Models for protein-protein interaction networks: what do age and (large) degree tell us?

Gesine Reinert

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

Kavli Centre, May 28-29, 2012

Outline

Joint work with Tiago Rito and Charlotte M. Deane, Oxford

Proteins

Proteins are

- large and complex organic molecules built of units called amino acids
- very versatile
- serve crucial functions in most biological processes

Most proteins function through **interactions** with other molecules, and often these are other proteins. These interactions are physical contacts.
Proteins are seen to form functional modules: cellular functions are carried out by modules consisting of a few interacting proteins.

Interaction detection

Experimental techniques for interaction detection include

- Crystalised complex: low through-put
- Co-immunoprecipitation: low through-put
- Yeast 2 Hybrid Assay: high through-put
- Purification of complex followed by Mass Spectrometry; TAP-MS: high through-put

The high through-put methods have a high false-positive and a high false-negative rate. Error rate estimates range from 20 to 70 % in Saeed and Deane, 2006. Newer datasets have potentially less error.

Available datasets

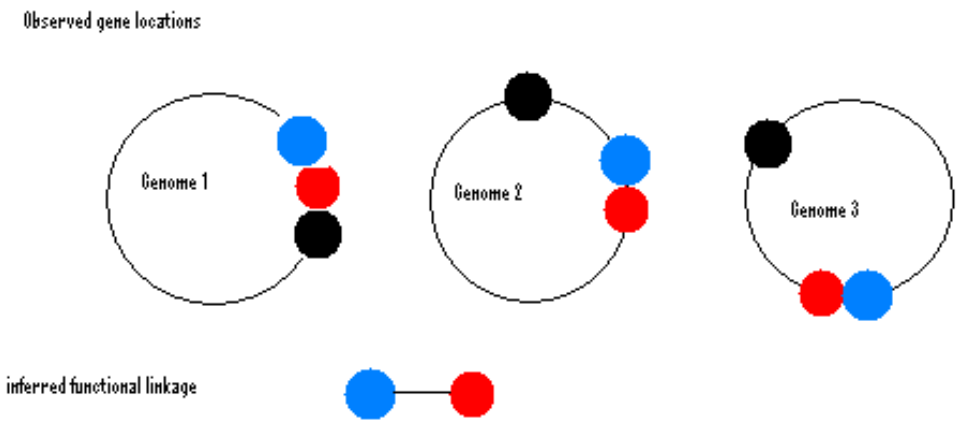
Protein interactions have been collected in a number of datasets, which have partial overlap:

- BIND Biomolecular Interaction Network Database (Bader et al. 2001)
- DIP Database of Interacting Proteins: about 23,000 Yeast interactions (Xenarios et al. 2002)
- GRID General Repository for Interaction Datasets (Breitkreutz et al. 2003)
- MINT Molecular INteractions Database (Zanzoni et al. 2002)
- HPRD Human Protein Reference Database: about 35,000 Human interactions (Keshava Prasad et al. 2009)

Predicting interactions

Interactions are also predicted or inferred, based on other biological data. Methods include

Gene Neighbour: If two genes are bound to be neighbours in several different genomes, then a functional linkage is inferred.



Rosetta Stone/ Gene Fusion

Some pairs of interacting proteins have homologs in other organisms which are fused to a single protein chain. This fused protein is called the fused domain or Rosetta Stone sequence. The two matches are functionally linked.

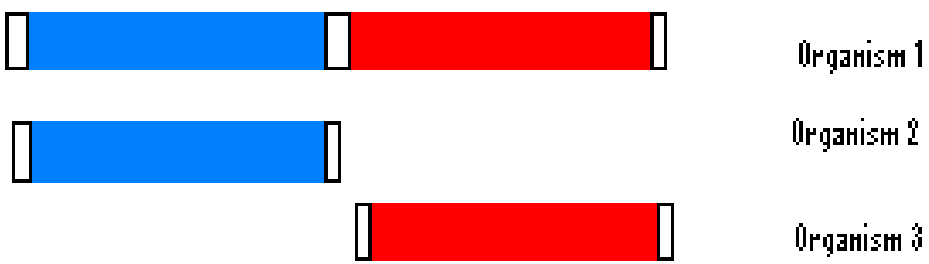


Figure: A sketch of the Rosetta Stone method

Phylogenetic profile

The phylogenetic profile has as number of entries the number of genomes which have been sequenced, and in each genome it is recorded whether or not the proteins in question are present. The proteins are then clustered based on the similarity of their phylogenetic profiles. See the following example:

EC	SC	BB	HP
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	0	1	1
P6	1	1	0

P2 and P6 are inferred to be functionally linked; P3 and P5 are inferred to be functionally linked.

Protein interaction networks

In protein interaction networks (PINs) proteins are nodes, interactions are edges, edges are undirected and may or may not have weights. Here are some global summaries of PINs:

Statistic	Yeast DIP	Human HPRD
Nodes	4823	12937
Edges	17471	43496
Avg. degree	6.10	6.72
Avg. Clustering Coeff.	0.1283	0.1419
Avg. Shortest Path	4.14	4.40

A typical hairball picture

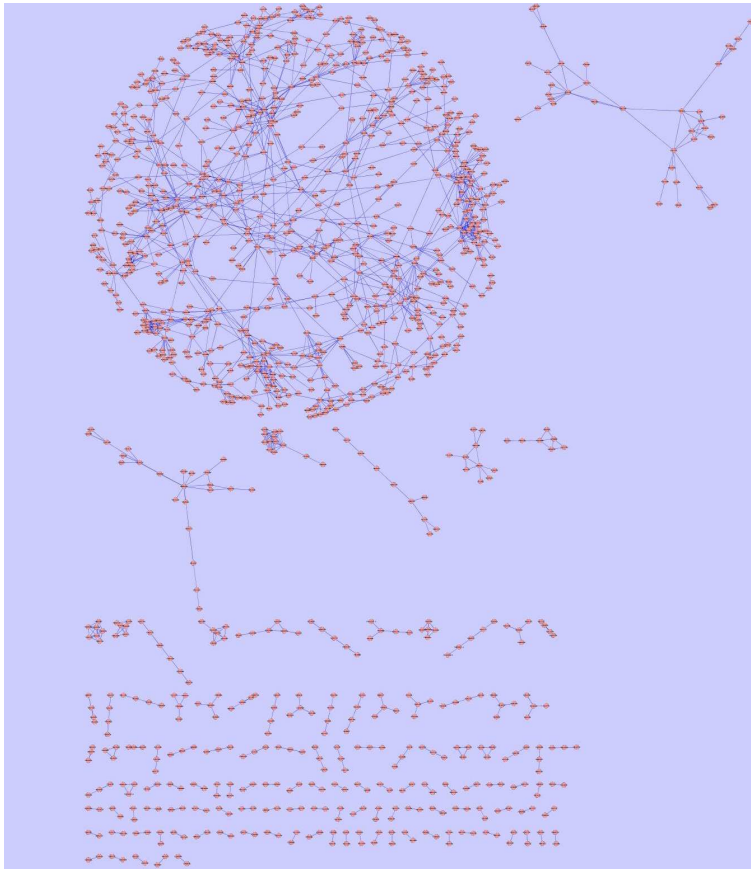


Figure: Yeast protein-protein interactions

Models for PINs

Motivated by functional motifs, we would like to find a statistical model for PINs which is suitable for small subgraph counts.
Here is a list of candidate small subgraphs, see *Przulj 2006*, where they are called *graphlets*. In *Przulj et al. 2004* their counts (and orbits) are used to construct a network comparison score called Graphlet Distribution Degree Agreement (GDDA).

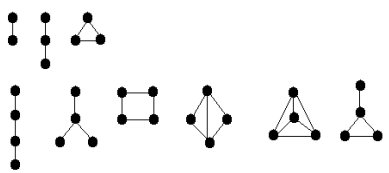


Figure: Small subgraphs on 2, 3 and 4 vertices

GDDA and Threshold behaviour

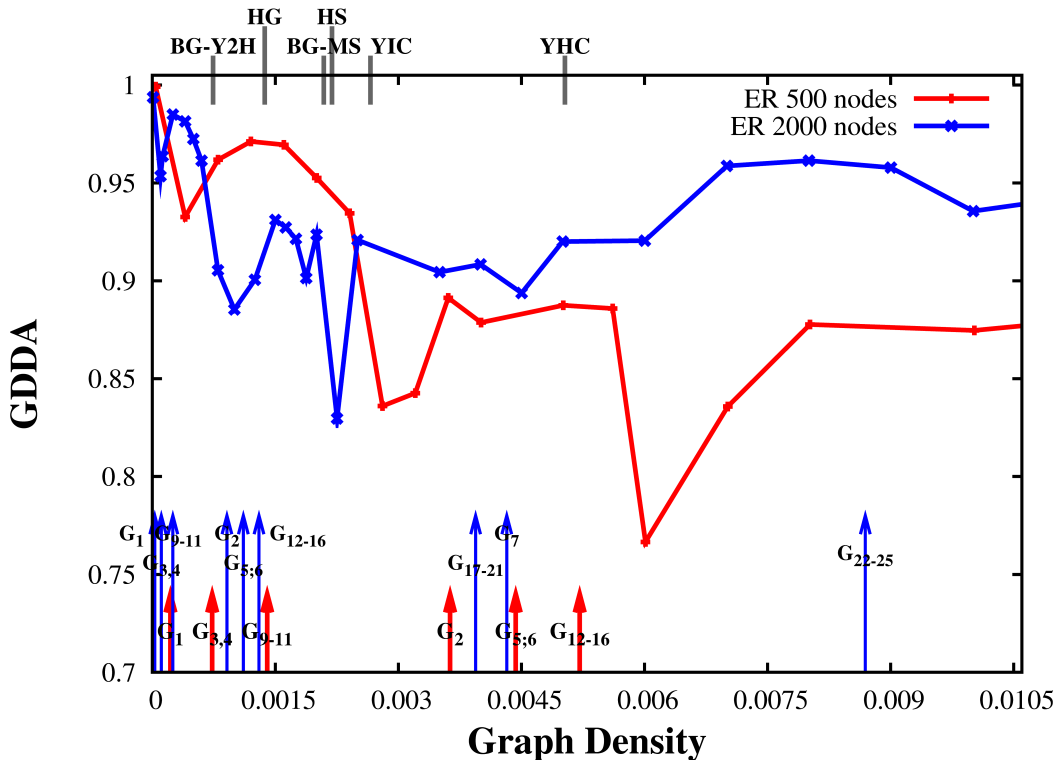


Figure: Threshold behaviour; from *Rito et al., 2010*

Issues with GDDA

- The score highly depends on the number of vertices in the networks
- For a network from a particular model with a fixed number of vertices we observe a strong non-monotone dependency with graph density in the region of interest for PINs
- Compared to a Bernoulli (ER) random graph, the PINs have graph densities which are close to the threshold of appearances of small subgraphs in the ER graphs

Why near threshold

Why would protein interaction networks operate near the threshold of the appearance of small graphlets? We conjecture:

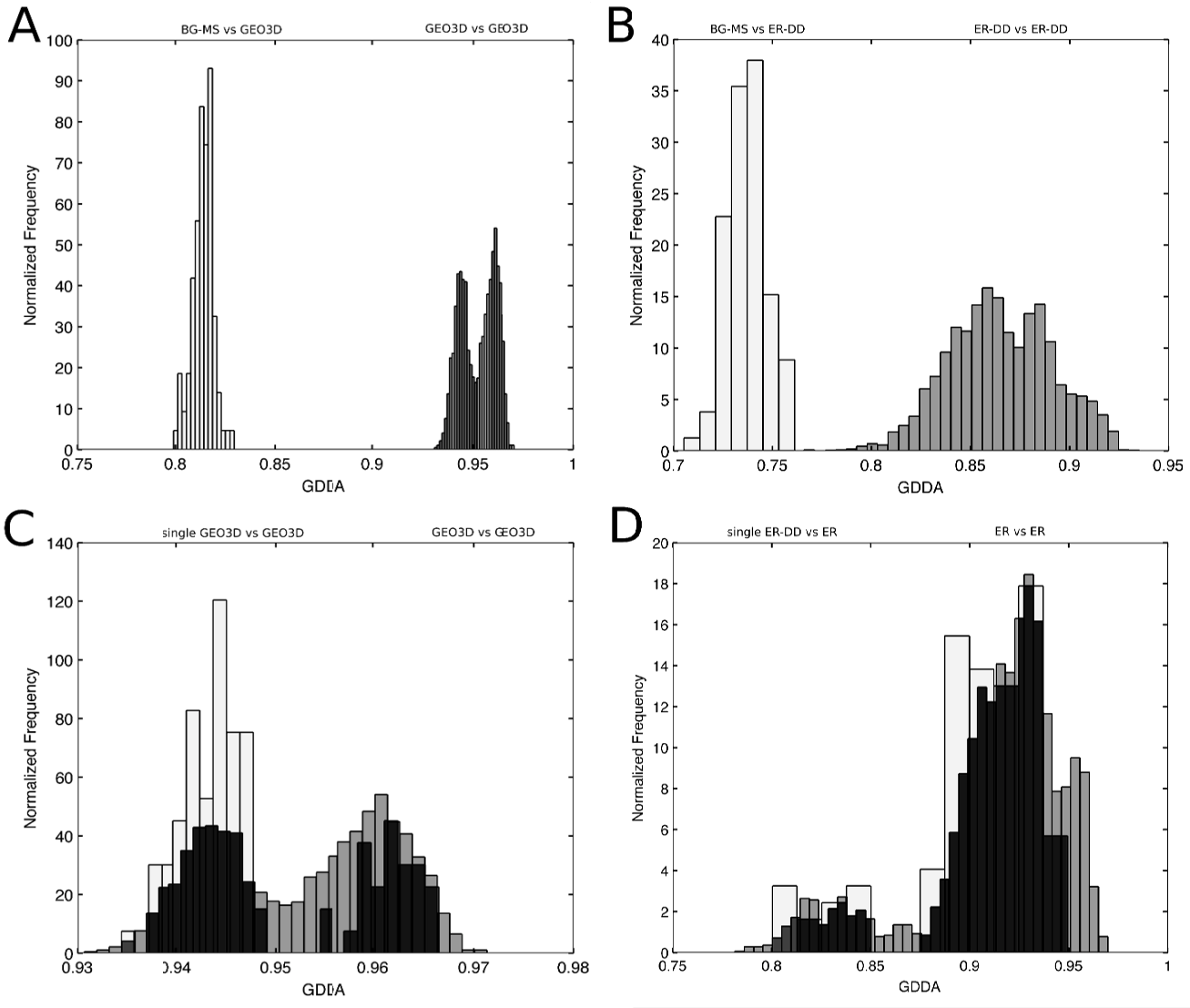
A small number of potential interactions, so that interactions are specific, makes the network efficient.

Some redundancy in the network makes the network robust against small errors.

Comparison to other models

- An ER random graph may not model the network well
- To assess model fit: a Monte-Carlo idea
  1. Generate graphs on 500, 1000, 2000 vertices with increasing graph density from the conjectured model
  2. Use these graphs as query networks
  3. Use GDDA to compare the graphs to 50 randomly generated networks from the same model: this gives us typical GDDA scores when the model is correct
  4. Now use PINs as input networks instead.

Histograms


A
BG-MS vs GEO3D
GEO3D vs GEO3D
Normalized Frequency
GDDA
B
BG-MS vs ER-DD
ER-DD vs ER-DD
Normalized Frequency
GDDA
C
single GEO3D vs GEO3D
GEO3D vs GEO3D
Normalized Frequency
GDDA
D
single ER-DD vs ER
ER vs ER
Normalized Frequency
GDDA

No model fits

- Panel D: the method gives a reasonable result for comparing ER with ER-DD (Bernoulli with fixed degree sequence)
- The observed GDDA values for our PINs are nowhere near the values we should see if the model was correct
- We also tried preferential attachment and gene duplication models: no fit
- Exponential random graph models were not able to reproduce number of edges and number of triangles.
- None of the standard models fit.

How do age and high degree come into play?

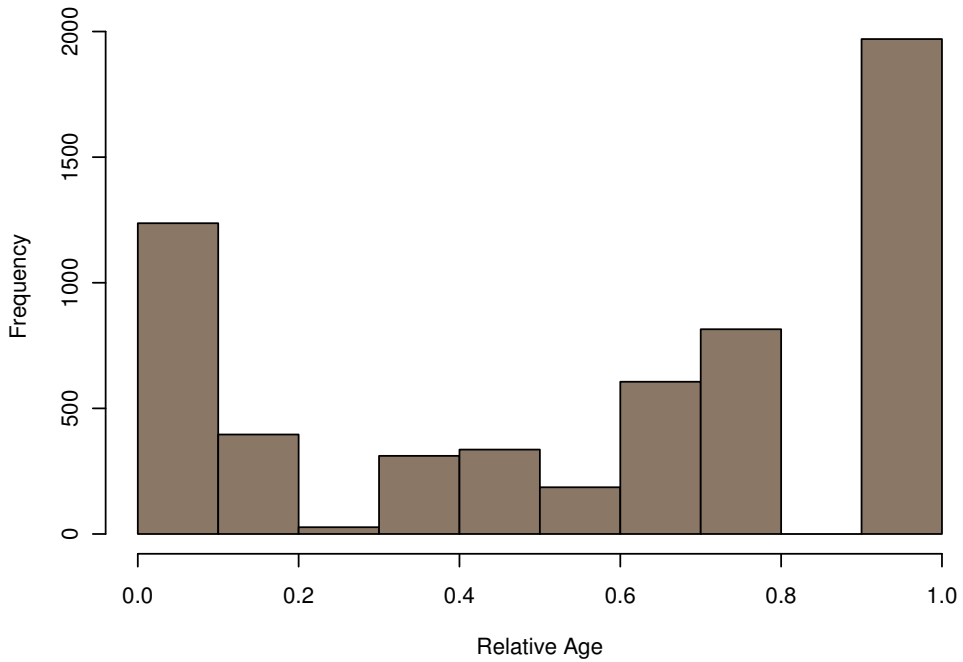
- The lack of model fit reflects our lack of basic knowledge about the network
- Use explanatory variables for interactions
- Examples could be: structural, functional classes, biochemical properties
- Here: protein age:

Protein age

Lineage specificity: ranges between 0 and 1
0: a protein which is only present in yeast, or in yeast and a few other highly related organisms on the same branch as yeast
1: a protein whose appearance can be traced to the most recent ancient branching of the tree

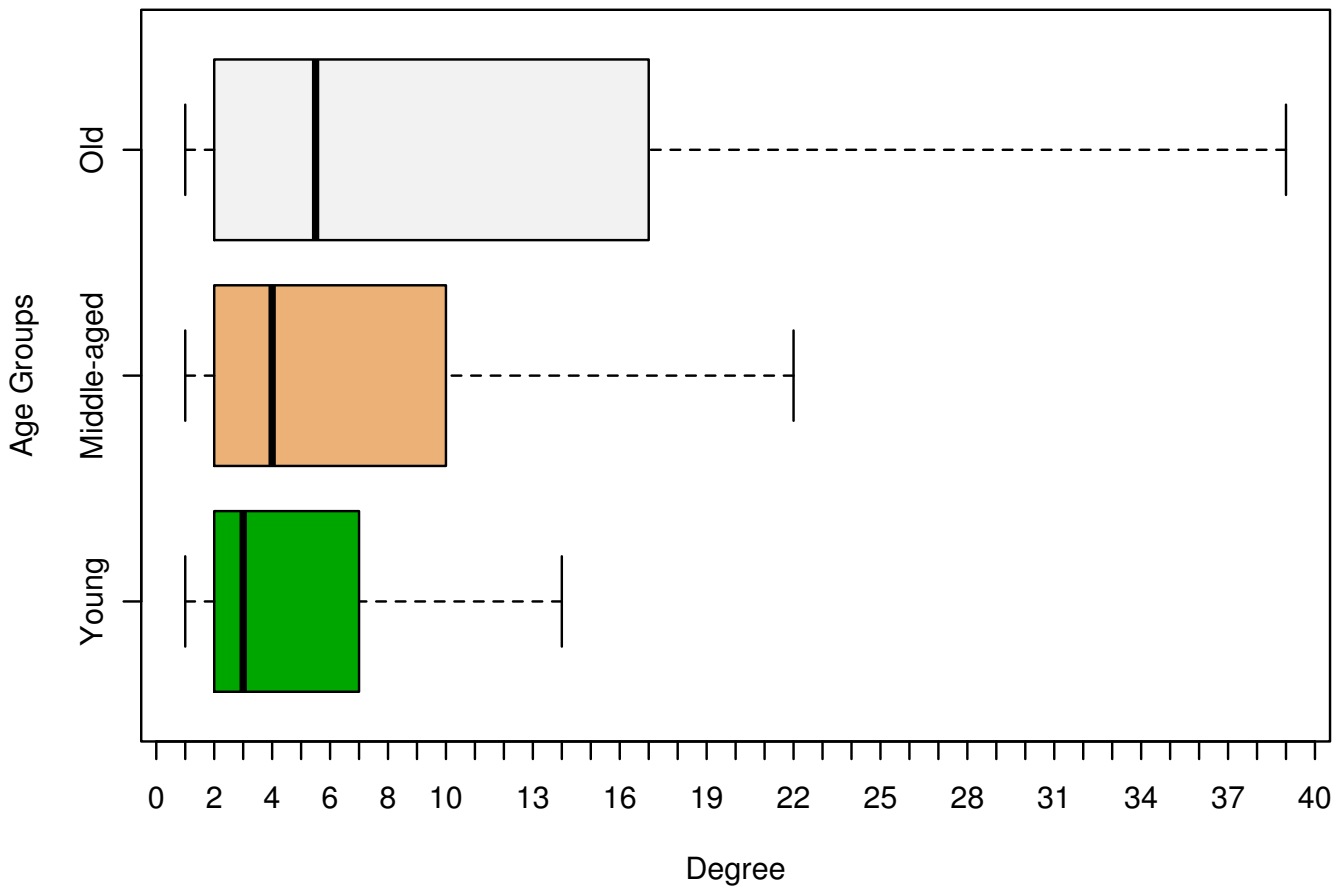
Here: we judge lineage specificity by occurrence pattern of orthologs in 99 eukaryotic species and E.Coli for which the whole genome has been sequenced

Protein age distribution in yeast



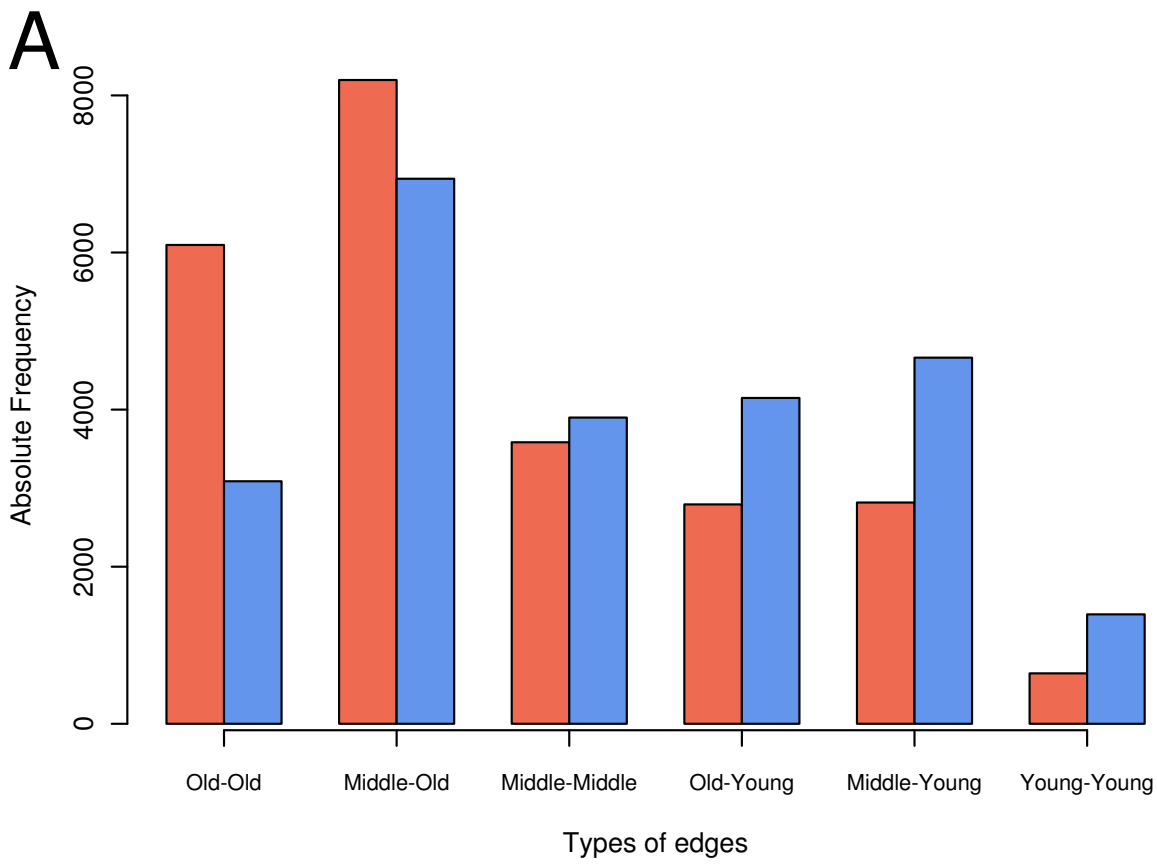
Young Middle-aged Old

Protein age and degree



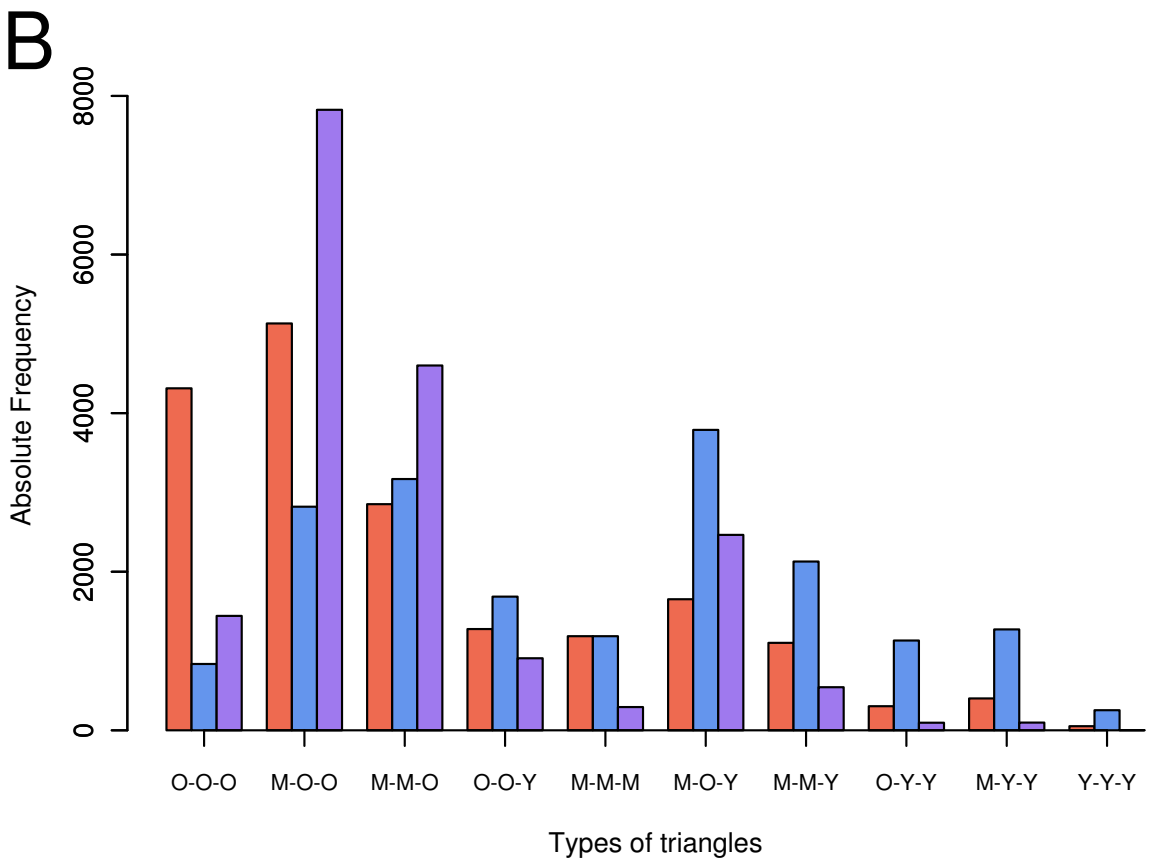
Protein age distribution: edges

Observed in yeast; expected given node frequencies



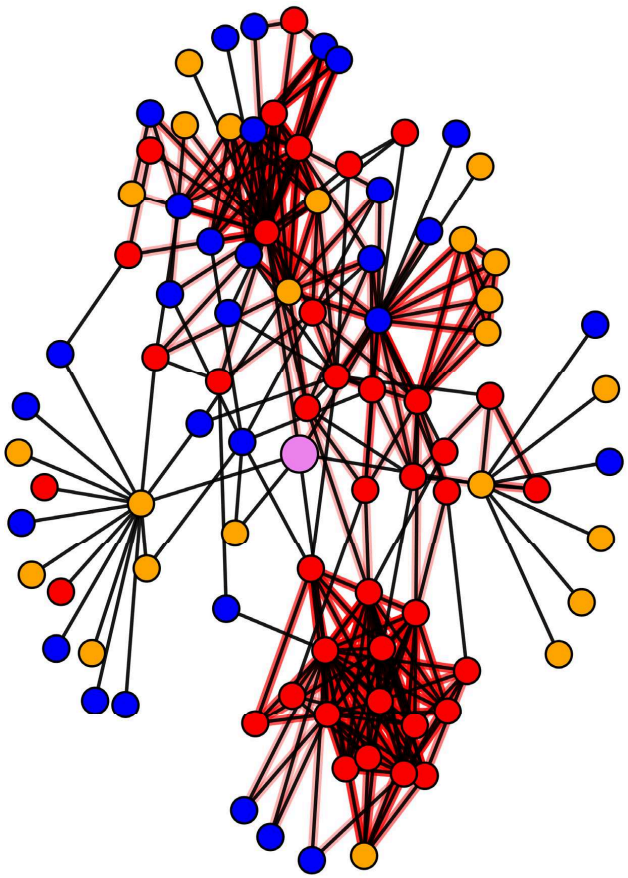
Protein age distribution: triangles

Observed in yeast; expected given node frequencies; expected given edge frequencies



Ego-network with protein ages highlighted

ego, Old, Middle-aged, Young



What we see

- Interactions between Old proteins are over-represented;
- Old proteins tend to have higher degrees;
- Triangles and edges involving a Middle-aged and Old protein are negatively selected despite being quite common;
- There are clumps of densely connected Old proteins but also many sparse regions.

What we don't see

- *Liu et al. 2011*: proteins like to interact with proteins of the same age? We find: Middle-Middle and Young-Young interactions are under-represented;
- Hub proteins? I.e. are high-degree proteins connected to many low-degree proteins in a star-shaped fashion? We find: 83 % of triangles are formed by nodes of degree at least 10.

Conclusions

- No tested model fits to the current PINs;
- The network is highly heterogeneous;
- The network shows beyond-pairwise dependence;
- Any homogeneous model is unlikely to fit.