

The ruin of gene network analysis by multifunctionality (And tips for coping)

Paul Pavlidis
paul@chibi.ubc.ca

UBC Dept. of Psychiatry &
Centre for High-Throughput Biology



Outline

Framing the issues

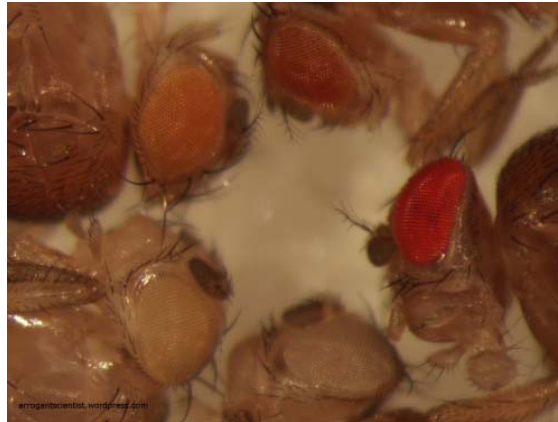
- Why and how is gene function determined?
- How is it scaled up in the post-genomic era?

Ways it can go wrong, thanks to multifunctionality

- Distorting effects of hubs
- Distorting effects of “critical edges”
- Broader impact on all “unbiased” studies

A central challenge

How do genes produce phenotypes?



Is this hard?

It's hard!

- The low number of genes in the genome
- The multifactorial nature of traits
 - Single genes cannot explain many interesting things.

It's easy!

- We can analyze all the genes
- Studies increasingly give lists of candidates
 - High throughput methods can give us answers

Applications of “function prediction”

Filling gaps in genome annotations

“About 40% of the predicted human proteins in the [draft genome] could be assigned to InterPro entries and functional categories” ... “74% of the proteins had significant [sequence similarity] to known proteins”

- Lander et al., *Nature* **409**, 860-921 (15 February 2001)

Ten years later, there are still thousands of poorly-annotated human genes.

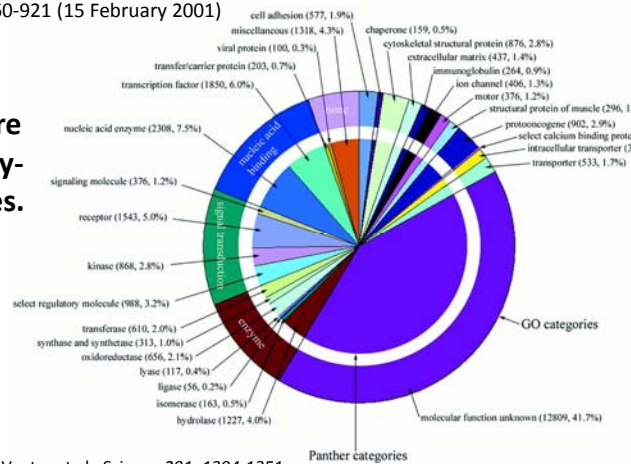


Figure from Venter et al., *Science* **291**: 1304-1351

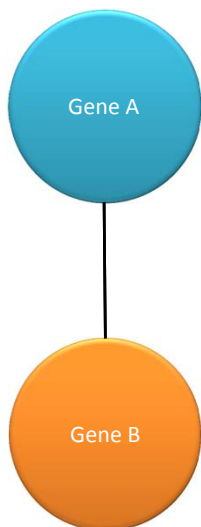
Determining gene function

De novo function analysis

- Genetic mapping studies showed which genes “control” which phenotypes
- Purification of activities let biochemists determine which proteins contribute to which molecular functions.
- Structural biology and fine mapping allowed dissection of domains and motifs that underlie the function of gene products.

These approaches continue to play a major role, but increasingly we build up on what is already known.

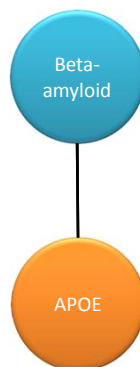
Classic approach of inference



“Discovering that a protein of unknown function interacts with one of known function provides a valuable clue to the role of the novel gene product, a concept that has been termed guilt-by-association”

- Stephen Oliver, “Guilt-by-association goes global”
Nature 403, 601-603 (10 February 2000)

A non-high-throughput example: APOE and Alzheimer’s disease



- APOE-epsilon 4 is a major risk factor for Alzheimer’s
- Discovered as a contaminant binding β A4 in isolates of cerebral spinal fluid

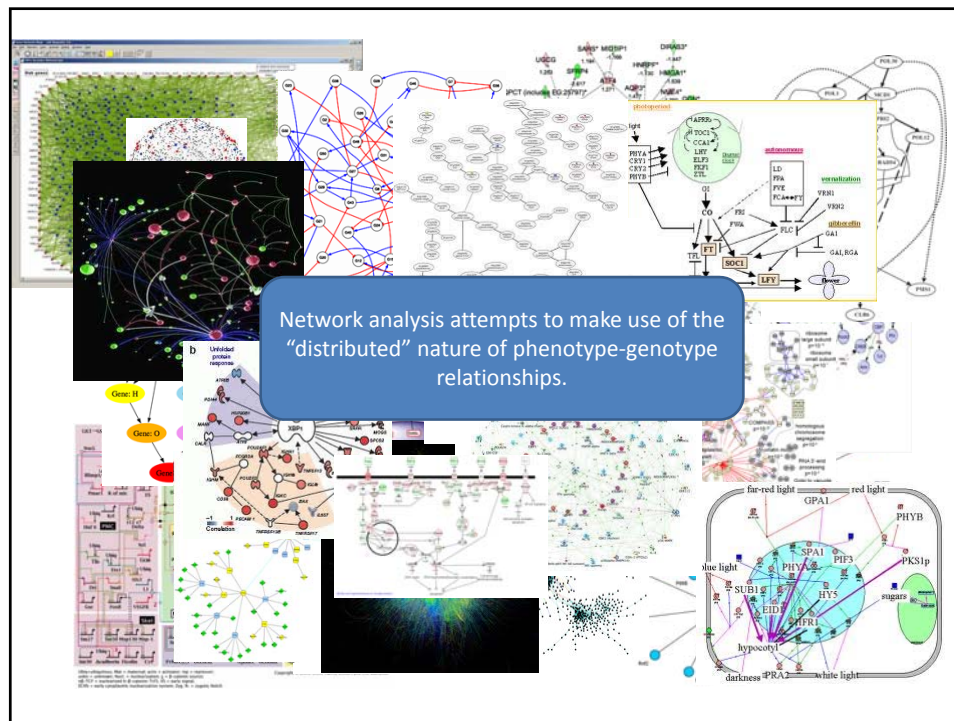
Strittmatter et al. Proc Natl Acad Sci U S A. 1993 Mar 1;90(5):1977-81.

Types of associations

- What else resembles its structure?
- What other genes have the same mutant phenotype?
- Which other genes have mutations which enhance or suppress its phenotypes?
- What genes are expressed in the same pattern?
- What proteins bind to its product?
- What genes are conserved in a similar pattern?

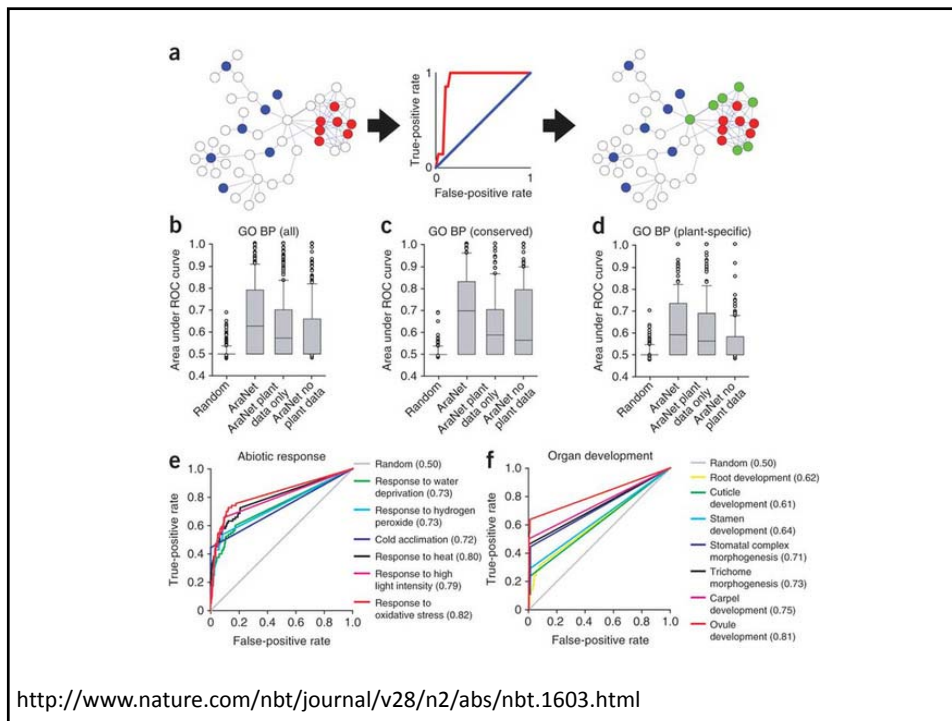


If it works for one gene ...



Scaling up GBA

- Use networks (or network-like data) to predict function or prioritize candidates
- Networks constructed from genes based on:
 - Coexpression
 - Protein interaction
 - Genetic interactions
 - Similar patterns of conservation (phylogenetic profile)
 - ...
- Use a training set to transfer “function” to other genes



“Many protein network-based predictions have recently been experimentally confirmed in yeast, worms, plants, and mice, and several successful approaches in model organisms have been directly translated to analyze human disease ...”

- Wang and Marcotte, J Proteomics. 2010 Oct 10;73(11):2277-89

“The GBA heuristic is broadly applicable across the population of nine hundred Gene Ontology categories”

- Wolfe et al., BMC Bioinformatics. 2005 Sep 14;6:227

“Computational approaches quickly provide accurate, unbiased predictions of protein function”

- Hess et al., PLoS Genet. 2009 Mar;5(3):e1000407

Summary so far

- A huge need for determining gene function and exploiting such information.
- GBA is a fundamental approach
- Computational methods have been repeatedly shown to successfully scale up.

Or have they?

Introducing multifunctionality

- Genes which have multiple functions
- Working definition: Number of Gene Ontology terms
- Not the same as (but might relate to):
 - Pleiotropy
 - Hub-ness
 - Promiscuity

Are multifunctional genes interesting?

- Genes which are likely to “come up” in assays tend to be multifunctional
- They cannot be specifically related to your question
- Algorithms tend to focus attention on those genes

The big question: Does real data encode this list?

If it does:

1. Computational gene function prediction will seem to work but be of little actual use.
2. Any experiment which “encodes” this gene list will seem to have yielded results but will tend to assign function to genes that already have them (the rich get richer)

Case study

- CNVs from intellectual disability (ID) study
- Task: Identify which genes in the CNV regions are most likely “responsible” for the phenotype.

Qiao Y, Harvard C, Tyson C, Liu X, Fawcett C, Pavlidis P, Holden JAA, Lewis MES, Rajcan-Separovic E (2010) Human Genetics, 128(2):179-94

Gene prioritization approach

- Tested multiple “candidate prioritization” programs (GBA)
 - Input 1: The genes covered by the CNV (genes of interest)
 - Input 2: A training set of genes
 - Experiment: A list of genes known to play a role in ID
 - Control: Random genes or genes related to other disorders
 - Input 3: A gene network (provided by the program)
 - Output: Ranking of genes by inferred “priority”

- Results: The experiment and controls yield similar rankings of our targets

Example: CNV at 19p13.3

- 97 genes covered by deletion

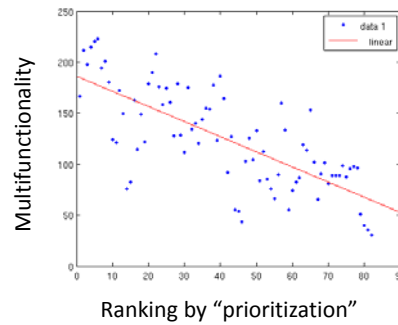
- Top 5 candidates from algorithm trained on “known ID genes”: SH3GL1, AES, EEF2, DAPK3, GNA11

- Top 5 candidates from control experiment: SH3GL1, AES, EEF2, DAPK3, CCDC94

Qiao Y, Harvard C, Tyson C, Liu X, Fawcett C, Pavlidis P, Holden JAA, Lewis MES, Rajcan-Separovic E (2010) Human Genetics, 128(2):179-94

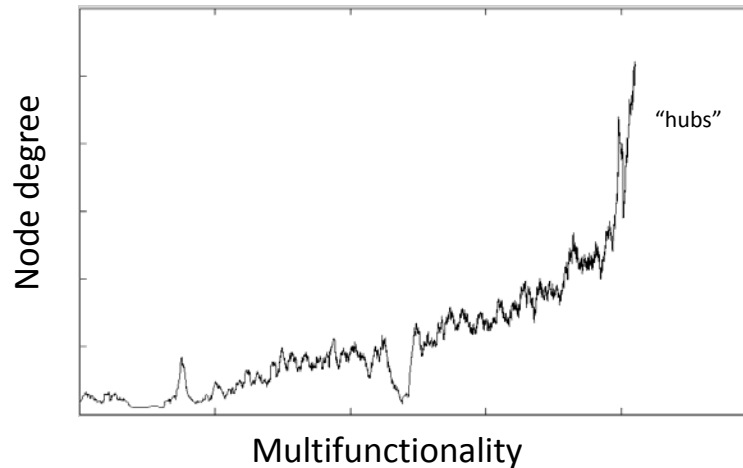
What is happening?

- Guilt-by-association will tend to give highest priority to well-annotated genes
- Doesn't mean predictions are *wrong*, but they are *too generic*.



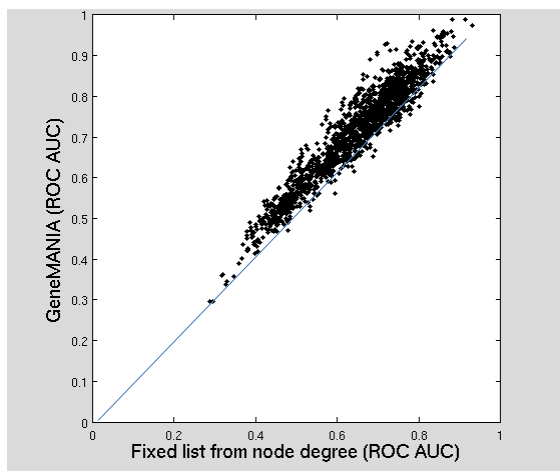
Jesse Gillis, Eloi Mercier

Multifunctionality is correlated with number of network neighbours



Smoothed to show trend

Functions that are learnable from a network are learnable by node degree



Holds true for a wide range of networks, algorithms, and evaluation metrics.

Gillis J, Pavlidis P, 2011 The Impact of Multifunctional Genes on "Guilt by Association" Analysis. PLoS ONE 6(2): e17258.

Is there anything left?

- About $\frac{1}{2}$ of prediction performance is accounted for by "generic" node-degree effects.
- What about the other half?

Functional and critical edges

- Functional edge: connects two “functionally-related genes”
- Critical edge: if removed from the network, ability to predict function decreases.
- We don’t want all “predictability” of a function to come from one edge

Critical edges

- Define: “critical edge” is a network edge important for learning a function.
 - Remove this edge, and prediction performance drops substantially
- Questions:
 - How common are critical edges?
 - What effect do they have on usefulness of GBA?

“Exceptional edges”

- Definition: An edge that is critical for predicting many GO groups
- If two genes are multifunctional, and an edge joins them, we predict that edge is more likely to be critical for many functions
- Prediction: exceptional edges are those that join the most multifunctional genes

Starting point: MouseFunc

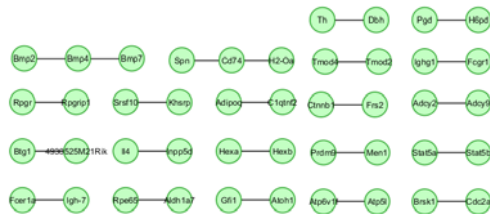
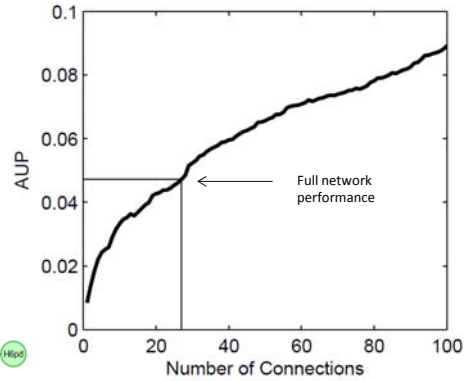
Research Open Access
A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence
 Lourdes Peña-Castillo¹, Murat Tasan², Chad L. Myers³, Hyunju Lee⁴,
 Trupti Joshi⁵, Chao Zhang⁶, Yuanfang Guan³, Michele Leone⁶,
 Andrea Pagnani⁸, Wan Kyu Kim⁷, Chase Krumpelman⁸, Weidong Tian⁷,
 Guillaume Obozinski⁹, Yanjun Qi¹⁰, Sara Mostafavi¹¹, Guan Ning Lin³,
 Gabriel F. Berriz², Francis D. Gibbons², Gert Lanckriet¹², Jian Qiu¹³,
 Charles Grant¹⁰, Zafer Barutcuoglu¹⁴, David P. Hill¹⁵, David Warde-Farley¹¹,
 Chris Grouios¹, Debajyoti Ray¹⁶, Judith A. Blake¹⁵, Minghua Deng¹⁷,
 Michael I. Jordan¹⁸, William S. Noble¹⁹, Quaid Morris^{1,12,20}, Judith Klein-
 Seetharaman²¹, Ziv Bar-Joseph¹⁹, Ting Chen²², Fengzhu Sun²²,
 Olga G. Troyanskaya³, Edward M. Marcotte², Dong Xu²,
 Timothy R. Hughes^{1,20} and Frederick P. Roth^{2,23}

Published: 27 June 2008
 Genome Biology 2008, 9:102

- Ten constituent networks
- 21.6k genes
- 4.5 million edges
- 774 GO terms tested
10-300 genes each
- Evaluate with precision-recall curves
- Nine competitors, sophisticated algorithms

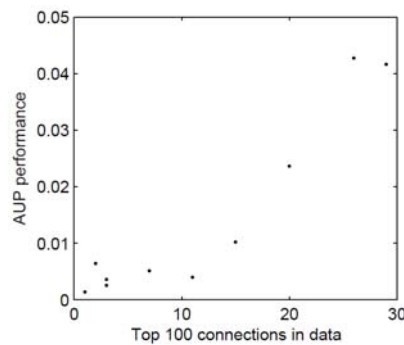
Exceptional edges in theory

- We can construct a network with just **23** edges that matches the performance of the full MouseFunc network



- AUP = Area under precision-recall curve, average for all GO groups tested
- Using a fast GBA method that performs nearly as well as the best method in MouseFunc.

The real Mousefunc networks contain exceptional edges

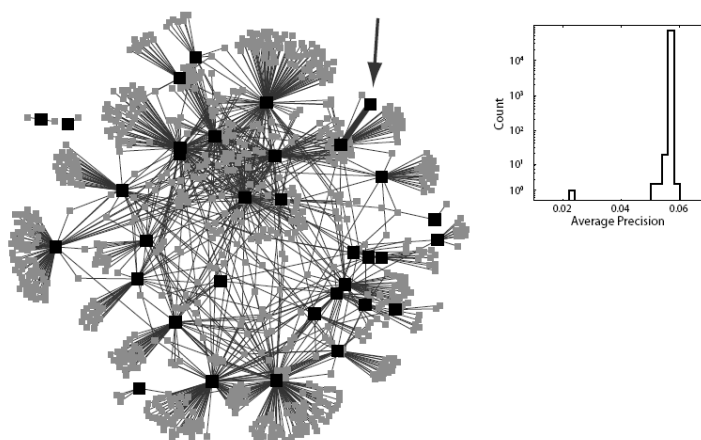


Each point is a network; Network contribution to performance is predicted by number of exceptional edges.

Exhaustively testing criticality

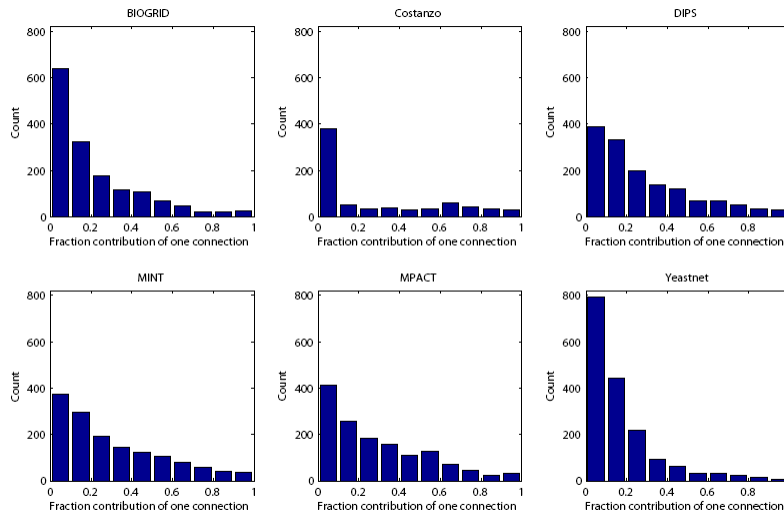
- Start with a high-quality yeast network + GO groups to test.
 1. Remove one edge
 2. Test cross-validation performance of GBA
- Repeat 1 & 2 for all edges

Critical connections in yeast protein interaction networks

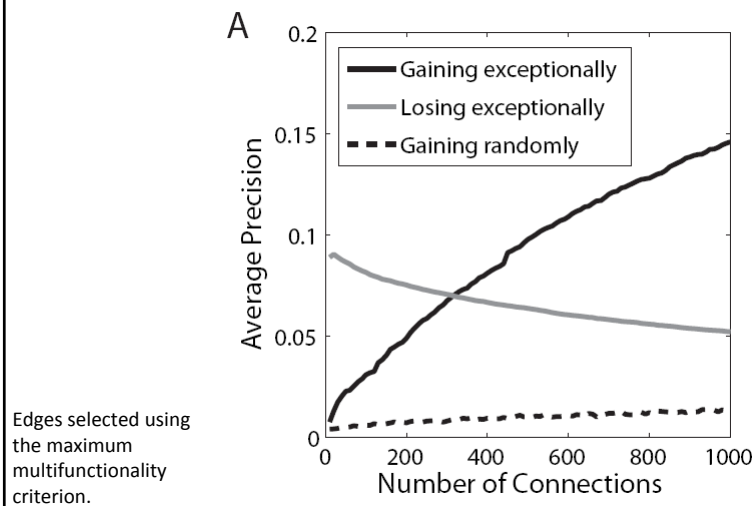


Aggregate yeast PPI from BioGRID, etc; Network is for "Cellular polysaccharide biosynthetic process" subnet; 27 edges in-group

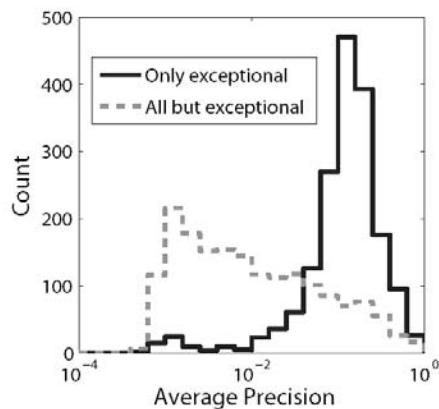
Most GO groups in most networks are affected by removing a single connection



Removing critical connections from the real network damages performance



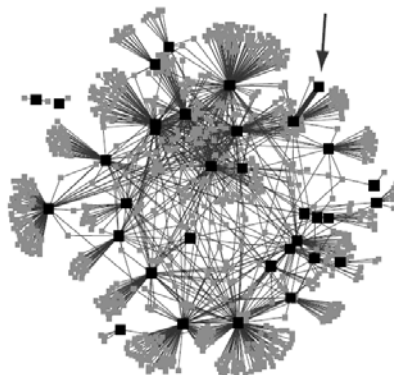
Functional information is not distributed throughout the network



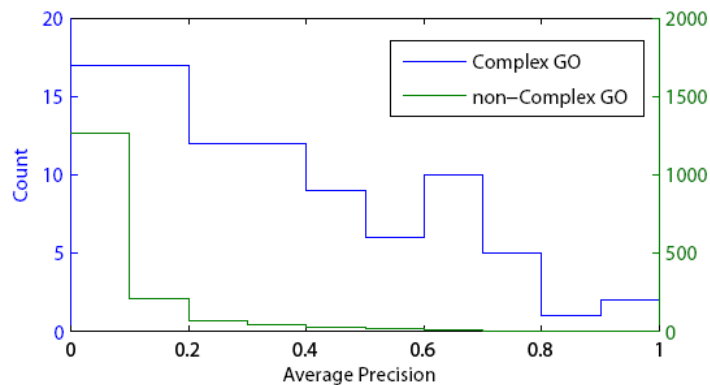
Removing all of the 4870 most critical edges from the network removes most of its performance (black solid line), while adding only those critical edges (grey dashed) yields high performance across all GO groups.

GBA cross-validation is untrustworthy

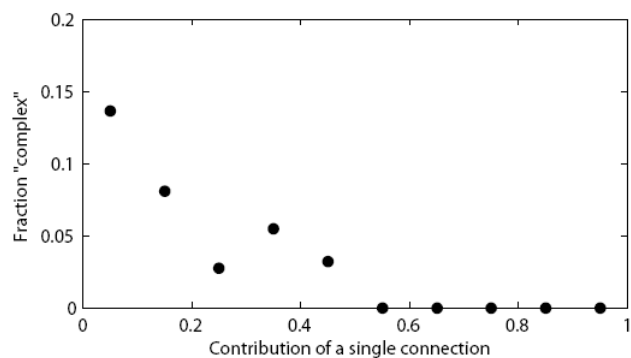
Because performance for most GO groups comes from a single connection, there can be no expectation that we can make good novel predictions (since they do not involve that connection)



The exception: protein complexes

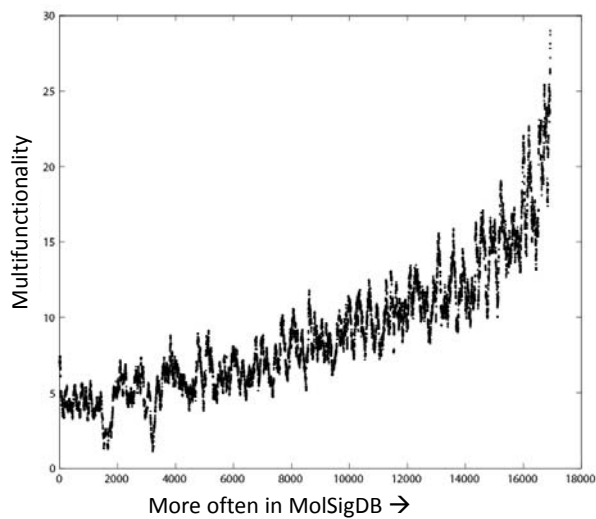


Protein complexes contribute very strongly to the “non-critical” GO



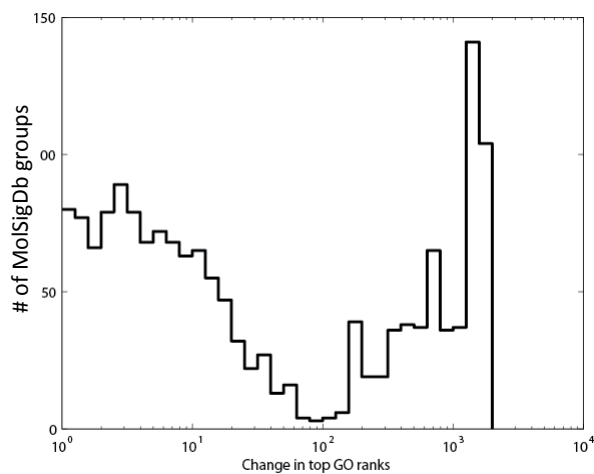
Prediction of protein complexes is, by their nature, resistant to the effects of removing one edge.

Multifunctional genes tend to be differentially expressed



Analysis of MolSigDB gene lists; lightly smoothed

GO enrichment is highly sensitive to multifunctional genes



Effect of removing the most multifunctional gene on top 10 GO enrichment results on MolSigDB lists.

Summing up

What makes a gene interesting?

Most obviously related to my favourite process
(But might be generic)

vs.

Specifically related my favourite process
(Maybe subtle)

Are networks still useful?

- Function is not obviously “distributed” in the networks.
 - Cross-validation GBA is highly misleading
- Analyses are strongly influenced by hubs and multifunctionality.
 - Removing ‘hubs’ not a very attractive option
- Local information is still of obvious value.

What do we do about it

- Beware: Is your gene ranking multifunctional?
- Adapt:
 - Predictions should be specific for a function.
 - Use data that is “function-specific” and choose appropriate controls
 - Use training sets that are as unifunctional as possible.

Conclusions

- Multifunctionality is lurking as a confound in the interpretation of many biological experiments.
- It has an especially pernicious effect on computational attempts to exploit known information and in the examination of gene networks.
- Correcting for it is challenging because multifunctional genes (probably) really are important, but detecting their effects is relatively easy.

Thanks

- **Jesse Gillis**
- **Eloi Mercier, Raymond Lim**
- **Vaneet Lotay, Gavin Ha, Anamaria Crisan, Fong Chan**
- **Evica Separovic and colleagues (CNVs in ID)**
- Quaid Morris and Sara Mostafavi for providing GeneMANIA

Relevant papers:

- Gillis J, Pavlidis P, 2011 The Impact of Multifunctional Genes on "Guilty by Association" Analysis. PLoS ONE 6(2): e17258.
- Gillis J, Pavlidis P, 2011 The role of indirect connections in gene networks in predicting function. Bioinformatics 27(13):1860-6.

