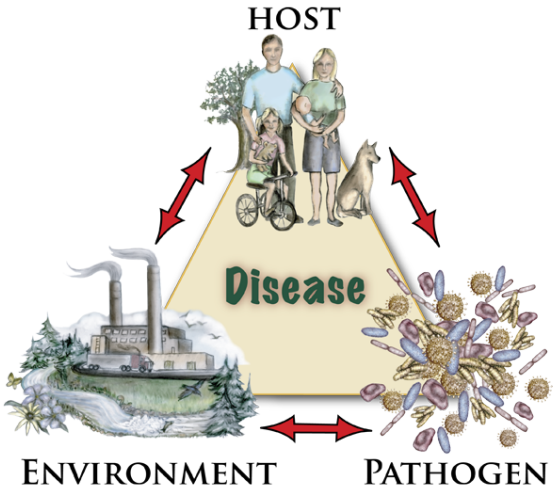
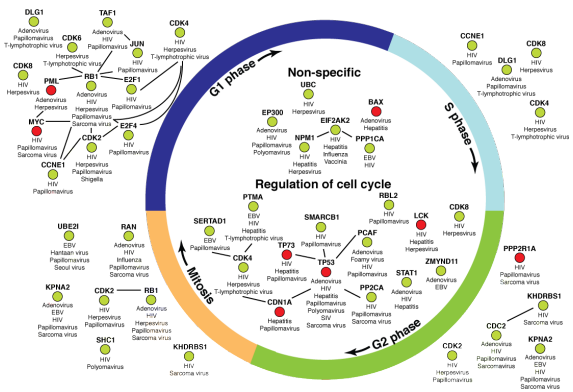




Host-Pathogen Protein Interaction Networks

Tutorial in 2008 ICSB

T. M. Murali

Virginia Polytechnic Institute and State University

August 22, 2008

Data, Data, Data

- Hundreds of genomes sequenced.
- Systematic gene knockouts.
- Gene expression, proteomic, metabolic measurements.
- Molecular interaction networks, metabolic pathways.
- Catalogues of gene and protein function.

Systems Biology of Model Organisms

- Highly-sophisticated modeling and analysis approaches applies to *E. coli* and *S. cerevisiae*.

Systems Biology of Model Organisms

- Highly-sophisticated modeling and analysis approaches applies to *E. coli* and *S. cerevisiae*.
- More recently, such ideas are being used for multi-cellular model organisms such as *C. elegans* and *D. melanogaster*.

Systems Biology of Model Organisms

- Highly-sophisticated modeling and analysis approaches applies to *E. coli* and *S. cerevisiae*.
- More recently, such ideas are being used for multi-cellular model organisms such as *C. elegans* and *D. melanogaster*.
- Also to study complex human diseases such as cancer and diabetes.

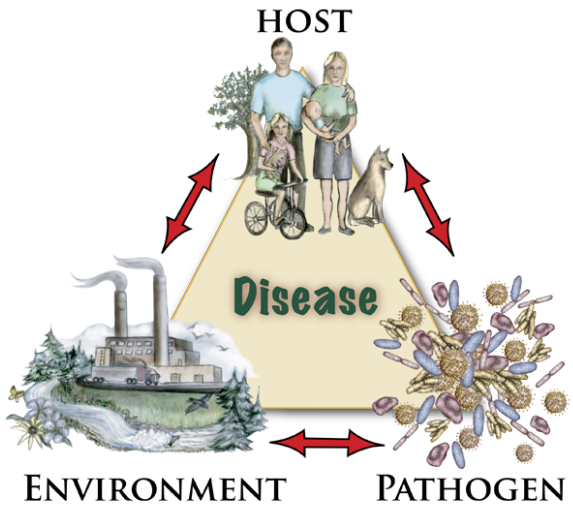
What about Infectious Diseases?

- Infectious diseases result in millions of deaths each year.
- Millions of dollars are spent annually to better understand how pathogens infect their hosts.
- Many deadly diseases have no effective vaccine.
- Several drugs have become obsolete due to acquired parasite resistance.

What about Infectious Diseases?

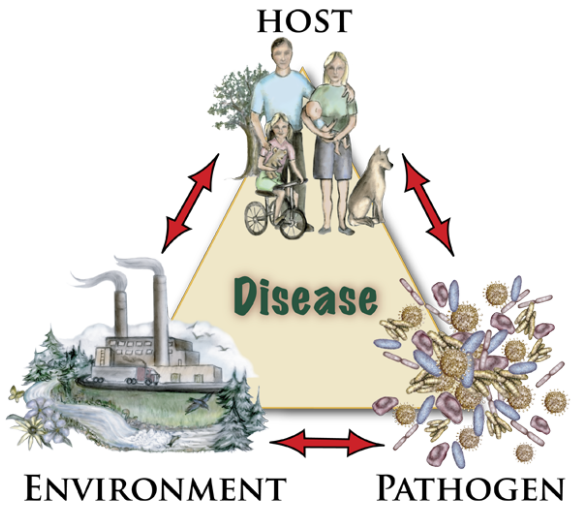
- Infectious diseases result in millions of deaths each year.
- Millions of dollars are spent annually to better understand how pathogens infect their hosts.
- Many deadly diseases have no effective vaccine.
- Several drugs have become obsolete due to acquired parasite resistance.
- Relatively less systems-level analysis than for model organisms or for complex diseases.
  - Functional genomic data for infectious diseases is sparse.
  - Systems biology of two interacting systems may be more complex than that of a single organism.

Focus of the Tutorial



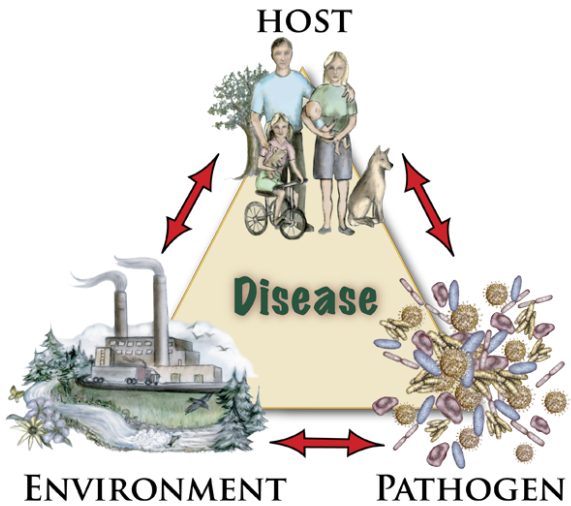
- Host-Pathogen Protein-Protein Interactions (HP PPIs).

Focus of the Tutorial



- Host-Pathogen Protein-Protein Interactions (HP PPIs).
- 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*: Merozoite surface proteins (MSP1s) allow the parasite to invade a red blood cell.

Focus of the Tutorial

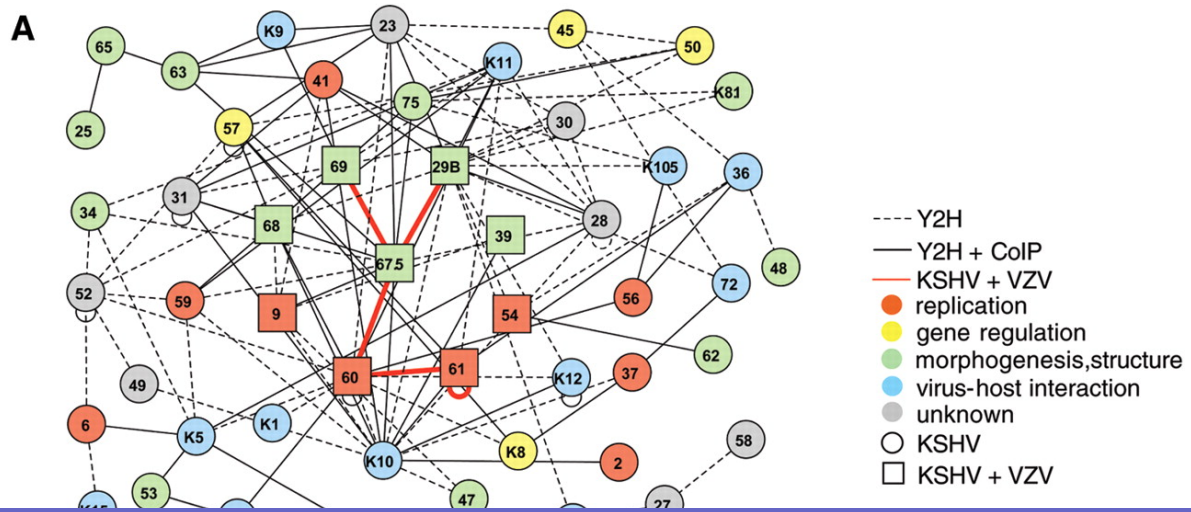


- Host-Pathogen Protein-Protein Interactions (HP PPIs).
- An important aspect of any host-pathogen system is the mechanism by which a pathogen infects its host.
  - Surface proteins and molecules form the foundation of communication between a host and pathogen.
  - Protein-protein interactions play a vital role in initiating infection.
  - *P. falciparum*: Merozoite surface proteins (MSP1s) allow the parasite to invade a red blood cell.
- Studying which interactions enable a pathogen to infect its host may provide us with potential targets for therapeutics.

Outline

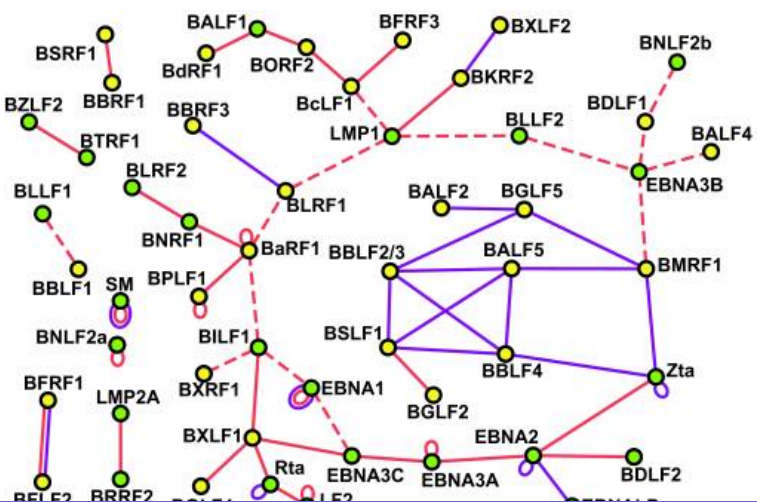
Large-scale Yeast Two-Hybrid Experiments

- Herpesviral protein networks and their interaction with the human proteome, Uetz et al., Science, 2006 Jan 13;311(5758):239-42.
  - 296 interactions between herpesviral proteins.
  - Predicted herpesviral-human PPIs and simulated infection.



Large-scale Yeast Two-Hybrid Experiments

- Herpesviral protein networks and their interaction with the human proteome, Uetz et al., Science, 2006 Jan 13;311(5758):239-42.
  - 296 interactions between herpesviral proteins.
  - Predicted herpesviral-human PPIs and simulated infection.
- Epstein-Barr virus and virus human protein interaction maps, Calderwood et al., Proc Natl Acad Sci U S A. 2007 May 1;104(18):7606-11
  - 43 EBV-EBV PPIs and 173 EBV-human PPIs.
  - Human proteins targeted by EBV were enriched for "hub" proteins and for proteins close to other human proteins.



Sources of Data

- Seven major databases: BIND, DIP, HPRD, IntAct, MINT, MIPS, and Reactome.
- Human-pathogen PPIs: 10,477 PPIs across 190 pathogen strains.
- Intra-human PPIs: 75,457 PPIs between human proteins.

Sources of Data

- Seven major databases: BIND, DIP, HPRD, IntAct, MINT, MIPS, and Reactome.
- Human-pathogen PPIs: 10,477 PPIs across 190 pathogen strains.
- Intra-human PPIs: 75,457 PPIs between human proteins.
- Intra-pathogen PPIs: MPIDB: the microbial protein interaction database, Goll et al., Bioinformatics. 2008 Aug 1;24(15):1743-4.

Distribution Across Pathogens

Group	#PPIs	#Strains	#Unique targeted human proteins	#proteins in human network
HIV	8,024	44	743	671
Hepatitis	1,244	16	109	93
Influenza	287	4	76	76
Papillomavirus	229	12	96	94
EBV	211	2	135	121
Adenovirus	80	9	60	59
Herpesvirus	64	20	54	54
Yersinia	57	3	56	45
Sarcoma virus	52	6	36	35
		...		
TOTAL	10,477	190	1,233	1,109

Distribution Across Pathogens

Group	#PPIs	#Strains	#Unique targeted human proteins	#proteins in human network
HIV	8,024	44	743	671
Hepatitis	1,244	16	109	93
Influenza	287	4	76	76
Papillomavirus	229	12	96	94
EBV	211	2	135	121
Adenovirus	80	9	60	59
Herpesvirus	64	20	54	54
Yersinia	57	3	56	45
Sarcoma virus	52	6	36	35
		...		
TOTAL	10,477	190	1,233	1,109

- Dominated by viruses: HIV (77%) and Hepatitis (12%).

Data Sources

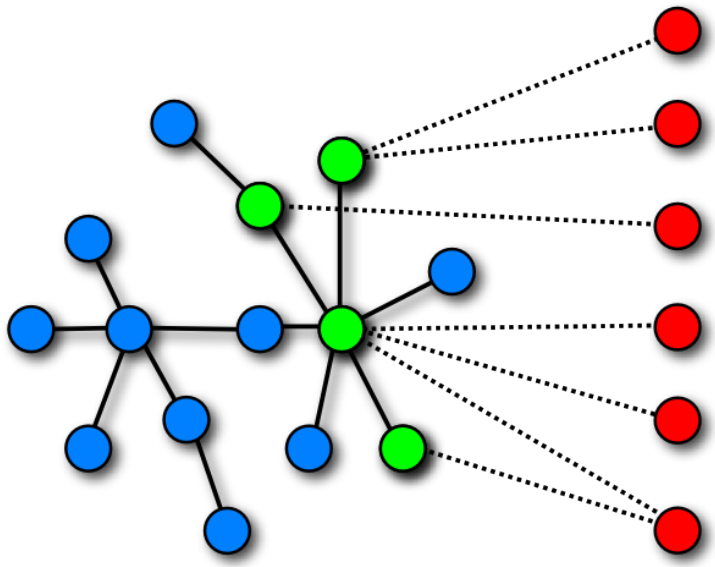
Support	Method	#interactions
Experimental	Reactome curated	7,229
	Not specified	2,305
	Yeast two-hybrid	419
	Pull down	314
	Co-immunoprecipitation	210
	Other technology (28 methods)	210
	Observed by $\geq$ two methods	260
Literature	Described in one paper	9,810
	Described in two papers	198
	Described in $\geq$ two papers	469

Outline

Goals of the Analysis

The Landscape of Human Proteins Interacting with Viruses and Other Pathogens, Dyer,Murali, and Sobral, *PLoS Pathogens*, volume 4, number 2, pp. e32, 2008.

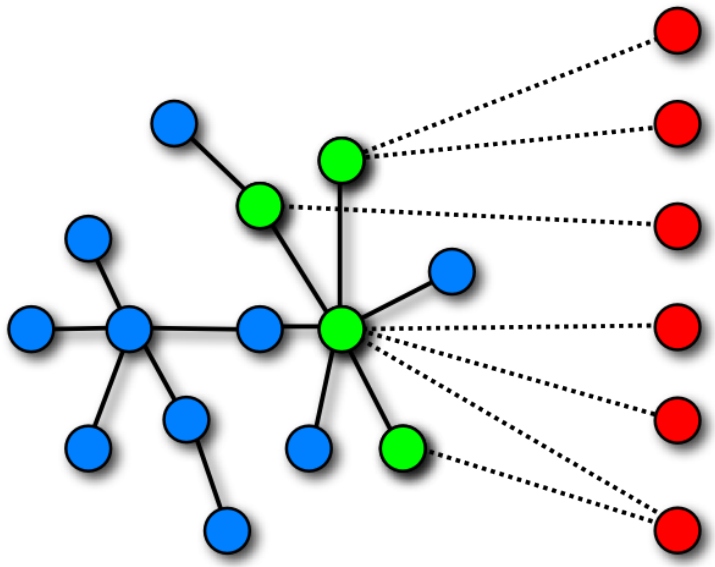
1. What are the properties of human proteins interacting with pathogens?
2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?



Goals of the Analysis

The Landscape of Human Proteins Interacting with Viruses and Other Pathogens, Dyer,Murali, and Sobral, *PLoS Pathogens*, volume 4, number 2, pp. e32, 2008.

1. What are the properties of human proteins interacting with pathogens?
2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?

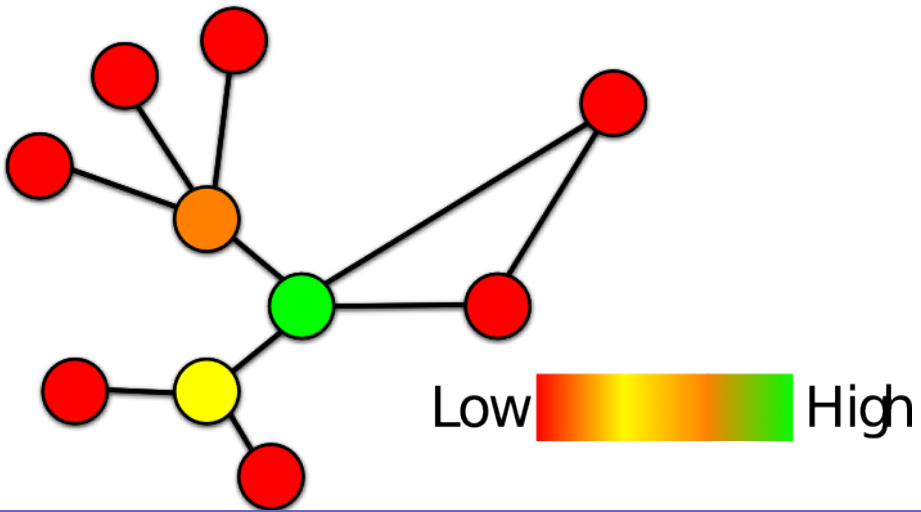


- We will attempt to study these questions primarily for viruses.
- We will focus on location of targeted human proteins in the human PPI network.

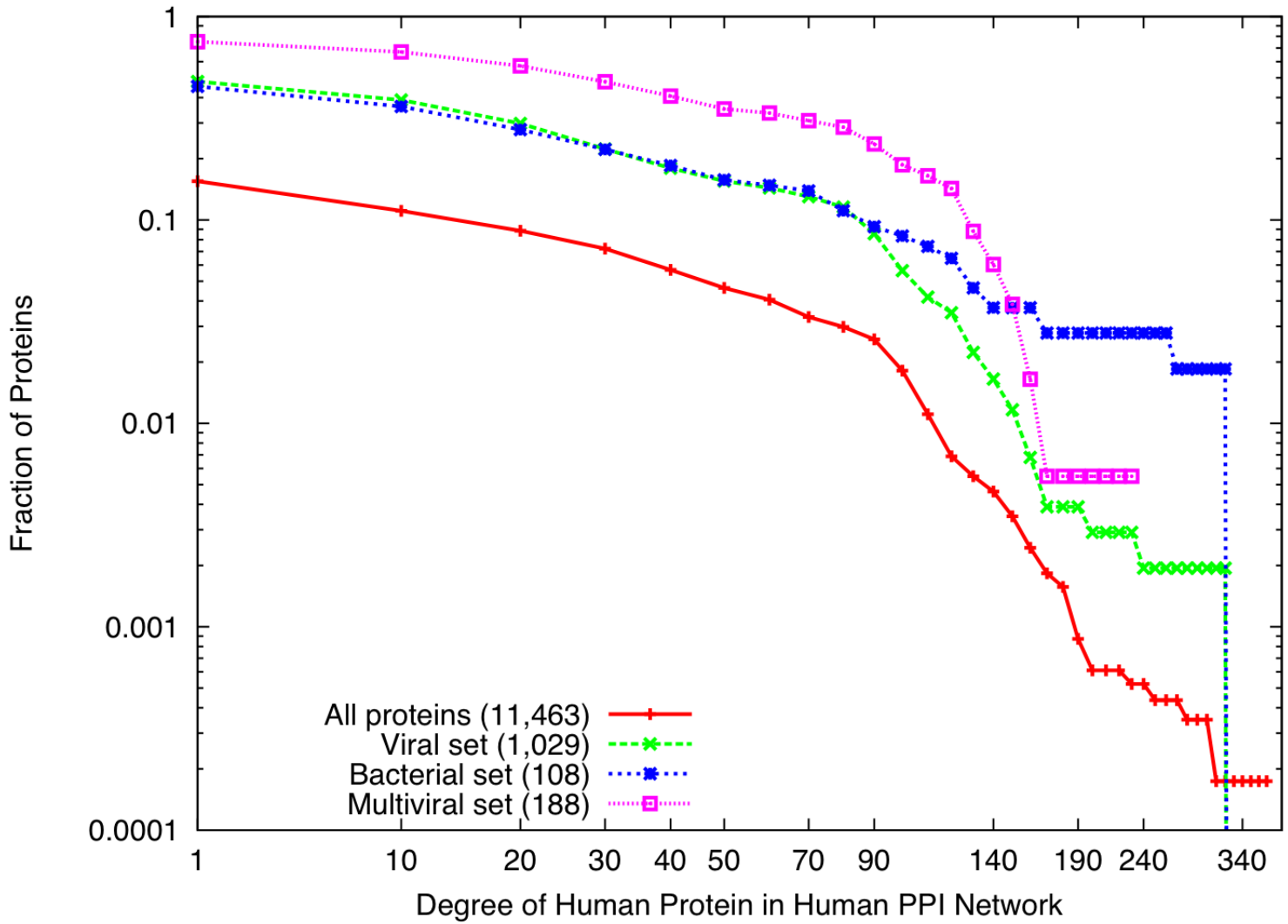
Definitions

- *Undirected graph* $G = (V, E)$ represents pairwise relationships (edges E) between objects (nodes V).
- *Degree* of a node $v \in V$ is number of edges incident on v.
- *Betweenness centrality* of a node $v \in V$ is the fraction of shortest paths that pass through v:

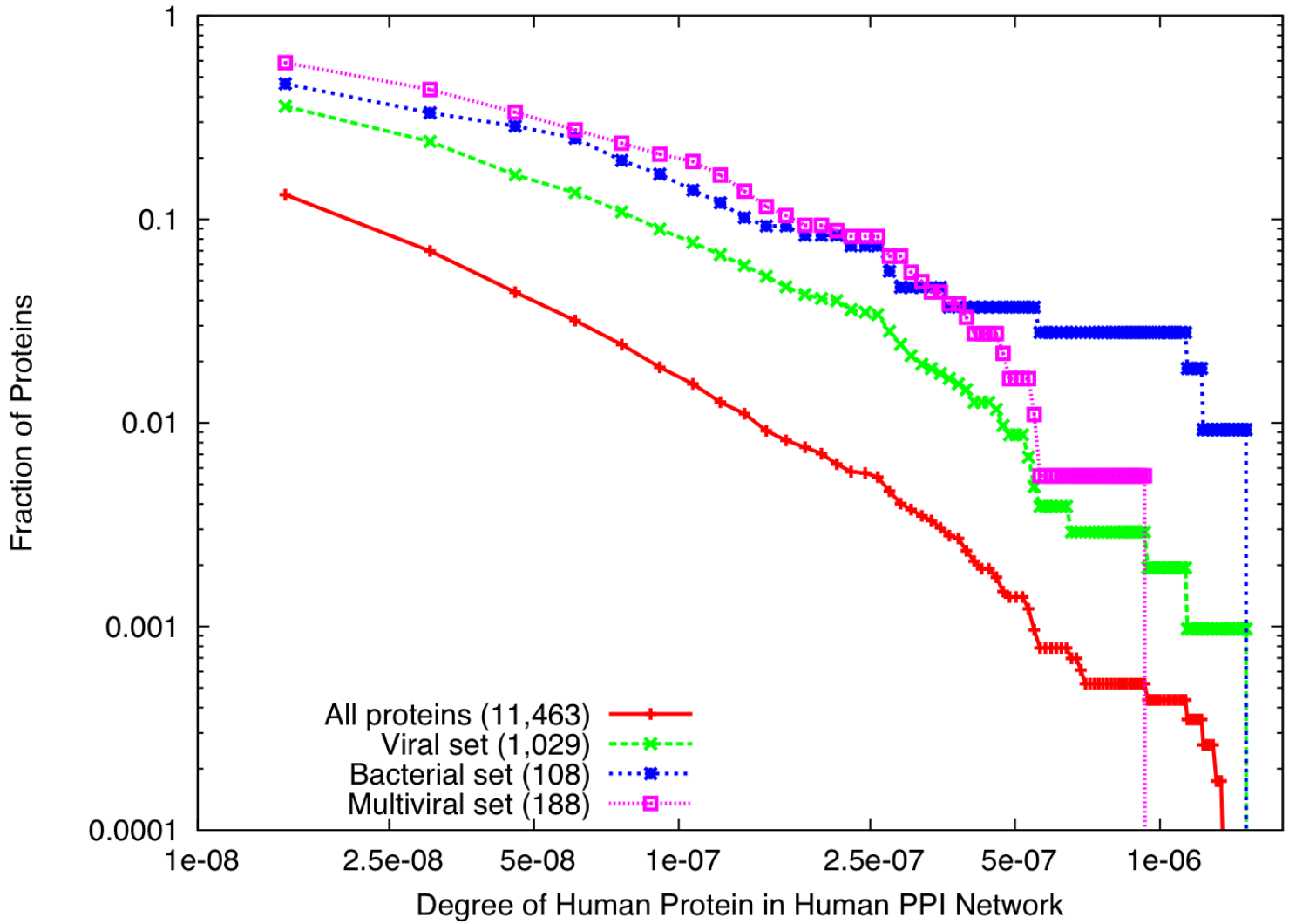
$$bc(v) = \sum_{\substack{u,w \in V \\ u,w \neq v}} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$$



Pathogens Interact with Human Hubs

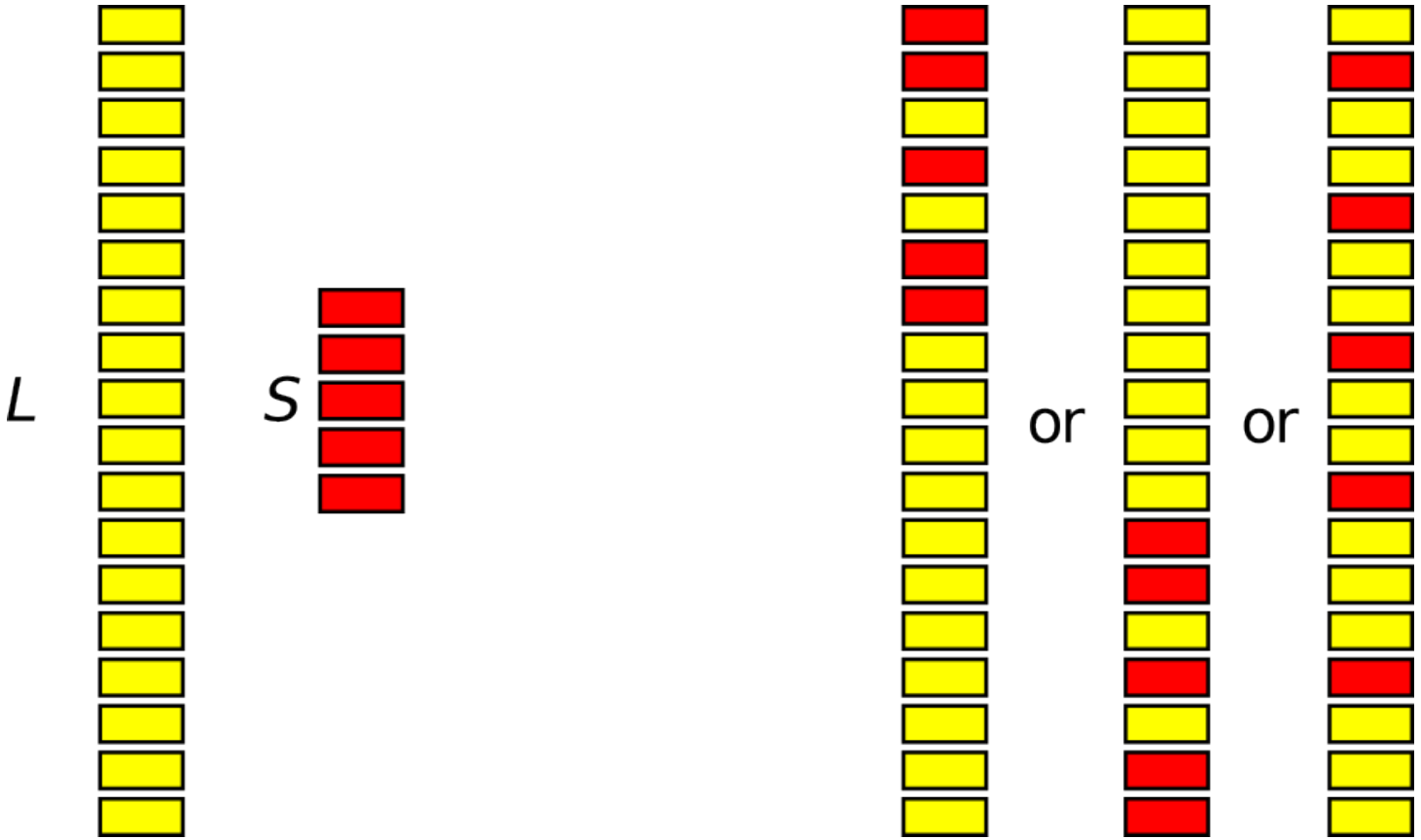


Pathogens Interact with Human Bottlenecks



Statistical Significance of Results

- Gene Set Enrichment Analysis (Subramanian et al., *PNAS* 2005).
- Given a ranked list L of proteins and a specific subset S of proteins, are the proteins in S randomly distributed in L or are they concentrated at the top of L?



Gene Set Enrichment Analysis

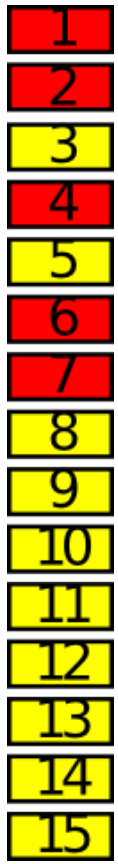
- Given a ranked list L of proteins and a specific subset S of proteins, are the proteins in S randomly distributed in L or are they concentrated at the top of L?

$$p_{hit}(S,i) = \sum_{j \in S, j \leq i} \frac{1}{|S|}$$

$$p_{miss}(S,i) = \sum_{j \notin S, j \leq i} \frac{1}{|L| - |S|}$$

$$es(S,i) = P_{hit}(S,i) - P_{miss}(S,i)$$

$$es(S,L) = \max_{1 \leq i \leq |L|} \left(\max\left(es(S,i), 0\right) \right)$$



Gene Set Enrichment Analysis

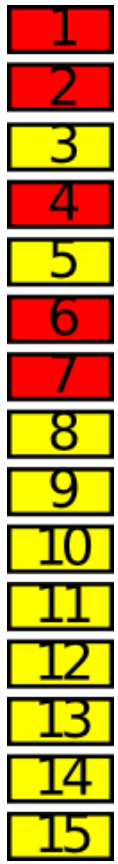
- Given a ranked list L of proteins and a specific subset S of proteins, are the proteins in S randomly distributed in L or are they concentrated at the top of L?

$$p_{hit}(S,i) = \sum_{j \in S, j \leq i} \frac{1}{|S|}$$

$$p_{miss}(S,i) = \sum_{j \notin S, j \leq i} \frac{1}{|L| - |S|}$$

$$es(S,i) = P_{hit}(S,i) - P_{miss}(S,i)$$

$$es(S,L) = \max_{1 \leq i \leq |L|} \left(\max\left(es(S,i), 0\right) \right)$$



- Assess the statistical significance of $es(S,L)$ from a distribution of scores for random sets of $|L|$ proteins.

GSEA Results

Network	Protein Set	#proteins in group	Degree ES	Degree *p*-value	Centrality ES	Centrality *p*-value
W	Virus	1,029	0.79	$< 1 \times 10^{-6}$	0.83	$< 1 \times 10^{-6}$
	Multivirus	182	0.84	$< 1 \times 10^{-6}$	0.86	1.2×10^{-4}
	Bacteria	108	0.76	3×10^{-5}	0.89	2.3×10^{-4}

GSEA Results

Network	Protein Set	#proteins in group	Degree ES	Degree *p*-value	Centrality ES	Centrality *p*-value
W	Virus	1,029	0.79	$< 1 \times 10^{-6}$	0.83	$< 1 \times 10^{-6}$
	Multivirus	182	0.84	$< 1 \times 10^{-6}$	0.86	1.2×10^{-4}
	Bacteria	108	0.76	3×10^{-5}	0.89	2.3×10^{-4}
HT	Virus	466	0.68	$< 1 \times 10^{-6}$	0.82	1.5×10^{-3}
	Multivirus	98	0.65	0.03	0.82	0.10
	Bacteria	43	0.79	2×10^{-3}	0.89	0.02
MC	Virus	958	0.78	$< 1 \times 10^{-6}$	0.80	$< 1 \times 10^{-6}$
	Multivirus	174	0.83	$< 1 \times 10^{-6}$	0.82	1.9×10^{-5}
	Bacteria	100	0.73	3.4×10^{-4}	0.85	9.2×10^{-4}

GSEA Results Without Human-HIV PPIs

Network	Protein Set	#proteins in group	Degree		Centrality	
			ES	p-value	ES	p-value
W	Virus	499	0.80	$< 1 \times 10^{-6}$	0.85	$< 1 \times 10^{-6}$
	Multivirus	81	0.83	$< 1 \times 10^{-6}$	0.88	3.6×10^{-3}
HT	Virus	267	0.70	1×10^{-6}	0.84	6.12×10^{-4}
	Multivirus	46	0.72	0.02	0.86	0.07
MC	Virus	958	0.78	$< 1 \times 10^{-6}$	0.80	$< 1 \times 10^{-6}$
	Multivirus	174	0.83	$< 1 \times 10^{-6}$	0.85	1.3×10^{-3}

Goals of the Analysis

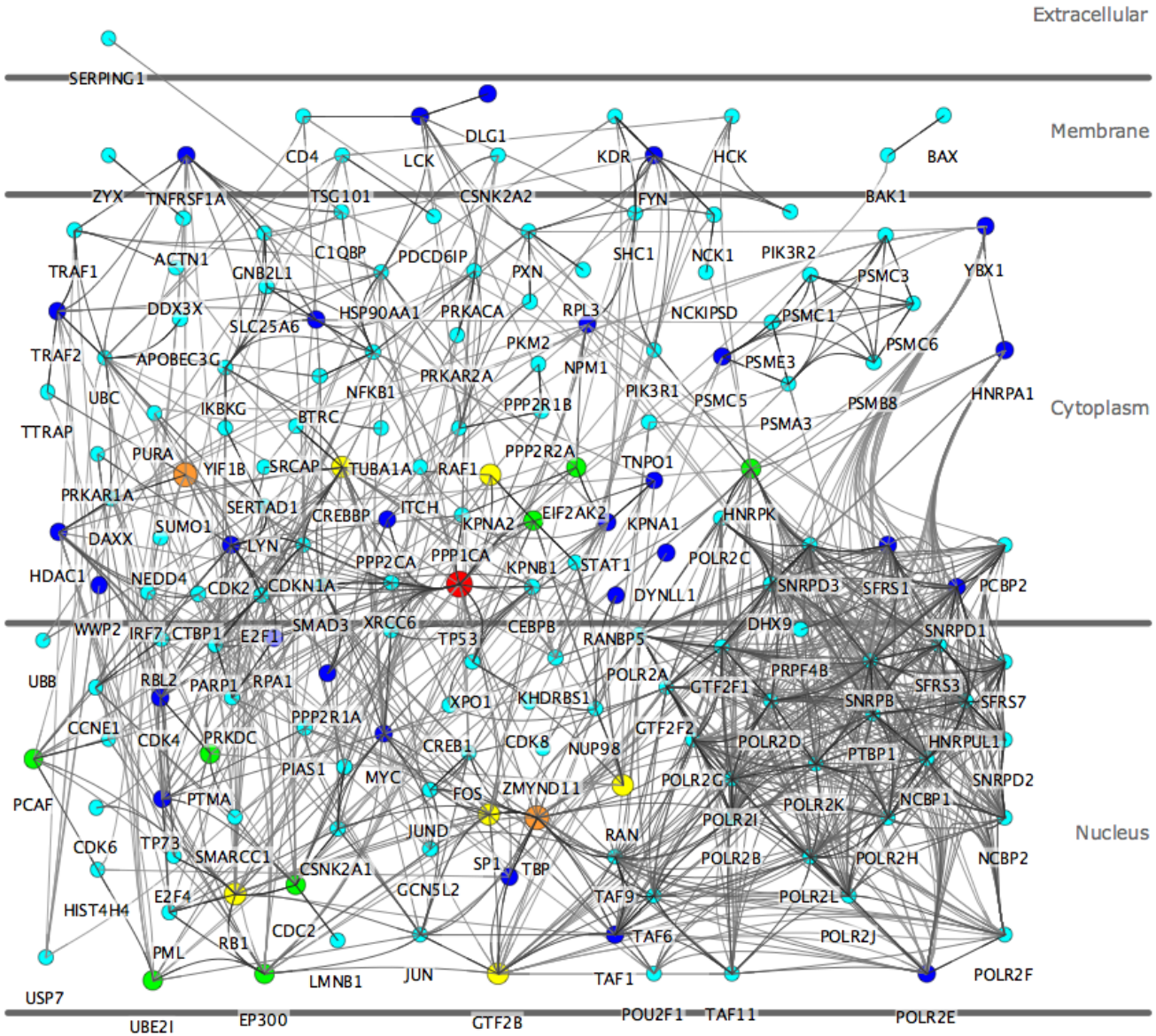
1. What are the properties of human proteins interacting with pathogens?

2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?

Goals of the Analysis

1. What are the properties of human proteins interacting with pathogens? Pathogens have evolved to interact with hubs and bottlenecks and with highly conserved and expressed proteins.
2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?

Multi-viral network



Functional Enrichment

- Use enrichment analysis to identify over-represented Gene Ontology (GO) functions annotating human proteins in the network.
- Account for multiple hypothesis testing using the method of Benjamini and Hochberg.
- Consider only functions enriched with a p-value of at most 0.05.

$$p_f(S, V) = \sum_{k=s_f}^{min(s_{pa(f)}, v_f)} \frac{\binom{v_f}{k}\binom{v_{pa(f)}-v_f}{s_{pa(f)}-k}}{\binom{v_{pa(f)}}{s_{pa(f)}}}$$

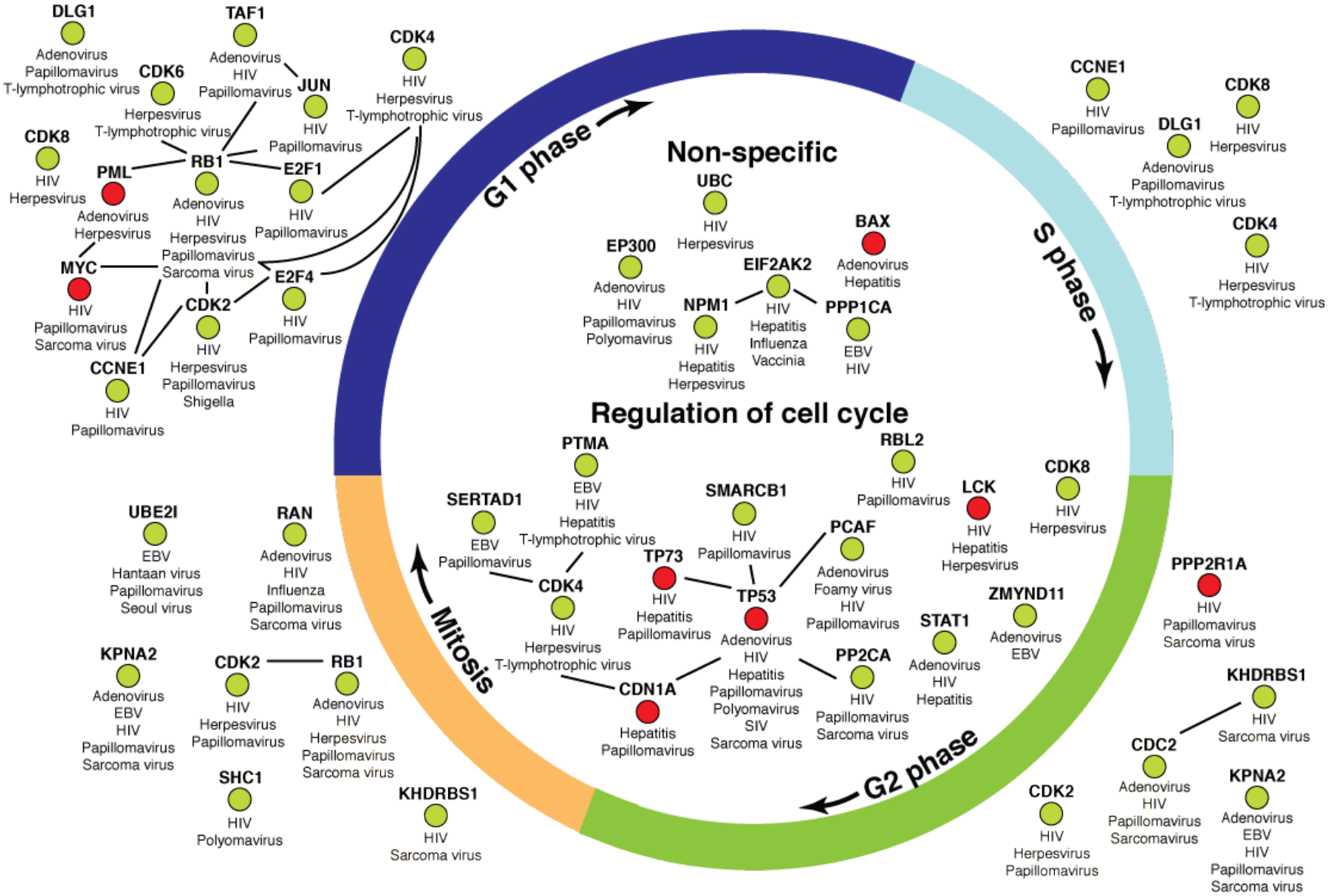
Summary of Enriched Functions

1. What are the properties of human proteins interacting with pathogens? Pathogens have evolved to interact with hubs and bottlenecks and with highly conserved and expressed proteins.
2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?

Summary of Enriched Functions

1. What are the properties of human proteins interacting with pathogens? Pathogens have evolved to interact with hubs and bottlenecks and with highly conserved and expressed proteins.
2. Do pathogens interact with certain functional classes of human proteins?
3. Which pathways are commonly activated by multiple pathogens?
   - Viral pathogens control host cell cycle program, regulate apoptotic machinery, and transport viral material across the nuclear membrane.
   - Bacterial pathogens: Trigger defence, immune, and wounding responses.

Cell Cycle


G1 phase
S phase
G2 phase
Mitosis
Non-specific
UBC
HIV
Herpesvirus
EP300
Adenovirus
HIV
Papillomavirus
Polyomavirus
EIF2AK2
HIV
Hepatitis
Influenza
Vaccinia
BAX
Adenovirus
Hepatitis
NPM1
HIV
Hepatitis
Herpesvirus
PPP1CA
EBV
HIV
Regulation of cell cycle
PTMA
EBV
HIV
Hepatitis
T-lymphotrophic virus
SERTAD1
EBV
Papillomavirus
CDK4
HIV
Herpesvirus
T-lymphotrophic virus
SMARCB1
HIV
Papillomavirus
TP73
HIV
Hepatitis
Papillomavirus
TP53
Adenovirus
HIV
Hepatitis
Papillomavirus
Polyomavirus
SIV
Sarcoma virus
PCAF
Adenovirus
Foamy virus
HIV
Papillomavirus
RBL2
HIV
Papillomavirus
LCK
HIV
Hepatitis
Herpesvirus
CDK8
HIV
Herpesvirus
CDN1A
Hepatitis
Papillomavirus
PP2CA
HIV
Papillomavirus
Sarcoma virus
STAT1
Adenovirus
HIV
Hepatitis
ZMYND11
Adenovirus
EBV
UBE2I
EBV
Hantaan virus
Papillomavirus
Seoul virus
RAN
Adenovirus
HIV
Influenza
Papillomavirus
Sarcoma virus
KPNA2
Adenovirus
EBV
HIV
Papillomavirus
Sarcoma virus
CDK2
HIV
Herpesvirus
Papillomavirus
RB1
Adenovirus
HIV
Herpesvirus
Papillomavirus
Sarcoma virus
SHC1
HIV
Polyomavirus
KHDRBS1
HIV
Sarcoma virus
DLG1
Adenovirus
Papillomavirus
T-lymphotrophic virus
CDK6
Herpesvirus
T-lymphotrophic virus
TAF1
Adenovirus
HIV
Papillomavirus
JUN
HIV
Papillomavirus
CDK4
HIV
Herpesvirus
T-lymphotrophic virus
CDK8
HIV
Herpesvirus
PML
Adenovirus
Herpesvirus
RB1
Adenovirus
HIV
Herpesvirus
Papillomavirus
Sarcoma virus
E2F1
HIV
Papillomavirus
MYC
HIV
Papillomavirus
Sarcoma virus
E2F4
HIV
Papillomavirus
CDK2
HIV
Herpesvirus
Papillomavirus
Shigella
CCNE1
HIV
Papillomavirus
CCNE1
HIV
Papillomavirus
DLG1
Adenovirus
Papillomavirus
T-lymphotrophic virus
CDK8
HIV
Herpesvirus
CDK4
HIV
Herpesvirus
T-lymphotrophic virus
PPP2R1A
HIV
Papillomavirus
Sarcoma virus
KHDRBS1
HIV
Sarcoma virus
CDC2
Adenovirus
HIV
Papillomavirus
Sarcomavirus
KPNA2
Adenovirus
EBV
HIV
Papillomavirus
Sarcoma virus
CDK2
HIV
Herpesvirus
Papillomavirus

Distribution Across Pathogens

Group	#PPIs	#Strains	#Unique targeted human proteins	#proteins in human network
HIV	8,024	44	743	671
Hepatitis	1,244	16	109	93
Influenza	287	4	76	76
Papillomavirus	229	12	96	94
EBV	211	2	135	121
Adenovirus	80	9	60	59
Herpesvirus	64	20	54	54
Yersinia	57	3	56	45
Sarcoma virus	52	6	36	35
		...		
TOTAL	10,477	190	1,233	1,109

Distribution Across Pathogens

Group	#PPIs	#Strains	#Unique targeted human proteins	#proteins in human network
HIV	8,024	44	743	671
Hepatitis	1,244	16	109	93
Influenza	287	4	76	76
Papillomavirus	229	12	96	94
EBV	211	2	135	121
Adenovirus	80	9	60	59
Herpesvirus	64	20	54	54
Yersinia	57	3	56	45
Sarcoma virus	52	6	36	35
		...		
TOTAL	10,477	190	1,233	1,109

- Can we fill in the blanks by predicting new PPIs?

Outline

Approaches for Predicting Intra-organism PPIs

- Sequence signature pairs: Sprinzak et al., *J Mol Biol.* (2001) 311(4):681-692.
- Protein domain profiles: Kim et al., *Genome Inform Ser.* (2002) 13:42-50; Ng et al, *Bioinformatics*, (2003) 19(8):923-929.
- Sequence homology: Yu et al., *Genome Res.* (2004) 14:1107-1118.
- Bayesian networks: Jansen et al., Science. (2003) 302:449-453.
- Decision tree: Zhang et al., *BMC Bioinformatics*. (2004) 5(38).
- Random forests and SVM: Qi et al. *Proteins*. (2006) 63(3):490-500.

Features Used in these Approaches

Feature	References
Protein Domains	Sprinzak *et al.* (2001), Ng *et al.* (2002), Deng *et al.* (2002), Kim *et al.* (2002)
Homology	Sprinzak *et al.* (2003), Haiyuan *et al.* (2004)
Gene Expression	Deng *et al.* (2002), Bar-Joseph *et al.* (2003), Jensen *et al.* (2003), Zhang *et al.* (2004)
Protein Expression	Ghaemmaghami *et al.* (2003)
Yeast Two-Hybrid	Utz *et al.* (2000), Ito *et al.* (2001), Zhang *et al.* (2004)
Synthetic Lethal	von Mering *et al.* (2002), Tong *et al.* (2004), Wong *et al.* (2004)
Tandem Affinity Purification	Gavin *et al.* (2002), Badger *et al.* (2003), Zhang *et al.* (2004)
Transcription Factor	Zhang *et al.* (2004)
Knockout Phenotype	Zhang *et al.* (2004)
Phylogenetic Analysis	von Mering *et al.* (2002), Zhang *et al.* (2004)

Challenges in Predicting HP PPIs

- Known (gold-standard) datasets of known PPIs are available for a very small number of host-pathogen systems.
- A number of datatypes used to train the previously-mentioned methods such as gene expression and knockout phenotype are not available for host-pathogen systems.
- For applying supervised methods, we need data such as simultaneous measurements of gene expression in both host and pathogen upon infection.

Unsupervised Methods

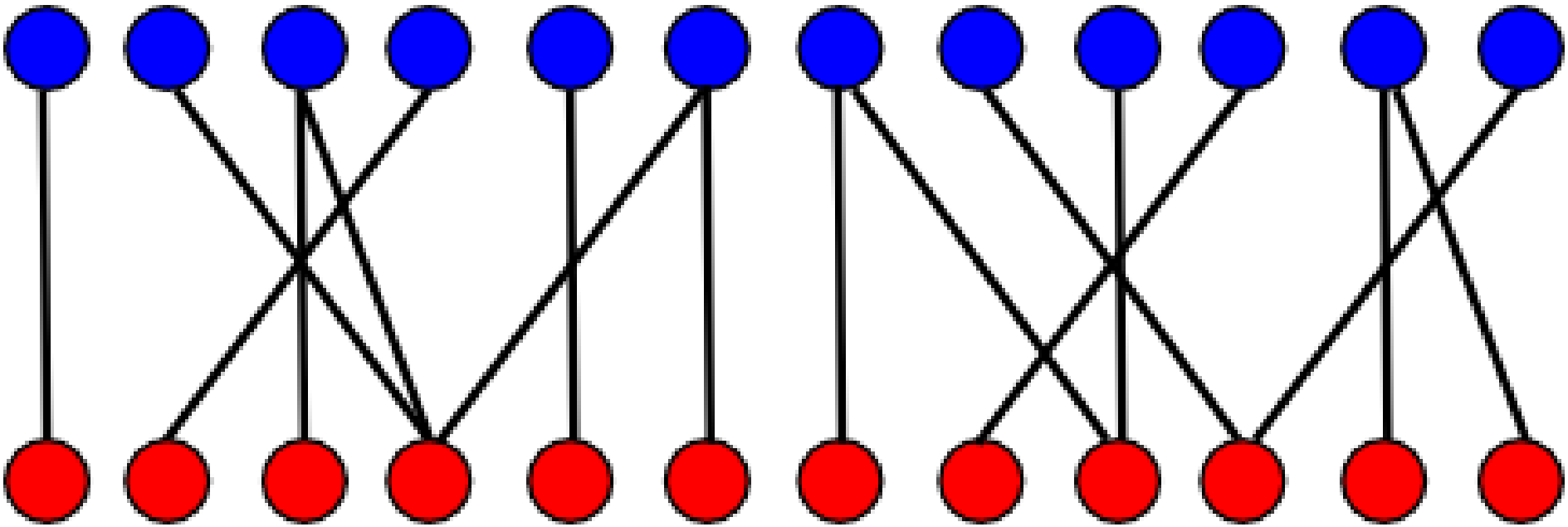
- Host pathogen protein interactions predicted by comparative modeling, Davis et al., *Protein Science*, 16: 2585-2596, 2007.
- Computational Prediction of Host-Pathogen Protein-Protein Interactions, Dyer, Murali, and Sobral, *Bioinformatics* volume 23, number 13, pp. i159-i166, 2007.
  - Integrate known intra-species PPIs datasets with protein-domain profiles to predict and study host-pathogen PPI networks.
  - Compute statistics of how often proteins containing specific domain pairs interact and use these statistics to make predictions.
  - Evaluate the validity of the predictions using three computational tests: Triplet proximity, triplet coexpression, and weighted functional enrichment.

Notation

- *PPI*: the association or physical interaction of two or more proteins (direct interaction or membership in the same complex).

Notation

- *PPI*: the association or physical interaction of two or more proteins (direct interaction or membership in the same complex).
- *Bipartite graph* $B = (V_1, V_2, E)$ represents pairwise relationships (edges E) between nodes in V_1 and nodes in V_2.



Approach

- $I(g, h)$: the event that proteins g and h interact.
- $D(g, d)$: the event that protein g contains domain d.

Approach

- $I(g, h)$: the event that proteins g and h interact.
- $D(g, d)$: the event that protein g contains domain d.
- We want to compute $\Pr(I(g, h) \mid D(g, d), D(h, e))$: the probability that proteins g and h interact given that g contains domain d and h contains e.

Approach

- $I(g, h)$: the event that proteins g and h interact.
- $D(g, d)$: the event that protein g contains domain d.
- We want to compute $\Pr(I(g, h) \mid D(g, d), D(h, e))$: the probability that proteins g and h interact given that g contains domain d and h contains e.
- What we can compute is $\Pr(D(g, d), D(h, e) \mid I(g, h))$: the probability that g contains domain d and h contains e given that proteins g and h interact.

Approach

- $I(g, h)$: the event that proteins g and h interact.
- $D(g, d)$: the event that protein g contains domain d.
- We want to compute $\Pr(I(g, h) \mid D(g, d), D(h, e))$: the probability that proteins g and h interact given that g contains domain d and h contains e.
- What we can compute is $\Pr(D(g, d), D(h, e) \mid I(g, h))$: the probability that g contains domain d and h contains e given that proteins g and h interact.
  - We can estimate these probabilities from input data.
- Use Bayes rule to "flip" $\Pr(D(g, d), D(h, e) \mid I(g, h))$ into $\Pr(I(g, h) \mid D(g, d), D(h, e))$:

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

Approach Continued

- P: set of (host or pathogen) proteins with at least one known domain and one known intra-species PPI.
- P_d: set of proteins containing domain d.
- S: set of intra-species interactions between proteins in P.
- $S_{d,e}$: set of interactions where one protein contains domain d and the other contains domain e.

Approach Continued

- P: set of (host or pathogen) proteins with at least one known domain and one known intra-species PPI.
- P_d: set of proteins containing domain d.
- S: set of intra-species interactions between proteins in P.
- $S_{d,e}$: set of interactions where one protein contains domain d and the other contains domain e.

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

Approach Continued

- P: set of (host or pathogen) proteins with at least one known domain and one known intra-species PPI.
- P_d: set of proteins containing domain d.
- S: set of intra-species interactions between proteins in P.
- $S_{d,e}$: set of interactions where one protein contains domain d and the other contains domain e.

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

- Compute $\Pr(g,h)$ by integrating probabilities from all pair of domains contained in g and h.

Evaluation

- We can use cross validation to evaluate supervised algorithms.

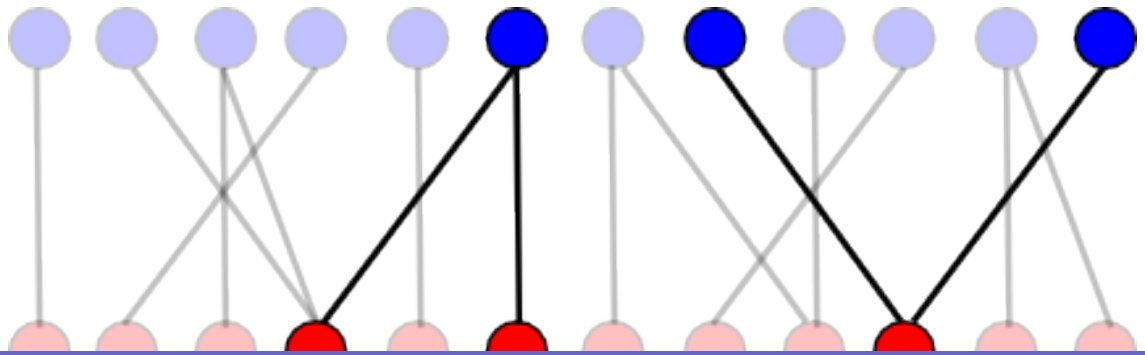
Evaluation

- We can use cross validation to evaluate supervised algorithms.
- We developed three tests to evaluate our approach:

Evaluation

- We can use cross validation to evaluate supervised algorithms.
- We developed three tests to evaluate our approach:

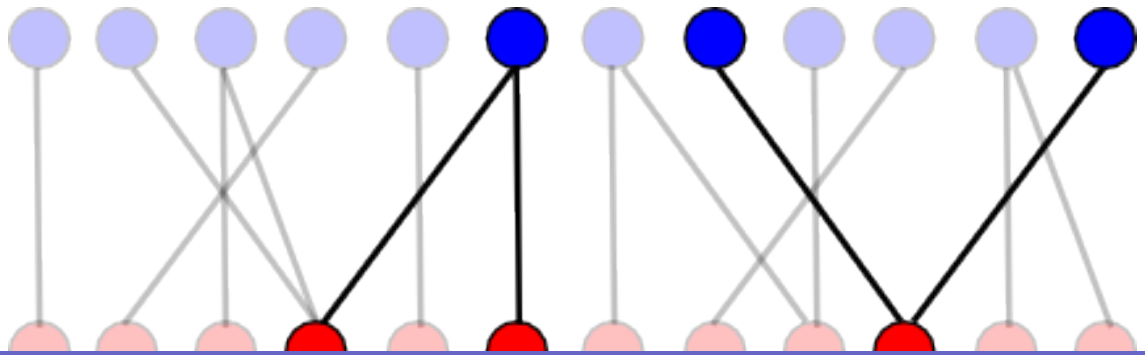
- *HHP Triplet*: two host proteins that interact with the same pathogen protein.
- *HPP Triplet*: two pathogen proteins that interact with the same host protein.



Evaluation

- We can use cross validation to evaluate supervised algorithms.
- We developed three tests to evaluate our approach:
  1. Triplet proximity: In an HHP triplet, how close to each other are the human proteins in the human PPI network?

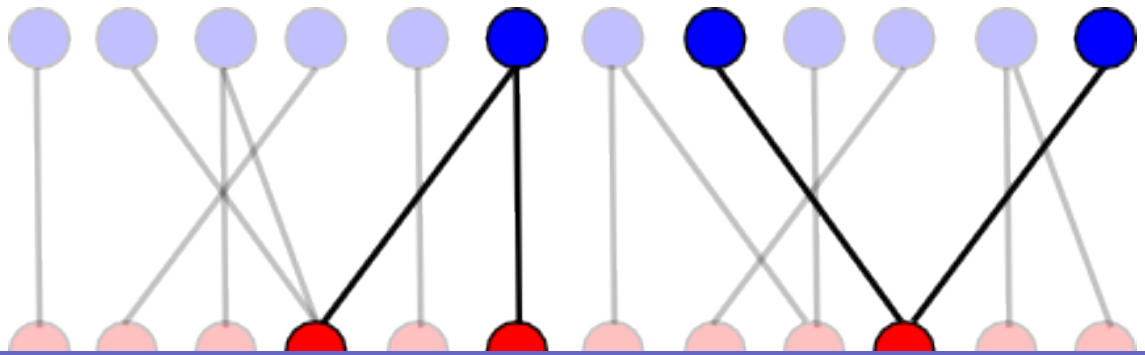
- *HHP Triplet*: two host proteins that interact with the same pathogen protein.
- *HPP Triplet*: two pathogen proteins that interact with the same host protein.



Evaluation

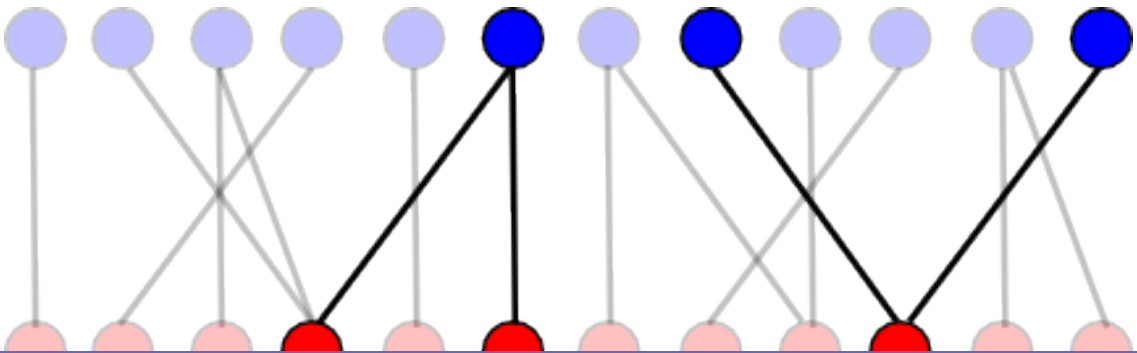
- We can use cross validation to evaluate supervised algorithms.
- We developed three tests to evaluate our approach:
  1. Triplet proximity: In an HHP triplet, how close to each other are the human proteins in the human PPI network?
  2. Triplet co-expression: In an HHP triplet, how co-expressed are the human proteins upon infection by the pathogen.

- *HHP Triplet*: two host proteins that interact with the same pathogen protein.
- *HPP Triplet*: two pathogen proteins that interact with the same host protein.



Evaluation

- We can use cross validation to evaluate supervised algorithms.
- We developed three tests to evaluate our approach:
  1. Triplet proximity: In an HHP triplet, how close to each other are the human proteins in the human PPI network?
  2. Triplet co-expression: In an HHP triplet, how co-expressed are the human proteins upon infection by the pathogen.
  3. Functional enrichment: Which GO functions are enriched in the predicted network?
- *HHP Triplet*: two host proteins that interact with the same pathogen protein.
- *HPP Triplet*: two pathogen proteins that interact with the same host protein.



Datasets Used For Prediction

- Applied our approach to predict interactions between human and *P. falciparum* proteins.
- Prediction:
  - Protein sequence information: Uniprot (Bairoch et al. 2005).
  - Protein domain profiles: InterProScan (Quevillon et al. 2005).
  - PPIs: same databases as before.

Datasets Used For Prediction

- Applied our approach to predict interactions between human and *P. falciparum* proteins.
- Prediction:
  - Protein sequence information: Uniprot (Bairoch et al. 2005).
  - Protein domain profiles: InterProScan (Quevillon et al. 2005).
  - PPIs: same databases as before.
- Trained system using only proteins that participate in at least one interaction and domains that are present in at least four such proteins.

Datasets Used For Prediction

- Applied our approach to predict interactions between human and *P. falciparum* proteins.
- Prediction:
  - Protein sequence information: Uniprot (Bairoch et al. 2005).
  - Protein domain profiles: InterProScan (Quevillon et al. 2005).
  - PPIs: same databases as before.
- Trained system using only proteins that participate in at least one interaction and domains that are present in at least four such proteins.
- Focus our predictions on proteins most likely involved in pathogenesis.
  - Discarded proteins annotated with the following functions: Nucleus, Proteolysis, Ribosome, Nucleic acid binding, Helicase activity,.
  - Ensured that we retained proteins with certain functions: Blood coagulation, Hemoglobin metabolism, Cell-cell communication, Cell death.

Datasets Used For Evaluation

- Functional annotations: Gene Ontology (Ashburner et al. 2000).
- 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.
  - *H. sapiens*:
    - Boldt et al. (Unpublished): Healthy, un/complicated symptoms, 15 samples.
    - Ockenhouse et al. (2006): Experimentally and naturally infected, 71 samples.

Predictions

- Training set included 7,876 human PPIs and 214 *Plasmodium* PPIs.
- Prediction set contained 39,107 human proteins and 1,502 *Plasmodium* proteins.

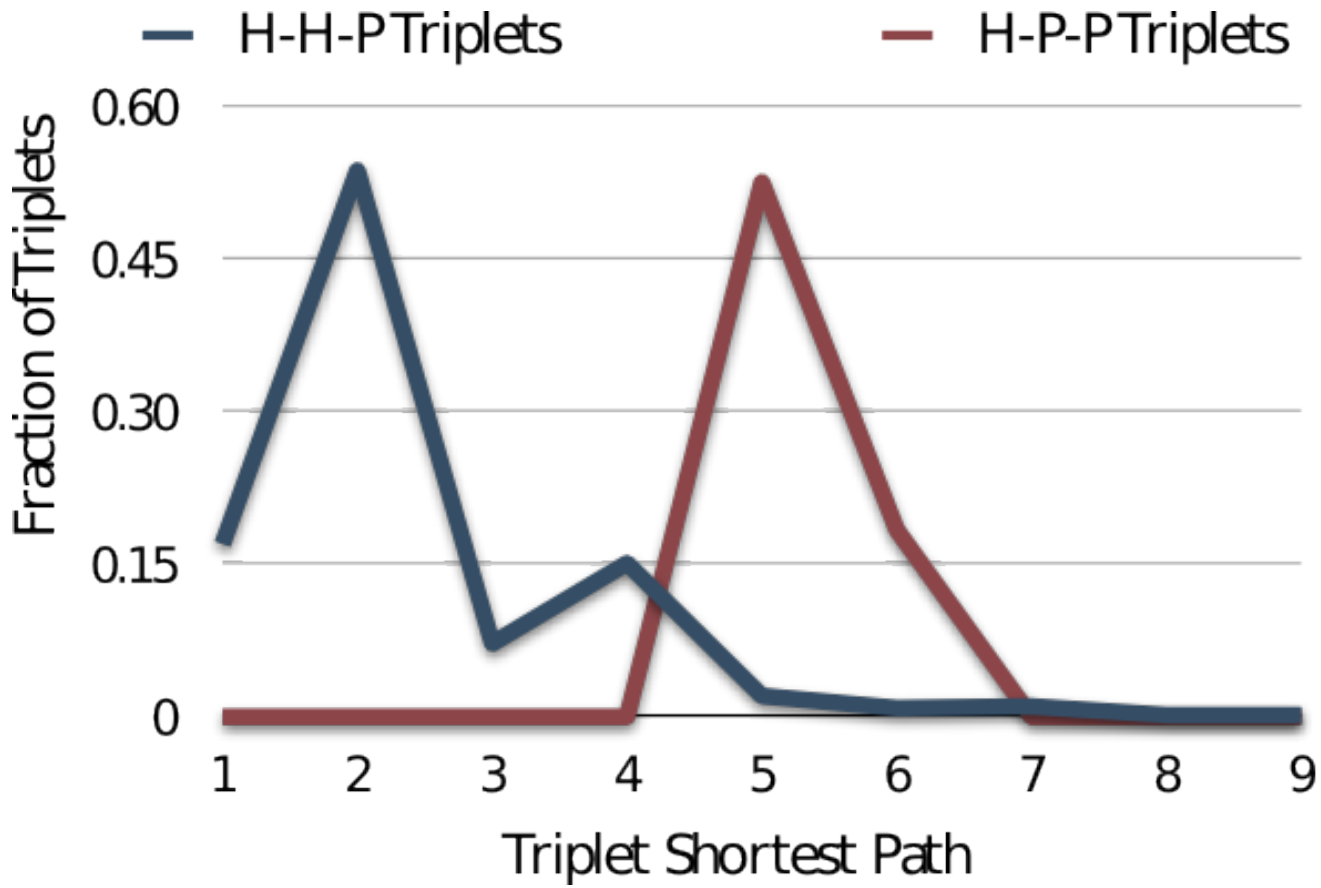
Predictions

- Training set included 7,876 human PPIs and 214 *Plasmodium* PPIs.
- Prediction set contained 39,107 human proteins and 1,502 *Plasmodium* proteins.
- We predicted 516 PPIs between 158 human proteins and 30 *Plasmodium* proteins.

PPI Probability	# human-*Plasmodium* PPIs	# fly-*Plasmodium* PPIs
0.50 – 0.55	185	6
0.55 – 0.60	175	15
0.60 – 0.65	31	11
0.65 – 0.70	61	12
0.70 – 0.75	16	0
0.75 – 0.80	48	0
Total	516	44

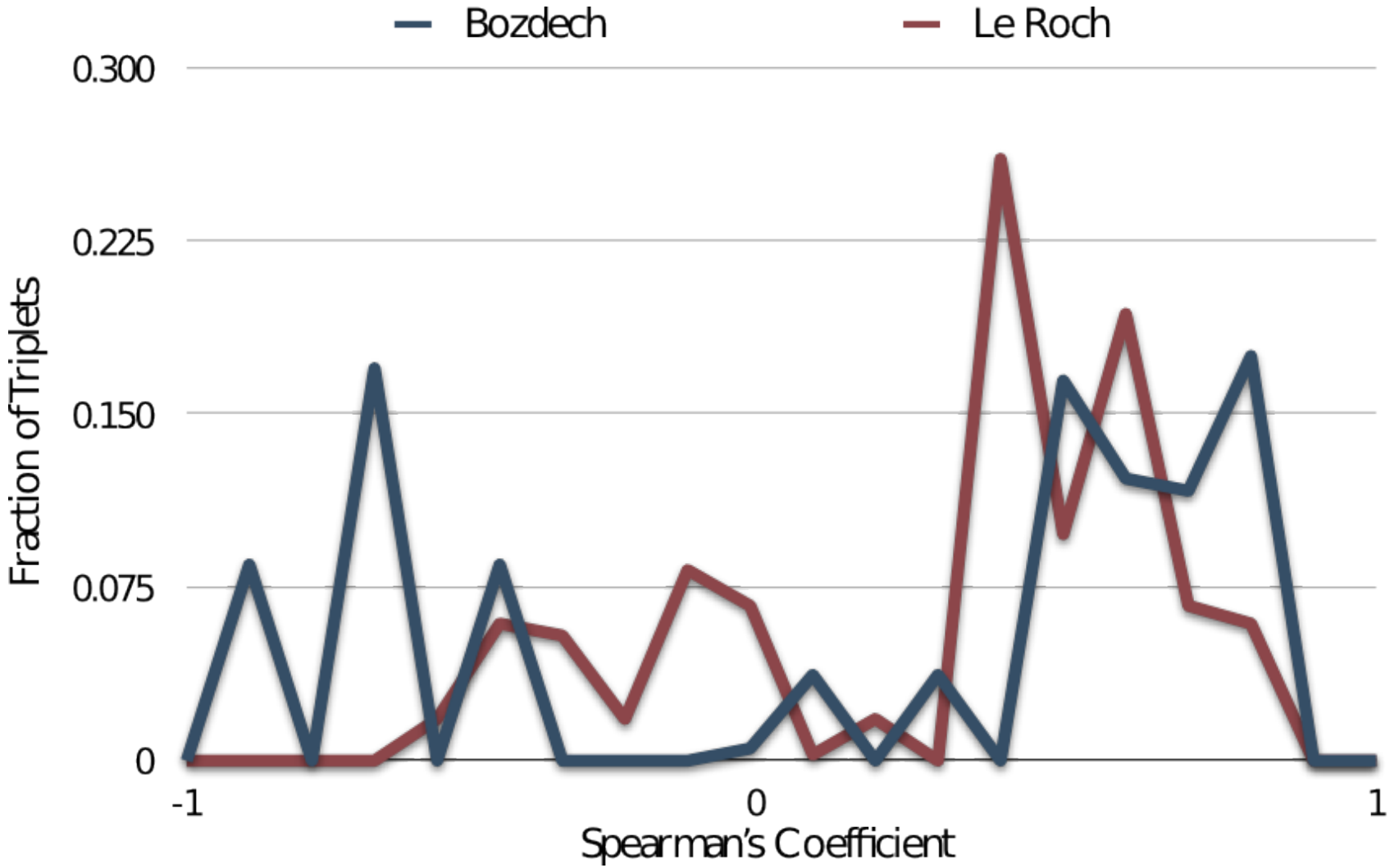
Triplet Proximity

- 72% of HHP triplets are at a distance of two or less in the human PPI network.
- The average distance between Plasmodium proteins (HPP) is 5.5.



Triplet Co-expression

- *Plasmodium* proteins in HPP triplets appear to be co-expressed.



Functional Enrichment

- Our method tests for the enrichment of pairs of functions.
- We retain only functions that are significant at the 0.05 level.
- 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 including Apoptosis (GO:0006915), Regulation of apoptosis (GO:0042981), Inflammatory response (GO:0006512), and Immune response (GO:0006955).

Examples of Predictions

- We find an enriched subnetwork between human proteins annotated with "blood coagulation" and *Plasmodium* proteins annotated with "integral to membrane"
- The network includes Plasmodium VAR, a known PfEMP1, which we predict to interact with human plasminogen and with with hepatocyte growth factors (HGFs).
  - An important step in merozoite release from the human RBC is the activation of plasminogen (Roggwiller et al. 1997).
  - HGF induction is required for hepatocyte invasion (Carrolo et al. 2003).

Summary

- What do we know?
  - We have discussed methods for analysing known human-pathogen PPI networks.
  - Comparative analysis might prove to be a powerful method for understanding pathogen infection in general and obtain clues for broad-spectrum medicines.

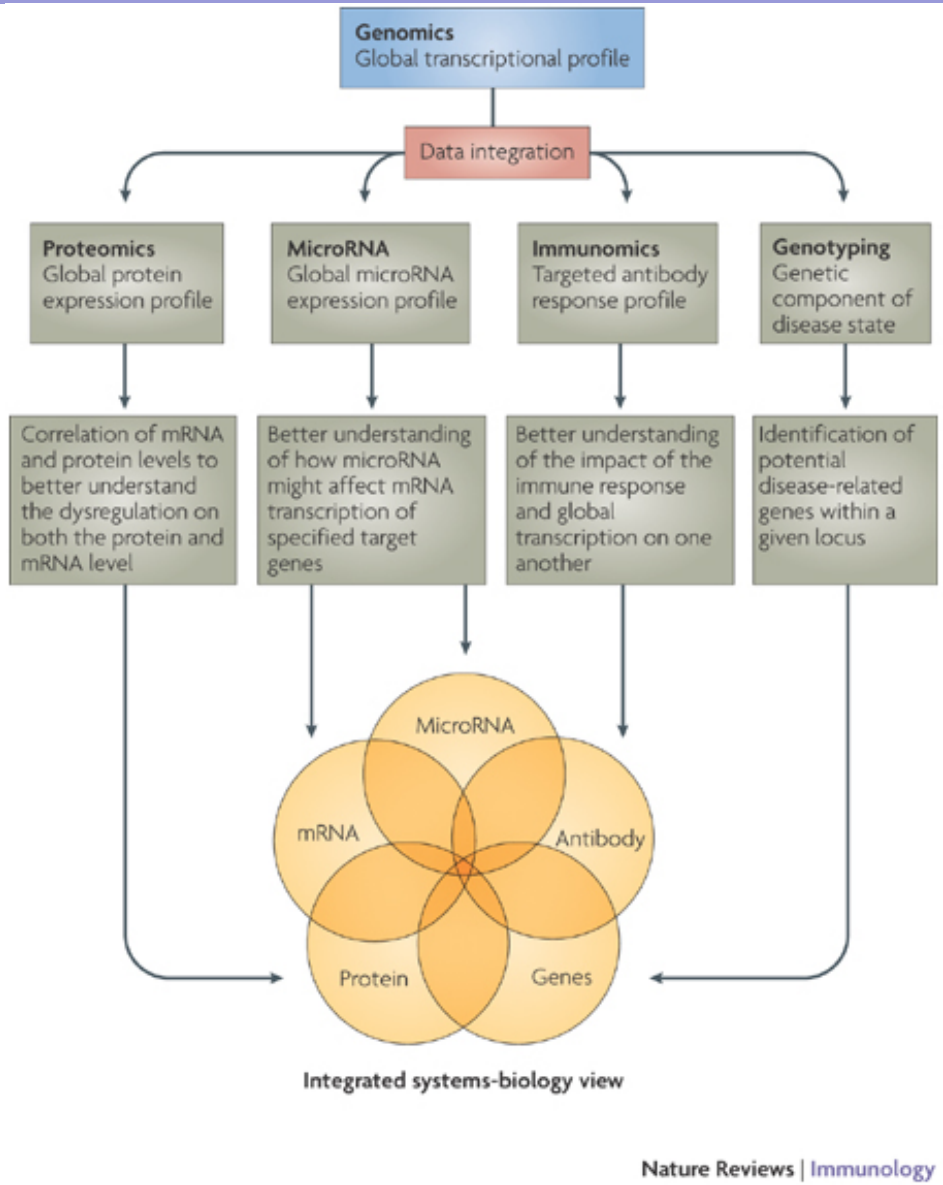
Summary

- What do we know?
  - We have discussed methods for analysing known human-pathogen PPI networks.
  - Comparative analysis might prove to be a powerful method for understanding pathogen infection in general and obtain clues for broad-spectrum medicines.
- What can we fill in?
  - We discussed unsupervised methods to predict host-pathogen PPIs.
  - Data for applying supervised methods is available for HIV and around the corner for some bacteria (work in progress).

Summary

- What do we know?
  - We have discussed methods for analysing known human-pathogen PPI networks.
  - Comparative analysis might prove to be a powerful method for understanding pathogen infection in general and obtain clues for broad-spectrum medicines.
- What can we fill in?
  - We discussed unsupervised methods to predict host-pathogen PPIs.
  - Data for applying supervised methods is available for HIV and around the corner for some bacteria (work in progress).
- What about other types of data?
  - Many compendia of host-response to gene expression.
  - siRNA analysis to identify "dependency factors" for HIV, influenza virus, and West Nile virus.

Systems Biology of Infectious Diseases



Innate immune modulation by RNA viruses: emerging insights from

functional genomics, Katze et al., Nature Reviews Immunology 8, 644-654 (August 2008) — doi:10.1038/nri2377

Acknowledgments

- Matthew D. Dyer and Bruno Sobral, Virginia Tech
- Simon Kasif and Esther Rheinbay, Boston University
- Donna Shattuck, Chris Neff, and Max Dufford, Myriad Genetics