Studying Evolutionionary Transitions By Comparing Protein Interaction Networks

Frank Dehne
School of Computer Science
Carleton University, Ottawa, Canada
www.dehne.net

About the speaker

- Chancellor's Professor of Computer Science, Carleton University, Ottawa, Canada
- Fellow, IBM Center For Advanced Studies Canada
- Research specialization:
  - parallel computing (multi-core, GPU, clusters)
  - parallel data analytics (OLAP)
  - parallel computational biology
- Research Community:
  - Program Committees: SPAA 2011, IEEE Cluster 2010 (Vice-Chair), IPDPS 2009, ...
  - IEEE Technical Committee on Parallel Processing (Vice-Chair: 2003-2006)
  - ACM SPAA Steering Committee (2000 - present)
- Journal Editing:
  - IEEE Transactions on Computers (2004 - 2009)
  - Journal of Bioinformatics Research and Applications (2004 - present)
  - Journal of Data Warehousing and Mining (2004 - present)
  - Information Processing Letters (1989 - 2008)
  - Journal of Parallel Algorithms and Applications (1992 - 2005)

Parallel Computing & Bioinformatics Lab

Carleton UNIVERSITY
Canada's Capital University

Lab Members


Prof. F. Dehne


Dr. S. Pitre


A. Amos-Binks


A. Barton


D. Robillard


A. Schoenrock


H. Zaboli

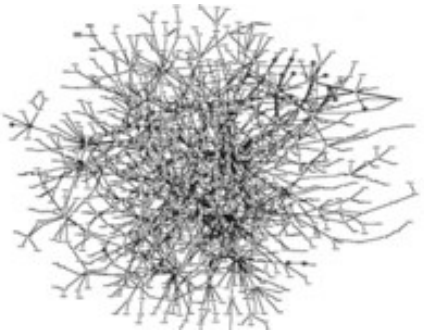

R. Zhou

Current Projects

2013: 1,900 core years (Compute Canada)

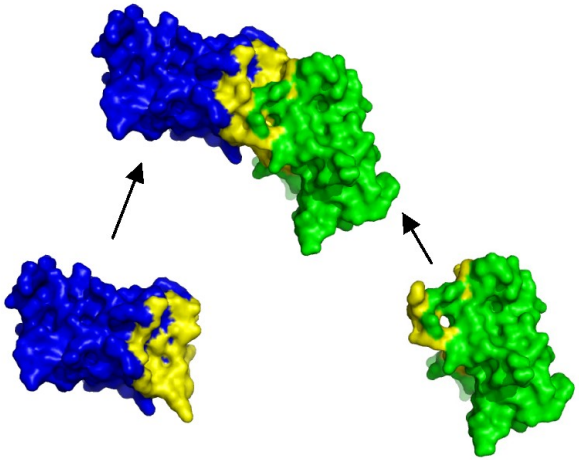
Parallel computing: auto-tuned parallel algorithms for multi-core processors, GPUs, clusters & clouds.

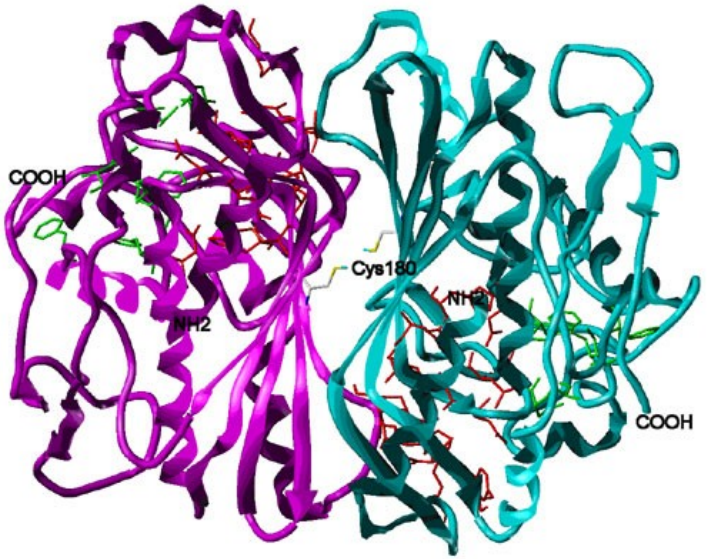
Parallel *big data* analytics: online analytical processing (OLAP).



Parallel computational biology: protein-protein interaction networks.

Protein-Protein Interactions





Important for many cellular functions, e.g.

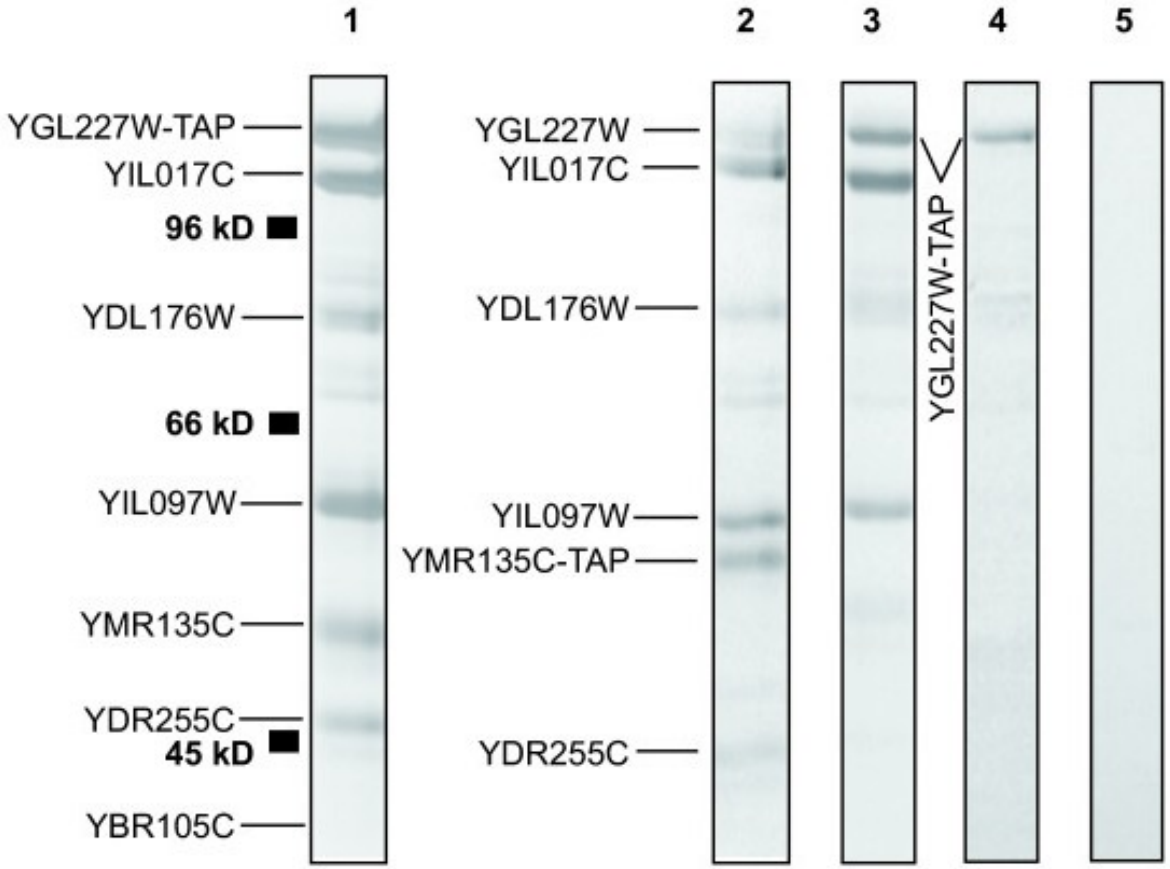
- DNA replication
- signal transduction
- ligand transport
- building structures

Detecting Protein-Protein Interactions

- Tandem affinity purification (TAP)
- Yeast two-hybrid screen
- Co-immunoprecipitation
- Bimolecular fluorescence complementation
- Affinity electrophoresis
- Pull-down assays
- Label transfer
- Phage display
- In-vivo crosslinking
- Chemical cross-linking
- Strepprotein interaction experiment
- Quantitative immunoprecipitation combined with knock-down
- Proximity ligation assay (PLA)

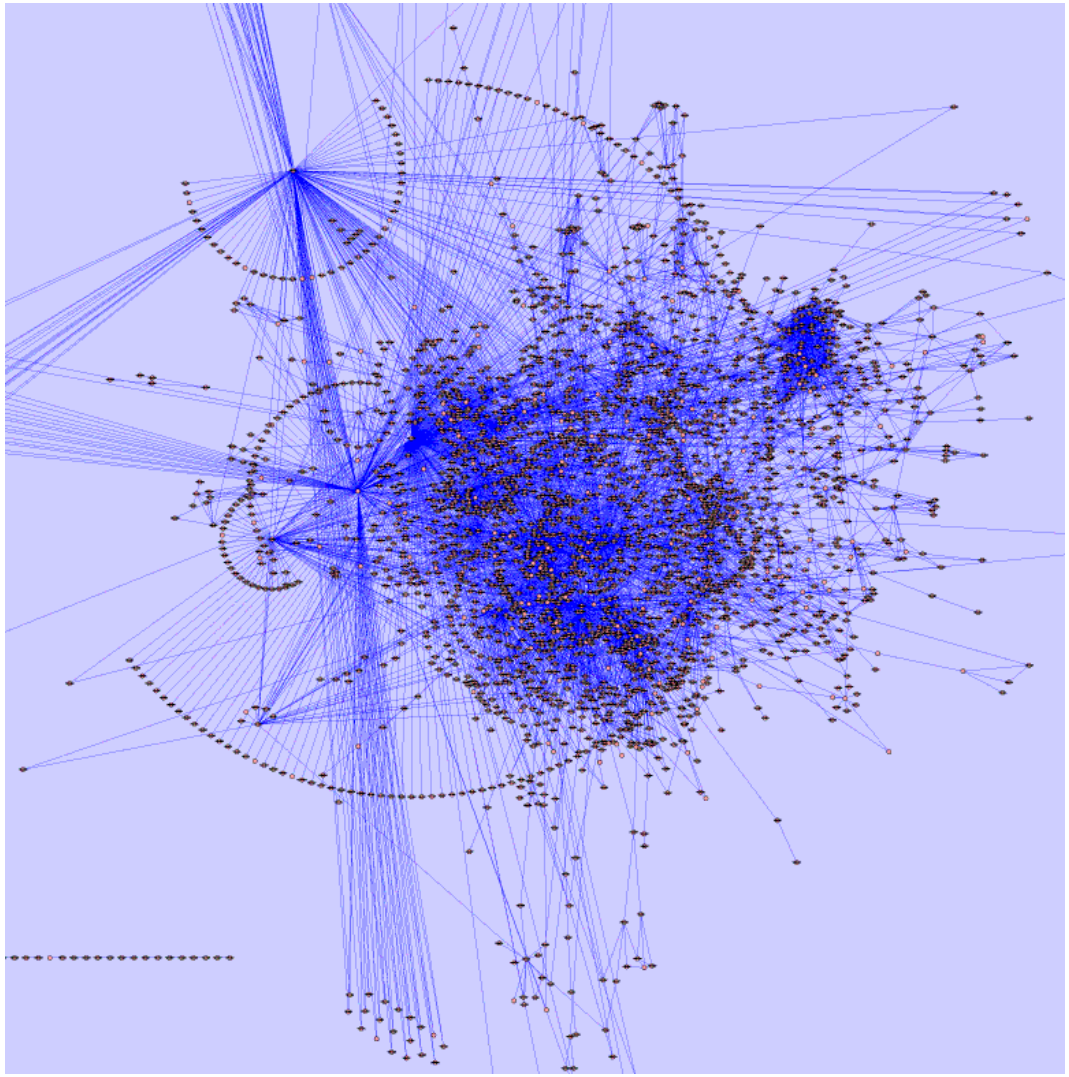
Tandem Affinity Purification (TAP)

Do YGL227W and YMR135C interact?



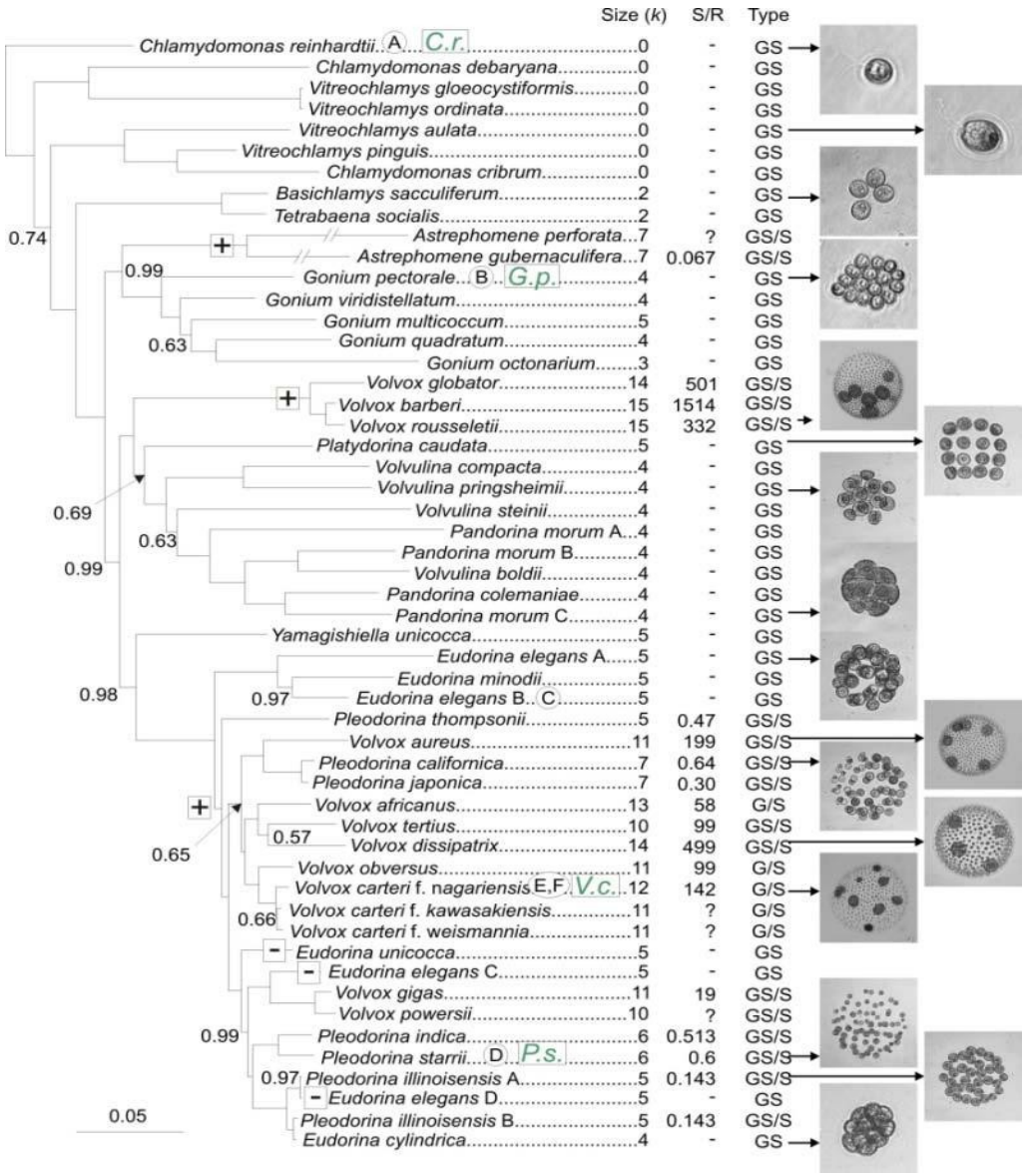
Yes.

Protein Interaction Network

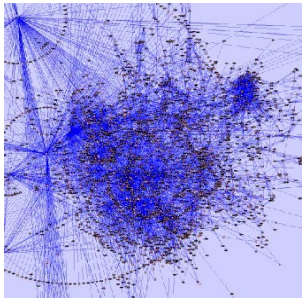


- **Pathways (chains)**
- **Functional Units (dense subgraphs, clusters)**

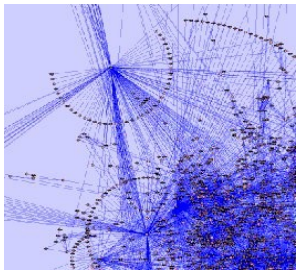
Evolutionary Transitions



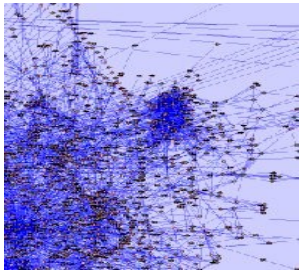
Comparison of Interactomes



V.c.



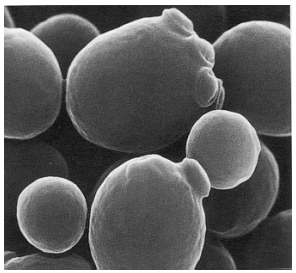
G.c.



C.r.

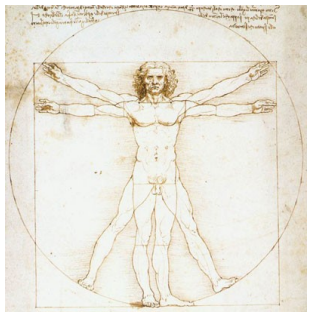
Known Protein Interactions

species	# proteins	# protein pairs	# known interactions	# unknown interactions
S. cerevisiae	6,300	19,867,056	15,151	???
C. elegans	23,684	280,454,086	6,607	???
H. sapiens	22,513	253,406,328	41,678	???



S. cerevisiae

C. elegans



H. sapiens

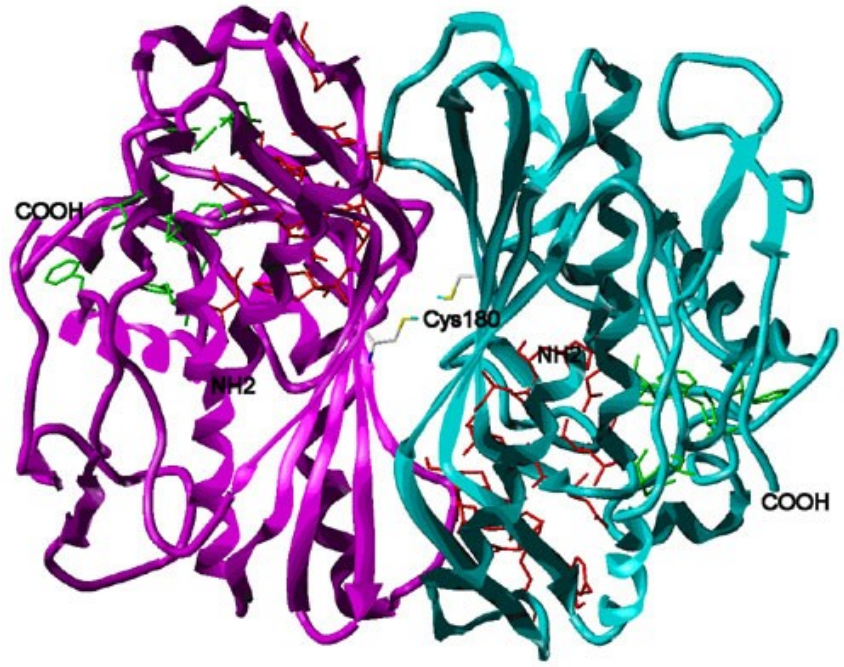
PIPE (Protein Interaction Prediction Engine)

- Project started in 2003
- Multi-disciplinary team
  - Computer Science
    - F.Dehne
  - Biochemistry
    - A.Golshani
    - J.Greenblatt
  - Biomed. Eng.
    - J. Green
  - Graduate Students / PostDocs
    - S.Pitre, C.North, A.Amos-Binks, A. Schoenrock, M.Alamgir, Bahram Samanfar, Mohsen Hooshyar, ...

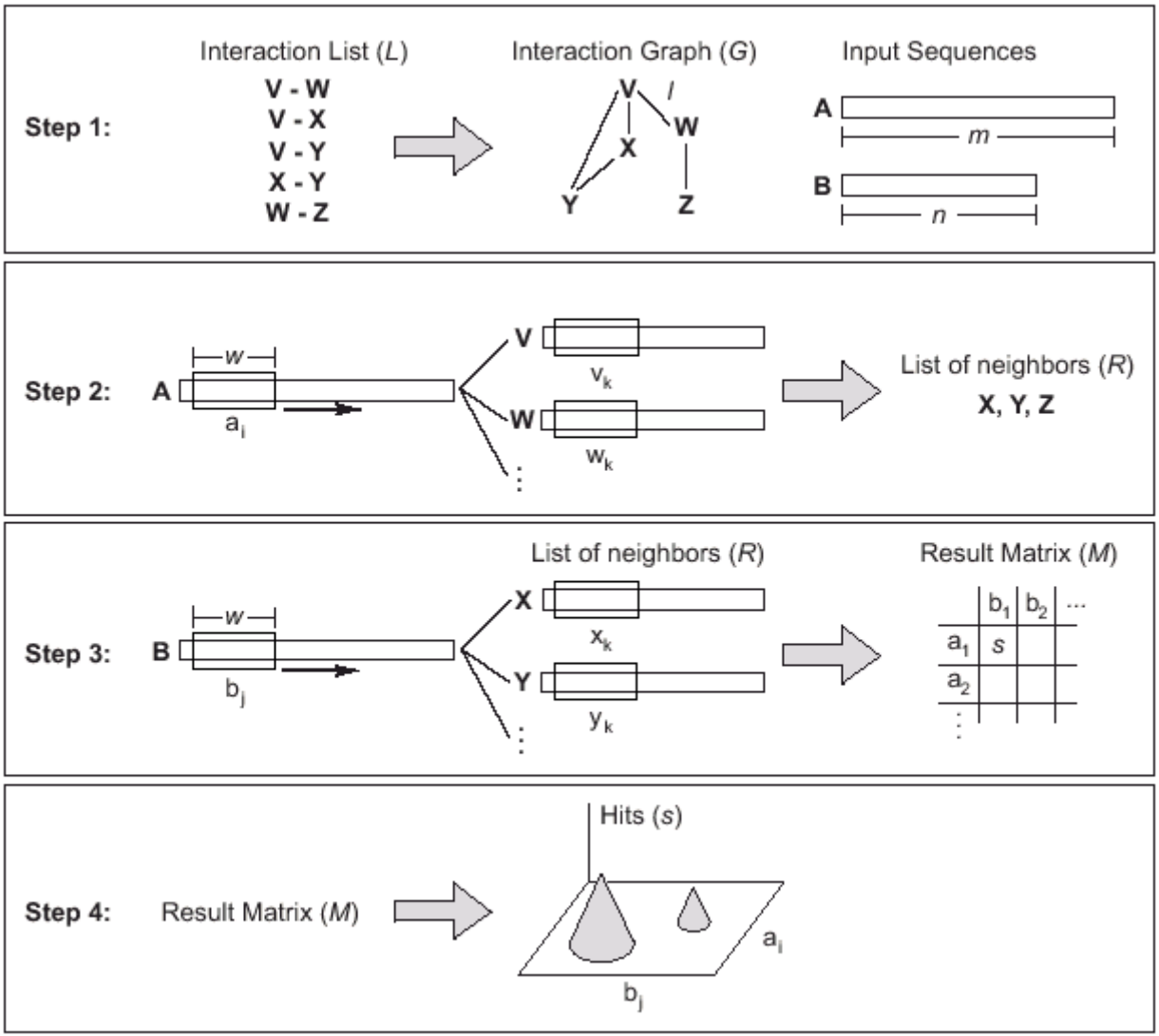
- Equipment used for this project:
  - 256 core PC Cluster
  - 1168 core Sun T2 “Victoria Falls” Cluster

Working Hypothesis

- Regions of interactions are usually small (20 - 40 amino acids)
- Interaction “Codes”



Basic PIPE Algorithm

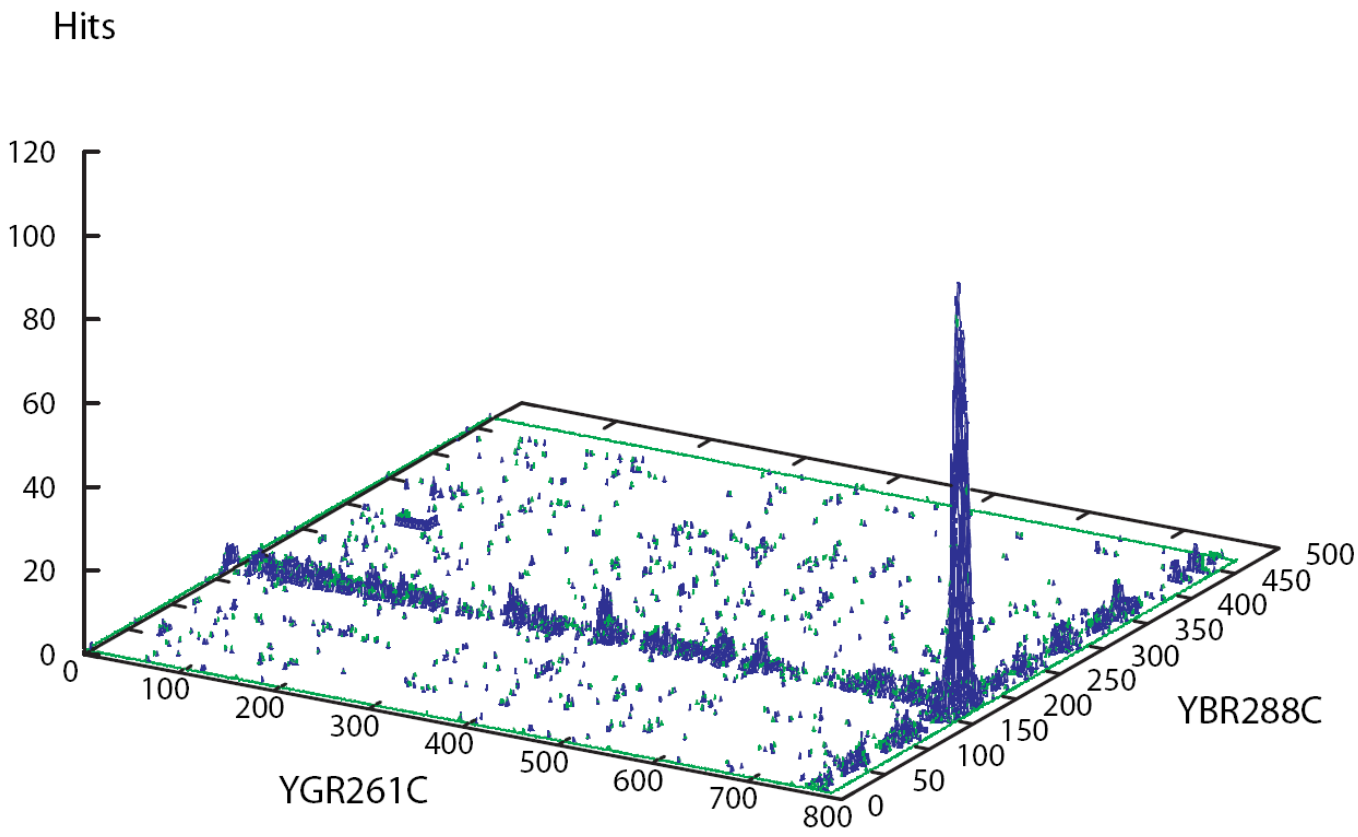


String comparison

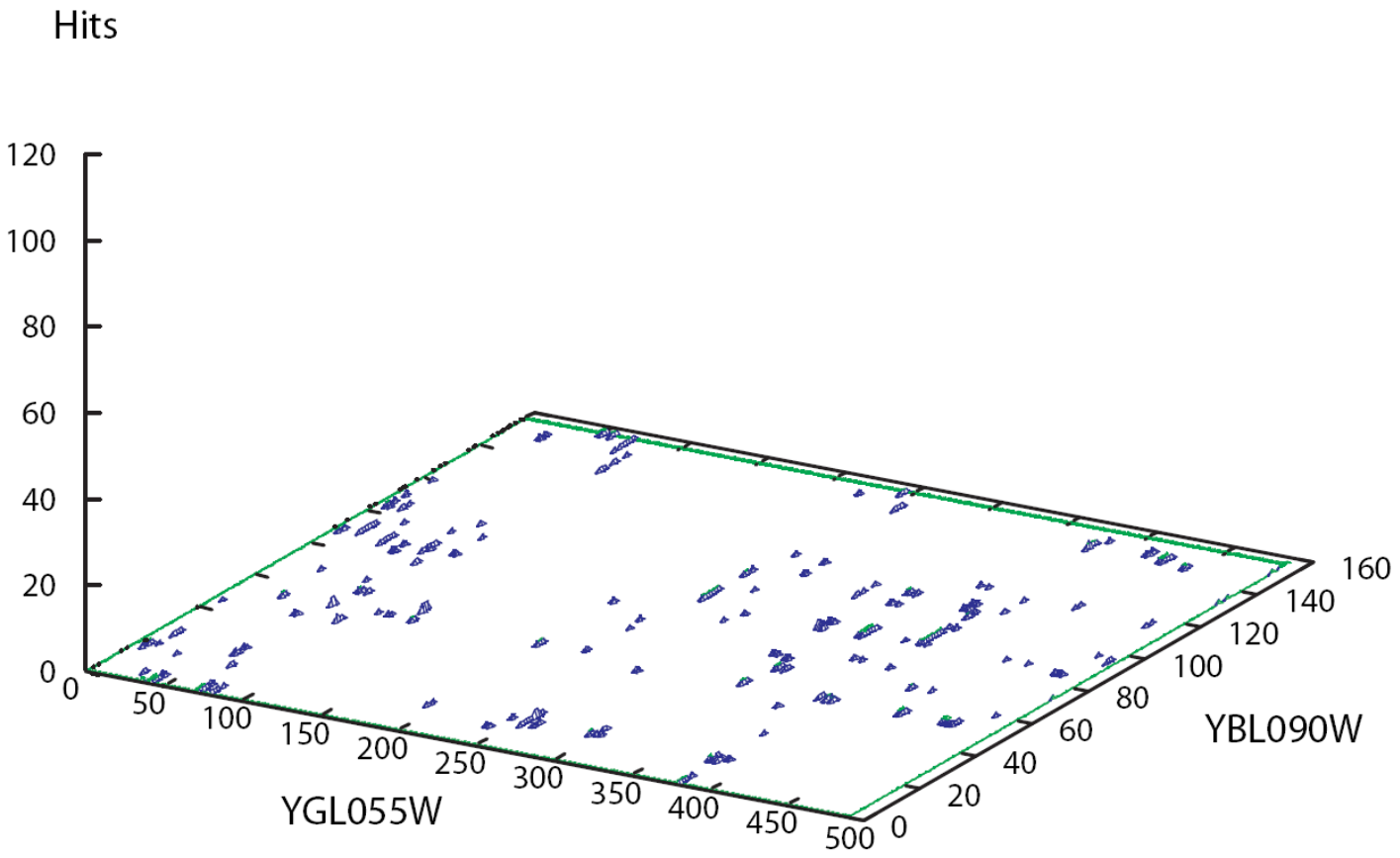

```
PAM 120
   A  R  N  D  C ...
A  3 -3 -1  0 -3
R -3  6 -1 -3 -4
N -1 -1  4  2 -5
D  0 -3  2  5 -7
C -3 -4 -5 -7  9
...
```

Match =
(Sum of pairwise PAM values > Threshold)

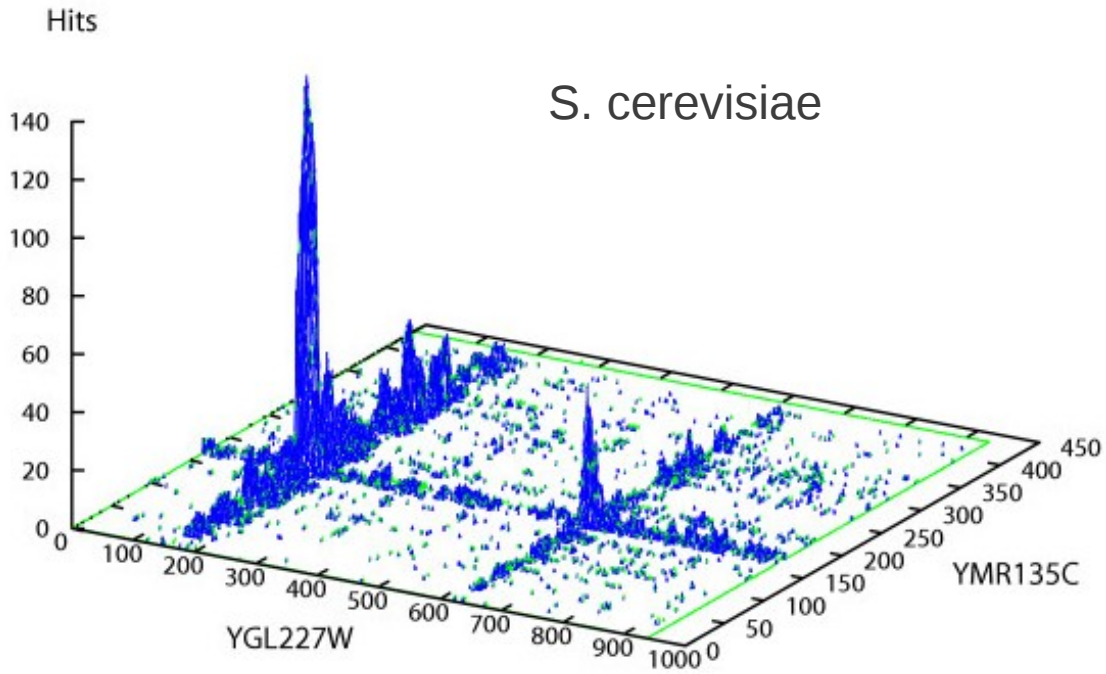
PIPE Output


Hits
120
100
80
60
40
20
0
0
100
200
300
400
500
600
700
800
YGR261C
0
50
100
150
200
250
300
350
400
450
500
YBR288C

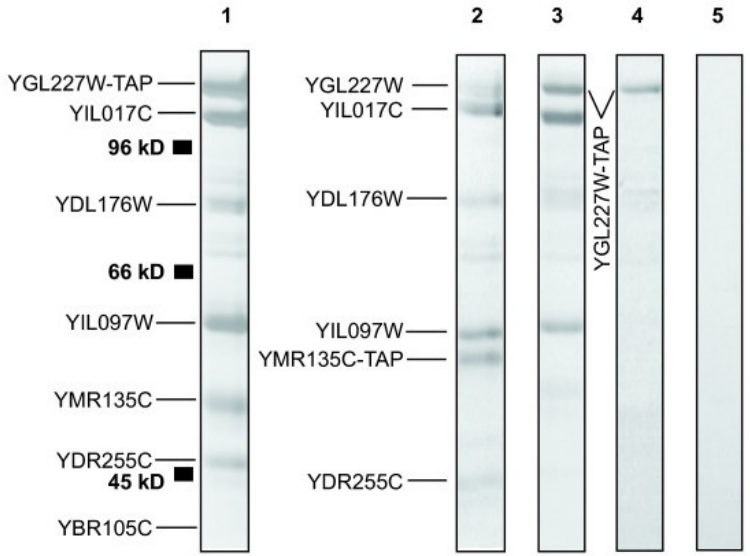
PIPE Output


Hits
120
100
80
60
40
20
0
0
50
100
150
200
250
300
350
400
450
500
YGL055W
0
20
40
60
80
100
120
140
160
YBL090W

PIPE: Detecting Novel Protein-Protein Interactions



S. cerevisiae

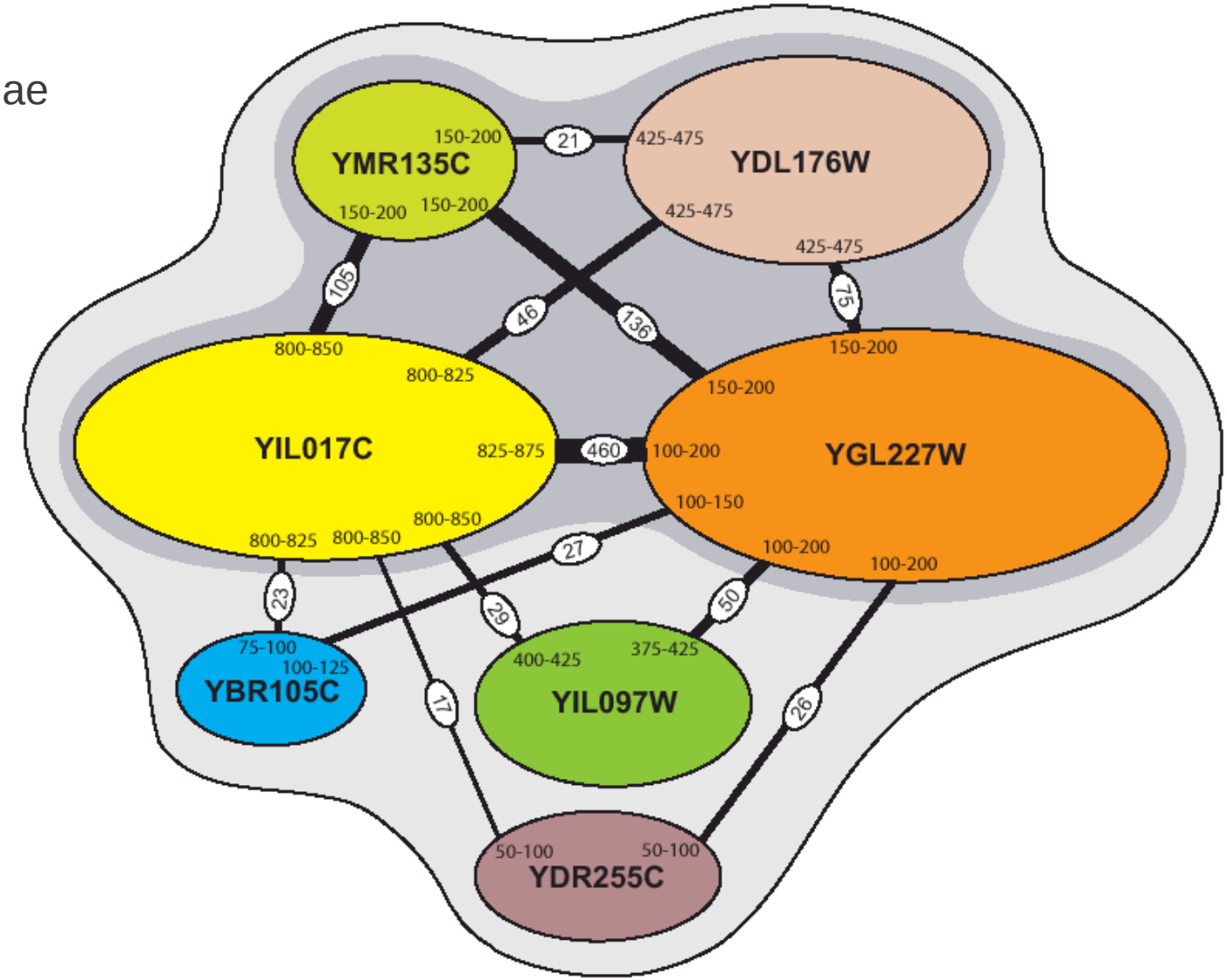


Banting and Best Institute of Medical Research, Toronto

Protein complex: YGL227W, YMR135C, YIL017C, YDL176W, YIL097W, YDR255C, YBR105C

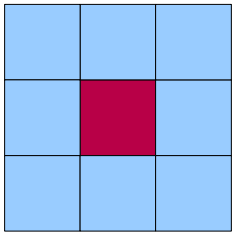
PIPE: Elucidating the Architecture of Protein Complexes

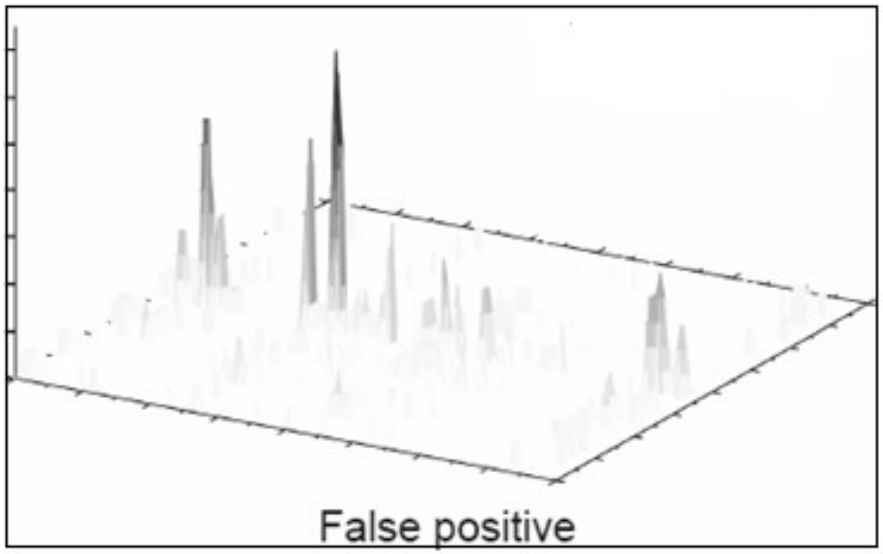
S. cerevisiae



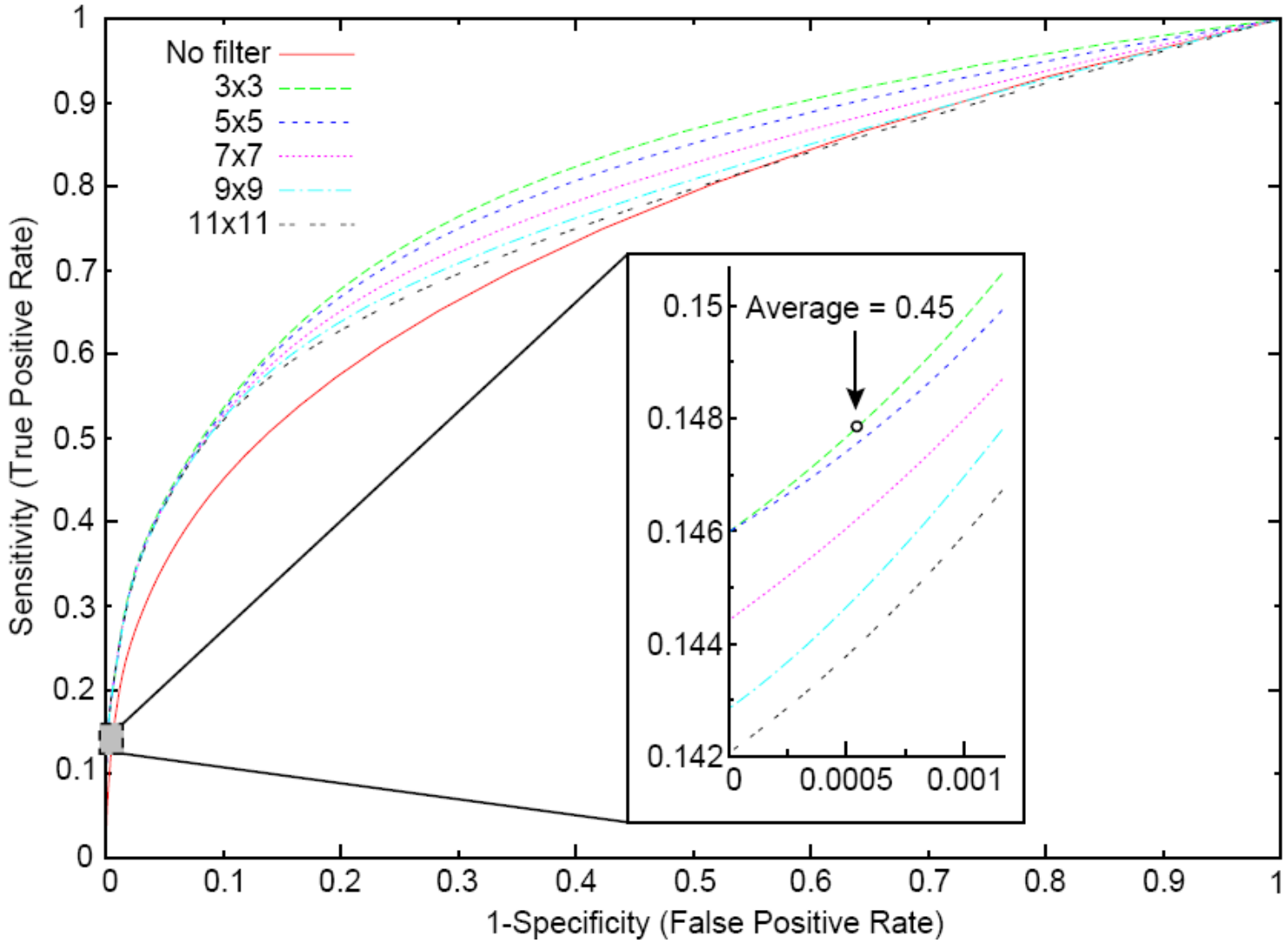
PIPE: Reducing False Positives

Eliminate “popular motifs” via median filter





PIPE's Prediction Accuracy



precision =	$\frac{TP}{TP+FP}$	specificity =	$\frac{TN}{TN+FP}$
recall =	$\frac{TP}{TP+FN}$	sensitivity =	$\frac{TP}{TP+FN}$

Many Other Methods...

Types of Protein Interaction Prediction Methods:

- Phylogenetic profiling
- Identification of homologous interacting pairs
- Identification of structural patterns (Van der Waals)
- Bayesian network modelling
- 3D template-based protein complex modelling
- Supervised learning (SVM)

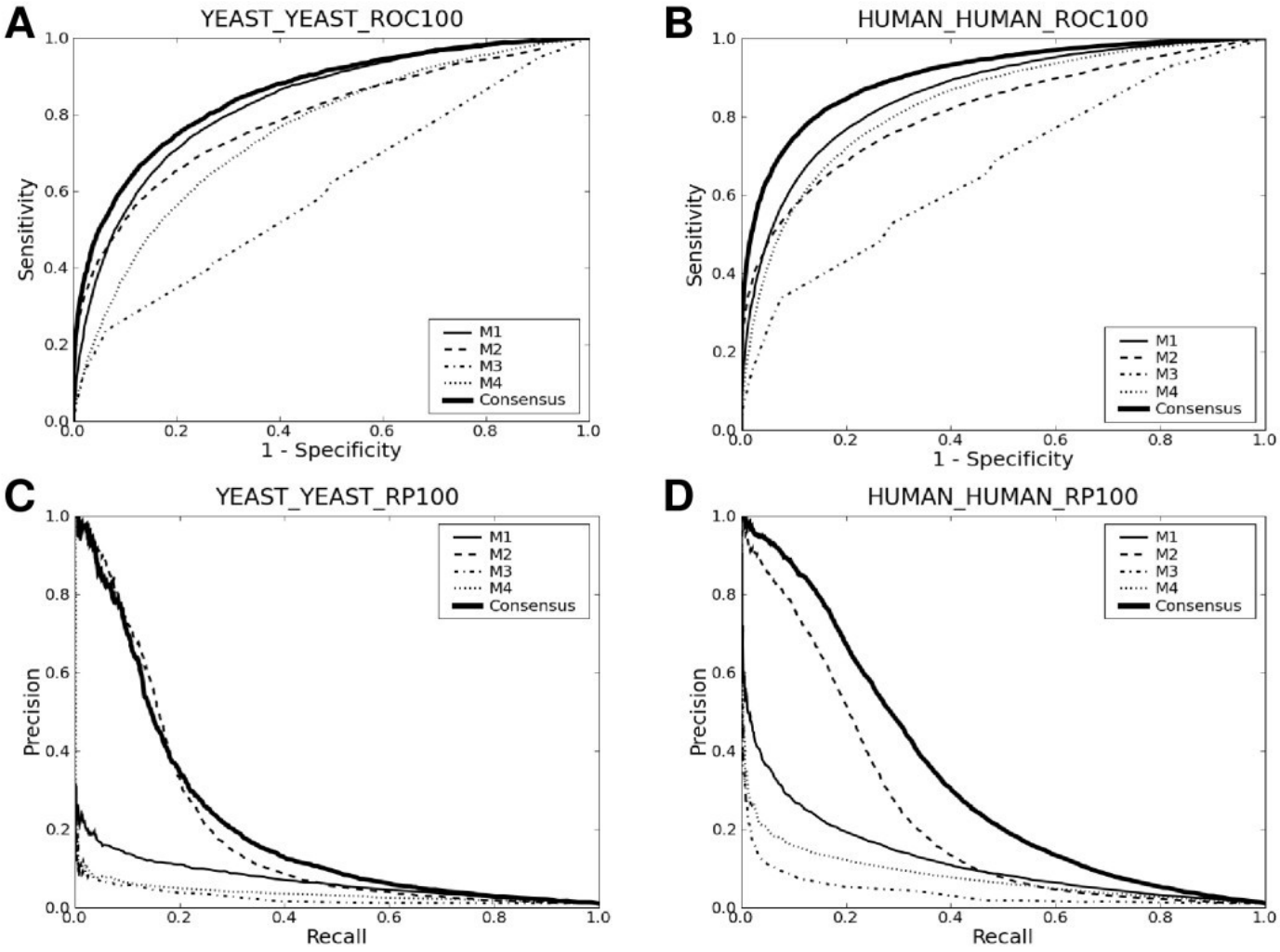
Park's Comparison Experiment

Park (BMC Bioinformatics, 2009, 10:419) compared the four *best* methods

- **[M1]** Martin et.al. (Bioinformatics 2005,21(2):218-226): protein pair is encoded by a product of signatures which is then classified by a support vector classifier
- **[M2]** PIPE
- **[M3]** Shen et.al. (Proc Natl Acad Sci USA 2007, 104(11):4337-4341): each protein sequence is encoded by a feature vector that represents the frequencies of 3 amino acid-long subsequences, and feature vectors are concatenated for a pair of proteins and classified by SVM.
- **[M4]** Guo et.al. (Nucl Acids Res 2008,36(9):3025-3030): each protein sequence is encoded by a feature vector that represents auto-correlation values of 7 different physicochemical scales, and feature vectors are concatenated for a pair of proteins and classified by SVM.

- Consensus Method: “Vote” among M1-M4.

Park's Comparison Experiment



From: Park, BMC Bioinformatics, 2009, 10:419

$\text{precision} = \frac{TP}{TP+FP}$		$\text{specificity} = \frac{TN}{TN+FP}$	
$\text{recall} = \frac{TP}{TP+FN}$		$\text{sensitivity} = \frac{TP}{TP+FN}$	

Global Scan of Entire Protein Interaction Network

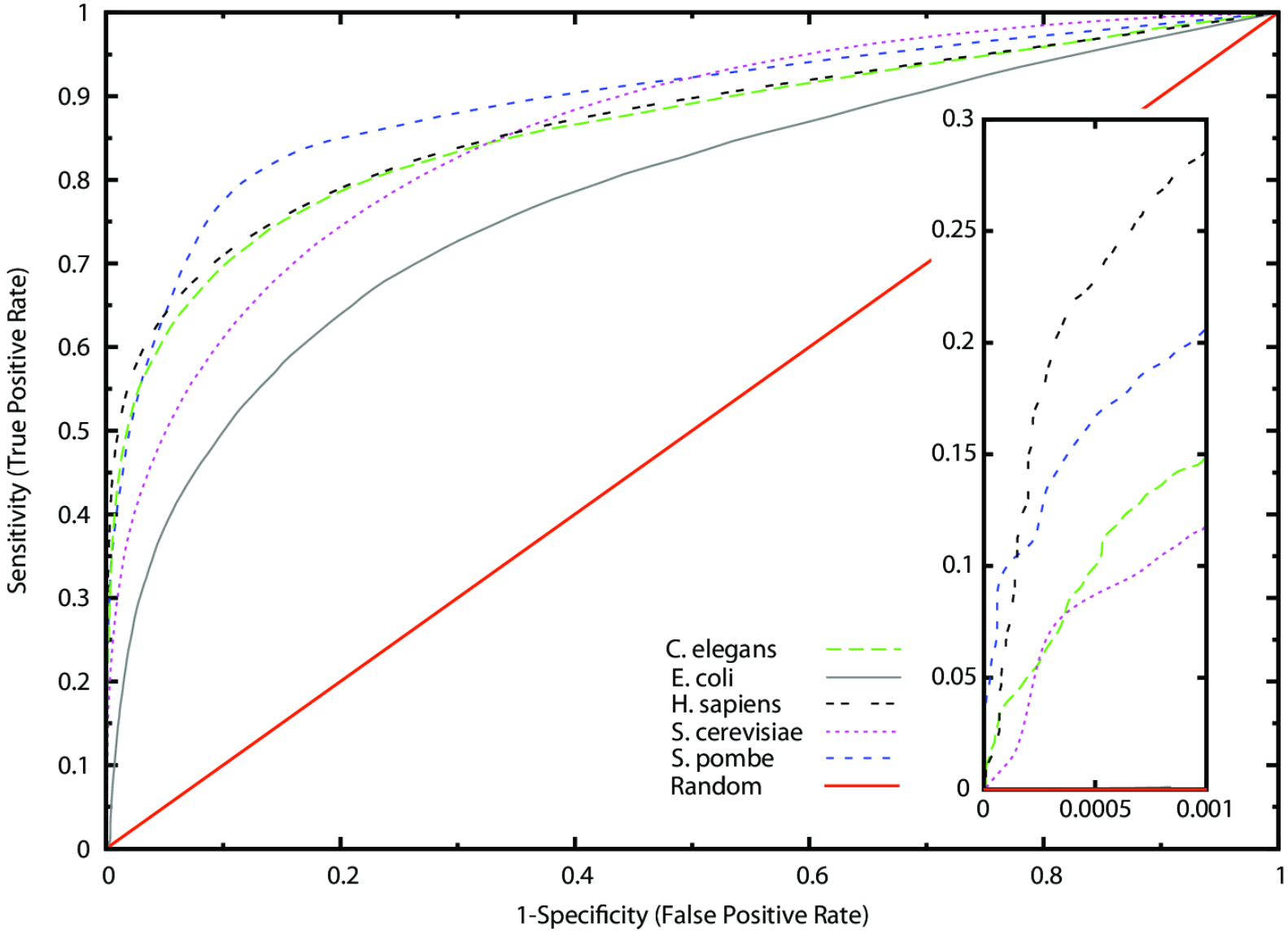
species	# proteins	# protein pairs	# known interactions
S. cerevisiae	6,300	19,867,056	15,151
C. elegans	23,684	280,454,086	6,607
H. sapiens	22,513	253,406,328	41,678

Open Problem...

Challenges:

- Large number of protein pairs (requires high speed, SVM not possible)
- Small number of true positives (very sparse, ~ 0.1 % density)
- Requires very high specificity ~99.95 % (i.e. less than 0.05% false positive rate) – Otherwise: #false positives > #true positives
- Massive computational challenge

PIPE's Prediction Accuracy



PIPE Sequential Performance Improvements

- Character based amino acid representation was converted into binary encodings. Removed need for character-to-index lookup in PAM120.
- “Sliding window” process was improved to use incremental updates.
- Pre-computed all possible protein fragment comparisons and stored all matches of similar fragments in a hash table.

Large Scale Parallelization: MP-PIPE

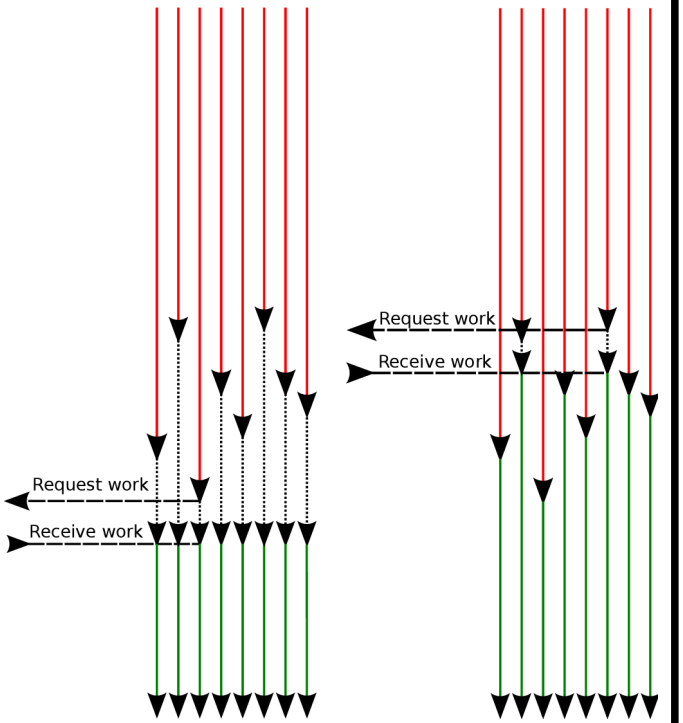
Algorithm 1: MP-PIPE Scheduler.

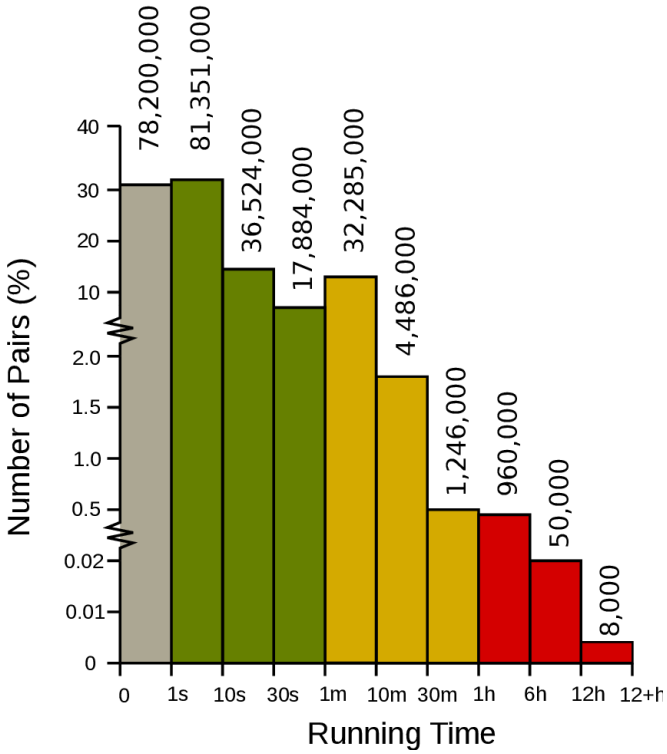
Split protein pairs into packets.
while *packets remain* **do**
  receive work request from worker x
  receive previous results from worker x
  send packet to worker x
  write results to output file
foreach *worker process* **do**
  receive work request from worker x
  receive previous results from worker x
  send KILL_SIGNAL to worker x
  write results to output file

Algorithm 2: MP-PIPE Worker.

Load PIPE database (interaction graph G and hash table of pre-computed protein fragment similarity matches).
current_packet ← ∅
current_results ← ∅
work_available ← TRUE
foreach *thread* **in parallel do**
  while *work_available* **do**
    if *current_packet* = ∅ **then**
      request work from scheduler
      send *current_results* to scheduler
      receive *message* from scheduler
      if *message* = *KILL_SIGNAL* **then**
        work_available ← FALSE
        BREAK
      else
        current_packet ← *message*
    retrieve pair from *current_packet*
    run PIPE algorithm on pair
    add results to *current_results*

Architecture:
- Cluster of multi-core processors
- One MP-PIPE worker per proc.
- Each worker with multiple threads





H.sapiens protein pairs

Summary of PIPE Results

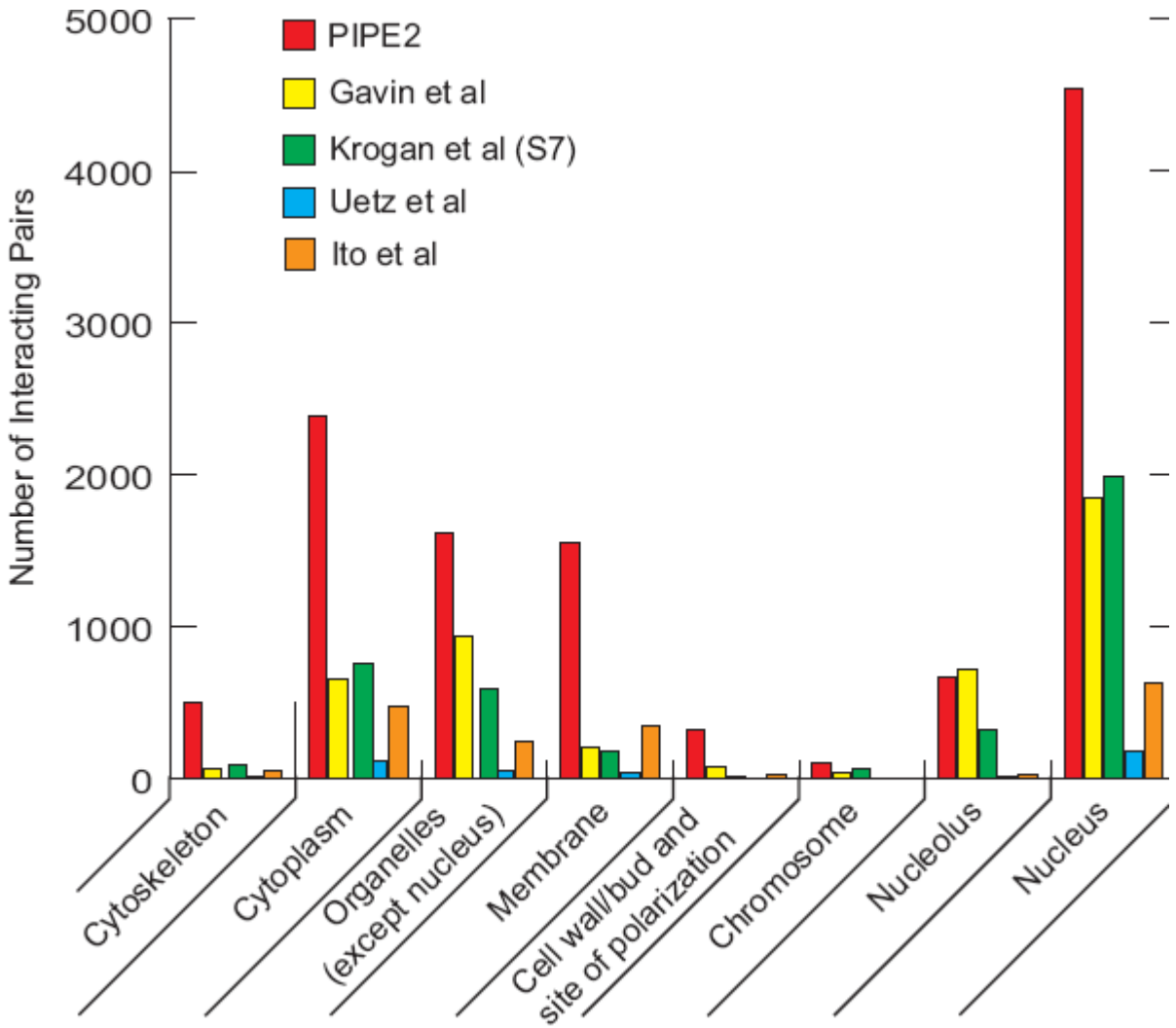
PIPE's superior performance and prediction accuracy enabled the first ever complete scan of entire protein interaction networks

species	# proteins	# protein pairs	# known interactions	# novel PIPE pred. *	Running time
S. cerevisiae	6,300	19,867,056	15,151	14,438	1 hour
C. elegans	23,684	280,454,086	6,607	32,548	1 week
H.sapiens	22,513	253,406,328	41,678	130,470	3 months

* False positive rate: 0.0001

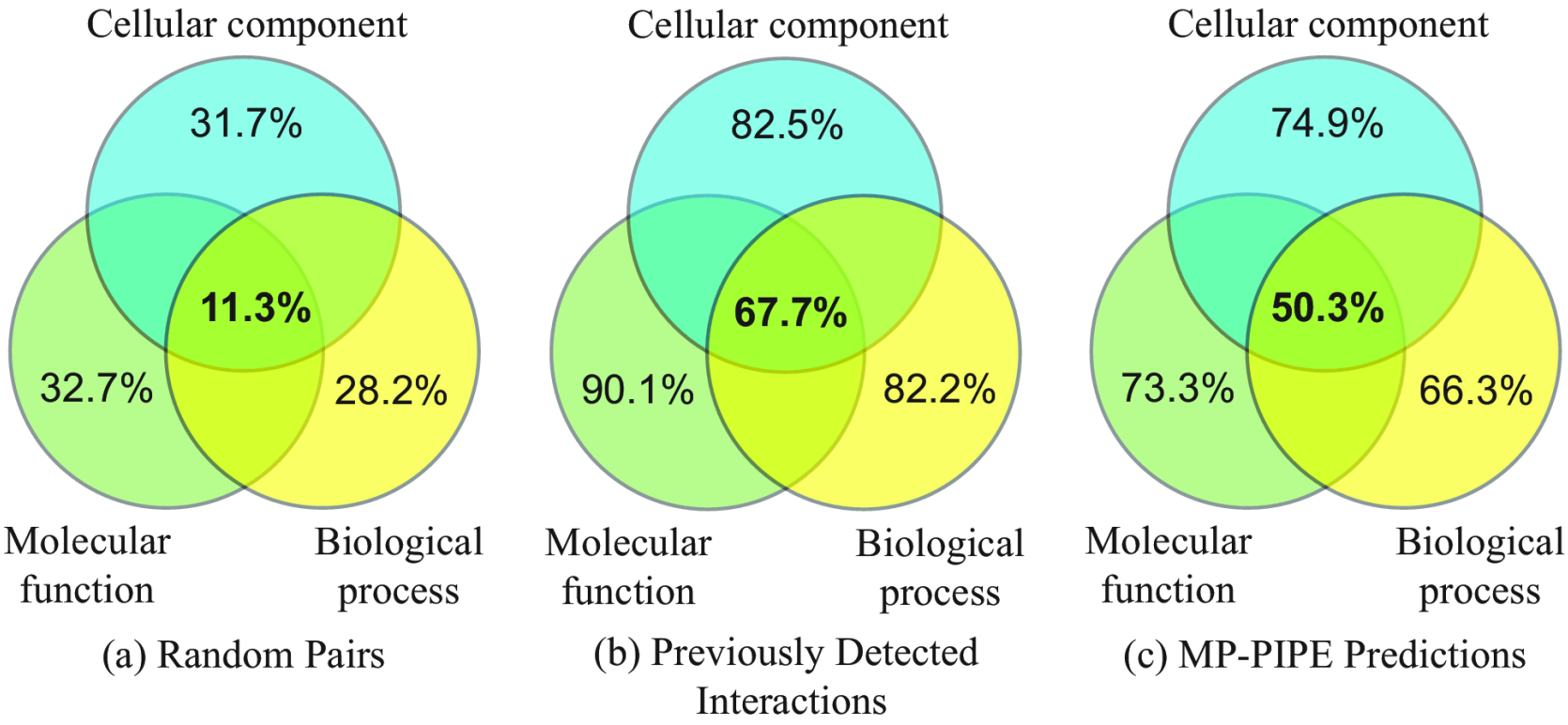
- 256 core PC Cluster
- 1168 core Sun T2 “Victoria Falls” Cluster

Cellular Co-Localization



S. cerevisiae

Cellular Co-Localization

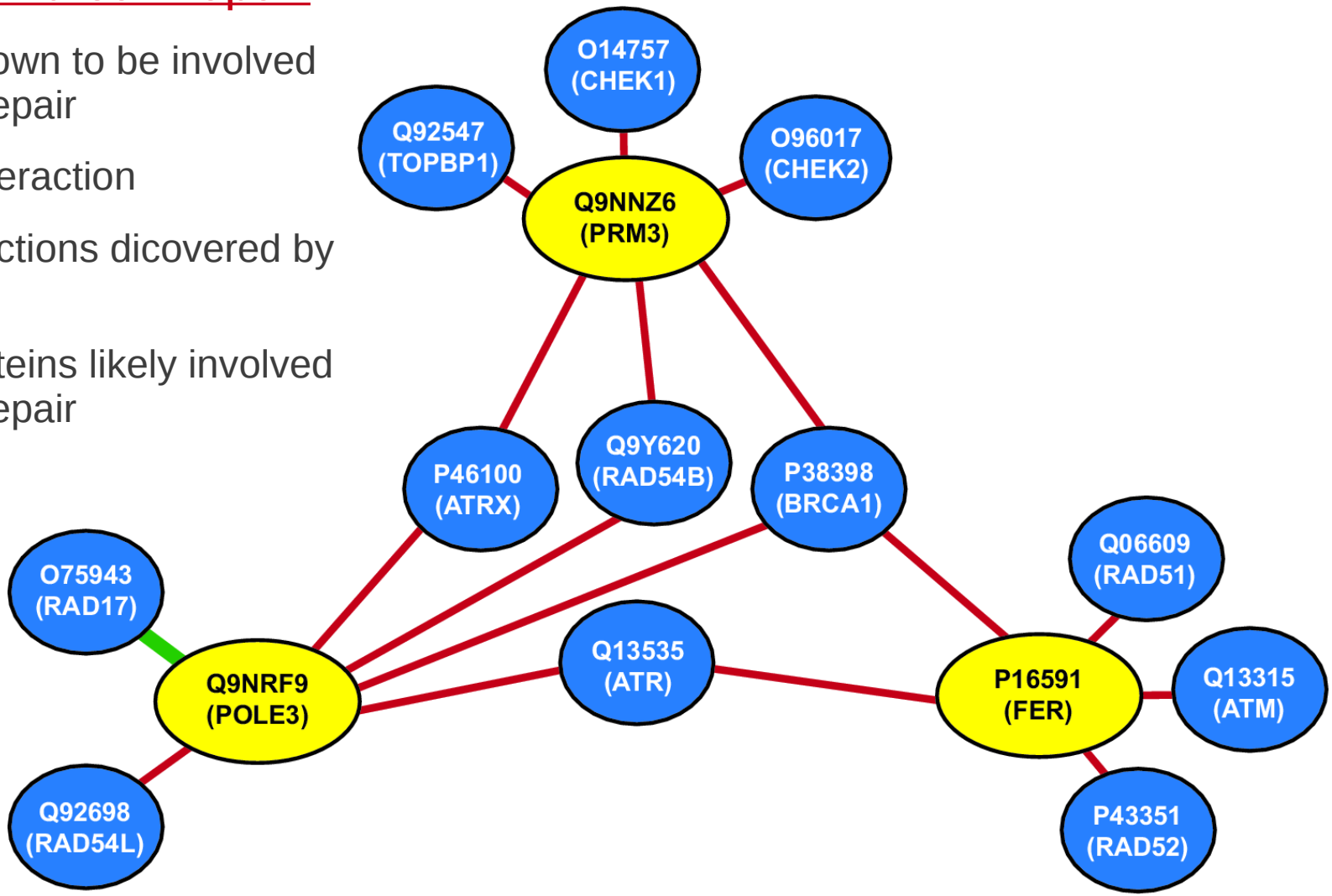


H.sapiens

PIPE Enabled Discoveries

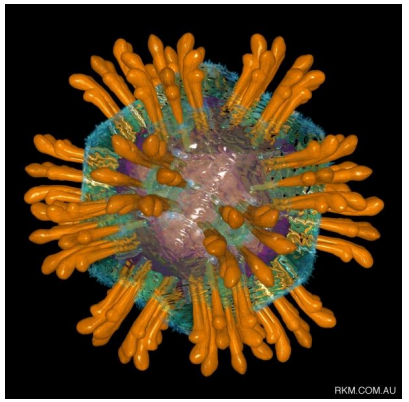
H.sapiens dsDNA break repair

- Blue: Proteins known to be involved in dsDNA break repair
- Green: Known interaction
- Red: Novel interactions dicovered by PIPE
- Yellow: Novel proteins likely involved in dsDNA break repair



Inter Species PIPE

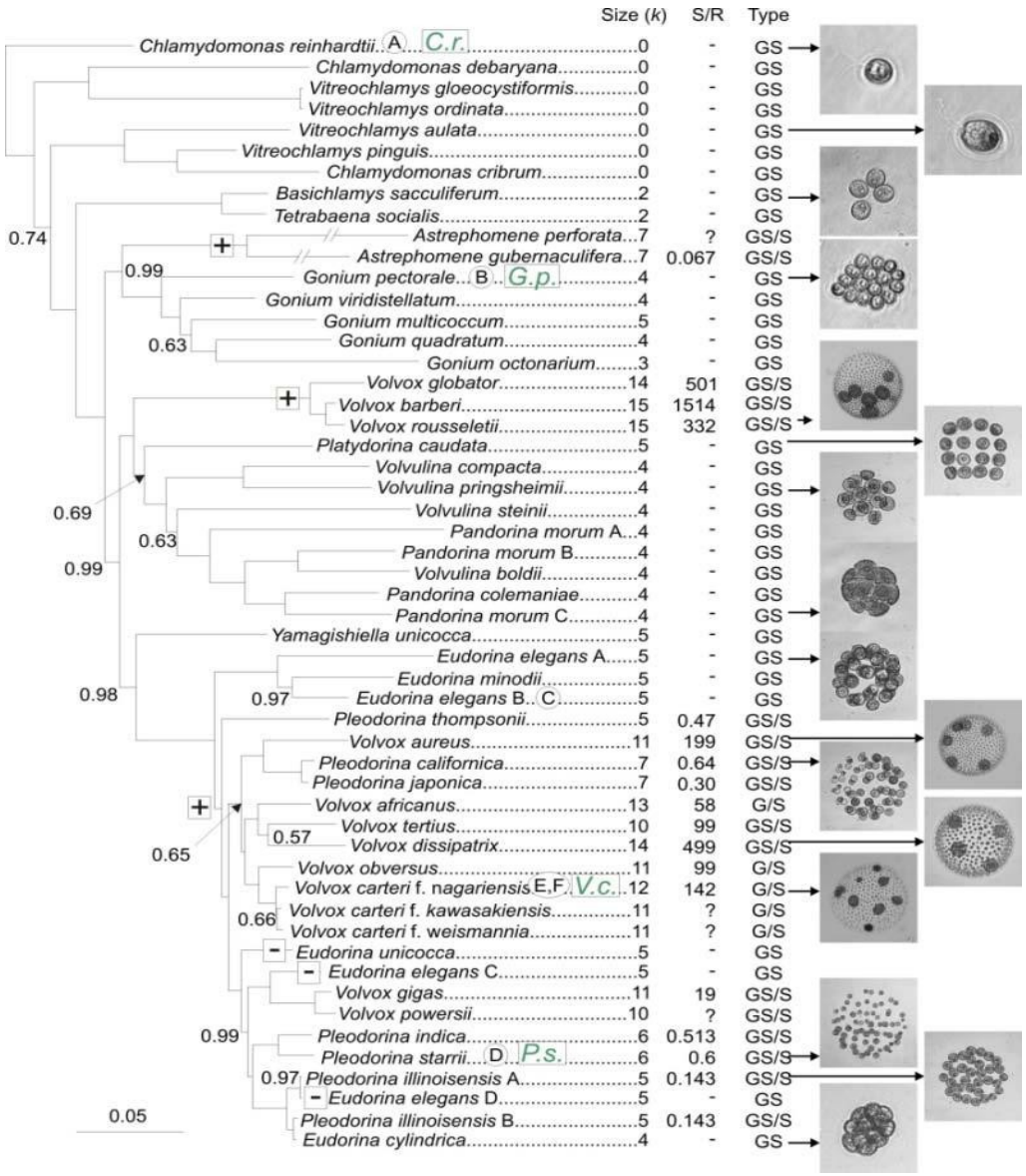
- Prediction of human-pathogen protein-protein interaction
- Used PIPE to predict Human vs. Hepatitis C (HCV), Influenza A, HIV-1 and HIV-2 interactions (using “all” database)
- Found novel interactions between HCV non-structural 5A (NS5A) protein and several Human proteins, includes human IPO5 which is known to be involved in HIV-1 transmission
- Found novel interactions between HIV-1 virion infectivity factor Vif and human APOBEC3A, APOBEC3B, APOBEC3D (family of APOBEC proteins are known to be deactivated during HIV infection)



Hepatitis C Virus

HIV Virus

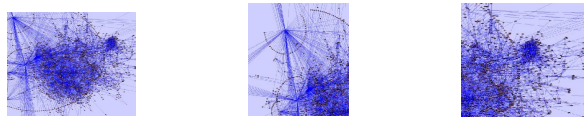
Evolutionary Transitions



New Project

- Frank Dehne et.al. (PIPE Group)
- Pierre Durand
- Richard Michod
- Bradley Olson

Comparison of Interactomes

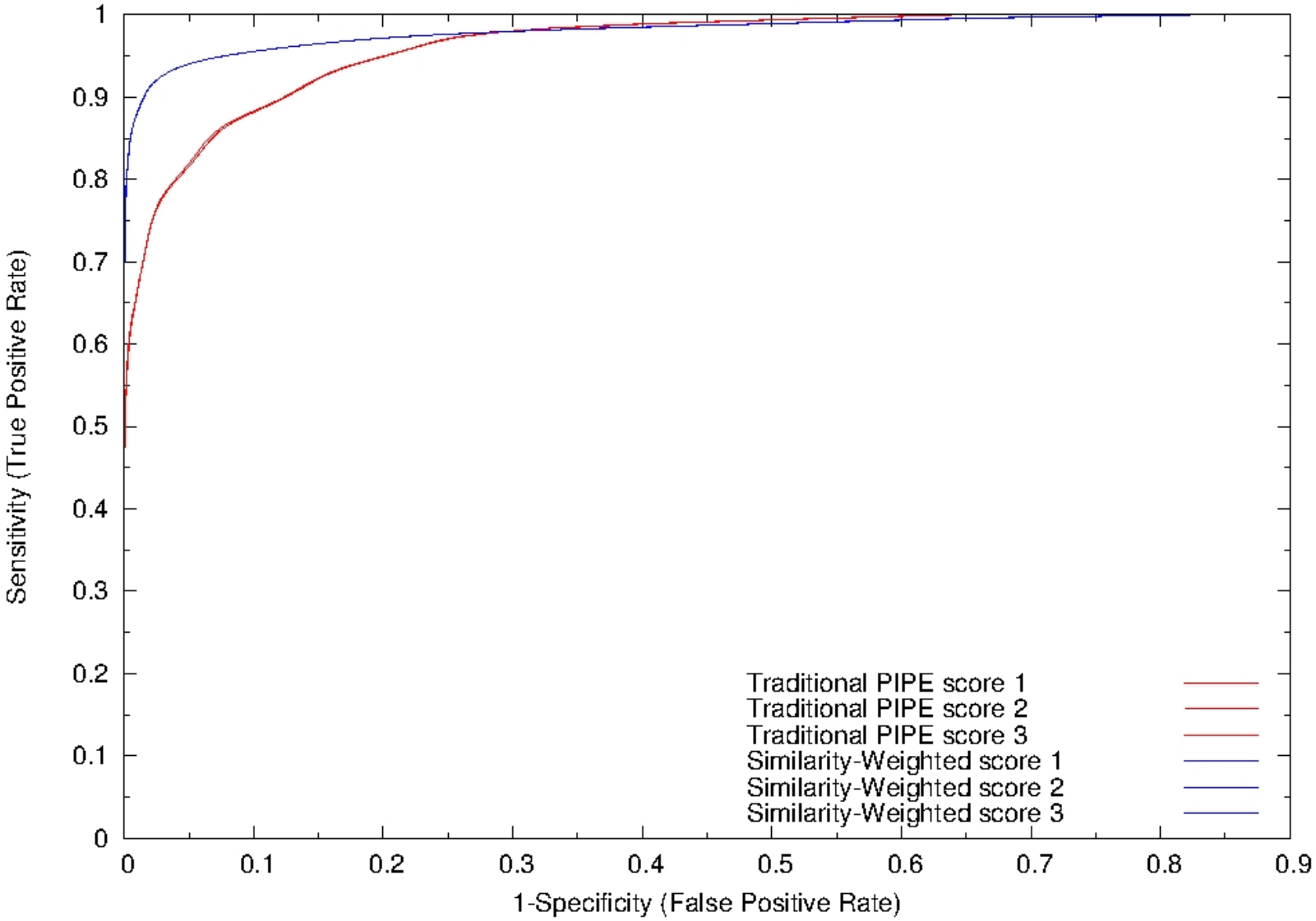


C.r. **G.p.** **V.c.**

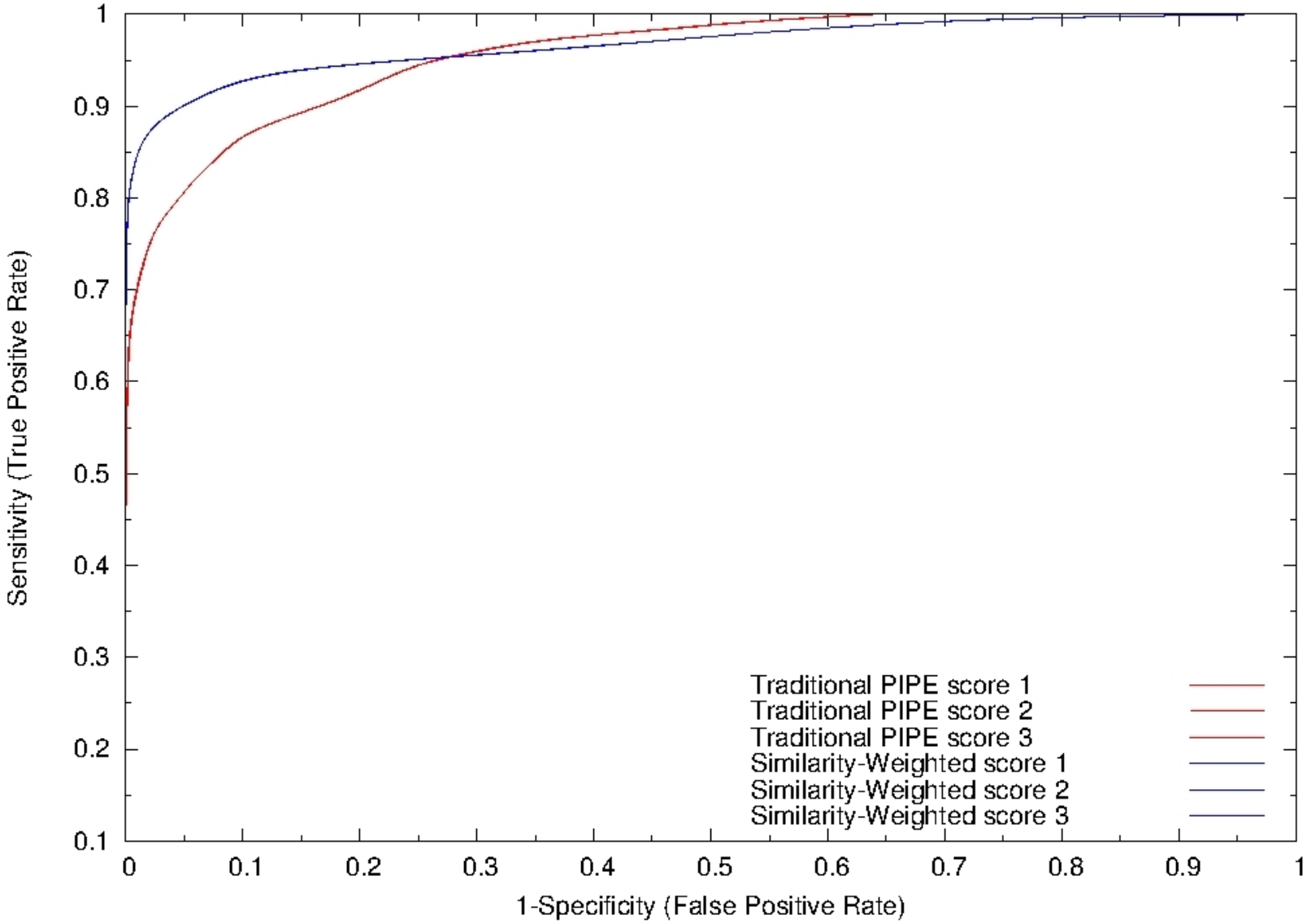
Volvox & Clamy Interactomes

- 8,885 mapped volvox and clamy proteins
- PIPE database: known arabidopsis interactions
- LOOCV test set:
  - POSITIVE: 509 arabidopsis interactions could be mapped into both chlamy and volvox
  - NEGATIVE: random pairs
- PIPE parameter tuning
  - Specificity: 99.95%
  - Sensitivity: 70%

PIPE parameter tuning: Volvox



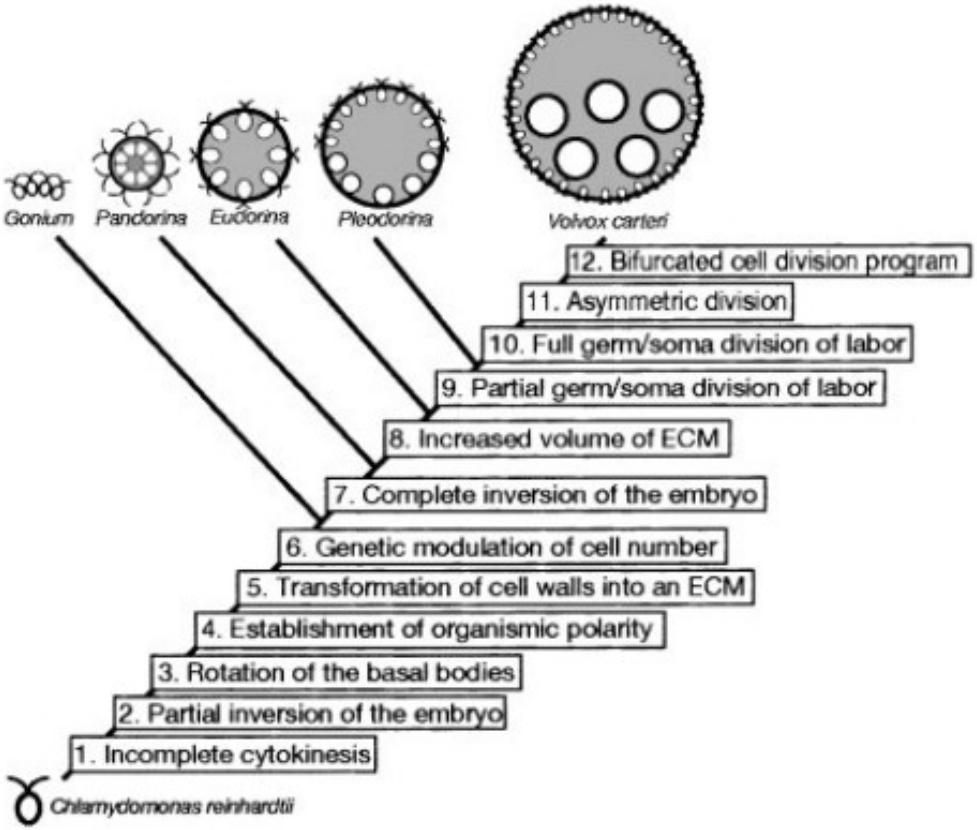
PIPE parameter tuning: Clamy



PIPE: Volvox & Clamy Interactomes

- 8.885 mapped volvox and clamy proteins
- PIPE parameter setting:
  - Sensitivity: 70%
  - Specificity: 99.95%
- Volvox interactome: 25,111 interactions
- Clamy interactome: 23,403 interactions

Comparison of Interactomes



- Compare networks around regA protein groups
- Compare networks around proteins of interest for evolutionary transitions;
  e.g. Kirk's 12 steps

Comparison of Interactomes

- Find functional units (“significant” interactome clusters) in volvox that are not present in clamy, and vice versa.
- Find pathways (“significant” interactome chains) in volvox that are not present in clamy, and vice versa.
- And more. Suggestions welcome.

Stay tuned...

SUMMARY

(1) PIPE can build high quality interactomes for species even with very little experimental PPI data available.

(2) Comparison of interactomes may provide new insights into evolutionary transitions. Work in progress...

Publications

Scientific Reports (Nature.com/srep), vol.2, art.239, 2012.

Short co-occurring polypeptide regions can predict global protein interaction maps

S.Pitre, M.Hooshyar, A.Schoenrock, B.Samanfar, M.Jessulat, J.R.Green, F.Dehne, A.Golshani

BMC Bioinformatics. 2011 Jun 2;12:225.

Highly accessed

Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences.

Amos-Binks A, Patulea C, Pitre S, Schoenrock A, Gui Y, Green JR, Golshani A, Dehne F.

Nucleic Acids Res. 2008 Aug;36(13):4286-94. Epub 2008 Jun 27.

Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences.

Pitre S, North C, Alamgir M, Jessulat M, Chan A, Luo X, Green JR, Dumontier M, Dehne F, Golshani A.

BMC Bioinformatics. 2006 Jul 27;7:365.

Highly accessed

PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.

Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, Gebbia M, Greenblatt J, Jessulat M, Krogan N, Luo X, Golshani A.