STAT 540 2017
Analysis of gene function II: Gene networks

Paul Pavlidis

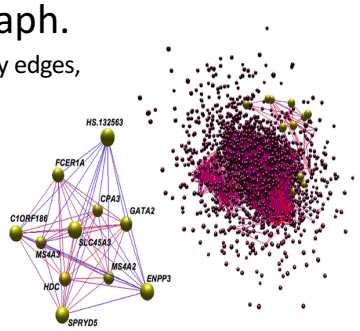
Outline

- What are gene networks
- Why are they used
- How are they constructed
- How are graphs analyzed in general
- Coexpression networks in more detail
- Combining gene function and networks: Guilt by association

What is a gene network?

Gene data *represented as* a graph.

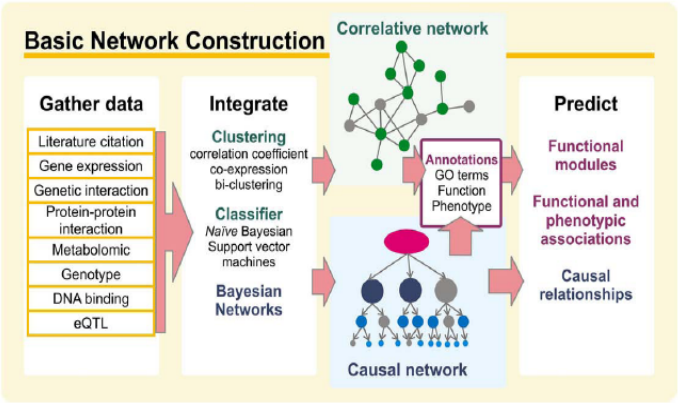
- A graph is a set of nodes (vertices) connected by edges, possibly with weights
- Here we consider only undirected graphs.



Why bother?

1. Computational convenience (e.g., sparse)
2. The graph *could* also have some relation to "reality" (e.g. physical relationships in the cell etc.) (but generally will not)
3. "It makes biological sense"

The gene network paradigm



General components:

- A network
- Functional annotations and/or genes of interest
- Algorithm to apply to the network (supervised or unsupervised)

Wang and Marcotte. It's the machine that matters: Predicting gene function and phenotype from protein networks. Journal of Proteomics, 2010.

Clustering+enrichment

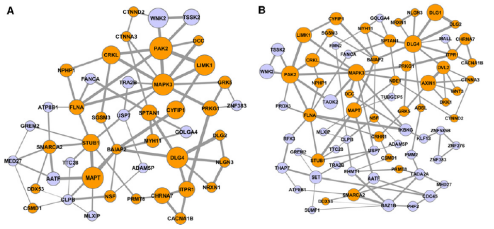


Figure 2. Gene Clusters Found using NETBAG Analysis of De Novo CNV Regions Observed in Autistic Individuals

(A) The highest scoring cluster obtained using the search procedure with up to one gene per each CNV region.

(B) The cluster obtained using the search with up to two genes per region. In the figure, genes (nodes) with known functions in the brain and nervous systems are colored in orange (see Table S2 for functional information about the genes forming the cluster). Node sizes represent the importance of each gene to the overall cluster score. Edge widths are proportional to the prior likelihood that the two corresponding genes contribute to a shared genetic phenotype. For clarity, we show only edges corresponding to the two strongest connections for at least one node.

See also Figure S2, Table S1, and Table S2.

Table 1. Gene Ontology (GO) Terms Highly Connected to the Functional Network in Figure 2A

Gene Ontology Term	GO Category	Q Value
GO:0007015: actin filament organization	Biological process	<0.01
GO:0030424: axon	Cellular component	<0.01
GO:0048469: cell maturation	Biological process	<0.01
GO:0007611: learning and or memory	Biological process	<0.01
GO:0044456: synapse part	Cellular component	<0.01
GO:0045202: synapse	Cellular component	<0.01
GO:0007163: establishment and or maintenance of cell polarity	Biological process	0.01

Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses

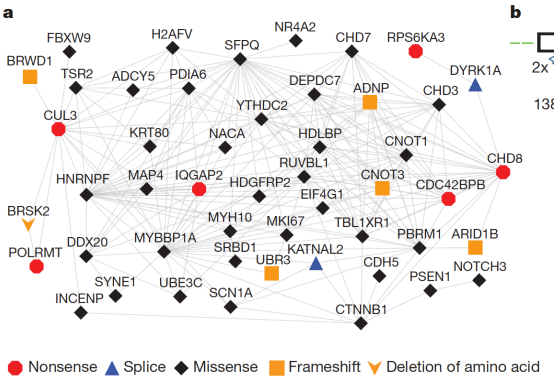
Sarah R. Gilman,[1,3] Ivan Iossifov,[2,3,*] Dan Levy,[2] Michael Ronemus,[2] Michael Wigler,[2] and Dennis Vitkup[1,*]

http://www.cell.com/neuron/abstract/S0896-6273%2811%2900439-9

Hit list interpretation

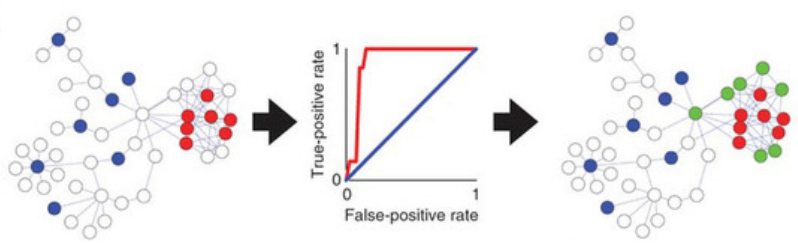
Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations

Brian J. O'Roak[1], Laura Vives[1], Santhosh Girirajan[1], Emre Karakoc[1], Niklas Krumm[1], Bradley P. Coe[1], Roie Levy[1], Arthur Ko[1], Choli Lee[1], Joshua D. Smith[1], Emily H. Turner[1], Ian B. Stanaway[1], Benjamin Vernot[1], Maika Malig[1], Carl Baker[1], Beau Reilly[2], Joshua M. Akey[1], Elhanan Borenstein[1,3,4], Mark J. Rieder[1], Deborah A. Nickerson[1], Raphael Bernier[2], Jay Shendure[1] & Evan E. Eichler[1,5]



We found 39% (49 of 126) of the genes mapped to a highly interconnected network wherein 92% of gene pairs in the connected component are linked by paths of three or fewer edges (Fig. 2a). We tested this degree of interconnectivity by simulation (n=10,000 replicates; Methods and Supplementary Fig. 9) and found that our experimental network had significantly more edges (P<0.0001) and a greater clustering coefficient (P<0.0001) than expected by chance.

Function prediction
("Candidate prioritization" et al.)



Red: "known genes"
Rest: "unknown" or "negative"
Train algorithm

Test performance with cross-validation

Classify (Green: new candidates "like red")

http://www.nature.com/nbt/journal/v28/n2/full/nbt.1603.html

Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*

Insuk Lee, Bindu Ambaru, Pranjali Thakkar, Edward M Marcotte & Seung Y Rhee
Affiliations | Corresponding authors

Nature Biotechnology **28**, 149–156 (2010) | doi:10.1038/nbt.1603

Building networks: Inference vs. relations

- In the majority of routine applications of gene networks, the edges are interpreted as meaning there is some *statistical* relationship between some features of the genes. Treat it as a distance or similarity measure.
- It does not mean there is a causal, regulatory, or physical relationship
  - Typically undirected
  - Often based on correlations (similarity/distance)

Types of gene associations (1/3)

- Protein interactions
  - Biased to well-studied and highly expressed genes
  - Mixes data from different conditions
  - Transient interactions lost (mostly get complexes?)
  - False negatives (not all possible interactions tested)
- Genetic interactions
  - Gene selection bias
  - Limited phenotypes tested ("viability" or "growth")
  - Choice of interpretation?

Types of gene associations (2/3)

- Functional similarity (e.g., GO functions)
  - Inherits all the problems of GO
  - Problematic to mix this with other types of data (information retrieval vs. discovery)
- Sequence similarity
  - Excellent way to infer function
  - Signal is very strong, swamps out others; best considered separately?
  - Related approaches: protein domain occurrence
- Coexpression
  - Comprehensive, but noisy
  - Can make "condition specific" (scant evidence this matters beyond changing the set of genes expressed)
  - "low resolution"?

Types of gene associations (3/3)

- Co-mentions (text mining)
  - Abundant (e.g. from PubMed abstracts)
  - Very noisy
  - Negation hard to detect
  - Information retrieval vs. discovery problem
- Co-conservation (phylogenetic profile)
  - Need lots of genomes at appropriate phylogenetic distances
  - Resolution limited by ability of protein sequence to distinguish distinct function relations
- Genomic proximity
  - Weak signal overall

Keep in mind

- What exactly is an edge or node in your network?
  - E.g. Don't interpret correlation as a physical interaction
  - Does this data have to be represented as a graph?
- Different types of networks might inform about different types of function
  - e.g. Sequence tells most about "molecular function"
- *Lack of an edge* may not be supported by evidence of disconnection
- Most networks are not state specific
  - Often a subset of all interactions that could occur

Basic graph (or node) properties

- Number of nodes, edges; Sparsity (or density)
- Node degree distribution
- Diameter, cluster coefficient, path length, betweenness centrality …

What do we expect for biological networks?

Commonly mentioned graph features

- "Hubs" – Nodes with lots of edges
- "Scale free" – node degree distribution follows a power law distribution
- "Small world" – Specific property of being sparse, but with short average path length – "cliquey"
  - "Clique" – Highly connected component

Cluster Coefficient

- The degree to which nodes tend to cluster together or form cliques

$$C_i = \frac{\lambda_G(v)}{\tau_G(v)}.$$

# of connected pairs between all neighbours of node

$$\tau_G(v) = C(k_i, 2) = \frac{1}{2}k_i(k_i - 1).$$

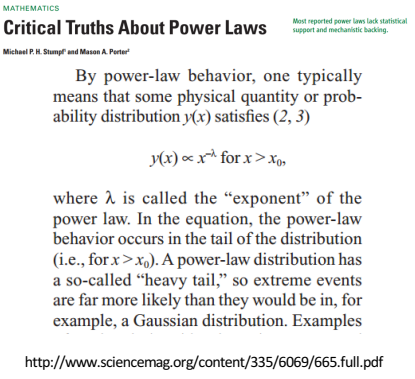
Total number of edges that can exist for a node with k neighbours

- Values range between 0 and 1
- Can be used to characterize sets of genes

(Many other measures can be used to characterize networks)

'Scale-free' property

- Power law describes the degree distribution
- Found in natural and man-made data (at least, arguably)
- "As a rule of thumb, a candidate power law should exhibit an approximately linear relationship on a log-log plot over at least two orders of magnitude in both the *x* and *y* axes."



MATHEMATICS

Critical Truths About Power Laws

Most reported power laws lack statistical support and mechanistic backing.

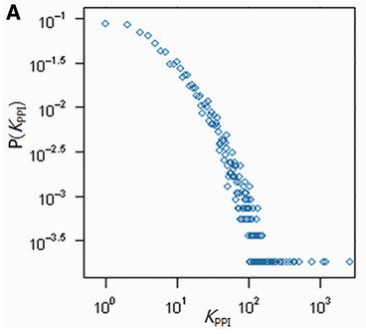
Michael P. H. Stumpf[1] and Mason A. Porter[2]

By power-law behavior, one typically means that some physical quantity or probability distribution $y(x)$ satisfies (*2*, *3*)

$$y(x) \propto x^{-\lambda} \text{ for } x > x_0,$$

where λ is called the "exponent" of the power law. In the equation, the power-law behavior occurs in the tail of the distribution (i.e., for $x > x_0$). A power-law distribution has a so-called "heavy tail," so extreme events are far more likely than they would be in, for example, a Gaussian distribution. Examples

http://www.sciencemag.org/content/335/6069/665.full.pdf



http://nar.oxfordjournals.org/content/41/20/9209/F1.large.jpg

Scale-freeness: biological?

- Cases can be made, but it's hard to tell the difference between "true" hubs and "human-made" hubs.
- Some biologically-motivated models of network "evolution" yield scale free networks
  - E.g. duplication-divergence. But these are mostly not the kind of networks we are discussing.
  - More general model is relevant: preferential attachment (rich get richer) might explain both artifactual and natural processes

For some discussion see:
http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000232

Properties of hubby genes

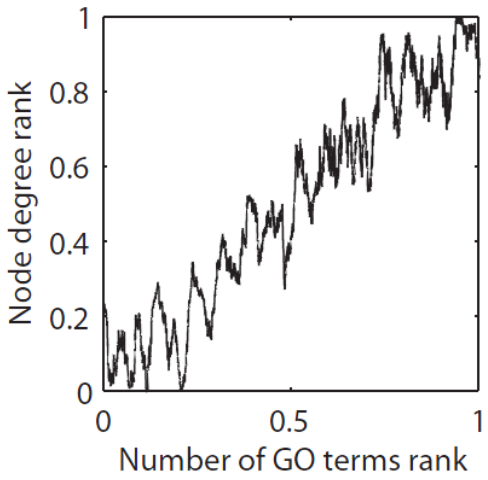
They tend to be:

- Multifunctional
- Essential
- Conserved
- Not conserved ("flexible" or disordered proteins)

Important to note:

- Being attached to a hub is not unusual
- … so hubs also tend to be attached to each other

Highly-annotated genes tend to be "more hubby" in networks



Gillis J, Pavlidis P, (2011) PLoS ONE 6(2): e17258

19

How do we decide if a network property is "significant"

- Typically done by creating random networks to generate a null distribution - Need to carefully construct the right null
- Re-shuffling the links between nodes to maintain individual node degree and the degree distribution
- Simply maintaining overall node degree distribution is not enough
- Maintaining per-gene node degree is important because node degree is such an important determinant of a gene's properties in the network

Specificity and Stability in Topology of Protein Networks

Sergei Maslov[1] and Kim Sneppen[2]

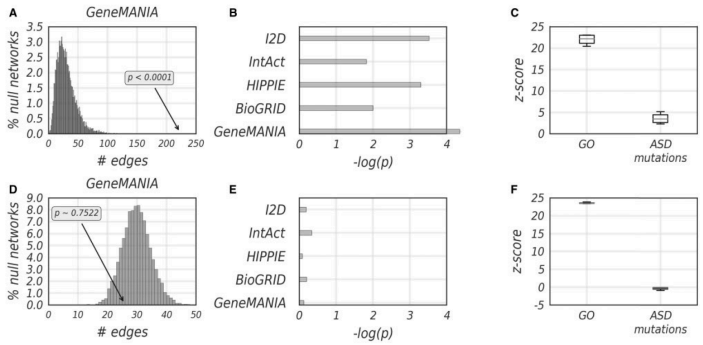


Fig. 3. Replicating network properties of de novo mutations in autism and edge permutation as a null removes all significance from the autism-derived gene list. (**A**) Replication of the Monte Carlo experiment performed in O'Roak *et al.* (2012) (see Supplementary Fig. S10). The number of edges between the genes in the autism (ASD) gene list in a network based upon GeneMANIA (version date August 3, 2011) physical interaction data is statistically different ($P < 0.0001$). The parsing of this network in the original analysis conflates 'tested' pairs with 'validated' pairs of interacting genes across much of the data. (**B**) Networks drawing on different network resources also show a statistically significant number of edges between the genes in the ASD gene list: BioGRID ($P \sim 0.007$), HIPPIE ($P \sim 0.0006$), IntAct ($P \sim 0.015$) and I2D ($P \sim 0.0004$). Because we ran 10 000 iterations to calculate *P* values, the maximum value on the graph is 4; the original GeneMANIA-based result is some point past this, indicated by its placement. (**C**) However, conducting the same analysis on gene lists derived from GO terms reveals that they are much more likely to exhibit significant linkage than the ASD gene list (GO: mean *z*-score = 21.96; ASD gene list = 4.41). (**D**) Using the corrected GeneMANIA data as well as holding node degree per each gene fixed in the null simulations removes all significant association between ASD genes. Note the null distribution of number of edges differs between this and (A); axis scale has also changed. (**E**) This property replicates across network data derived from multiple sources. (**F**) Functional sets of genes defined by GO remain learnable even after accounting for node degree in this way

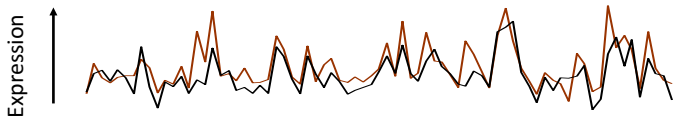
Positive and negative forms of replicability in gene network analysis

W. Verleyen, S. Ballouz and J. Gillis*

Gene coexpression

- Genes that are coexpressed *tend* to have related function; Needed at the same place at the same time

Two ribosomal protein genes.

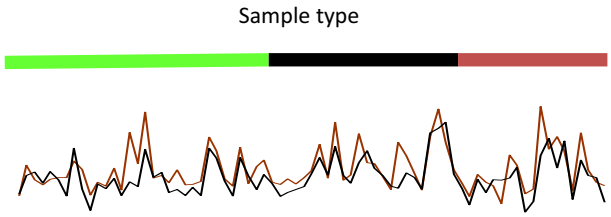


Samples

Unlike many types of networks, it's more common to construct them for a specific study

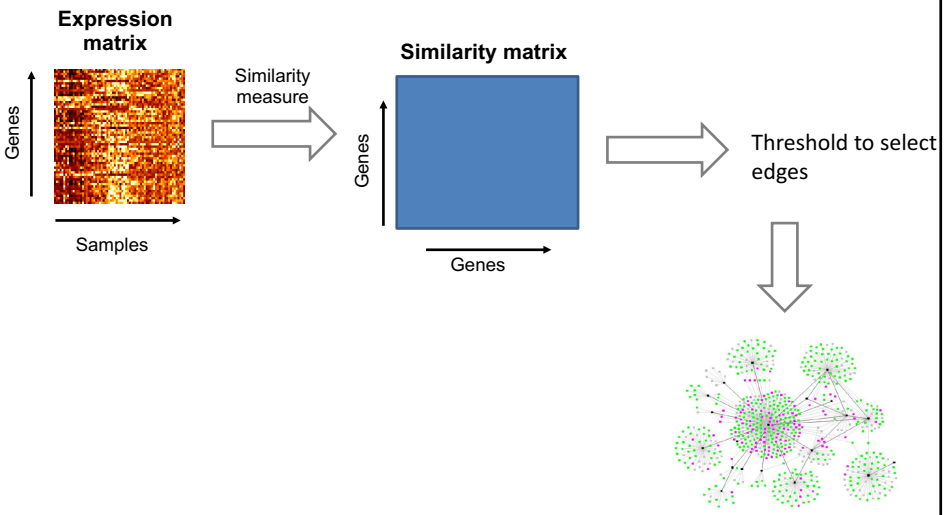
Biological noise

- Go beyond gene expression effects associated with the experimental design
- Gene expression varies between "*replicates*" in biologically-meaningful ways.
- Experiment does not have to be a time course



Potential presence of technical noise means it might be a good idea to combine data sets.

Basic coexpression network construction



Coexpression with WGCNA

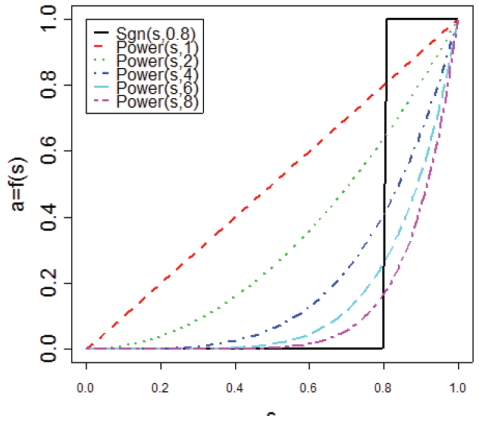
- A popular coexpression network package, R implementation
- Focus: module identification
- Many extra features
  - Network comparison
  - Visualization
  - Functional analysis

Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.

Some slides adapted from http://labs.genetics.ucla.edu/horvath/MTOM/

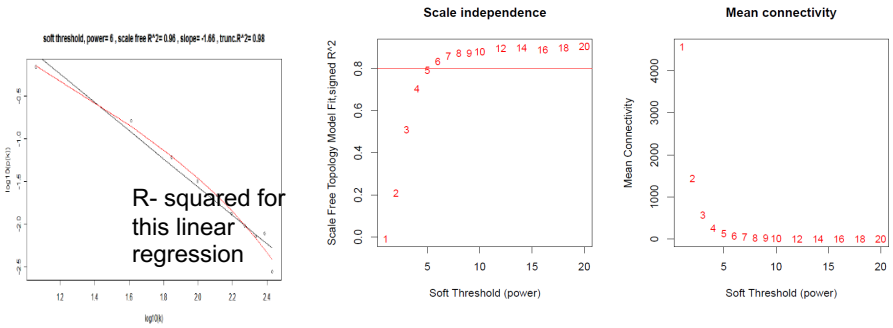
"Soft-thresholding"

- Transforms correlation in a particular way: Instead of thresholding, raise corr. to a power β



Choosing β

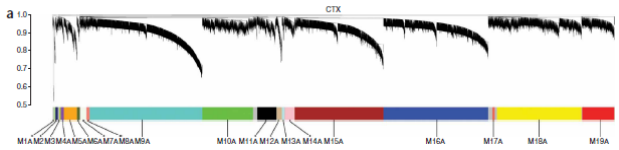
- Deciding how to threshold is a perennial problem in coexpression networks.
  - For correlation et al., can put into a hypothesis testing framework ("significant" correlation)
  - Or can choose a threshold based on "known" coexpression
  - Or choose arbitrarily. WCNA solution: choose parameter that makes the resulting network "scale free"

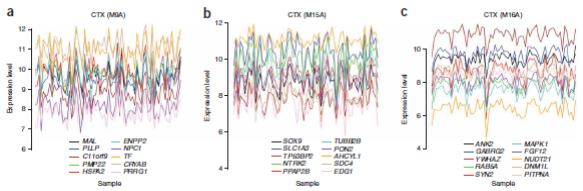


Clustering a network

- Same basic goal as clustering the gene profiles – look for patterns.
- Initial step is still computing a distance matrix
- But moves to a graph representation
  - Sparsified – low correlations eliminated
  - "Modules" == "clusters"
- Picking right number of modules still a problem
  - Where to cut the tree

WGCNA uses hierarchical clustering of the "topological overlap", which is a derived measure of coexpression after thresholding (based on how many "partners" the genes share)





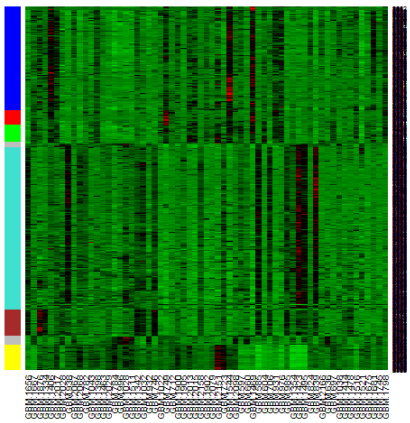
Functional organization of the transcriptome in human brain

Michael C Oldham[1,2], Genevieve Konopka[2], Kazuya Iwamoto[3], Peter Langfelder[4], Tadafumi Kato[3], Steve Horvath[4–6] & Daniel H Geschwind[2,4,6]

Non-network view of gene co-expression modules

Columns= samples

Rows=Genes
Color band indicates module membership



Horvath

Function prediction

How is gene function determined?

- 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 on what is already known using inference or "guilt by association"

Applying inference: APOE and Alzheimer's disease

"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"

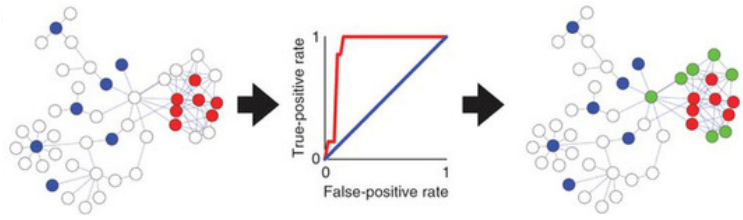
- Stephen Oliver, "Guilt-by-association goes global" *Nature* 403, 601-603 (10 February 2000)

- APOE-epsilon 4 is a major risk factor for Alzheimer's
- Discovered as a "contaminant" binding βA4 in isolates of cerebral spinal fluid

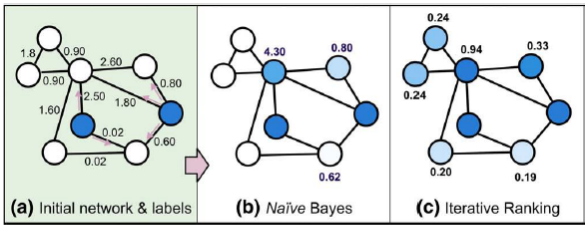
Strittmatter et al. Proc Natl Acad Sci U S A. 1993 Mar 1;90(5):1977-81.

"Guilt by association"

- Using the properties of the existing genes in a functional group (e.g. GO group) to determine inclusion of new candidate genes



More specifics: label propagation



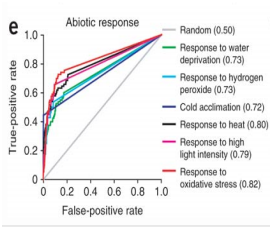
A comparison of diffusion algorithm-based methods for predicting genes associated with a function. In (a) two proteins are known to have a particular function, indicated by their color. The strength of association between proteins is indicated along each edge. (b) Naïve Bayes assigns scores to neighboring nodes. The ranking of scores is indicated by the shade of color: higher ranked proteins are more darkly colored. Note that several proteins have no score because they are not directly linked. In (c), all proteins are assigned to a score, but the overall rankings differ.

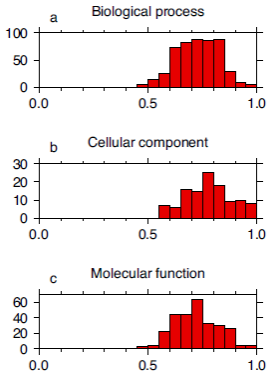
Adapted from Wang and Marcotte. It's the machine that matters: Predicting gene function and phenotype from protein networks. Journal of Proteomics, 2010.

Are many functions predictable?

Possible interpretation:

1) Different genes are learnable from the network for each category
2) The same genes are learnable from the network for each category: predict the same group of genes that are involved in all functions.



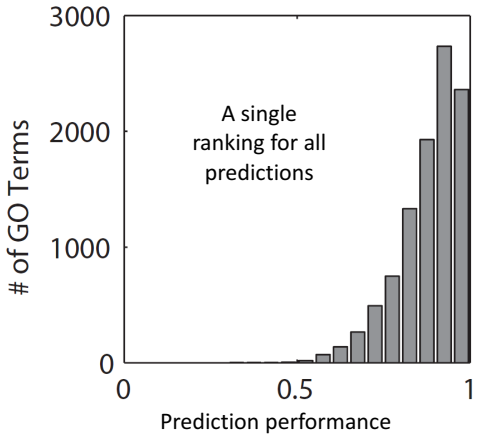


Distributions of AU-ROC based on coexpression
http://www.biomedcentral.com/1471-2105/6/227

Multifunctionality + GBA

- We suspected that the rankings provided by GBA were biased.
- We realized that the ideal bias would be to rank genes by multifunctionality
- If we want to be right about predicting a gene's function, we'd do best to choose the most multifunctional genes.
- How would we do in the ideal case?

Ranking genes by multifunctionality predicts most GO groups

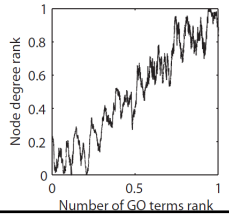


Why this matters for GBA

1. If a gene has many functions, it is more likely to have a given function.

2. A gene with more neighbours tends to have more functions

3. Should be able to predict function only by node degree

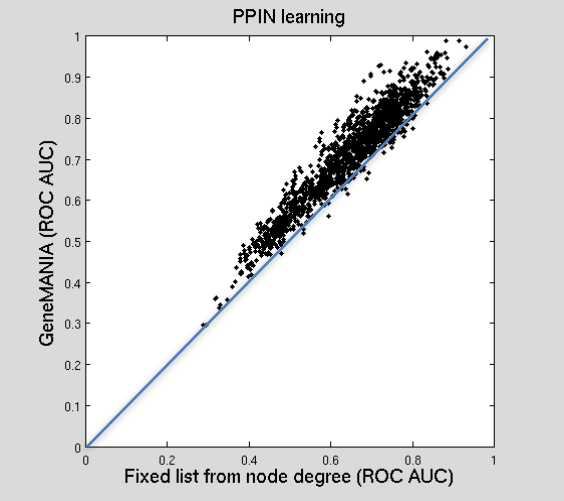


The big question: Does real data encode this list?

Consequences if true:

1. GBA will seem to work, but be of too little use.
   - There has been research into problems with GBA (training set specificity, algorithms, network specificity, promiscuity), but this isn't the same thing.
2. Any experiment which encodes this gene list will seem to have yielded "relevant" results but will not predict interesting genes (the rich get richer)

Functions that are learnable are learnable by node degree



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

Application to genetics of intellectual disability

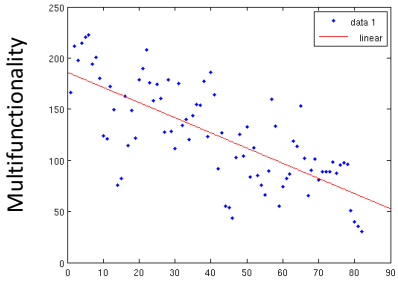
- Deletions and duplications (CNVs) implicated in intellectual disability (ID)
- Task: Identify which genes in the CNV regions are most likely "responsible" for the phenotype.
- Example: Deletion of 97 genes at 19p13.3
- Top 5 candidates from algorithm trained on "known ID genes": SH3GL1, AES, EEF2, DAPK3, GNA11
- Control experiment: For training set, use genes for another disorder, or random genes. Top 5 candidates: 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

42

What is happening?

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



Jesse Gillis, Eloi Mercier

43

Summary

- Analysis of gene networks is surprisingly tricky: it's easy for something to seem interesting
- Algorithms tend to be "black boxes": it's hard to see why a result was obtained (network huge, etc.)
- Multifunctionality causes a lot of problems in telling the "specifically interesting" from the "generically important" in networks.
  - http://f1000research.com/articles/1-14/v1 for review

Thank you for your attention!