediction of Host-Pathogen Interactions

Analysis Flow

Decompose Domain Pairs And Increment Counts

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image of D1 and D2 domains]</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

Host
Pathogen

P1
P2

D1
D2
Decompose Domain Pairs And Increment Counts
Computational Prediction of Host-Pathogen Interactions

Analysis Flow

Decompose Domain Pairs And Increment Counts

\[
\text{Pr}\{I(g, h) | D(g, d) D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}
\]
Computational Prediction of Host-Pathogen Interactions

**Analysis Flow**

Decompose Domain Pairs And Increment Counts

\[
\Pr\{I(g, h) \mid D(g, d)D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}
\]

\[
Pr\{I(g, h)\} = 1 - \prod_{d \in M_g} \prod_{e \in M_h} (1 - \Pr\{I(g, h) \mid D(g, d)D(h, e)\})
\]
Intraspecies predictors can use gold standard datasets.
Intraspecies predictors can use use gold standard datasets.

No such datasets for host-pathogen PPIs.
Evaluating Predictions

- Intraspecies predictors can use gold standard datasets

- No such datasets for host-pathogen PPIs

- We developed three computational tests to evaluate our predictions
  - Triplet proximity
  - Triplet coexpression
  - Weighted functional enrichment
Computational Prediction of Host-Pathogen Interactions

Evaluation (Triplet Proximity)

Identify triplets in the predicted network
Identification of triplets in the predicted network.
Identify triplets in the predicted network

Compute shortest path in triplets and plot distributions

H-H-P Triplet

H-P-P Triplet
Identify triplets in the predicted network

Compute shortest path in triplets and plot distributions

Expect a negative correlation between the number of such pairs at a particular distance and the distance itself

H-H-P Triplet

H-P-P Triplet
Computational Prediction of Host-Pathogen Interactions

Evaluation (Triplet Coexpression)

Identify triplets in the predicted network
Compute Speaman’s correlation and plot distributions

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]

H-H-P Triplet

H-P-P Triplet
Dyer, Murali, and Sobral

Computational Prediction of Host-Pathogen Interactions

Evaluation (Triplet Coexpression)

- Identify triplets in the predicted network
- Compute Speaman’s correlation and plot distributions
- Expect pairs of proteins to show correlated expression

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]

H-H-P Triplet

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]

H-P-P Triplet
Given a pair of GO functions $c$ and $d$, let $G_{c,d}$ be the subgraph of $G$ induced by the host proteins annotated with $c$ and pathogen protein annotated with $d$. 

Evaluation (Functional Enrichment)
Given a pair of GO functions $c$ and $d$, let $G_{c,d}$ be the subgraph of $G$ induced by the host proteins annotated with $c$ and pathogen protein annotated with $d$.
Given a pair of GO functions \( c \) and \( d \), let \( G_{c,d} \) be the subgraph of \( G \) induced by the host proteins annotated with \( c \) and pathogen protein annotated with \( d \).

\[ w_{c,d} \] is the sum of the weighted edges.
Given a pair of GO functions $c$ and $d$, let $G_{c,d}$ be the subgraph of $G$ induced by the host proteins annotated with $c$ and pathogen protein annotated with $d$.

- $w_{c,d}$ is the sum of the weighted edges.

- Generate random graphs to determine a p-value for $w_{c,d}$:

$$p_{c,d} = \frac{\text{num graphs} \geq w_{c,d}}{\text{num random graphs}}$$
Functions in the GO hierarchy are specified at multiple levels of detail so a pair of enriched functions may contain closely related pairs of functions.
Functions in the GO hierarchy are specified at multiple levels of detail so a pair of enriched functions may contain closely related pairs of functions.

Collapse enrichments based on the following criteria given two sets of enrichments \((f, g)\) and \((l, m)\):

1. \(p_{l, m} < p_{f, g}\) i.e. \((l, m)\) is more statistically significant than \((f, g)\), and
2. \(l\) is either an ancestor or descendant of \(f\), and
3. \(m\) is either an ancestor or descendant of \(g\).
Computational Prediction of Host-Pathogen Interactions

Data Sources

- Protein sequence information
  - Uniprot (Bairoch et al. 2005)

- Protein domain profiles
  - InterProScan (Quevillon et al. 2005)

- Functional annotations
  - Gene Ontology (Ashburner et al. 2000)

- Protein interactions
  - IntAct (Hermjakob et al. 2004)
  - MIPS (Guldener et al. 2006)
  - DIP (Salwinski et al. 2004)
  - BIND (Gilbert et al. 2005)
  - Reactome (Joshi-Tope et al. 2005)

- Gene expression
  - NCBI GEO (Edgar et al. 2002)
*P. falciparum*

- Bozdech et al. (2003)
  - Merozoite invasion of human red blood cell
  - 46 samples

- Le roch et al. (2003)
  - Merozoite invasion of human red blood cell
  - 17 samples

Gene Expression Data

- **P. falciparum**
  - Bozdech et al. (2003)
    - Merozoite invasion of human red blood cell
    - 46 samples
  - Le roch et al. (2003)
    - Merozoite invasion of human red blood cell
    - 17 samples

- **H. sapiens**
  - Boldt et al. (Unpublished)
    - Healthy, un/complicated symptoms
    - 15 samples
  - Ockenhouse et al. (2006)
    - Experimentally and naturally infected
    - 71 samples

To focus predictions on proteins most likely involved in pathogenesis, we prepare our training set as follows.
To focus predictions on proteins most likely involved in pathogenesis we prepare our training set as follows:

1. Discard proteins with certain functions
   - Nucleus
   - Ribosome
   - Proteolysis
   - Nucleic acid binding
   - Helicase activity
To focus predictions on proteins most likely involved in pathogenesis, we prepare our training set as follows:

1. Discard proteins with certain functions
   - Nucleus
   - Ribosome
   - Proteolysis
   - Nucleic acid binding
   - Helicase activity

2. Retain proteins with certain functions
   - Hemoglobin metabolism
   - Cell-cell communication
   - Cell death
   - Blood coagulation
   - Subtilase activity
To focus predictions on proteins most likely involved in pathogenesis, we prepare our training set as follows:

1. Discard proteins with certain functions
   - Nucleus
   - Ribosome
   - Proteolysis
   - Nucleic acid binding
   - Helicase activity

2. Retain proteins with certain functions
   - Hemoglobin metabolism
   - Cell-cell communication
   - Cell death
   - Blood coagulation
   - Subtilase activity

3. Remove proteins which do not participate in at least one interaction
To focus predictions on proteins most likely involved in pathogenesis, we prepare our training set as follows:

1. Discard proteins with certain functions
   - Nucleus
   - Ribosome
   - Proteolysis
   - Nucleic acid binding
   - Helicase activity

2. Retain proteins with certain functions
   - Hemoglobin metabolism
   - Cell-cell communication
   - Cell death
   - Blood coagulation
   - Subtilase activity

3. Remove proteins which do not participate in at least one interaction

4. Remove proteins which contain only infrequent domains (appears in < 4 proteins)
Pr\{I(g, h)\|D(g, d)D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}

Pr\{I(g, h)\} = 1 - \prod_{d \in M_g} \prod_{e \in M_h} (1 - Pr\{I(g, h)\|D(g, d)D(h, e)\})
## Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Training (PPIs)</th>
<th>Prediction (Proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>4,177</td>
<td></td>
</tr>
<tr>
<td>Fly</td>
<td>9,384</td>
<td></td>
</tr>
<tr>
<td><em>Plasmodium</em></td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>*<em>Total Human-<em>Plasmodium</em></em></td>
<td>4,304</td>
<td></td>
</tr>
<tr>
<td>*<em>Total Fly-<em>Plasmodium</em></em></td>
<td>9,511</td>
<td></td>
</tr>
</tbody>
</table>
## Data Sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Training (PPIs)</th>
<th>Prediction (Proteins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>4,177</td>
<td>27,371</td>
</tr>
<tr>
<td>Fly</td>
<td>9,384</td>
<td>11,924</td>
</tr>
<tr>
<td><em>Plasmodium</em></td>
<td>127</td>
<td>938</td>
</tr>
<tr>
<td>Total Human-<em>Plasmodium</em></td>
<td>4,304</td>
<td>&gt; 25 million</td>
</tr>
<tr>
<td>Total Fly-<em>Plasmodium</em></td>
<td>9,511</td>
<td>&gt; 11 million</td>
</tr>
</tbody>
</table>
We predict 516 PPIs between 158 human proteins and 30 *Plasmodium* proteins.

<table>
<thead>
<tr>
<th>PPI Probability</th>
<th>Human-<em>Plasmodium</em></th>
<th>Fly-<em>Plasmodium</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 - 0.55</td>
<td>185</td>
<td></td>
</tr>
<tr>
<td>0.55 - 0.60</td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>0.60 - 0.65</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>0.65 - 0.70</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>0.70 - 0.75</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>0.75 - 0.80</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>516</strong></td>
<td></td>
</tr>
</tbody>
</table>
We predict 516 PPIs between 158 human proteins and 30 *Plasmodium* proteins and 44 PPIs between 29 fly proteins and 8 *Plasmodium* proteins.

<table>
<thead>
<tr>
<th>PPI Probability</th>
<th>Human-<em>Plasmodium</em></th>
<th>Fly-<em>Plasmodium</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 - 0.55</td>
<td>185</td>
<td>6</td>
</tr>
<tr>
<td>0.55 - 0.60</td>
<td>175</td>
<td>15</td>
</tr>
<tr>
<td>0.60 - 0.65</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>0.65 - 0.70</td>
<td>61</td>
<td>12</td>
</tr>
<tr>
<td>0.70 - 0.75</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>0.75 - 0.80</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>516</strong></td>
<td><strong>44</strong></td>
</tr>
</tbody>
</table>
72% of H-H-P triplets are at distance < 3 in the human PPI network

• Predictions are likely connecting two human proteins with a functional relationship
72% of H-H-P triplets are at distance < 3 in the human PPI network.

- Predictions are likely connecting two human proteins with a functional relationship.

The average distance between *Plasmodium* proteins (H-P-P) is 5.5.

- *Plasmodium* PPI network is sparse (2,643 PPIs).
Pairs of *Plasmodium* proteins seem to be coexpressed (H-P-P-P).

- **Spearman's Coefficient**
  - Fraction of Triplets
  - **Bozdech (2003)**
  - **Le Roch (2003)**
Coexpression of human proteins show more variance (H-H-P)

- Ockenhouse (2006)
- Boldt (Unpublished)
Generate 1 million random graphs to calculate p-values for enrichments (discard when $p_{c,d} > 0.05$)
Generate 1 million random graphs to calculate p-values for enrichments (discard when $p_{c,d} > 0.05$)

Collapse enrichments and remove any pairs in which one function does not have a depth of at least three in the GO hierarchy
Functional Enrichment Results

- Generate 1 million random graphs to calculate p-values for enrichments (discard when $p_{c,d} > 0.05$)

- Collapse enrichments and remove any pairs in which one function does not have a depth of at least three in the GO hierarchy

- Ockenhouse et al. (2006) report that genes up-regulated in infected individuals are enriched for fifteen GO terms.
  - We identify ten of these functions in our analysis
    - Apoptosis (GO:0006915)
    - Regulation of apoptosis (GO:0042981)
    - Inflammatory response (GO:0006512)
    - Immune response (GO:0006955)
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes Q8IAS3, a known PfEMP1
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes Q8IAS3, a known PfEMP1
- Predicted interaction with Q5TEH4 (plasminogen)
  - Important step in merozoite release from the human RBC is activation of plasminogen (Roggwiller et al. 1997)
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes Q8IAS3, a known PfEMP1
- Predicted interactions with hepatocyte growth factors (HGFs)
- HGF induction is required for hepatocyte invasion (Carrolo et al. 2003)
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes Q8IAS3, a known *PfEMP1*
- Also includes Q8IAL6 and Q8I339, hypothetical proteins
  - Each contains a transmembrane domain
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes Q8IAS3, a known PfEMP1
- Also includes Q8IAL6 and Q8I339, hypothetical proteins
  - Predicted interactions with thrombospondin
  - Mature parasitized RBCs have an affinity for thrombospondin (Baruch et al. 1996)
Our analysis also finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “subtilase activity” and “dense granule”

- Dense granule is a specialized secretory organelle important in RBC invasion and degradation (O’Donnell *et al.* 2005)
- Includes Q8IHZ5, a known subtilisin-like protease
- Involved in blood platelet degradation
Our analysis also finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “subtilase activity” and “dense granule”

- Dense granule is a specialized secretory organelle important in RBC invasion and degradation (O’Donnell et al. 2005)
- Includes Q8IHZ5, a known subtilisin-like protease
- Involved in blood platelet degradation
- Additionally, includes Q8IKP8, hypothetical protein
Predicting interactions between host and pathogen proteins is an unsolved problem with important implications in biomedicine.
Predicting interactions between host and pathogen proteins is an unsolved problem with important implications in biomedicine.

We have presented an algorithm that integrates known domain profiles with interactions between proteins from the same organism to predict interactions between host and pathogen proteins.
Predicting interactions between host and pathogen proteins is an unsolved problem with important implications in biomedicine.

We have presented an algorithm that integrates known domain profiles with interactions between proteins from the same organism to predict interactions between host and pathogen proteins.

When applied to the *H. sapiens* - *P. falciparum* system our method identifies several biologically important sub-networks.
Predicting interactions between host and pathogen proteins is an unsolved problem with important implications in biomedicine.

We have presented an algorithm that integrates known domain profiles with interactions between proteins from the same organism to predict interactions between host and pathogen proteins.

When applied to the *H. sapiens - P. falciparum* system our method identifies several biologically important sub-networks.

Additional predictions are made with known *PfEMP1s* and *MSP1s* with probabilities less than 0.50 suggesting that integrating additional data sources into our system may enable us to predict more PPIs involved in pathogenesis with increased confidence.
Experimental Verification

Myriad Data Results If Available In Time
Matt Dyer (Virginia Bioinformatics Institute)

T. M. Murali (Department of Computer Science, Virginia Tech)

Dharmendar Rathore (Virginia Bioinformatics Institute)

This work was supported by the Department of Defense under grant #DAAD 13-02-C-0018 to B. Sobral, PI


Matt’s presentation is Wednesday, July 25, 10:15 - 10:50 AM