Computational Prediction of Host-Pathogen Protein-Protein Interactions

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

$^1$Genetics, Bioinformatics, and Computational Biology Program, $^2$Virginia Bioinformatics Institute, and $^3$Department of Computer Science.
Virginia Polytechnic Institute and State University.
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
Computational Prediction of Host-Pathogen Interactions

Malaria Life-cycle

Mosquito Stages
- Mosquito infection by ingestion of gametocytes
- Development of sporozoites in mid gut
- Ookinetes penetrate the midgut wall
- Macrogamete
- Microgamete
- Zygote formation
- Oocyst development
- Invasion of salivary glands

Human Infection
- by sporozoites in salivary secretion

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

Erythrocytic Stages
- Release of erythrocytic merozoites
- Asexual Erythrocyte Stages
- Schizont
- Trophozoite
- Ring stage
- Macrogametocyte
- Microgametocyte

Sexual Stages
- Mosquito stages
Computational Prediction of Host-Pathogen Interactions

Malaria Life-cycle

Mosquito Stages
- Oocyst development
- Ookinet penetration of midgut wall
- Development of sporozoites in mid gut
- Mosquito infection by ingestion of gametocytes

Invasion of salivary glands
- Invasion of hepatic cells by sporozoites

Erythrocytic Stages
- Merozoites
- Release of erythrocytic merozoites

Asexual Erythrocyte Stages
- Schizont
- Invasion

Liver Stages
- Formation of exo-erythrocytic merozoites in hepatic schizont

Release of exo-erythrocytic merozoites
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
Predicting PPIs

Organism X
Computational Prediction of Host-Pathogen Interactions

Predicting PPIs

Organism X
Literature Review

**Domain Profiles**

2001

Domain Profiles

2007
Domain Profiles
Domain Profiles

Computational Prediction of Host-Pathogen Interactions

2001

- **Domain Profiles**

- **Sequence Homology**

2007

Literature Review

Source Organisms
Computational Prediction of Host-Pathogen Interactions

**Domain Profiles**

**Sequence Homology**

Target Organism

Identify Orthologs And Map Interactions
Computational Prediction of Host-Pathogen Interactions

**Domain Profiles**

**Sequence Homology**

**Literature Review**

**Source Organisms**

**Identify Orthologs And Map Interactions**

**Target Organism**
2001

- **Domain Profiles**

- **Sequence Homology**
Computational Prediction of Host-Pathogen Interactions

2001

- **Domain Profiles**

- **Sequence Homology**

- **Machine Learning**

2007

**Literature Review**


- huang et al. bmc bioinformatics (2007) 8(1):152


---

**Domain Profiles**


**Sequence Homology**


**Machine Learning**

Literature Review

- **Domain Profiles**

- **Sequence Homology**

- **Machine Learning**
Literature Review

2001

- **Domain Profiles**

- **Sequence Homology**

- **Machine Learning**

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

2007
Predicting PPIs

Host

Pathogen
Computational Prediction of Host-Pathogen Interactions

Host

Pathogen

Predicting PPIs
Experimental studies test only a small number of interactions at a time.
Research Challenges

- 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)
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)
$P$ - set of proteins with at least one domain and one interaction
$P$ - set of proteins with at least one domain and one interaction

$P_d$ - subset of proteins 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 $Pr\{I(g,h) \mid D(g,d) \cap D(h,e)\}$ using Bayes rule
Probability that two proteins (g and h) interact given they contain domains d and e

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

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

Do not correct for possible correlation between domain occurrences
- Current available data is too sparse to be useful
Decompose Domain Pairs And Increment Counts

D1

D2

Analysis Flow
Computational Prediction of Host-Pathogen Interactions

Analysis Flow

Decompose Domain Pairs And Increment Counts

Host

Pathogen

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Computational Prediction of Host-Pathogen Interactions

Analysis Flow

Decompose Domain Pairs And Increment Counts

Host
Pathogen
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></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></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
Analysis Flow

Decompose Domain Pairs And Increment Counts

\[
\Pr\{I(g, h) | D(g, d) D(h, e)\} = \frac{|S_{d,e}|}{|P_d||P_e| - |P_d \cap P_e|}
\]
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 use gold standard datasets
Intraspecies predictors use gold standard datasets.

No such datasets for host-pathogen PPIs.
Intraspecies predictors 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
Identify triplets in the predicted network
Computational Prediction of Host-Pathogen Interactions

Identify triplets in the predicted network

Evaluation (Triplet Proximity)
Computational Prediction of Host-Pathogen Interactions

- Identify triplets in the predicted network
- Compute shortest path in triplets and plot distributions

**Evaluation (Triplet Proximity)**

1. H-H-P Triplet
2. H-P-P Triplet

**Diagram:**
- Nodes represent entities: Host (blue), Pathogen (red)
- Connections (edges) represent interactions
- Dashed lines indicate shortest paths in triplets
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
Identify triplets in the predicted network

Evaluation (Triplet Coexpression)
Identify triplets in the predicted network

Compute Speaman’s correlation and plot distributions

Evaluation (Triplet Coexpression)

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

H-H-P Triplet

H-P-P Triplet
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

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 proteins 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 proteins 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 proteins 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 proteins annotated with $d$.

- $w_{c,d}$ is the sum of the weighted edges.
- Generate random graphs to determine a p-value for $w_{c,d}$.

$$p_{c,d} = \frac{\text{num graphs } \geq w_{c,d}}{\text{num random graphs}}$$
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)
<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><em>Plasmodium</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Plasmodium</em></td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>Total Human-Plasmodium</td>
<td>4,304</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><strong>Human</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4,177</td>
<td>27,371</td>
</tr>
<tr>
<td><strong>Plasmodium</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127</td>
<td>938</td>
</tr>
<tr>
<td><strong>Total Human-Plasmodium</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4,304</td>
<td>&gt; 25 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>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 - 0.55</td>
<td>185</td>
</tr>
<tr>
<td>0.55 - 0.60</td>
<td>175</td>
</tr>
<tr>
<td>0.60 - 0.65</td>
<td>31</td>
</tr>
<tr>
<td>0.65 - 0.70</td>
<td>61</td>
</tr>
<tr>
<td>0.70 - 0.75</td>
<td>16</td>
</tr>
<tr>
<td>0.75 - 0.80</td>
<td>48</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>516</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

![Graph showing the distribution of H-H-P and H-P-P triplets vs. intraspecies distance.](image-url)
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).

<table>
<thead>
<tr>
<th>Fraction of Triplets</th>
<th>Spearman’s Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.300</td>
<td>-1</td>
</tr>
<tr>
<td>0.225</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.150</td>
<td>0</td>
</tr>
<tr>
<td>0.075</td>
<td>0.50</td>
</tr>
<tr>
<td>0.00</td>
<td>1</td>
</tr>
</tbody>
</table>
Coexpression of human proteins show more variance (H-H-P)

- Ockenhouse (2006)
- Boldt (Unpublished)

Spearman’s Coefficient

Fraction of Triplets

Spearman’s Coefficient
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
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 VAR (Q8IAS3), a known PfEMPI
Our analysis finds an enriched subnetwork between human proteins annotated with “blood coagulation” and *Plasmodium* proteins annotated with “integral to membrane”

- Includes VAR (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 VAR (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 VAR (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 VAR (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](#))
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 biologically important sub-networks.
T. M. Murali (Department of Computer Science, Virginia Tech)

Bruno Sobral (Virginia Bioinformatics Institute, Virginia Tech)

Dharmendar Rathore (Virginia Bioinformatics Institute, Virginia Tech)

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