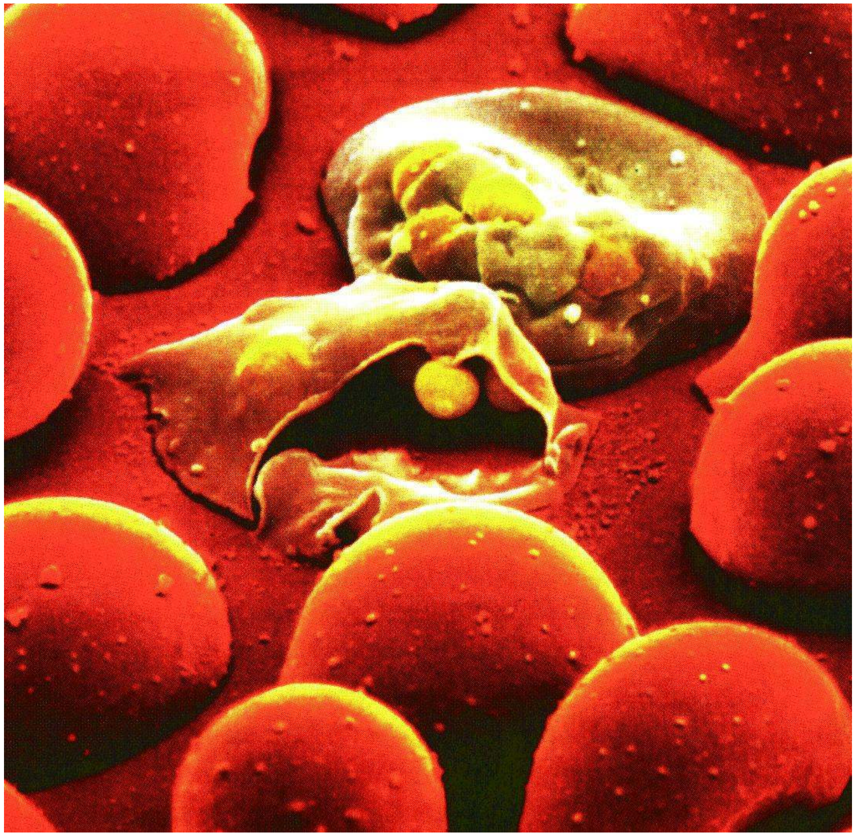
Distant homology, networks, and *Plasmodium falciparum*

Alejandro Ochoa
Singh lab
Colloquium 2007-10-25

Plasmodium falciparum

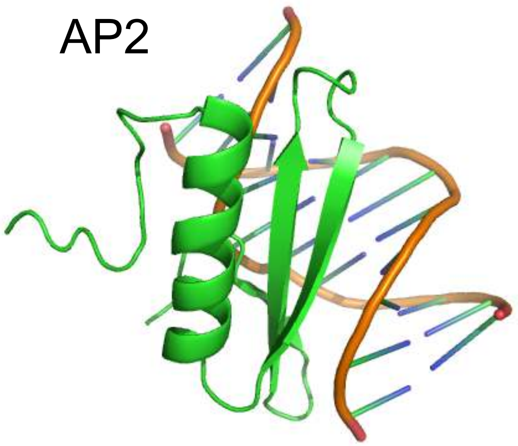
- Malaria:
  - Deadly on children
  - One of oldest human diseases
  - No effective cure or treatment
- Eukaryote
- 5.5k proteins, 43% unknown function
  - 36% unknown in yeast
- Some plant-like genes (AP2)

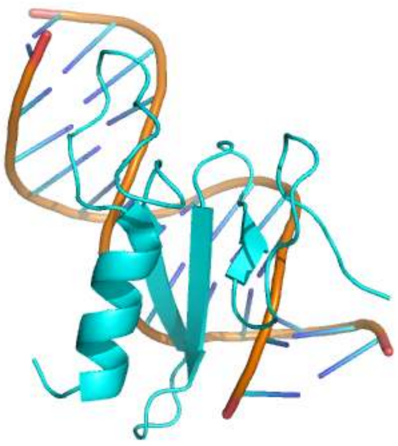


Contrast: structural biology

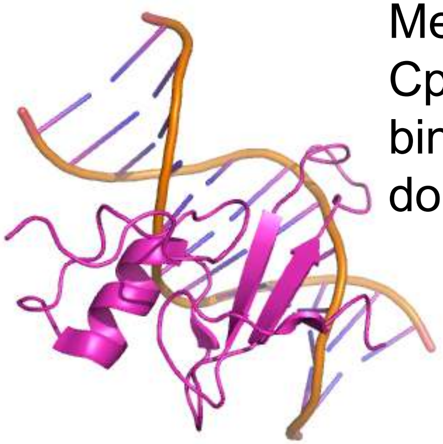
- Few structures in nature
- Reuse of domains rather than reinvention
- Function is conserved beyond sequence homology

AP2





Tn916 Integrase

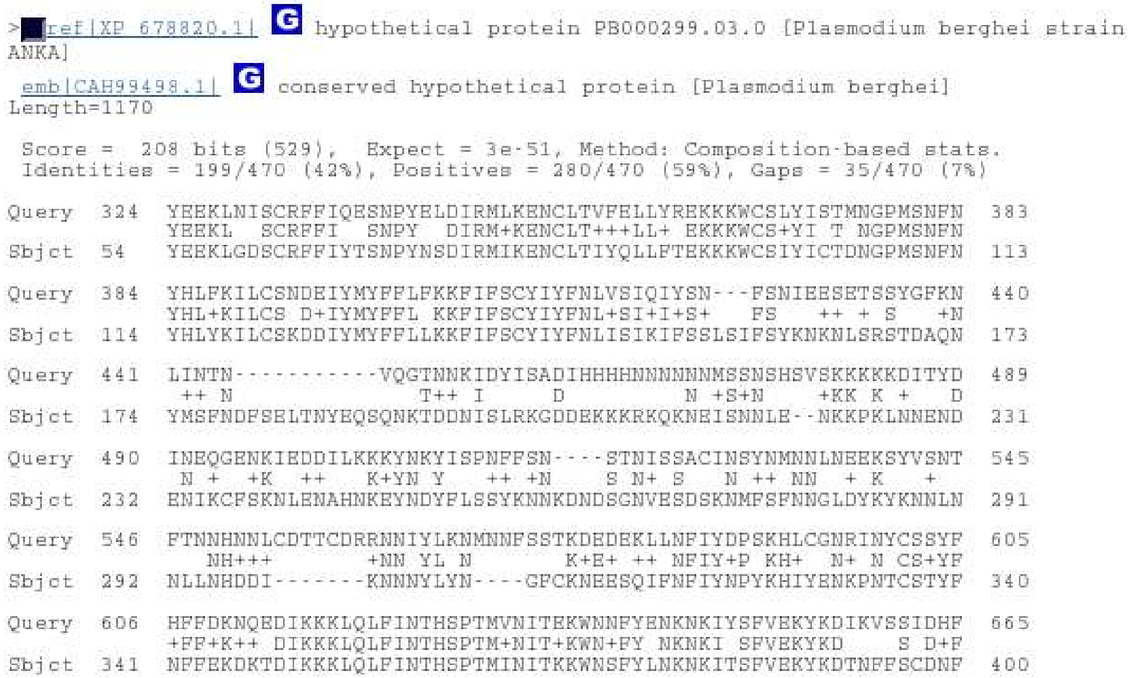


Methyl-CpG binding domain

P. falciparum has to have more homology, we just need to try harder finding it

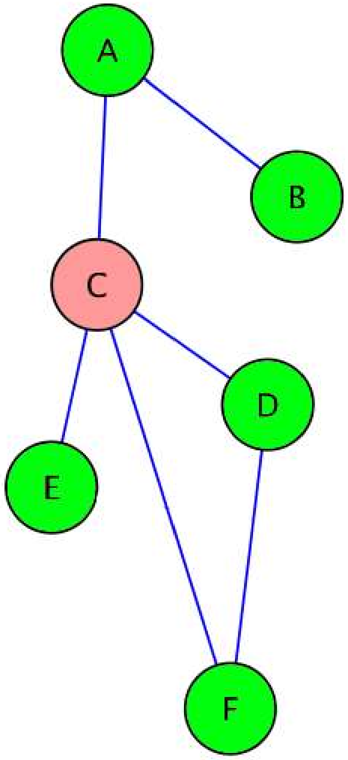
Common computational annotation tools

- BLAST
  - Sequence based
  - No other data required



```
>ref|XP_678820.1|  hypothetical protein PB000299.03.0 [Plasmodium berghei strain
ANKA]
 emb|CAH99498.1|  conserved hypothetical protein [Plasmodium berghei]
Length=1170

 Score =  208 bits (529),  Expect = 3e-51, Method: Composition-based stats.
 Identities = 199/470 (42%), Positives = 280/470 (59%), Gaps = 35/470 (7%)

Query  324  YEEKLNISCRFFIQESNPYELDIRMLKENCLTVFELLYREKKKWCSLYISTMNGPMSNFN  383
            YEEKL  SCRFFI  SNPY  DIRM+KENCLT+++LL+ EKKKWCS+YI T NGPMSNFN
Sbjct  54   YEEKLGDSCRFFIYTSNPYNSDIRMIKENCLTIYQLLFTEKKKWCSIYICTDNGPMSNFN  113

Query  384  YHLFKILCSNDEIYMYFFLFKKFIFSCYIYFNLVSIQIYSN---FSNIEESETSSYGFKN  440
            YHL+KILCS D+IYMYFFL KKFIFSCYIYFNL+SI+I+S+   FS   ++ + S   +N
Sbjct  114  YHLYKILCSKDDIYMYFFLLKKFIFSCYIYFNLISIKIFSSLSIFSYKNKNLSRSTDAQN  173

Query  441  LINTN-----------VQGTNNKIDYISADIHHHHNNNNNNMSSNSHSVSKKKKKDITYD  489
             ++ N              T++ I     D         N +S+N    +KK K +   D
Sbjct  174  YMSFNDFSELTNYEQSQNKTDDNISLRKGDDEKKKRKQKNEISNNLE--NKKPKLNNEND  231

Query  490  INEQGENKIEDDILKKKYNKYISPNFFSN----STNISSACINSYNMNNLNEEKSYVSNT  545
             N +  +K  ++   K+YN Y   ++ +N    S N+ S   N ++ NN  + K   +
Sbjct  232  ENIKCFSKNLENAHNKEYNDYFLSSYKNNKDNDSGNVESDSKNMFSFNNGLDYKYKNNLN  291

Query  546  FTNNHNNLCDTTCDRRNNIYLKNMNNFSSTKDEDEKLLNFIYDPSKHLCGNRINYCSSYF  605
               NH+++       +NN YL N       K+E+ ++ NFIY+P KH+  N+ N CS+YF
Sbjct  292  NLLNHDDI-------KNNNYLYN----GFCKNEESQIFNFIYNPYKHIYENKPNTCSTYF  340

Query  606  HFFDKNQEDIKKKLQLFINTHSPTMVNITEKWNNFYENKNKIYSFVEKYKDIKVSSIDHF  665
            +FF+K++ DIKKKLQLFINTHSPTM+NIT+KWN+FY NKNKI SFVEKYKD    S D+F
Sbjct  341  NFFEKDKTDIKKKLQLFINTHSPTMINITKKWNSFYLNKNKITSFVEKYKDTNFFSCDNF  400
```

- Guilt by association
  - Network based
  - Requires annotations



PFAM

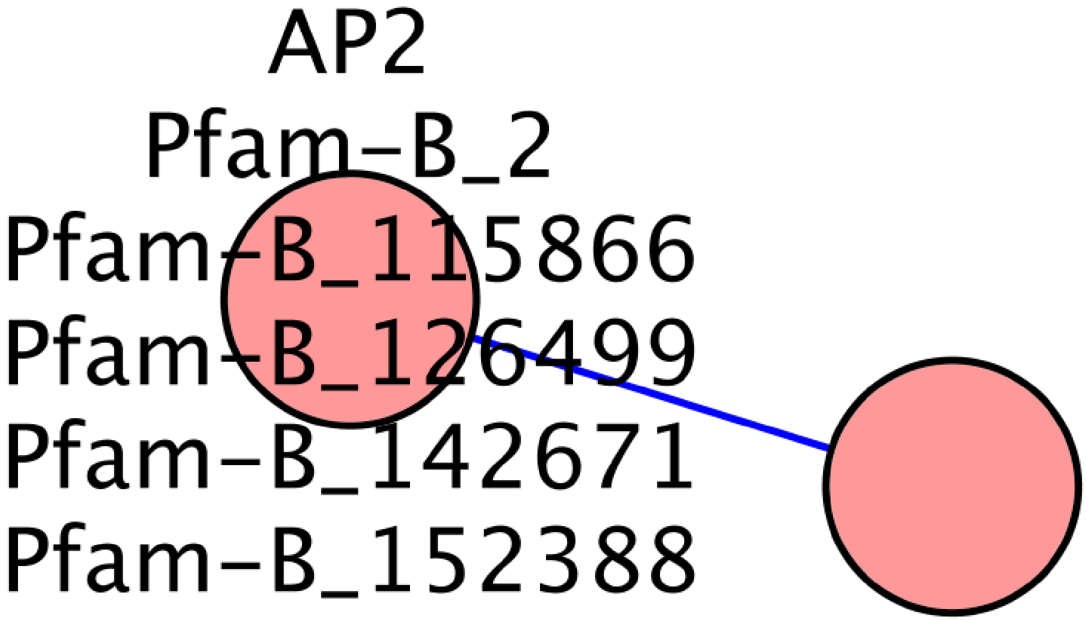
- Protein families and domains
- Collection of Hidden Markov Models (HMM)
  - Alignments -> fancy profiles/PWM
- PFAM-A
  - Curated alignments, 9318 domains, 74% protein coverage of Uniprot
- PFAM-B
  - Automatic PSI-BLAST alignments, 182k domains, 13% additional coverage
- 52% coverage in *P. falciparum*

Three computational approaches

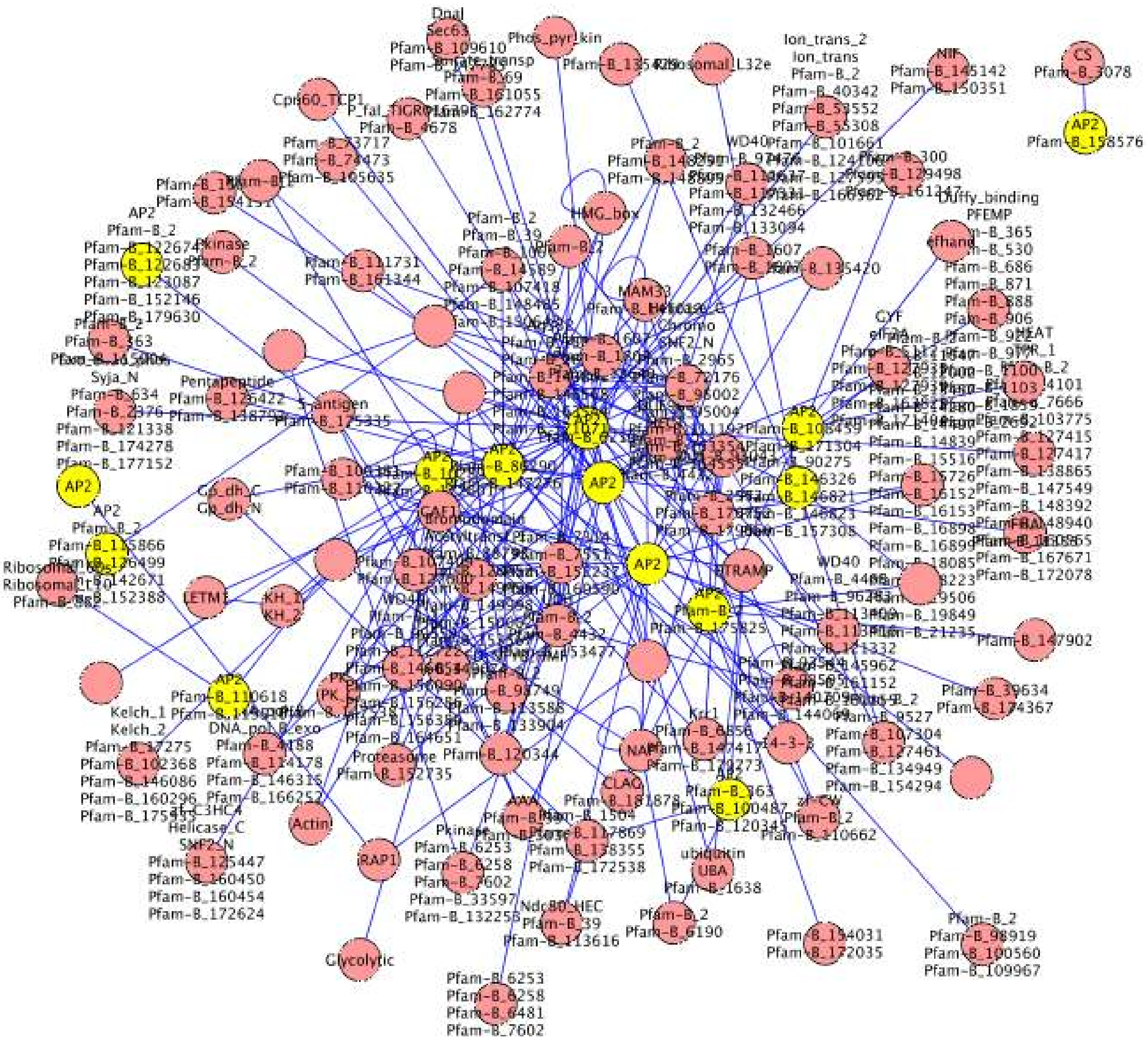
- Domain pairs on a protein network
  - With results!
- Homology through profiles, domain architecture
- Network alignments on weak signal

Domain pairs on a protein network

- Your favorite protein and its first shell
- Domain annotations

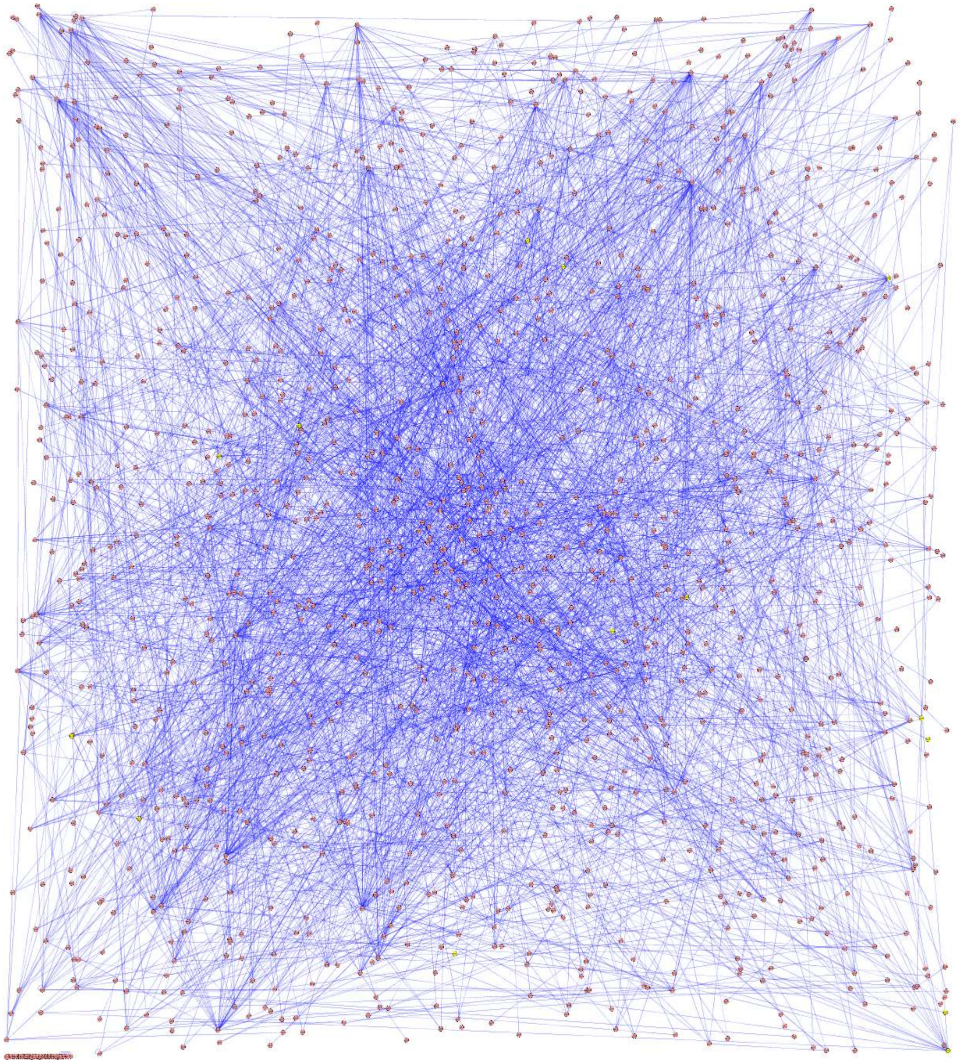


Domain pairs on a protein network



Domain pairs on a protein network

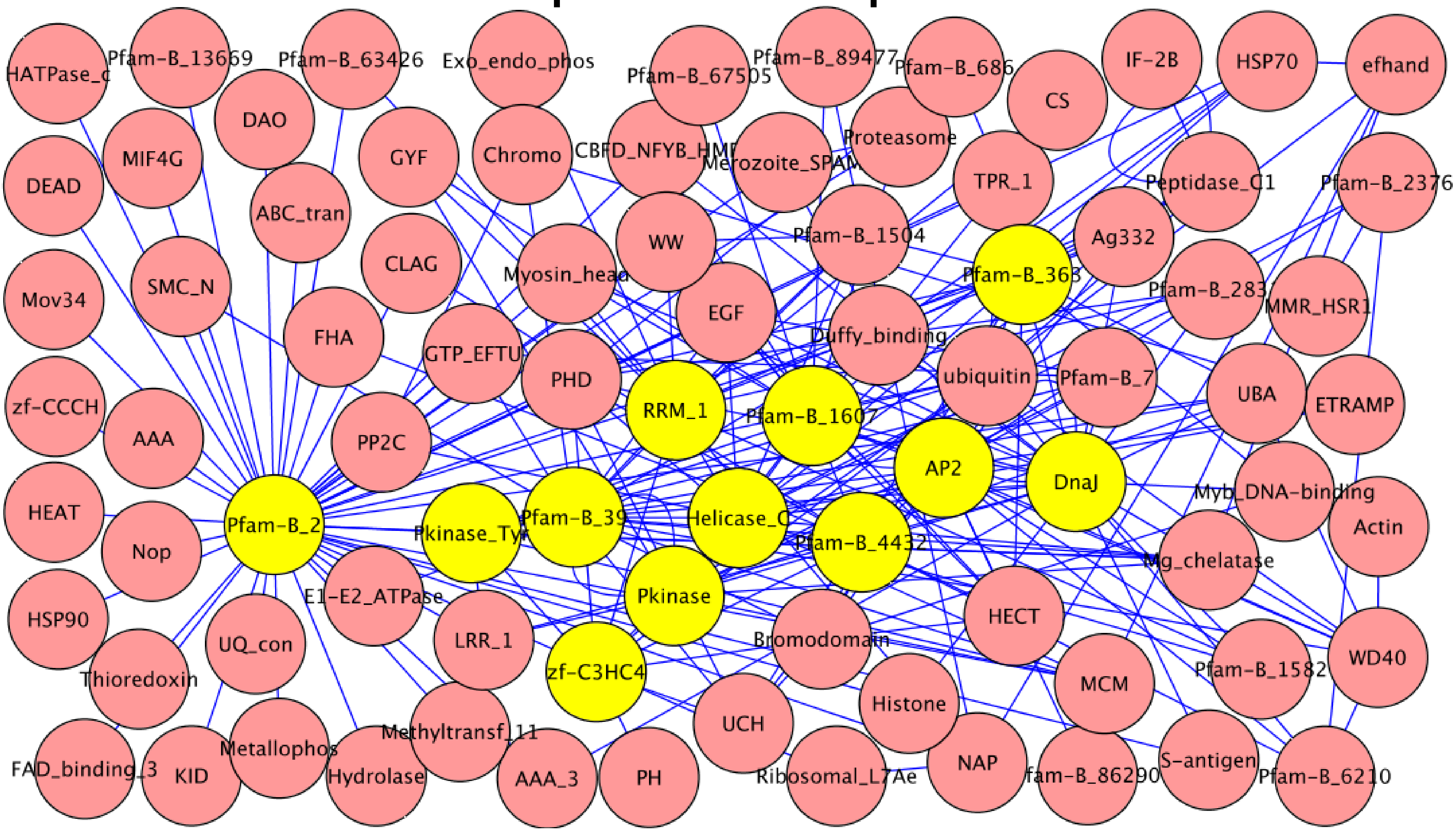
- For each domain pair, count* occurrences in net
- Count* in 1000 randomized* nets
- Estimate false discovery rate
- Get overrepresented pairs



LaCount, et al. Nature 2005

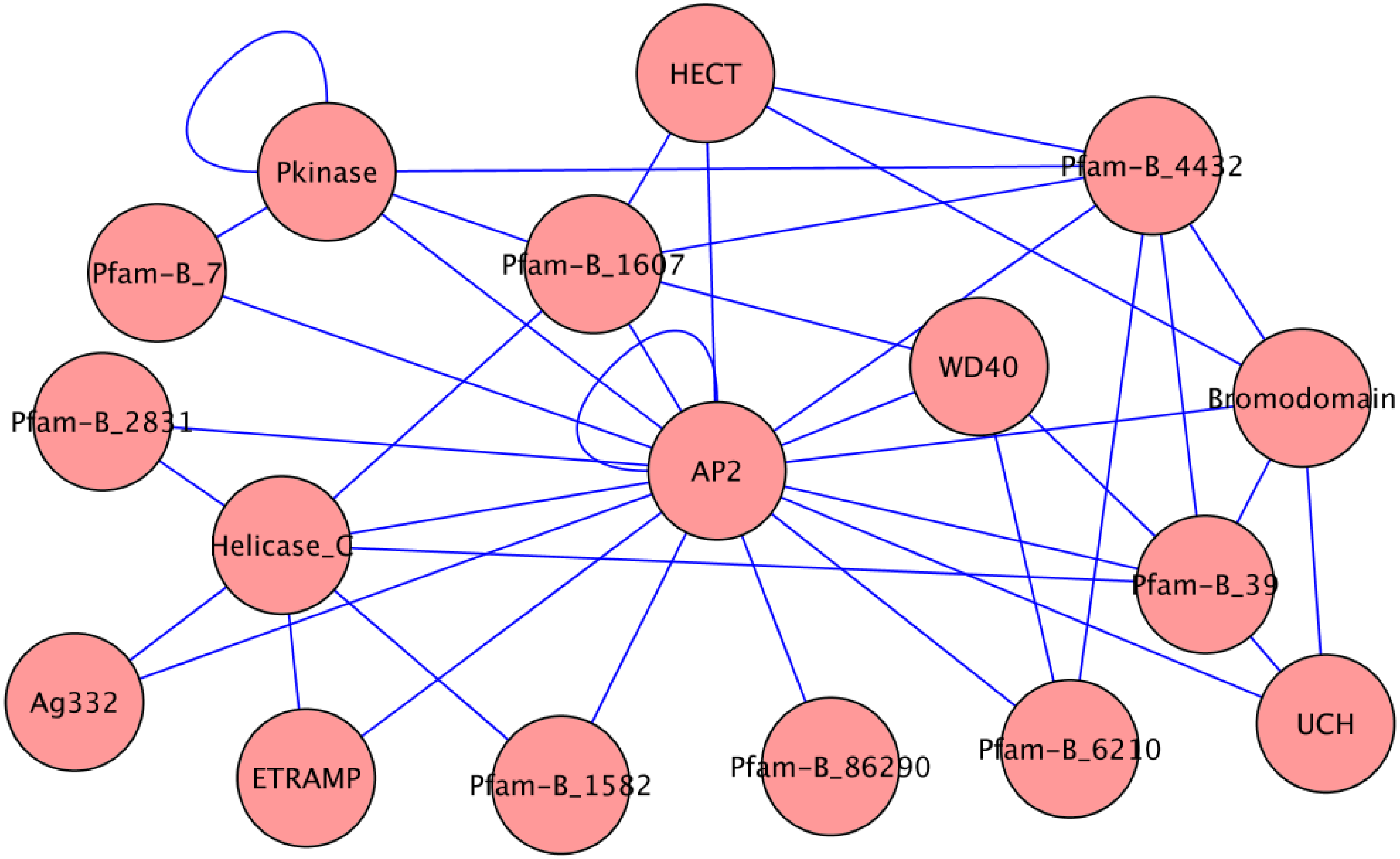
Domain pairs on a protein network

- Results: overrepresented pairs



Domain pairs on a protein network

- Results: AP2 pair subnetwork



Three computational approaches

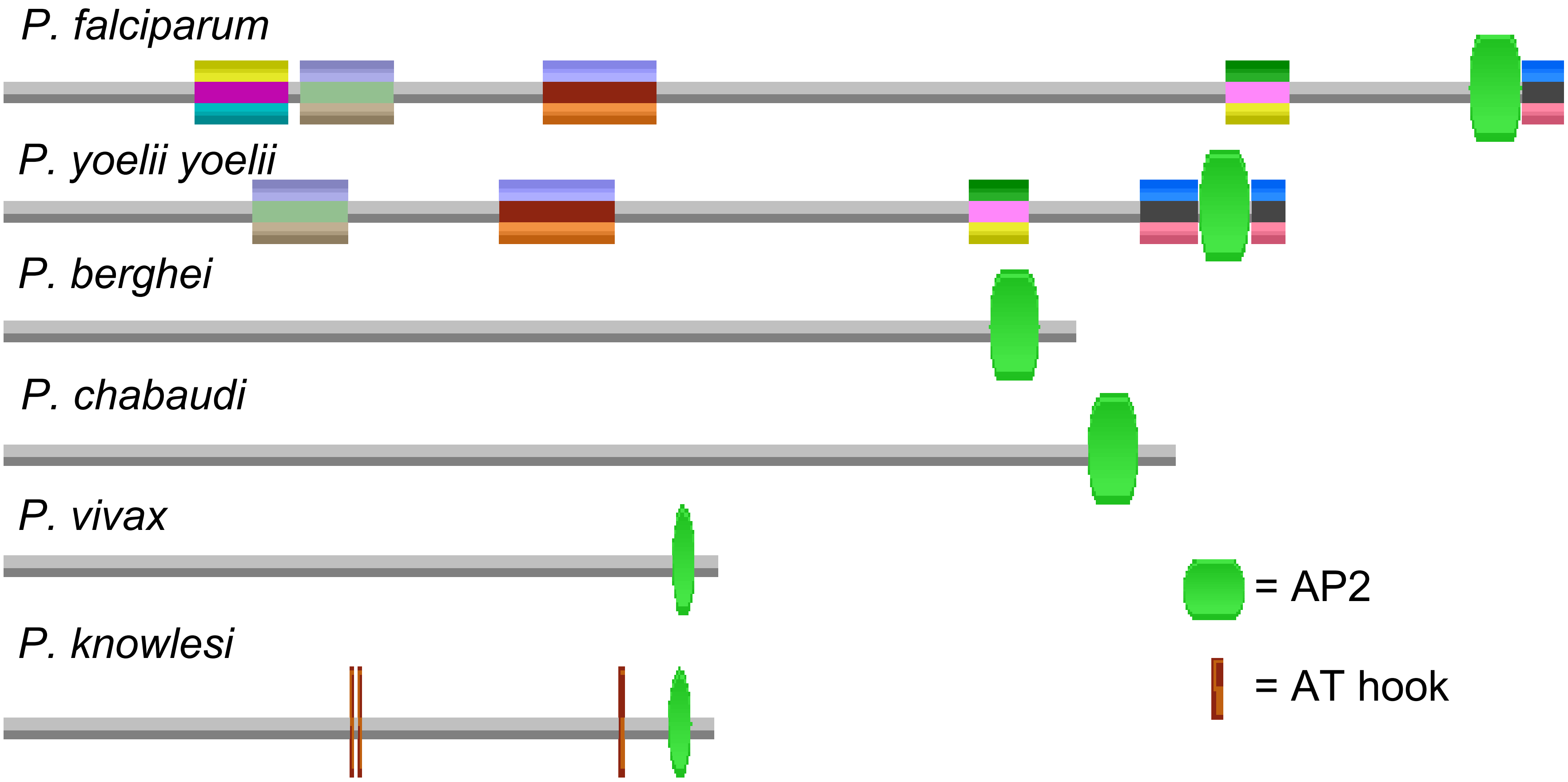
- Domain pairs on a protein network
  - With results!
- Homology through profiles, domain architecture
- Network alignments on weak signal

Homology through profiles, domain architecture

- Get a *P. falciparum* gene: PF14_0079
- Fish out orthologs in other Plasmodia:
  - – *P. vivax*: human
  - – *P. knowlesi*: macaque, human
  - – *P. yoelii yoelii*: mouse
  - – *P. berghei*: thicket rat
  - – *P. chabaudi*: shiny thicket rat

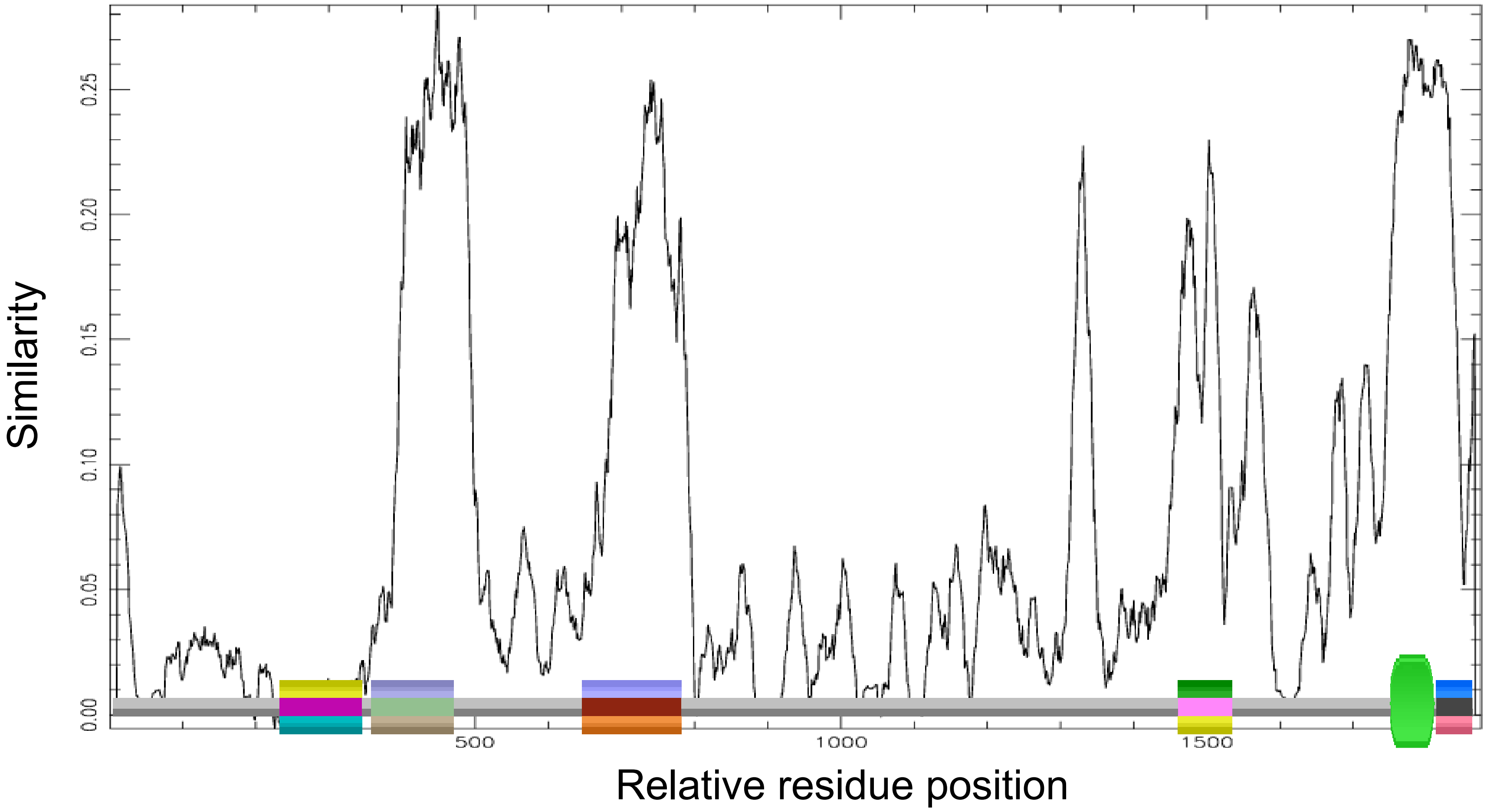
Homology through profiles, domain architecture

- PFAM domain architecture of PF14_0079 and orthologs



Homology through profiles, domain architecture

- Sequence profile of PF14_0079 and orthologs



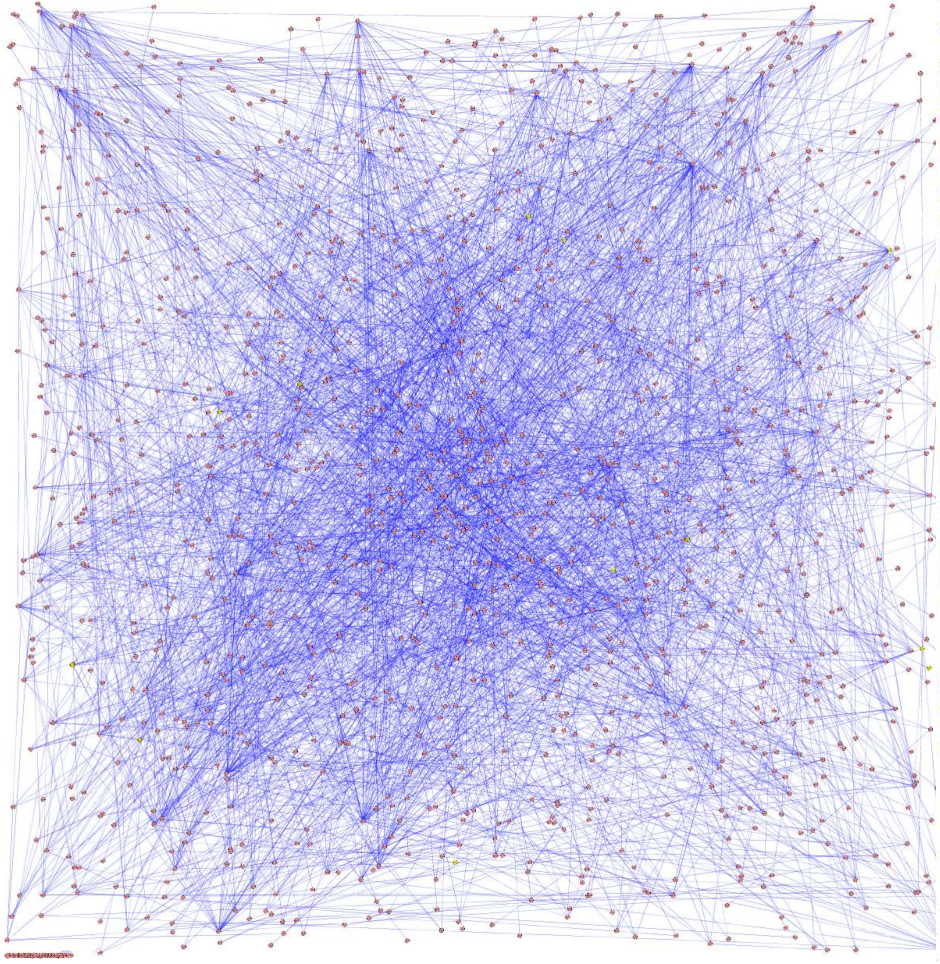
Three computational approaches

- Domain pairs on a protein network
  - With results!
- Homology through profiles, domain architecture
- Network alignments on weak signal

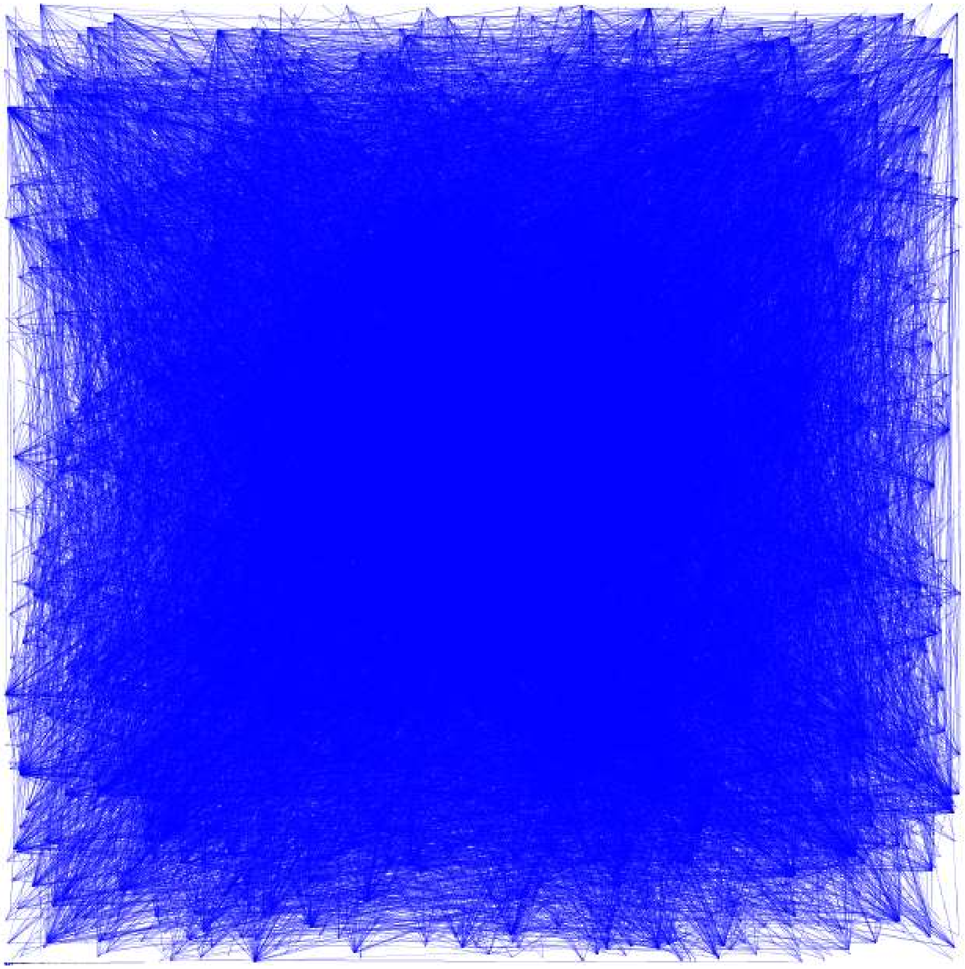
Network alignments on weak signal

- *P. falciparum* Y2H
- *A. thaliana* curated and predicted nets

LaCount, et al. Nature 2005

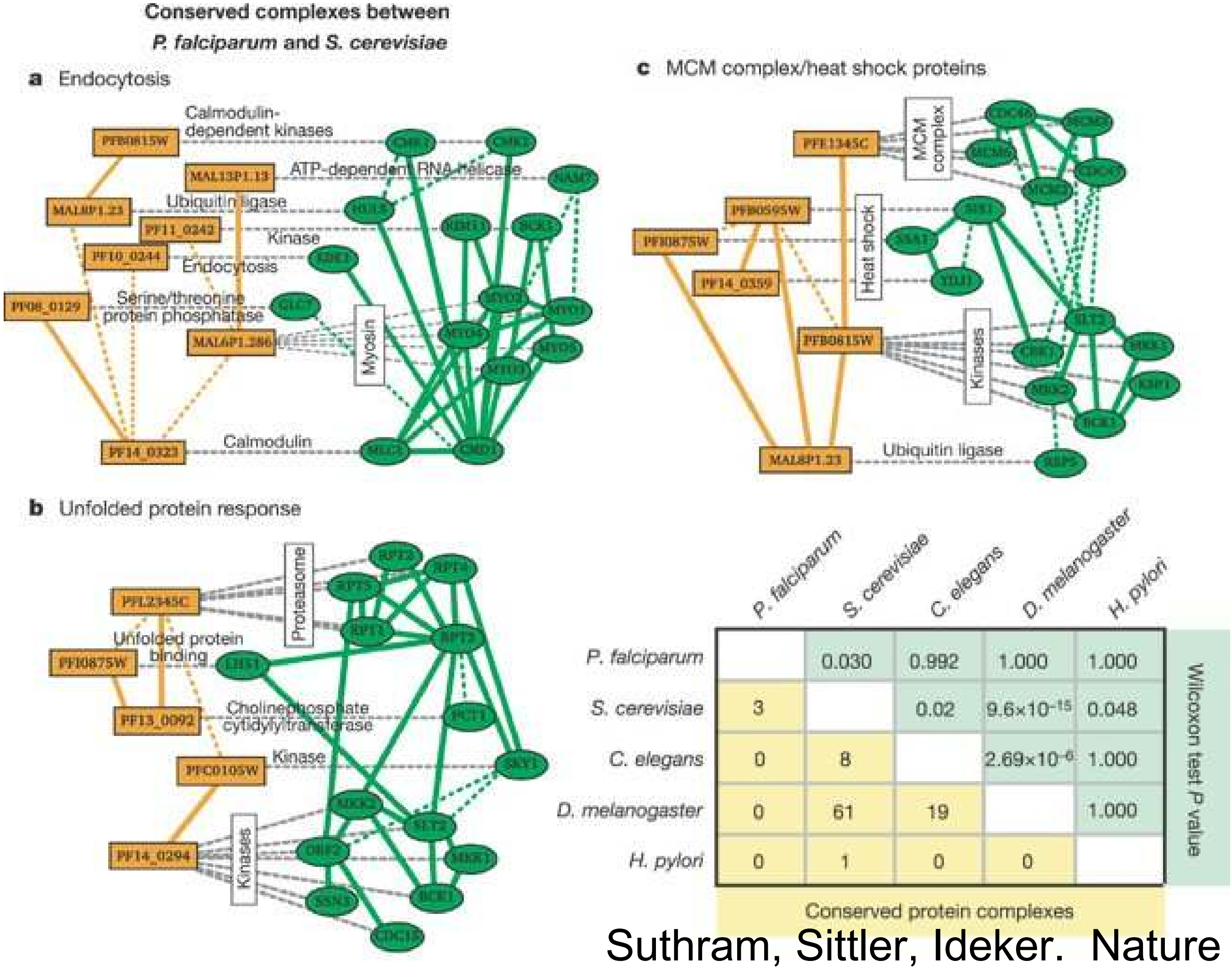


TAIR



Network alignments on weak signal

- Precedent: net align with strong signal



	P. falciparum	S. cerevisiae	C. elegans	D. melanogaster	H. pylori
P. falciparum		0.030	0.992	1.000	1.000
S. cerevisiae	3		0.02	9.6×10^{-15}	0.048
C. elegans	0	8		2.69×10^{-6}	1.000
D. melanogaster	0	61	19		1.000
H. pylori	0	1	0	0	

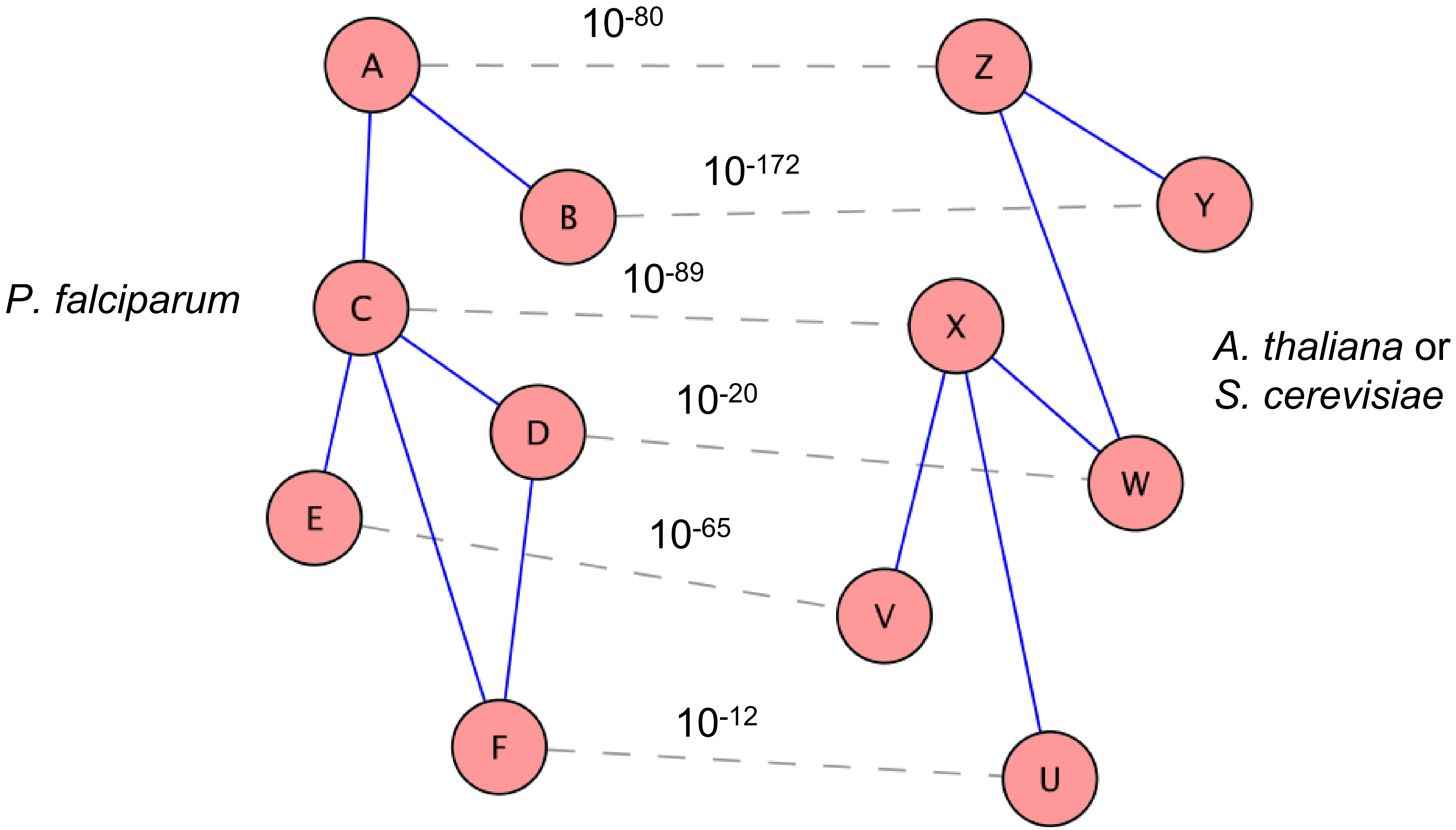
Wilcoxon test *P* value

Conserved protein complexes

Suthram, Sittler, Ideker. Nature 2005

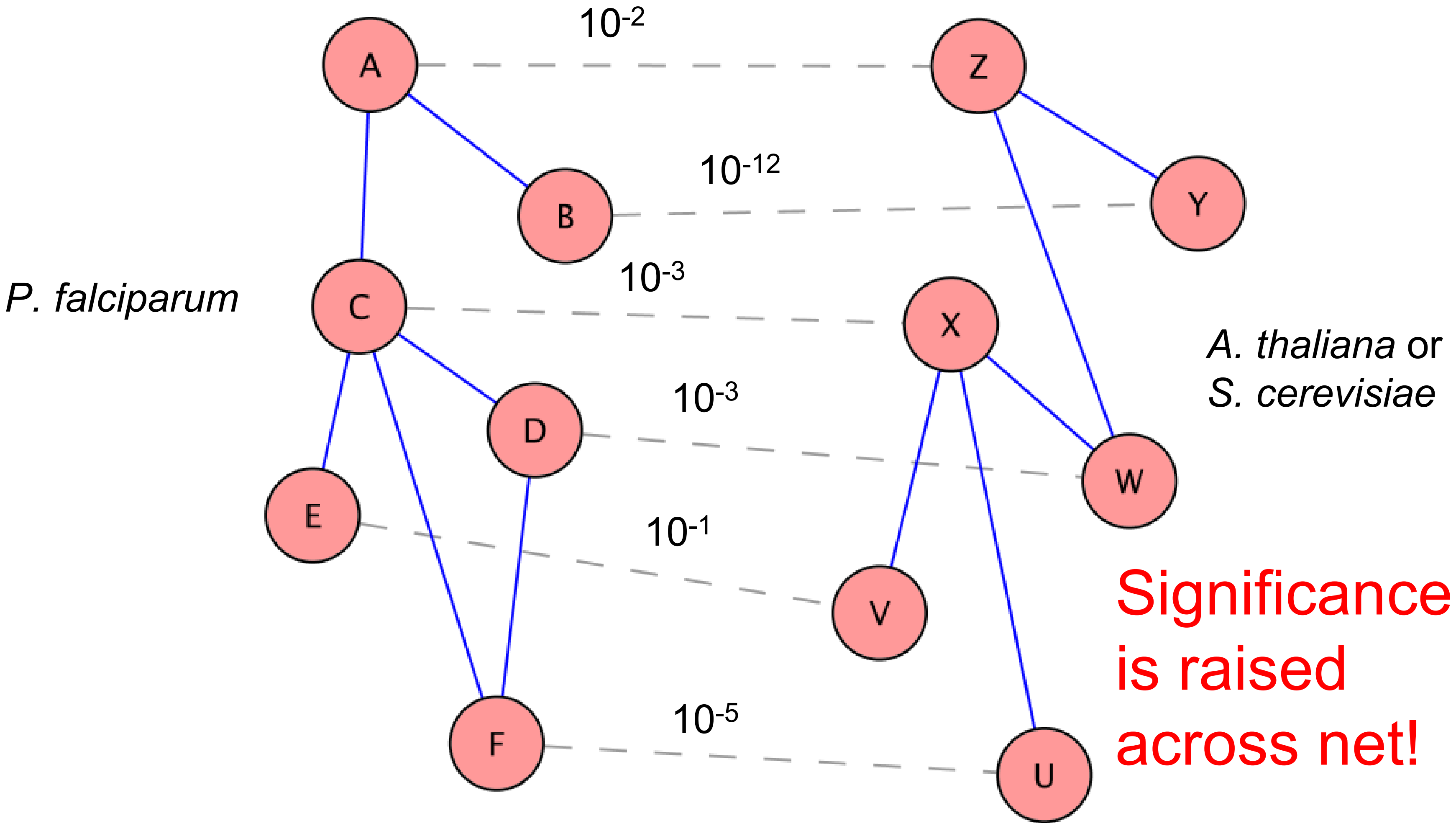
Network alignments on weak signal

- Strong BLAST hits (other methods)



Network alignments on weak signal

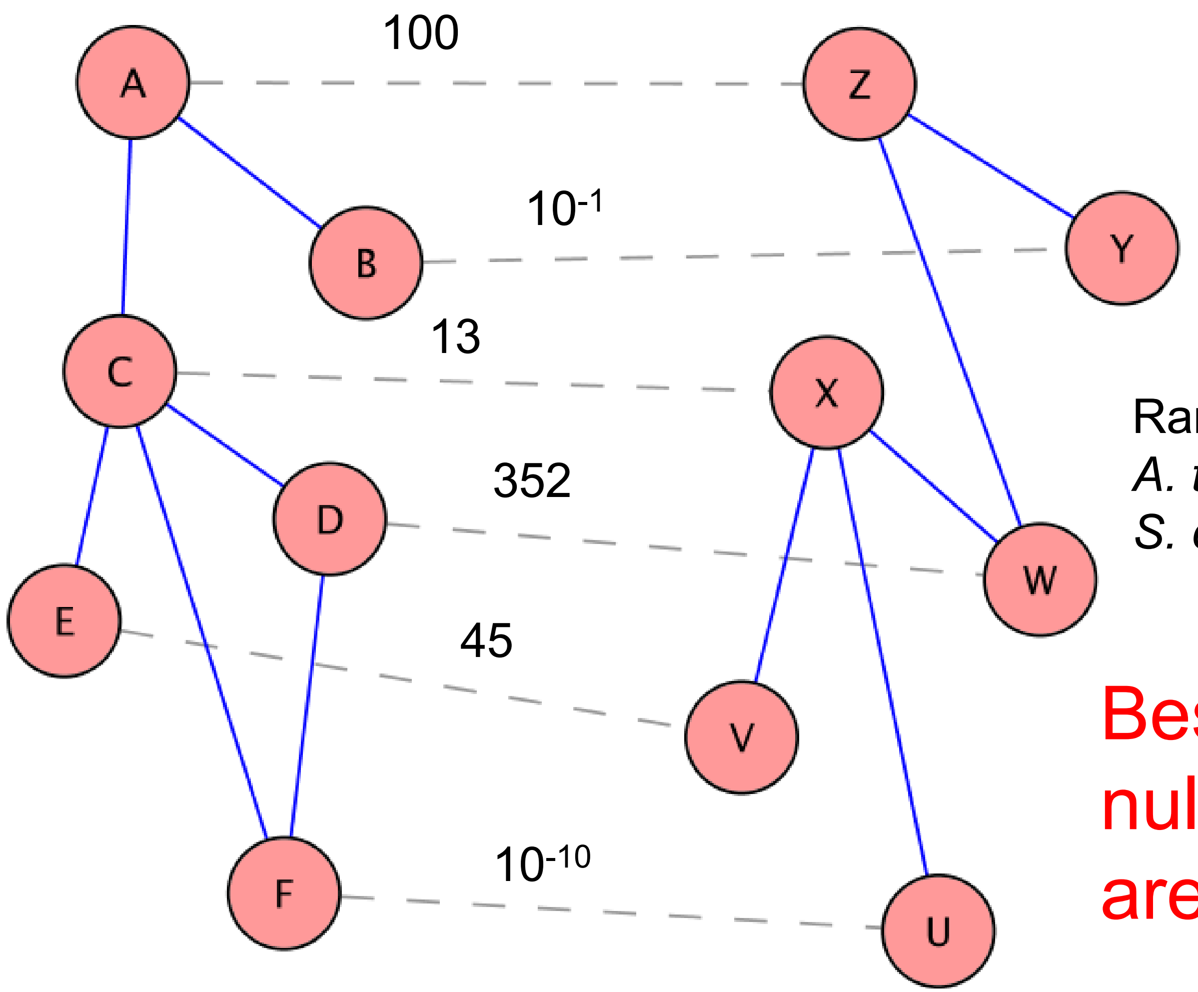
- Weak BLAST hits (our method)



Significance is raised across net!

Network alignments on weak signal

- Random BLAST hits



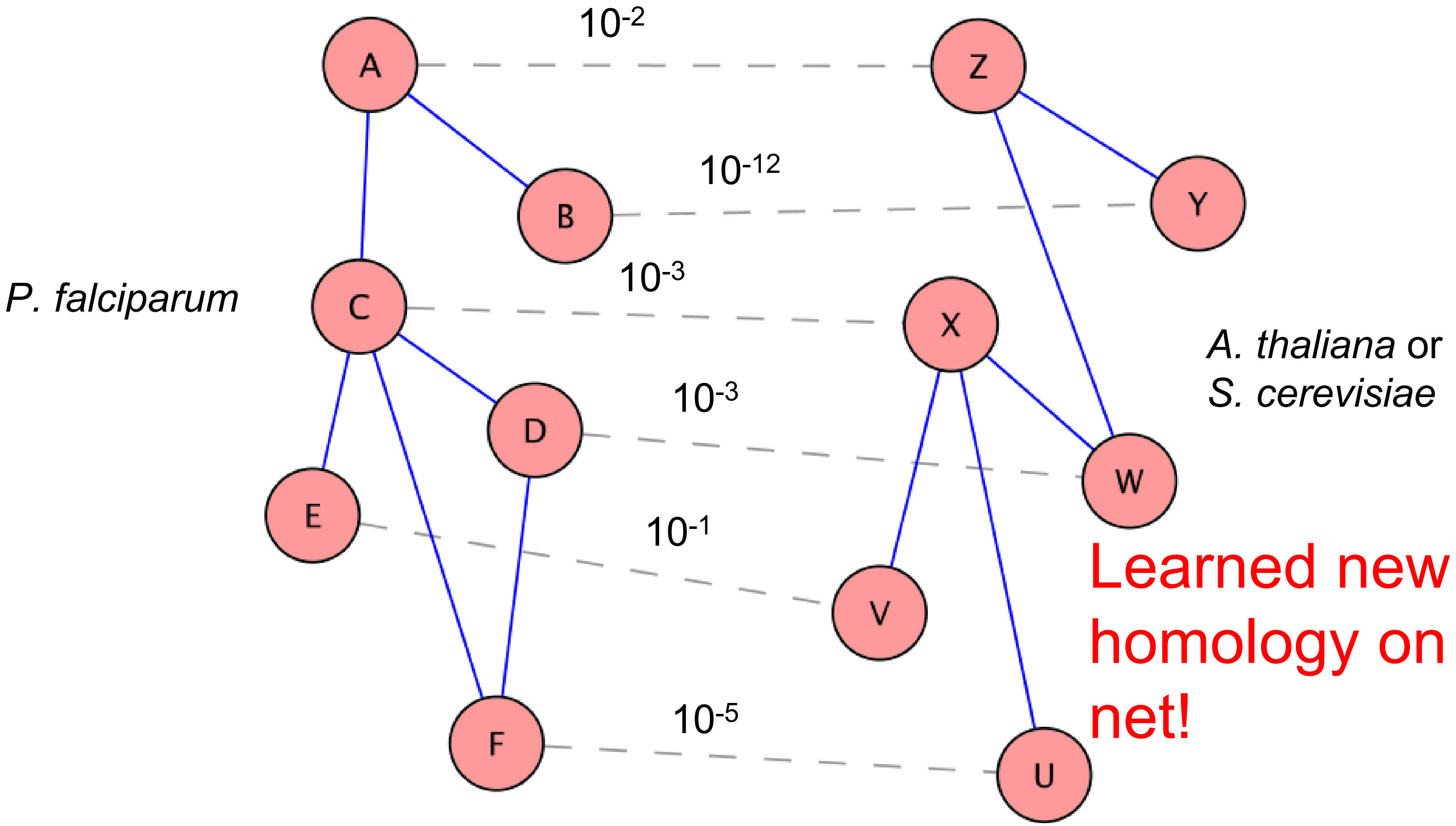
P. falciparum

Randomized
A. thaliana or
S. cerevisiae

Best hits in
null model
are poor!

Network alignments on weak signal

- Weak BLAST hits (our method)



Learned new homology on net!

Three computational approaches

- Domain pairs on a protein network
  - With results!
- Homology through profiles, domain architecture
- Network alignments on weak signal

Acknowledgements

- Singh Lab
  - Mona Singh
  - Zia Khan
  - Elena Nabieva
  - Eric Banks
  - Elena Zaslavsky
  - Tony Capra
  - Anton Persikov
  - Jimin Song
- Llinás Lab
  - Manuel Llinás
  - Erandi De Silva
  - Tracey Campbell

Domain pairs on a protein network

- Results: AP2 pair subnetwork

Proteins in net

Proteins in edges, Num edges