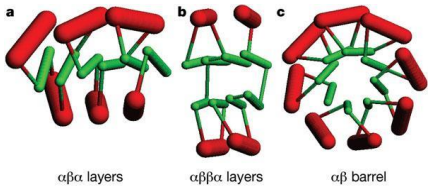
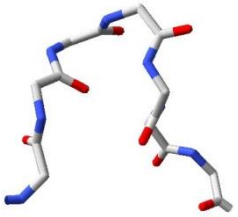
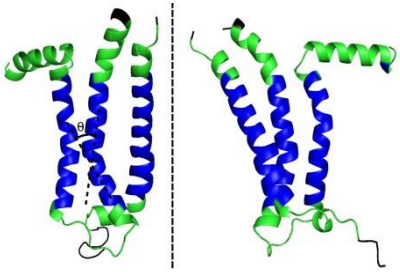
opig - Oxford Protein Informatics Group



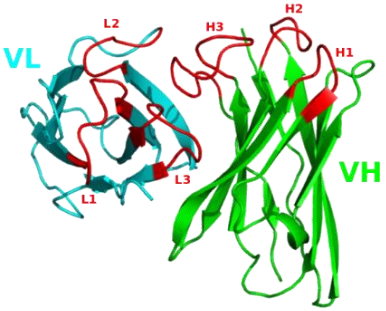
Fold space & structural evolution



Loop structure prediction



Membrane proteins



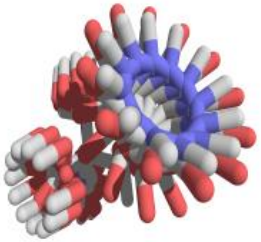
Antibodies

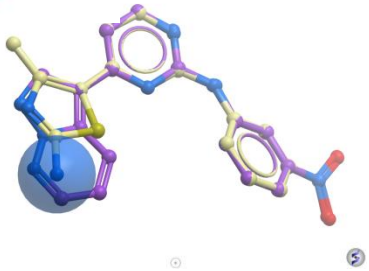
Protein Structure

Protein Evolution

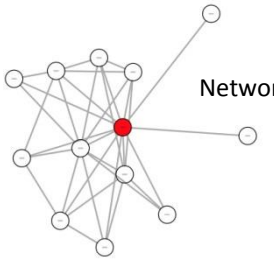
Protein-Protein interactions

Small molecules

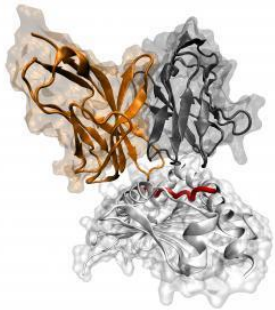




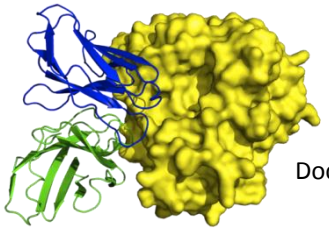
Fragment based screening



Network Evolution



Immunology – TCR and MHC



Docking

UNIVERSITY OF OXFORD

Extracting evolutionary and functional information from protein interaction networks

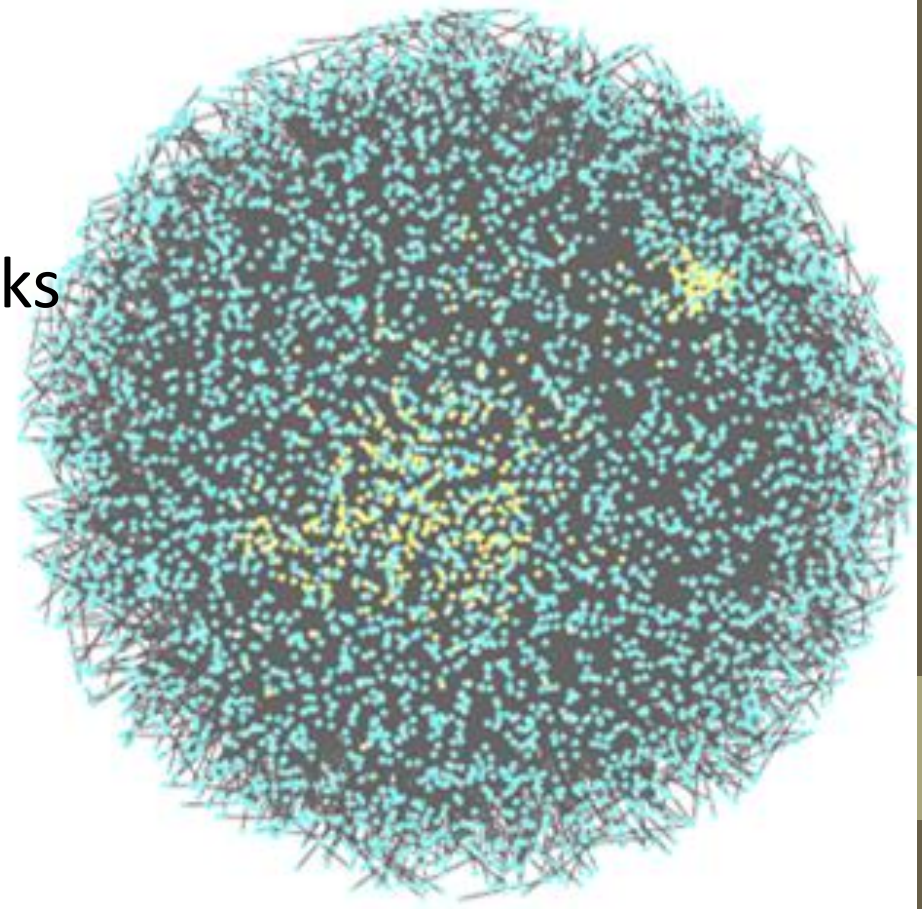
Charlotte Deane

Department of Statistics
University of Oxford
deane@stats.ox.ac.uk

Protein Interaction Networks (PINs)

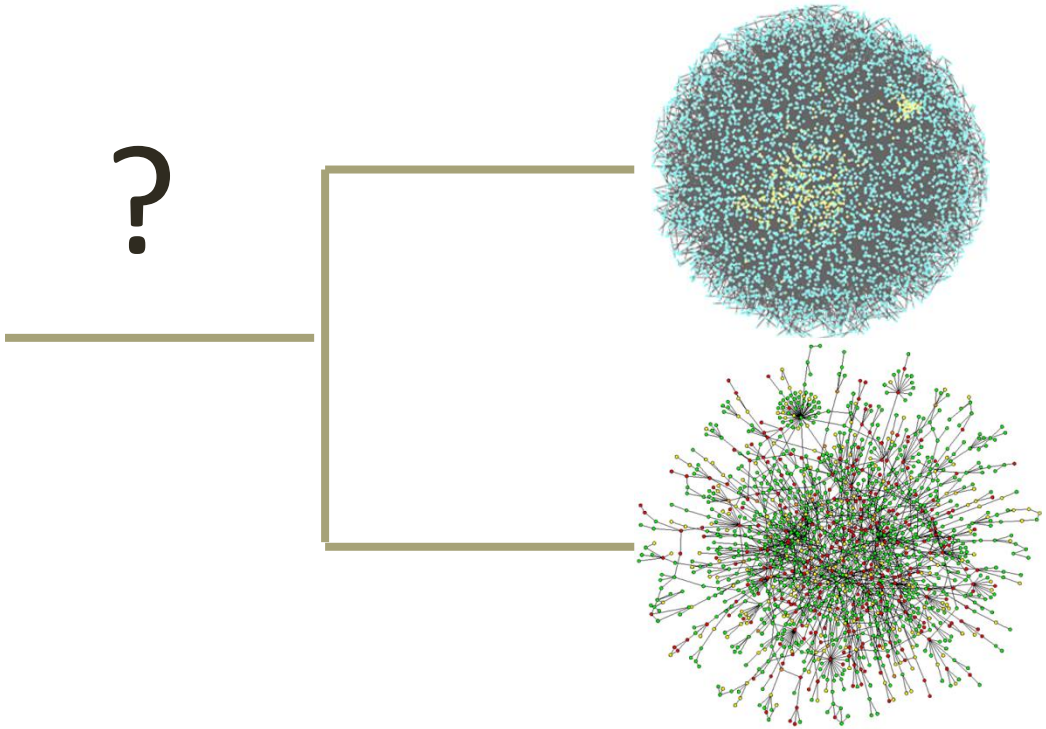
- Proteins
  - Large organic molecules
  - The main actors within the cell
  - Carry out duties encoded by genes

- Protein Interaction Networks
  - Proteins are nodes
  - Interactions are edges
  - undirected
  - Edges may/may not have weights



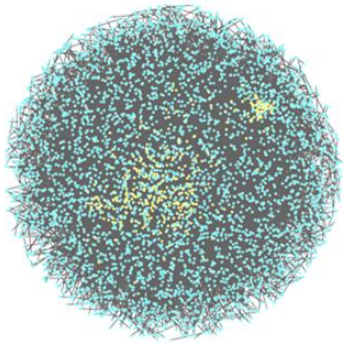
The Questions

- How do protein interaction networks evolve?
- How do we compare networks which are of different size and which may contain different vertices, yet which are related?

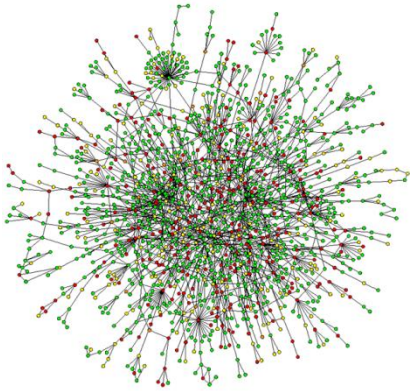


Comparing/Aligning networks

- Standard network comparison uses statistics that describe global properties of the network (e.g. Average degree, clustering coefficient, characteristic path length, diameter).
  - Not sensitive enough to be able to reconstruct phylogeny or shed light on evolutionary processes.



7.7



4.7

Comparing/Aligning networks

- Standard network alignment methods link specific nodes across two networks

- Identify “identical” nodes” using
  - local network similarity, sequence similarity (homologs)

- Usually computationally intensive.
- Not generally suitable for different network types.

Comparing protein interaction networks from different species

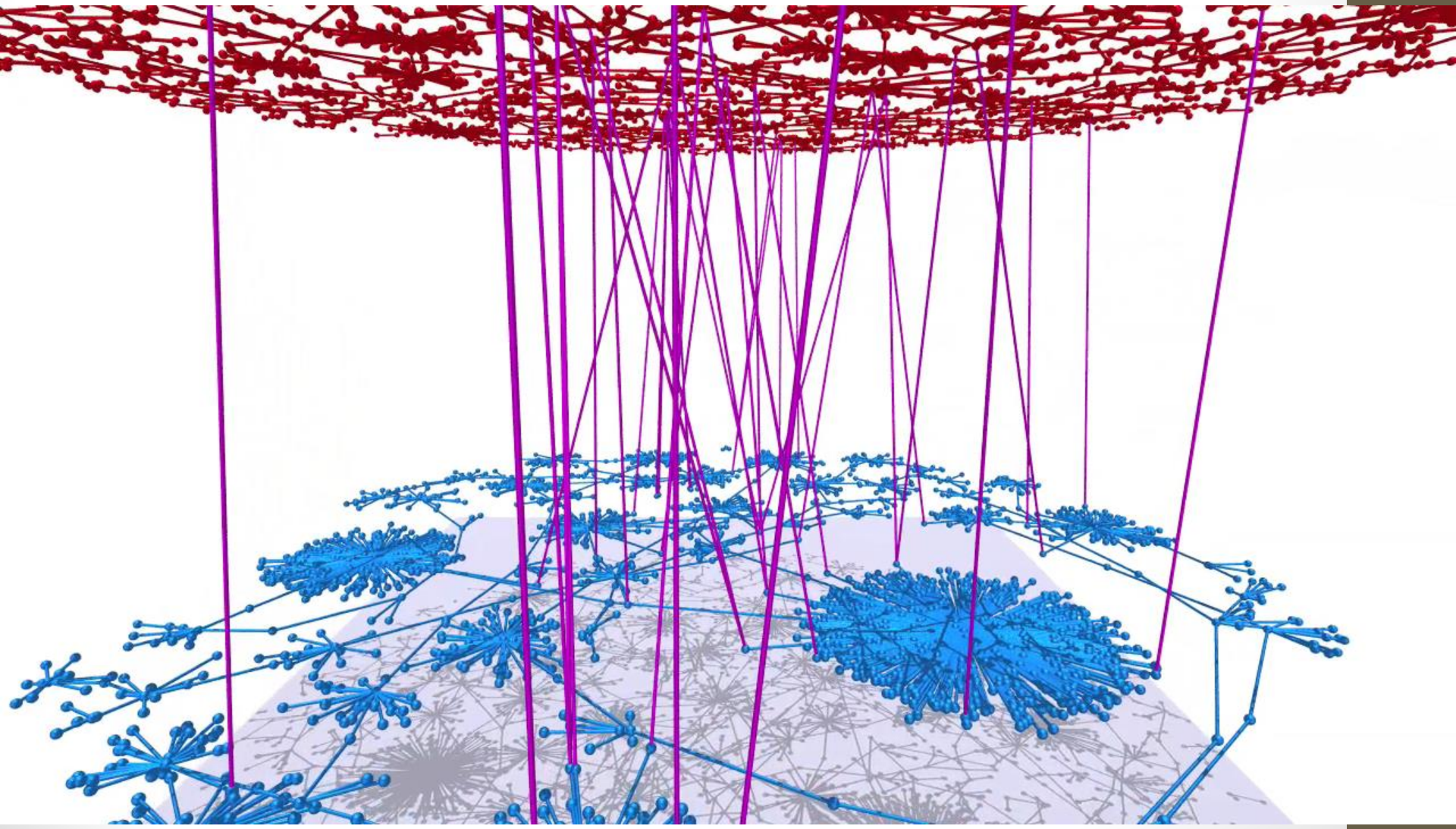
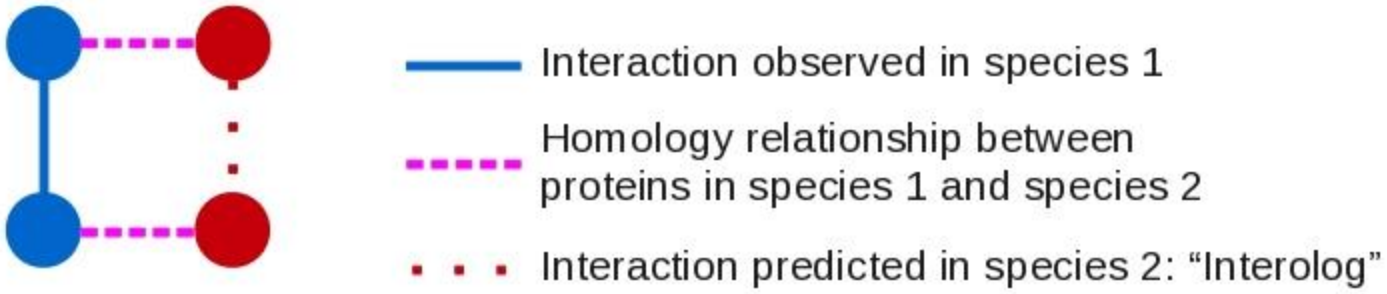


Image from Stumpf et al.

Interologs

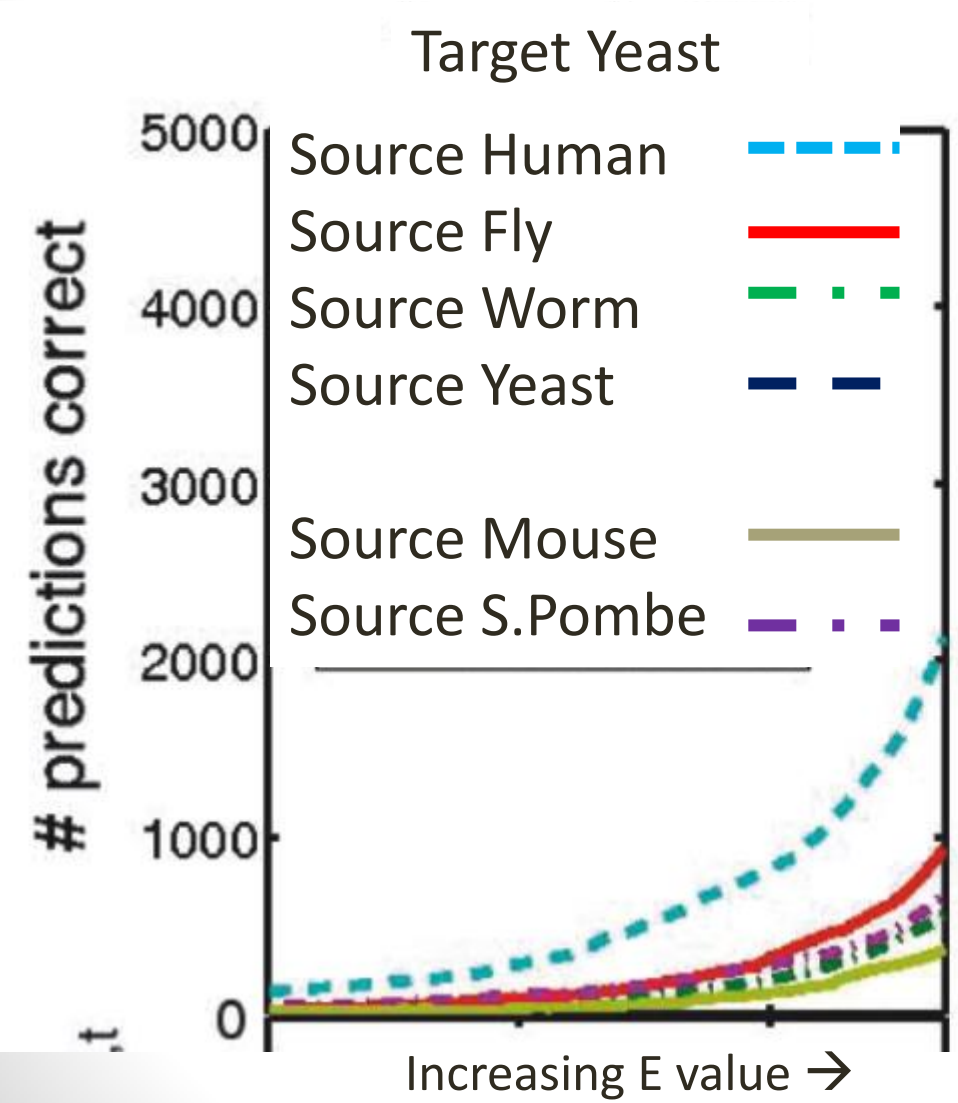


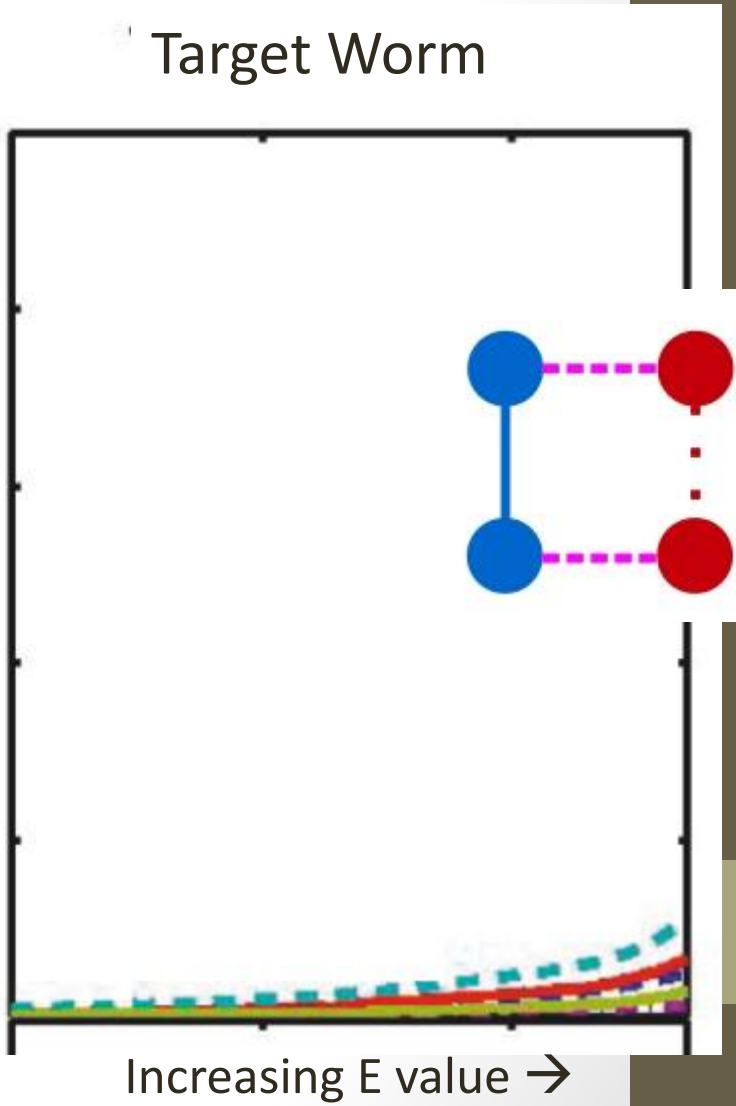
- Homology Measure
  - Sequence match between the proteins’ BLAST E-value

Protein interaction data

Species	Nodes (proteins)	Edges (interactions)	Number of proteins in Genome
Yeast (SC)	5782	44266	~6000
Fly (DM)	6514	20334	~13000
Human (HS)	9597	45695	~21000
Worm (CE)	3988	7275	~19500

Predict interactions via Interologs

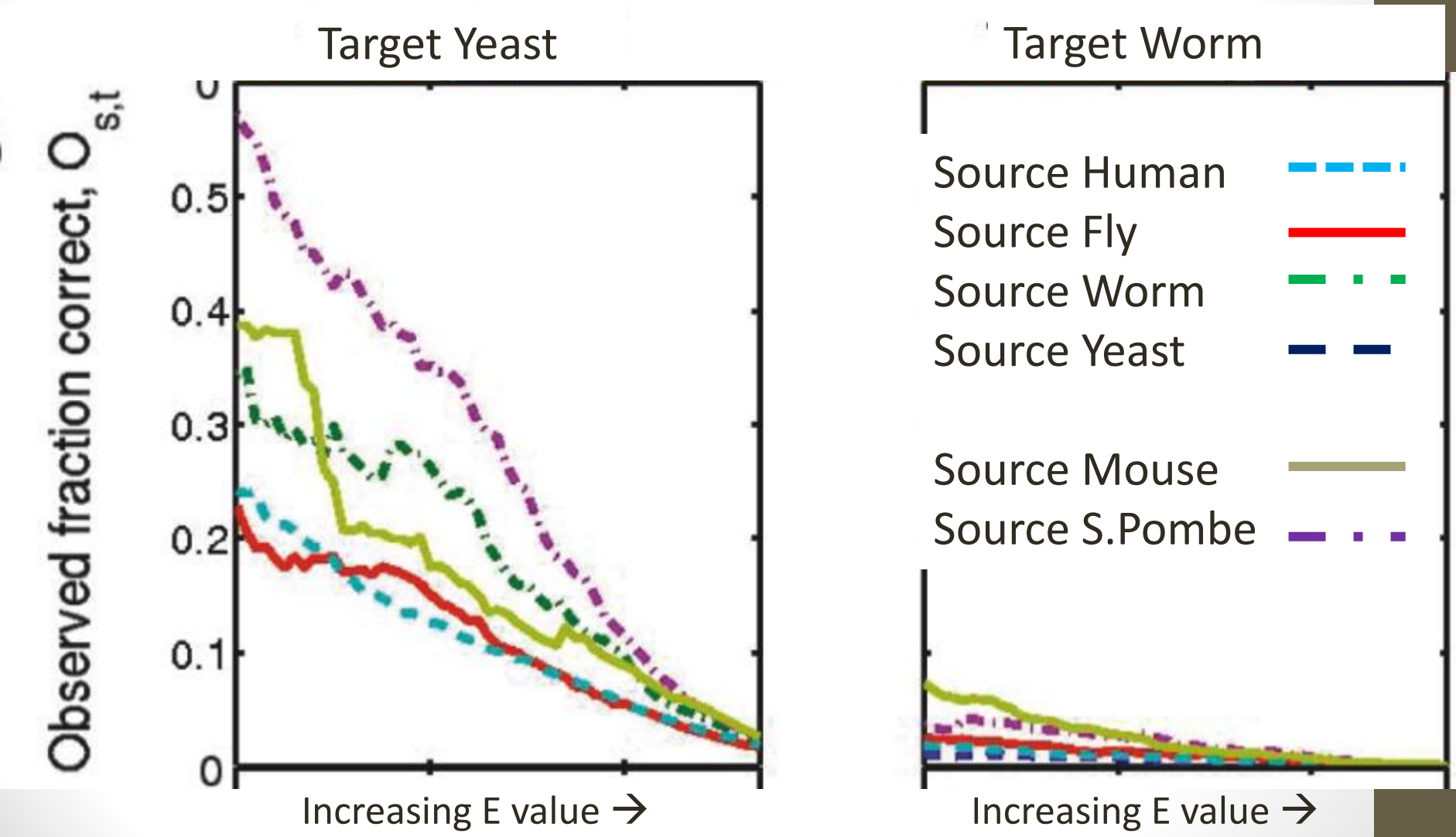




Number of interactions that are conserved (Interologs)

	Highly similar E-value 10^ -60	Similar E-value 10^-5
Fly (DM) to Yeast (SC)	19	464
Human (HS) to Yeast (SC)	141	1711

Fraction of correct interologs



Why don't interologs exist?

Species	Nodes (proteins)	Edges (interactions)	Number of proteins in Genome	Estimated number of interactions*
Yeast (SC)	5784	45045	~6000	35000-13500
Fly (DM)	6514	20334	~13000	71000 – 248000
Human (HS)	9597	45695	~21000	564000 - 722000

- Is it just network coverage?
- Is it just network error?

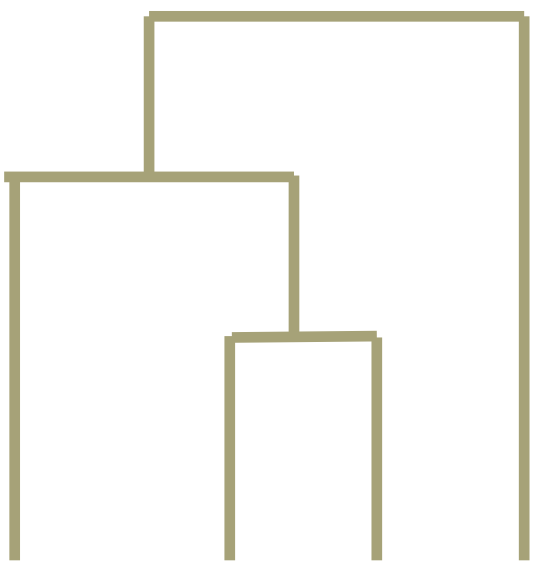
*From Stumpf et al 2008

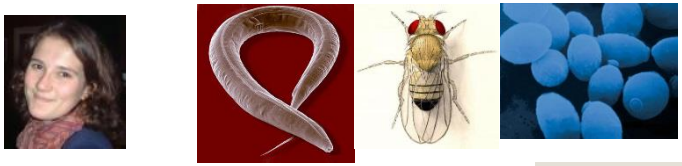
Coverage or error?

$$O(s,t) = E(s,t) * c(t)$$

- s – source species and t – target species
- O ~ fraction of inferred interactions observed to be correct
- E ~ fraction of inferred interactions estimated to be correct
- c ~ coverage of the target-species interactome

Coverage/error or evolutionary divergence?





161,000 343,000 110,000 45,000 ⇦ Estimated size of the networks

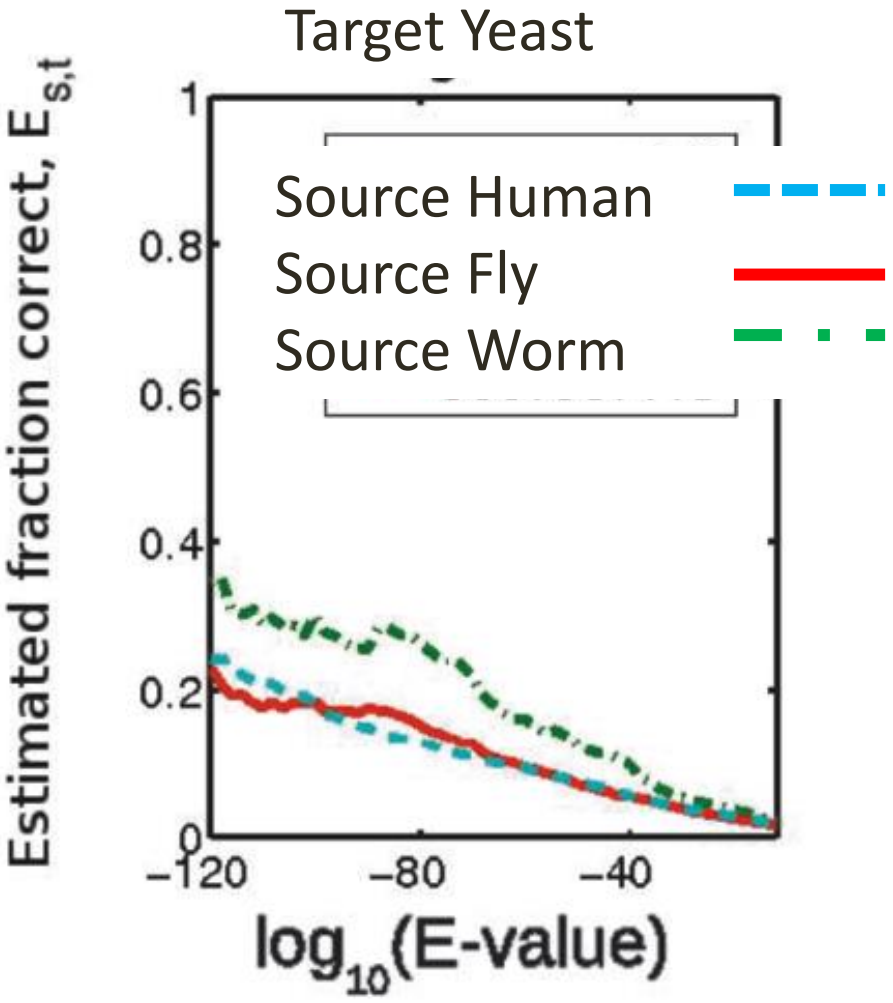
Solve O(s,t)=E(s,t)c(t) for each pair of species

Two assumptions
Yeast interactome is complete c(yeast) = 45,000

Fraction of conserved interactions between any species and Yeast is the same as from Yeast to that species E(Yeast, x) = E(x, Yeast)

Use E(s,t) to estimate a species tree

Fraction of correct predictions (taking error/coverage into account)



- Even taking into account coverage
- Interactions aren't conserved…..

Fraction of correct predictions (taking error into account)

- For example, to achieve a 50% success rate for transferring interactions between Yeast and Human at an E-value cut-off of 10^{-70} there would need to be over 400,000 interactions in Yeast and over two million in Human

- Remember there are only ~6000 proteins in yeast and

~21000 in human

Comparing Protein interaction networks

- Very few interactions are conserved between different species.

- Network alignment between species networks may therefore not be useful

- Develop a new methodology based on alignment free comparison
  - generally applicable for different network types.
  - Less computationally expensive

Based on alignment-free sequence comparison

- Alignment-free methods compare sequences through k-tuple content.
  - Designed to deal with large and/or noisy datasets
- For two sequences of letters R and S from an alphabet A (all finite), and for a word *w* of length k (k-tuple)
- Let X*w* and Y*w* be the centred number of occurrences of *w* in R and S, obtained by subtracting their expectations.
- Compare R and S through

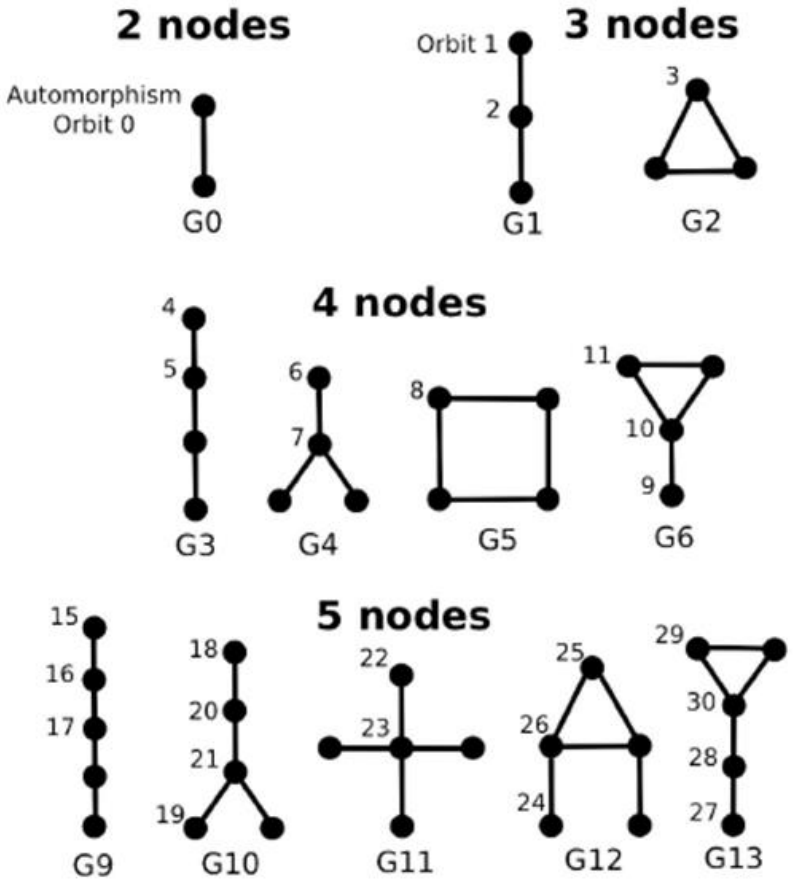
$$D_2^S = \sum_{w \in A^k} \frac{\tilde{X}_w \tilde{Y}_w}{\sqrt{\tilde{X}_w^2 + \tilde{Y}_w^2}}.$$

- Without subtracting the expectation the statistic would tend to measure single-sequence background noise.
- This has been applied to construct trees from sequence data.

Generalise to networks

- The obvious generalisation to networks is to replace length-k words by k-node sub graphs

K*
Complete sub graphs
Triangles
All shapes of *nodes



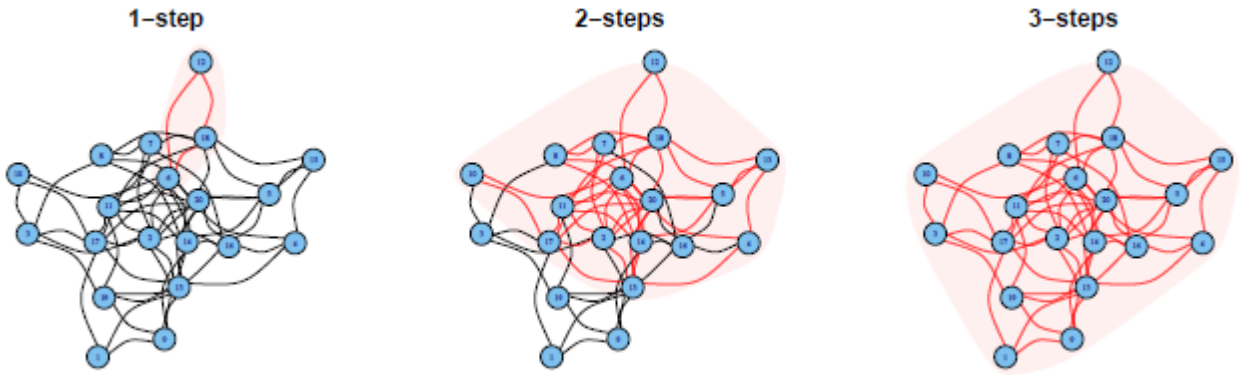
A null model for Protein interaction networks?

- Next need the background expectation of sub-graph counts:
  - Subgraph content of individual networks can be volatile (Rito et al. 2010, 2012).
  - For each species we only have one realisation of the network available

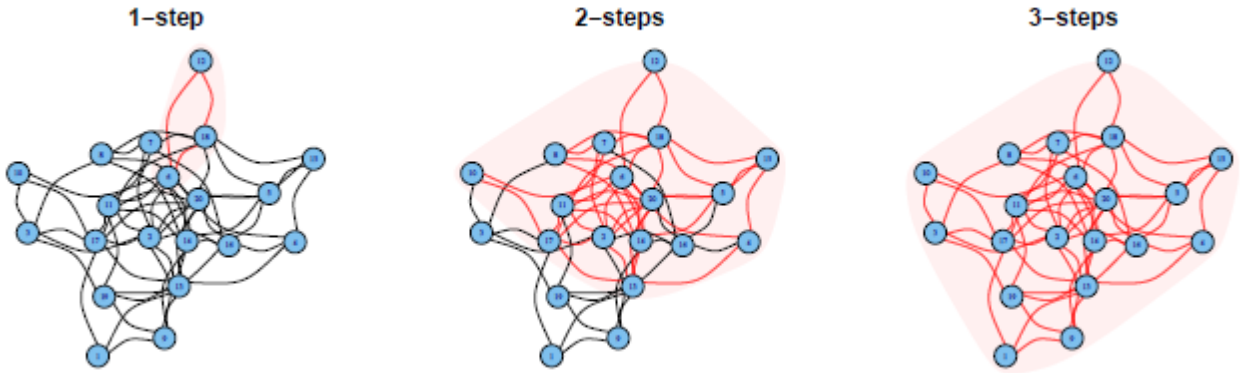
- Compare local protein neighbourhoods
  - Every network contains a large number of subnetworks
  - Use these local neighbourhoods, for better statistical behaviour

Ego-networks through snowball sampling

- A single node is picked in the network (the ego), then all nodes which are directly connected to it are picked, as well as the edges between them.
- The process can be extended to k-step ego networks

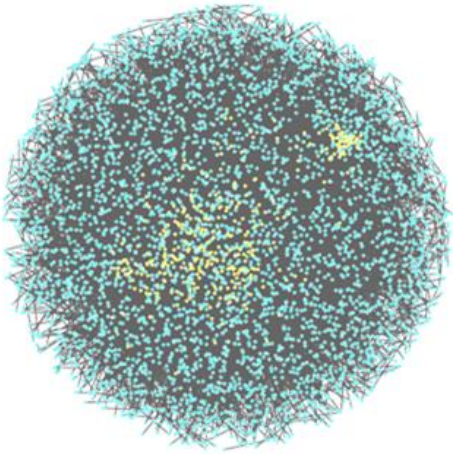


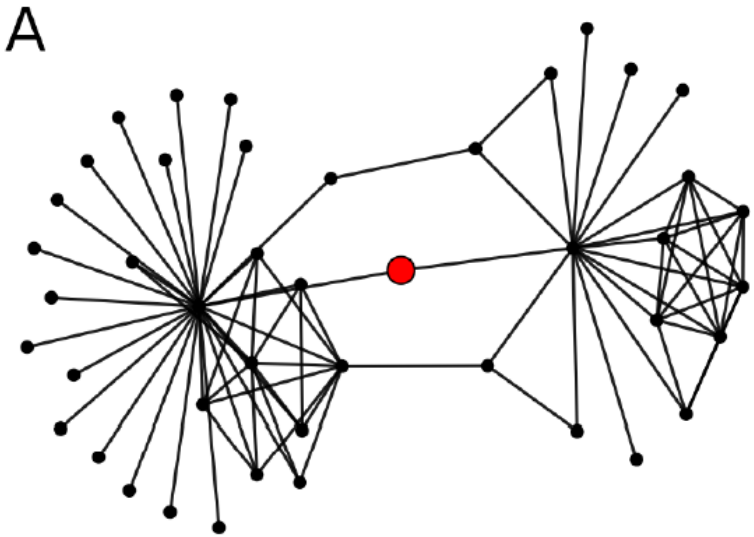
Ego-networks through snowball sampling

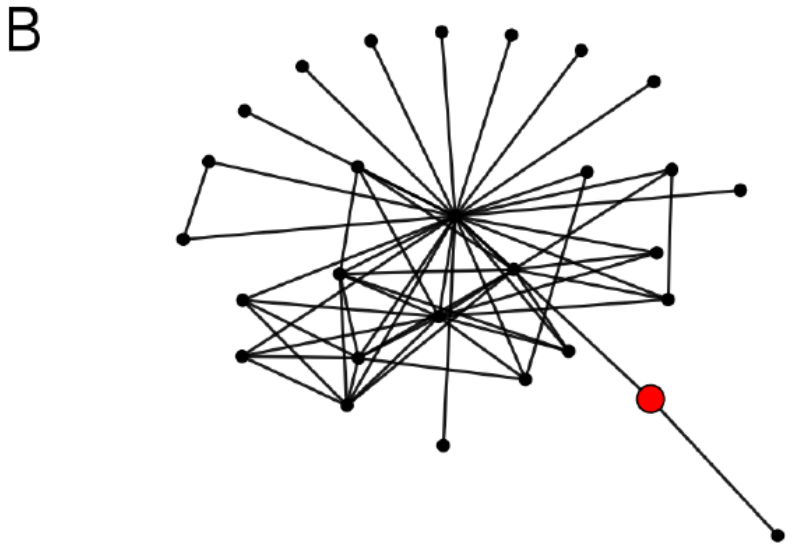


- From a single n-node network, this process generates an ensemble of n smaller sub-networks
- Two networks can be compared based on the subgraph content of their ego-network ensemble

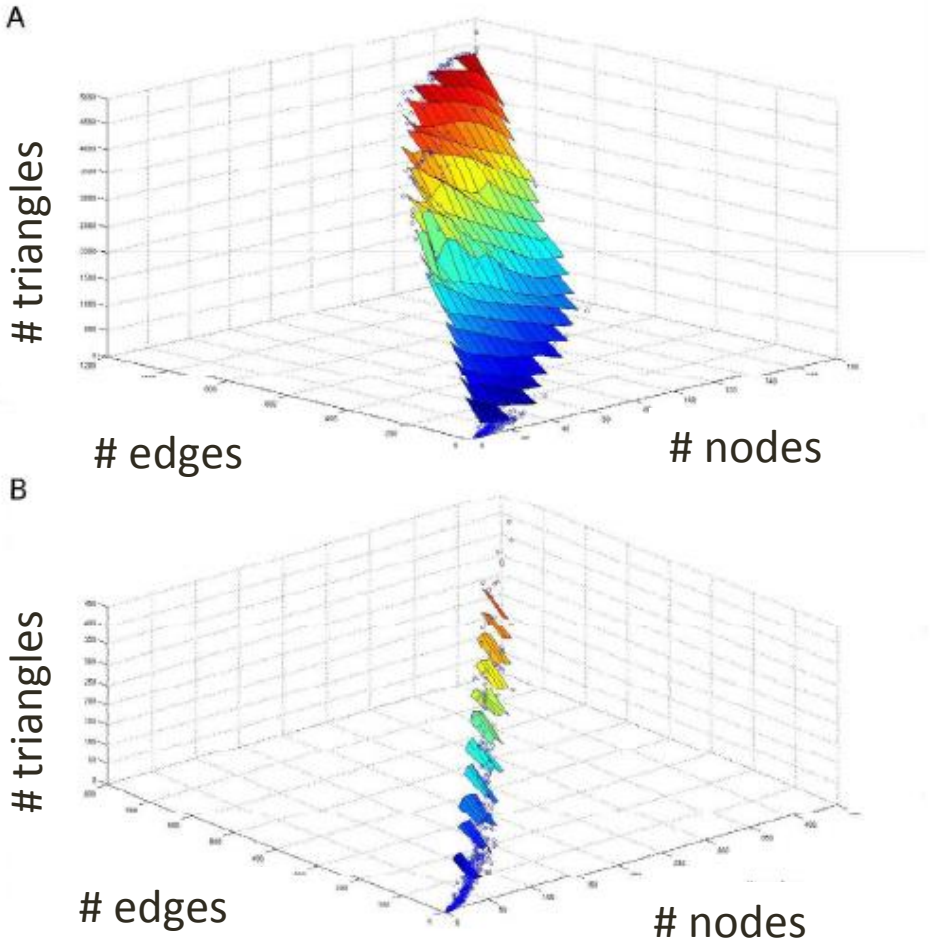
Two step ego networks from a PIN







Protein interaction networks have a rich ego network space



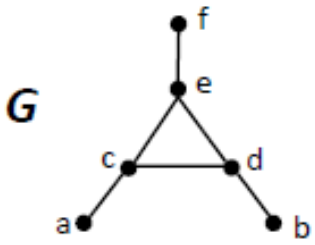
- Ego-networks from
  - (A) Yeast
  - (B) a rewired Yeast network.
- The surface from the real data is much richer than the one resulting from re-wiring.
- Random graphs give an even poorer surface.

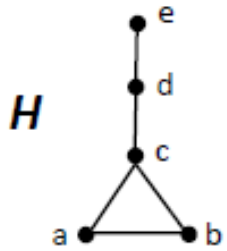
The ensemble of Ego-networks

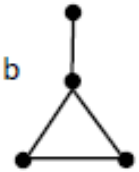
- A Protein Interaction network of 5000 nodes gives rise to 5000 (possibly overlapping) ego-networks.
- Compare two networks through the small graph counts in their ensembles of ego-networks.
- Example all small graphs on 3 nodes.
  - For each of the protein ego-networks we count the number of occurrences of each of the 2 possible small graphs on 3 nodes.
  - We estimate the expected small graph counts from the ensemble of ego-networks with similar density from a gold-standard network.
  - Then for each small graph w on 3 nodes we calculate

$$S_w(G) = \sum_{i \text{protein in } G} \left(\text{No. of } w \text{ in ego-network of } i - \text{expected no.} \right)$$

Subgraph counts in ego-networks





Ego_a		ρ = 2/3	= 2	= 1	
Ego_b		ρ = 2/3	= 2	= 1	
Ego_f		ρ = 2/3	= 2	= 1	
Ego_c, d, e	*G*	ρ = 2/5	= 6	= 1	
Ego_a, b		ρ = 2/3	= 2	= 1	
Ego_e		ρ = 2/3	= 1	= 0	
Ego_c, d	*H*	ρ = 1/2	= 3	= 1	

Background expectation of counts

- Estimate expectations from a gold-standard network.
  - Bin the ego-networks of the gold-standard by their densities.
  - Average the counts of a sub-graph in a density bin to get the expected count.

- This density-specific estimated expected count serves as our background expectation.
- The expectation for query ego-networks are retrieved from the relevant density bin

Netdis: Comparing two networks

For two networks G and H, we define three statistics by

$$netD_2^S(k) = \frac{1}{M(k)} \sum_{w \in A(k)} \left(\frac{S_w(G) S_w(H)}{\sqrt{S_w(G)^2 + S_w(H)^2}} \right), \quad k = 3, 4, 5,$$

where $M(k)$ is a normalising constant so that $netD_2^S(k) \in [-1, 1]$.

The corresponding distance statistic, which we call *Netdis*, is defined as

$$netd_2^S(k) = \frac{1}{2}(1 - netD_2^S(k)) \in [0, 1].$$

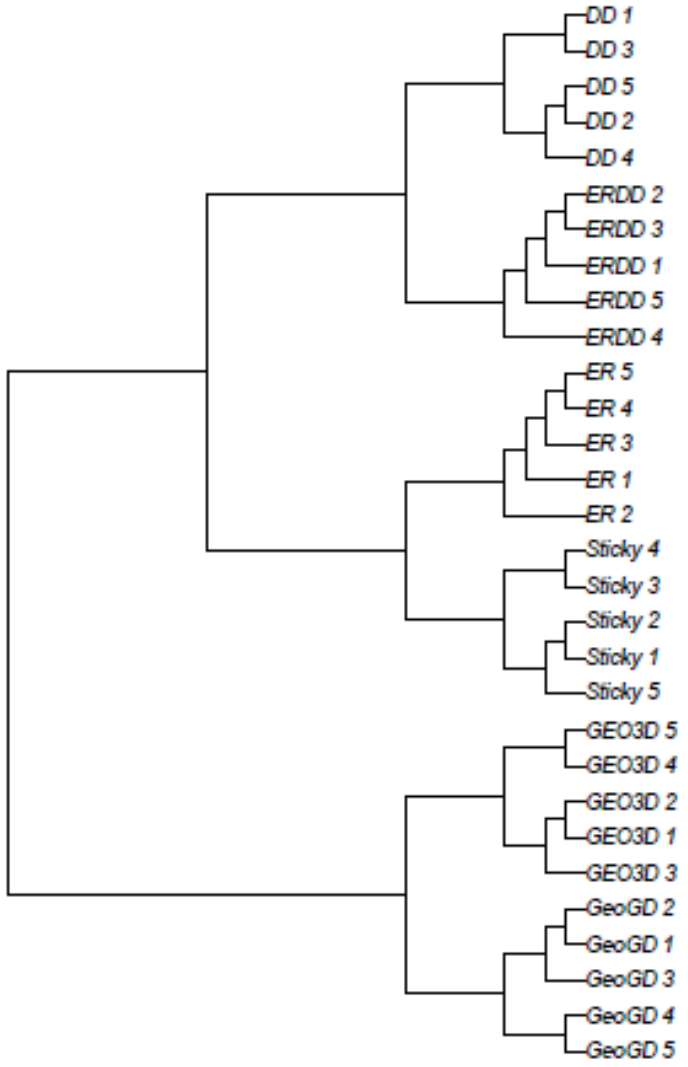
This distance is used to build the distance matrix for all query networks.

Results

- Pair-wise Netdis values from a set of networks can be used to generate a distance matrix.
- Distance matrices can be used in existing tree-building methods to cluster networks by similarity.
- All results with DIP [Salwinski et al.(2004)] yeast core network as the gold standard.

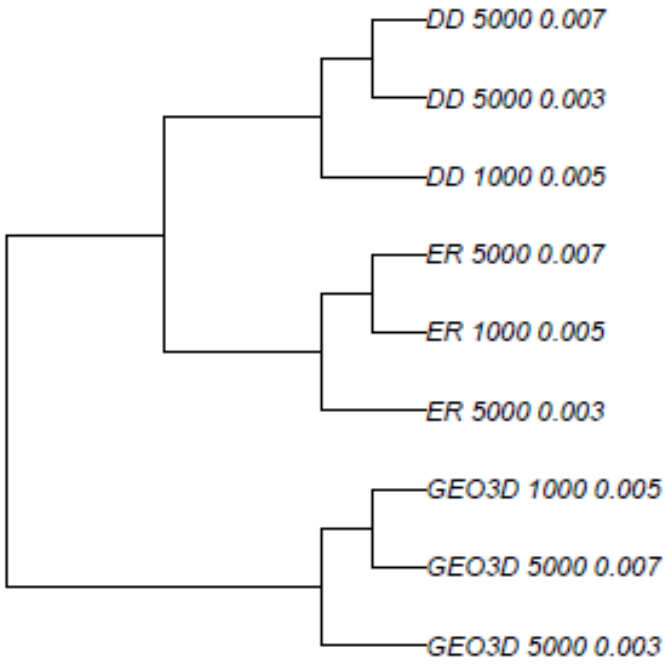
Simulated data - 1

- Five networks simulated from each of six models.
- Parameter choice: Simulated networks closely match DIP-yeast network.
- The networks are clustered perfectly by model type.
- In this case, dropping the expectation from Netdis performs equally well.



Simulated data - 2

- Models can be distinguished despite introducing highly variable network size and density.
- This would not be possible using raw sub-graph counts.
- Removal of the normalization from Netdis fails to generate the correct clustering.



Protein interaction networks

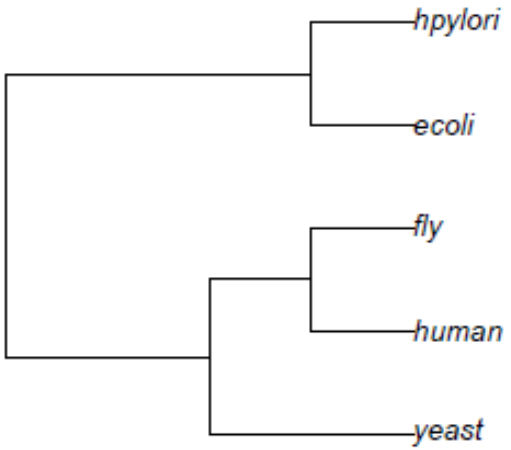
- Model species having at least 500 interactions.

Species	# Genes	Nodes	Edges	Coverage	density $\times 1000$
Hsap	21,224	9,223	36,631	43.9	0.8
Dmel	13,917	7,565	22,800	54.3	0.8
Scer	6,692	5,078	22,103	86.2	1.7
Ecoli	4,303	2,968	11,604	68.9	2.6
Hpyl	1,553	714	1,361	45.9	5.3

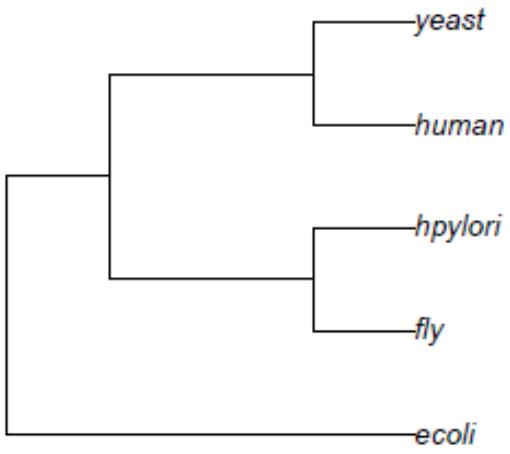
Protein Interaction Networks

- Netdis obtains correct tree (A) with fly next to human and yeast, and the two bacterial networks in a separate clade.
- Removing background expectation and/or normalization results in an incorrect tree (B).

(A) ✓



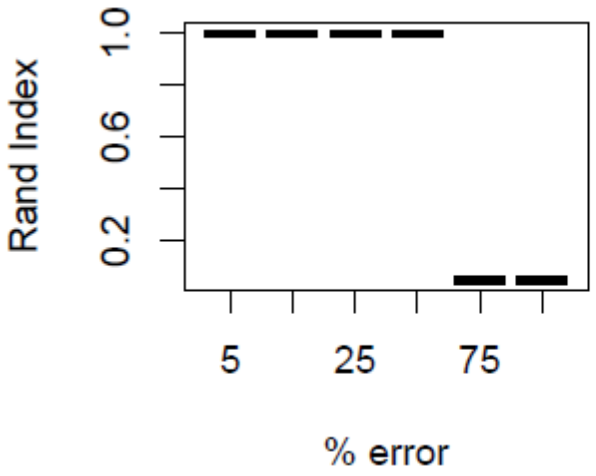
(B) ✗



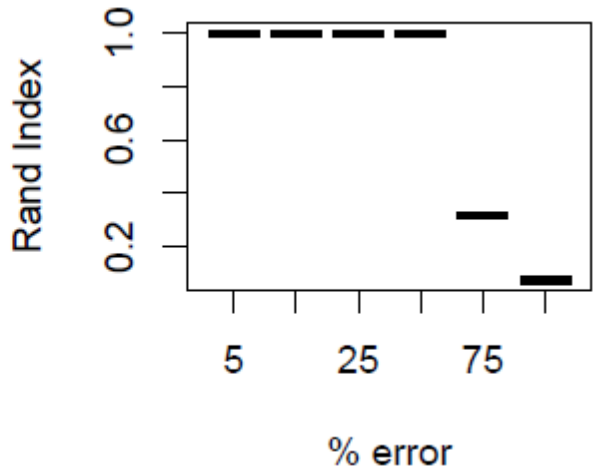
Robustness of error

- The method is robust to random error in the networks.
  - Until false positive and negative rate >50% get correct clustering

Protein Interaction networks



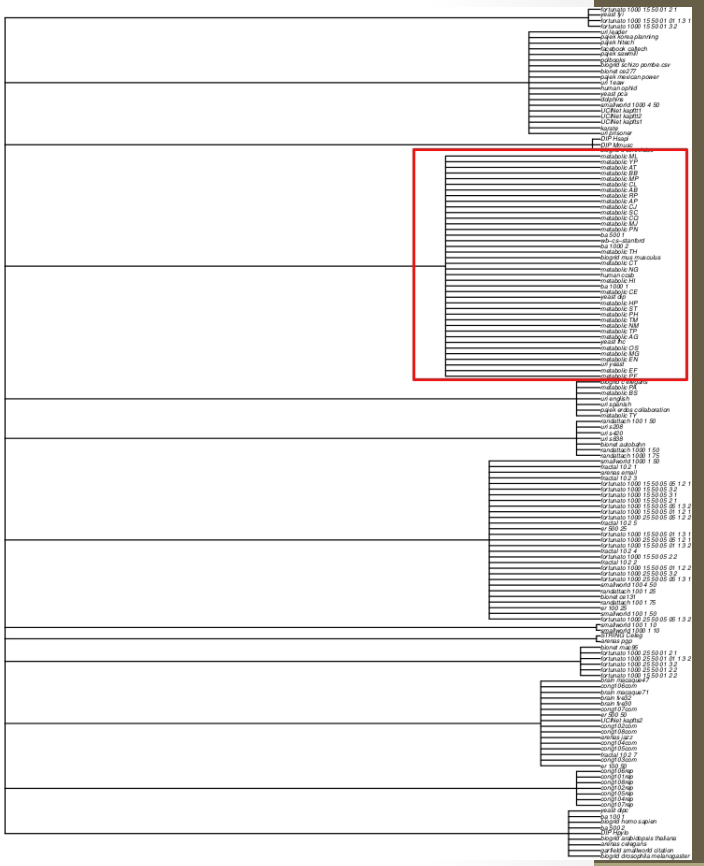
Simulated networks



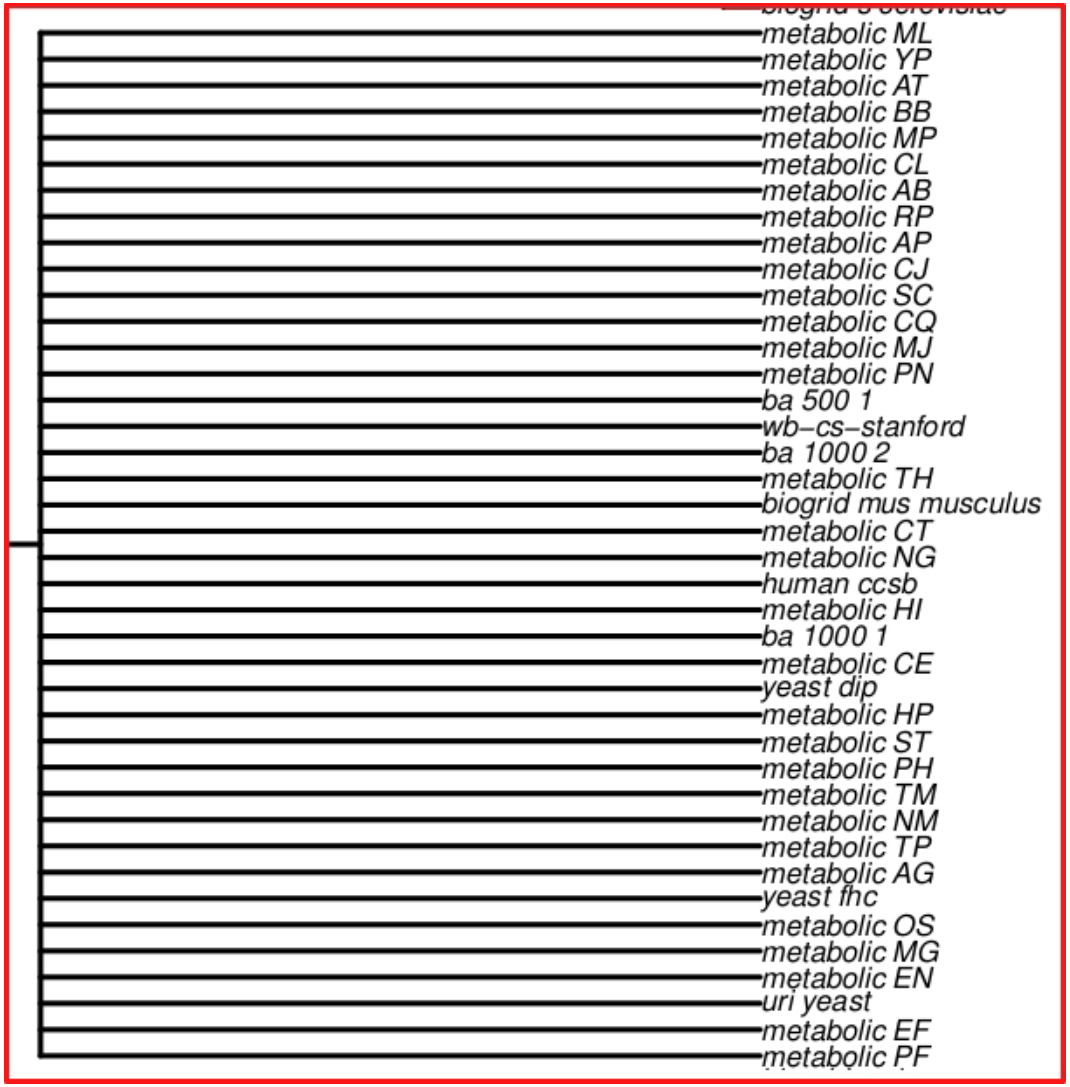
Diverse networks

- 151 networks from recent study [Onnela et al.(2012] were grouped into 13 clusters manually based on type.

- The best clustering is generated by Netdis with expectation and normalization.
- Using raw sub-graph counts leads to a clustering not significantly better than random.



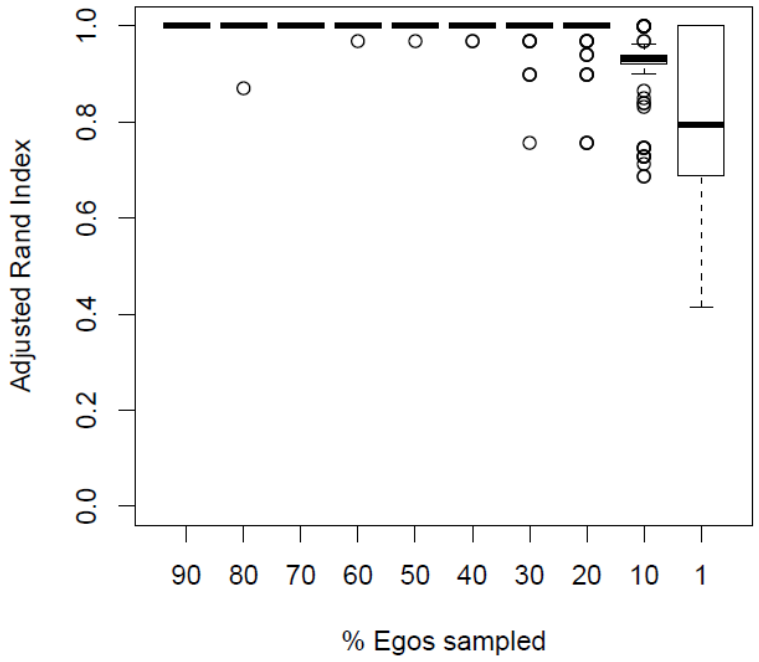
Diverse networks



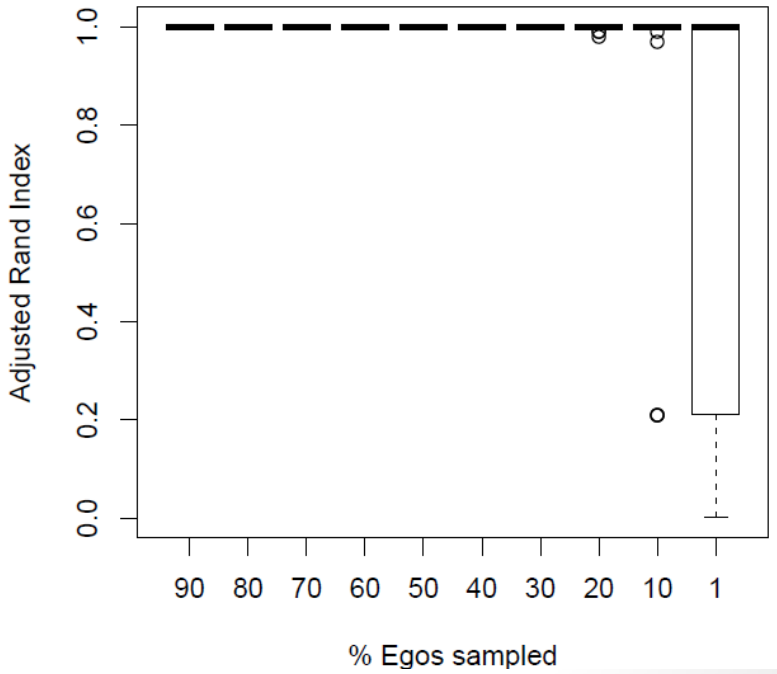
Scaling up

- One potential bottle-neck is the need to analyse all ego-networks.
- Random sampling of ego-networks faster for very large networks
- Also negates the need to know the entire network

Simulated networks



Protein Interaction networks



Conclusions

- Netdis is a fast and scalable alternative to alignment where a quantitative measure of network similarity is sought.
- As only network data is used, many types of network data can be analysed together.
- The method can be also be used to test competing models for a particular network.

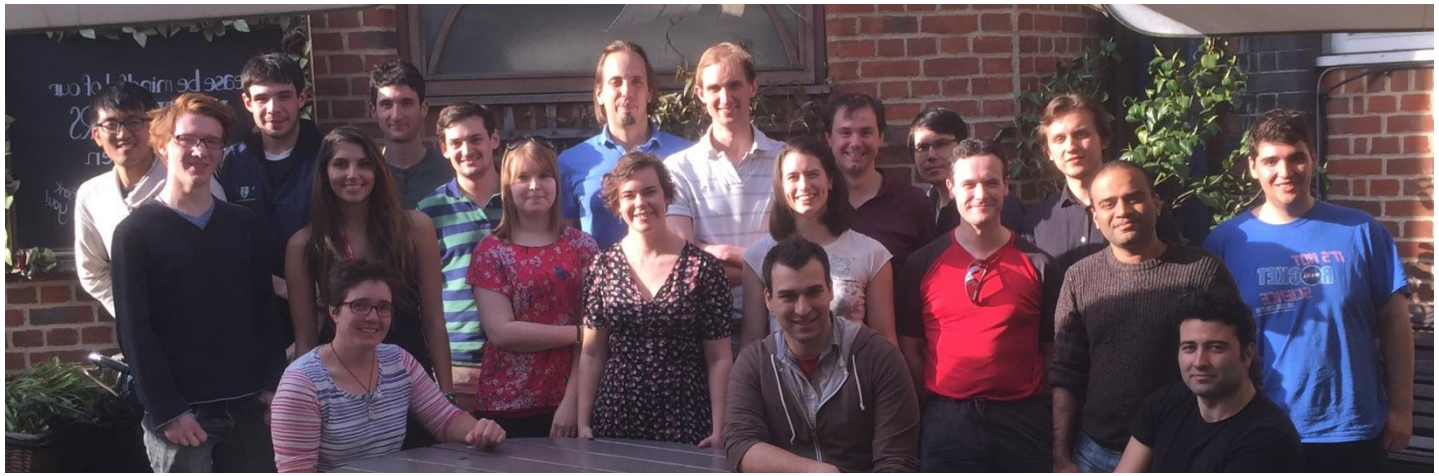
- In terms of Proteins
  - The underlying assumption for Netdis is that species that are more related will on average share more protein interaction network neighbourhoods which are topologically similar than unrelated species do.

  - The interaction neighbourhoods may play a crucial role in the evolution of proteins.

Future directions

- Developing a sample based version of Netdis
- Calculating a more robust background expectations of sub-graph counts.
- Extend to directed networks

ACKNOWLEDGEMENTS

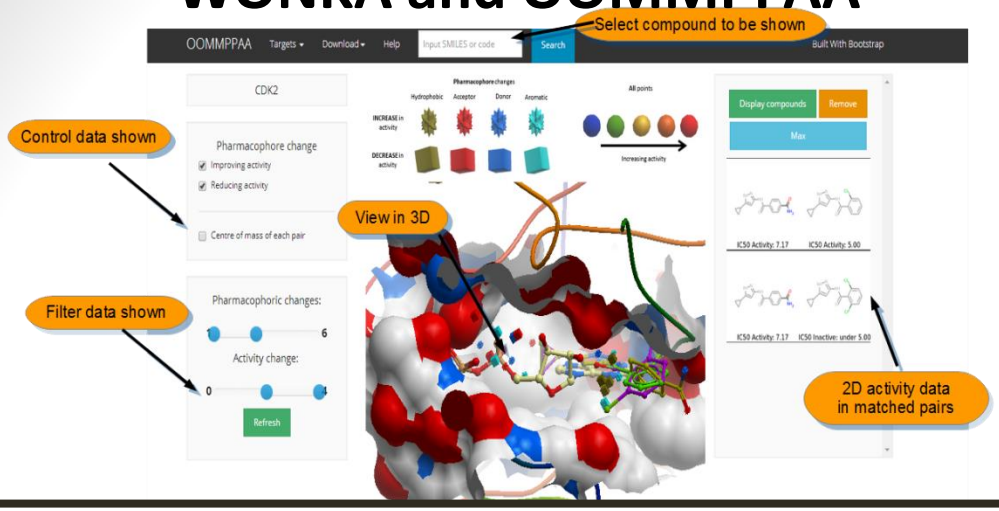


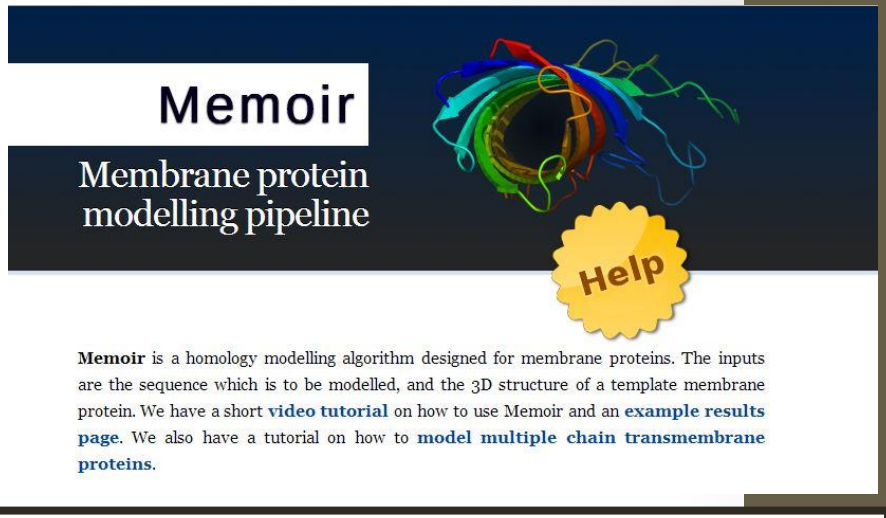
Waqar Ali
Malte Lucken

Gesine Reinert
Mason Porter

Anna Lewis
Tiago Rito
Fengzhu Sun (USC, Los Angeles)
Nick Jones (Imperial)

WONKA and OOMMPPAA





http://www.stats.ox.ac.uk/proteins/resources

NETDIS