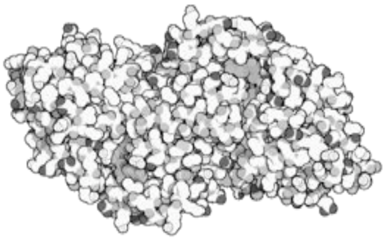
Bioinformatics 2

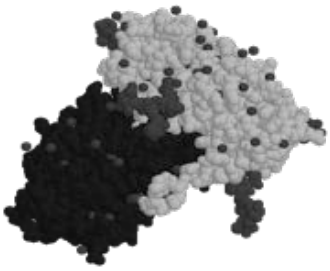
Protein (Interaction) Networks

- Biological Networks in general
- Metabolic networks
- Briefly review proteomics methods
- Protein-Protein interactions
- Protein Networks
- Protein-Protein interaction databases
- An example

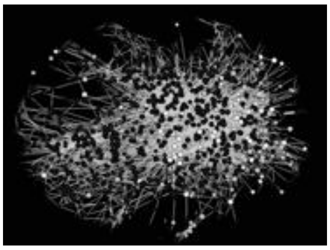
alcohol dehydrogenase



ricin (A and B)

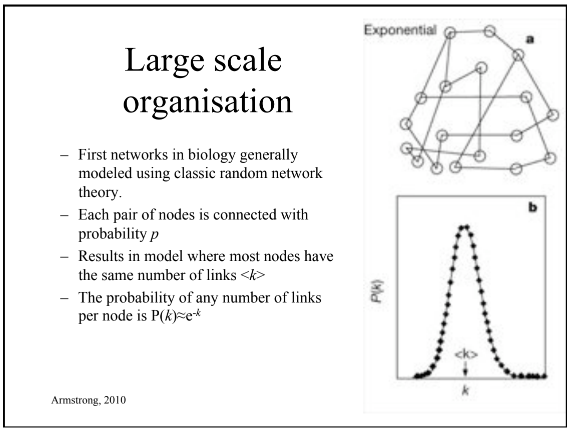


synaptic proteome



Biological Networks

- Genes - act in cascades
- Proteins - form functional complexes
- Metabolism - formed from enzymes and substrates
- The CNS - neurons act in functional networks
- Epidemiology - mechanics of disease spread
- Social networks - interactions between individuals in a population
- Food Chains

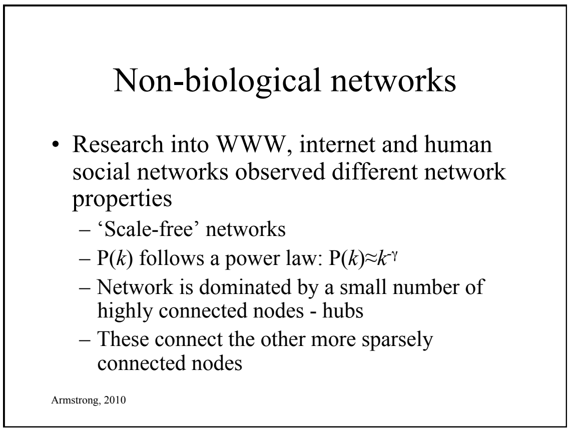


Large scale organisation

- First networks in biology generally modeled using classic random network theory.
- Each pair of nodes is connected with probability *p*
- Results in model where most nodes have the same number of links <*k*>
- The probability of any number of links per node is $P(k) \approx e^{-k}$

Armstrong, 2010

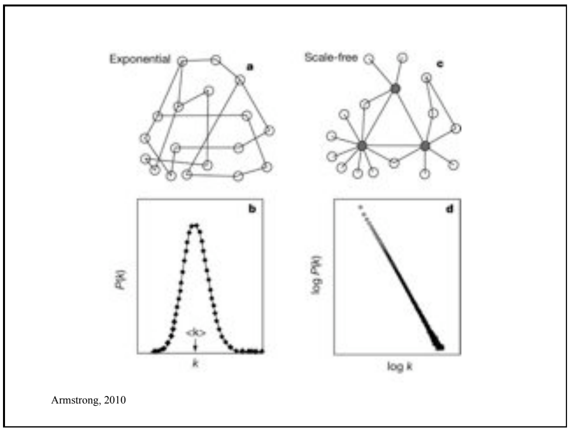
Exponential
a
b
P(k)
<k>
k



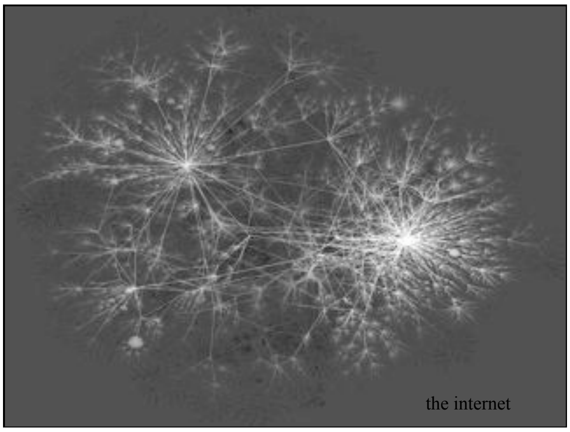
Non-biological networks

- Research into WWW, internet and human social networks observed different network properties
  - ‘Scale-free’ networks
  - P(*k*) follows a power law: $P(k) \approx k^{-\gamma}$
  - Network is dominated by a small number of highly connected nodes - hubs
  - These connect the other more sparsely connected nodes

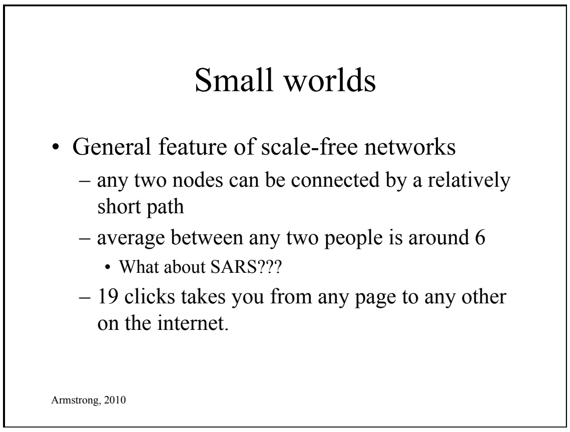
Armstrong, 2010



Armstrong, 2010



the internet



Small worlds

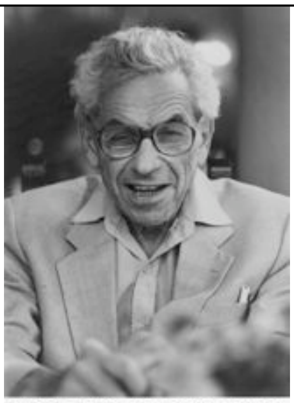
- General feature of scale-free networks
  - any two nodes can be connected by a relatively short path
  - average between any two people is around 6
    - What about SARS???
  - 19 clicks takes you from any page to any other on the internet.

Armstrong, 2010

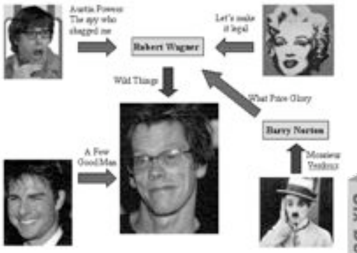
6 degrees of separation..?

- Stanley Milgram’s work in late 1960’s
- Sent letters to people in Nebraska
- Target unknown person in Massachusetts
- Average 6 ‘jumps’ to reach target

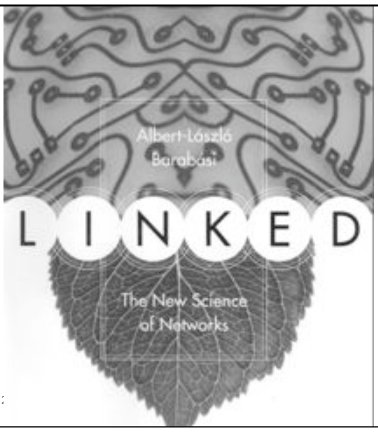
(only 5% got there)



Paul Erdős, the most prolific mathematician who ever lived, has no home and no job, but he has wandered the world for over fifty years, inspiring other mathematicians. From the documentary N is a Number: A Portrait of Paul Erdős © 1993 by George Csicsery



http://oracleofbacon.org/
(real centre is Rod Steiger)



Biological organisation

Jeong et al., 2000 The large-scale organisation of metabolic networks. Nature 407, 651-654

- Pioneering work by Oltvai and Barabasi
- Systematically examined the metabolic pathways in 43 organisms
- Used the WIT database
  - ‘what is there’ database
  - http://wit.mcs.anl.gov/WIT2/
  - Genomics of metabolic pathways

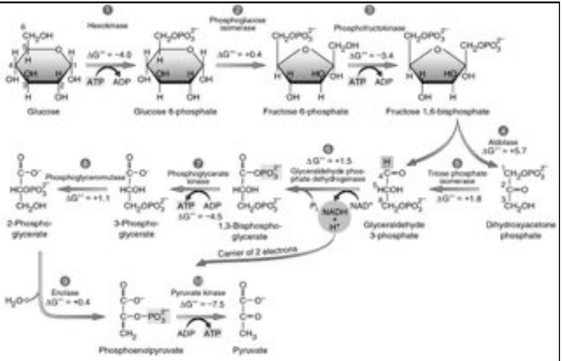
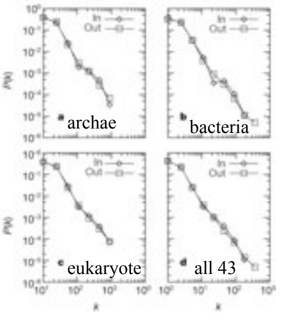


Image taken from http://fig.cox.miami.edu/~cmallery/255/255atp/255makeatp.htm

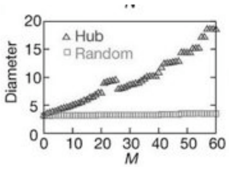
Using metabolic substrates as nodes



=scale free!!!

Random mutations in metabolic networks

- Simulate the effect of random mutations or mutations targeted towards hub nodes.
  - Measure network diameter
  - Sensitive to hub attack
  - Robust to random



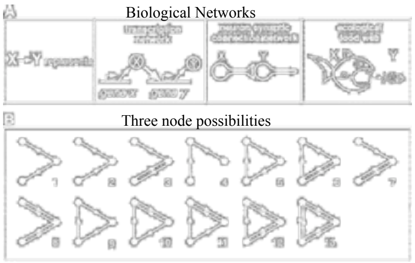
Consequences for scale free networks

- Removal of highly connected hubs leads to rapid increase in network diameter
  - Rapid degeneration into isolated clusters
  - Isolate clusters = loss of functionality
- Random mutations usually hit non hub nodes
  - therefore robust
- Redundant connectivity (many more paths between nodes)

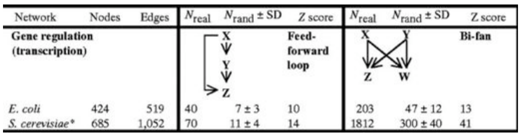
Network Motifs

- Do all types of connections exist in networks?
- Milo et al studied the transcriptional regulatory networks in yeast and E.Coli.
- Calculated all the three and four gene combinations possible and looked at their frequency

Milo et al. 2002 Network Motifs: Simple Building Blocks of Complex Networks. Science 298: 824-827

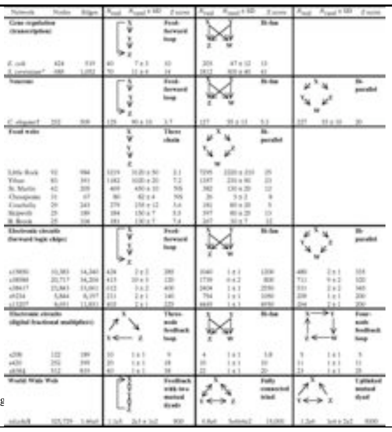


Gene sub networks



Network	Nodes	Edges	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Heavy bias in both yeast and E.coli towards these two sub network architectures



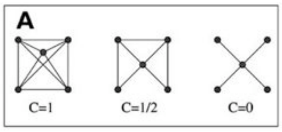
Network	Nodes	Edges	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score	N_{real}	N_{rand} ± SD	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4 ± 4e2	15,000	1.2e6	1e4 ± 2e2	5000

What about known complexes?

- OK, scale free networks are neat but how do all the different functional complexes fit into a scale free proteome arrangement?
  - e.g. ion channels, ribosome complexes etc?
- Is there substructure within scale free networks?
  - Examine the clustering co-efficient for each node.

Clustering co-efficients and networks.

- $C_i=2n/k_i(k_i-1)$
- n is the number of direct links connecting the k_i nearest neighbours of node *i*
- A node at the centre of a fully connected cluster has a C of 1



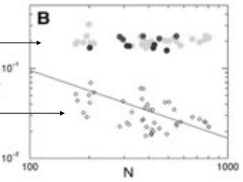
Clustering co-efficients and networks.

Ravasz et al.,(2002) Hierarchical Organisation of Modularity in Metabolic Networks. Science 297, 1551-1555

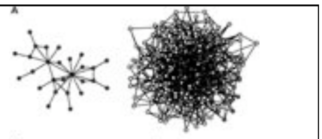
- The modularity (ave C) of the metabolic networks is an order of magnitude higher than for truly scale free networks.

Metabolic network

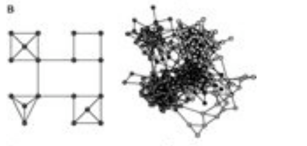
Non modular network



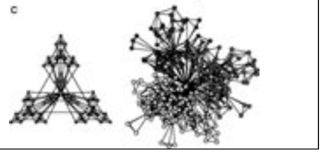
No modularity
Scale-free



Highly modular
Not scale free

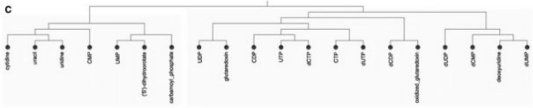


Hierarchical network
Scale-free



Clustering on C

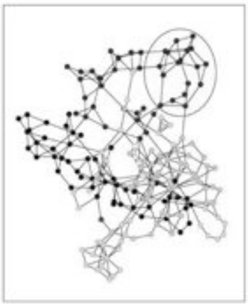
- Clustering on the basis of C allows us to rebuild the sub-domains of the network



- Producing a tree can predict functional clustered arrangements.

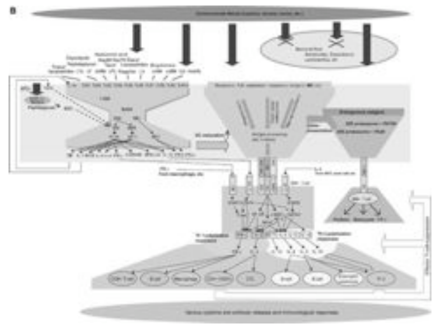
Armstrong, 2010

Cluster analysis on the network



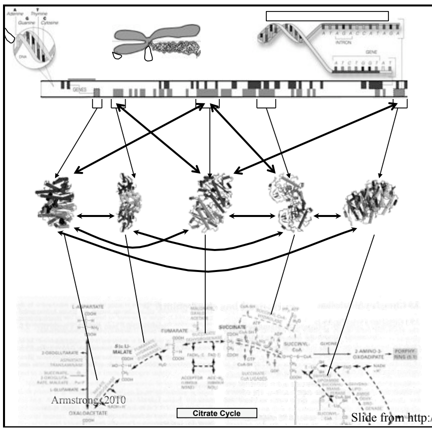
Armstrong, 2010

Bow-tie and nested bow-tie architectures



Armstrong, 2010

http://www.nature.com/msb/journal/v2/n1/fig_tab/msb4100039_F2.html



GENOME

protein-gene interactions

PROTEOME

protein-protein interactions

METABOLISM

Bio-chemical reactions

Armstrong, 2010

Slide from http://www.nd.edu/~networks/

Common Biological Networks

- Genes - Microarrays
  - cDNA arrays
  - oligonucleotide arrays
  - whole genome arrays
- Proteins - Proteomics
  - yeast two hybrid
  - PAGE techniques
  - Mass Spectrometry (Lecture 2)

Armstrong, 2010

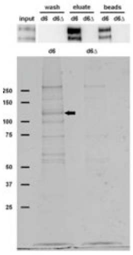
Proteomics

- What is Proteomics?
  - Protein profiling in a sample
  - Reveal protein interactions
  - Current state of proteins in sample
- What is there?
  - 2D PAGE, DiGE & Mass Spec (Juri)
- How is it connected together?

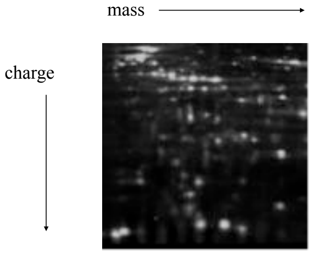
Armstrong, 2010

Proteomics - PAGE techniques

- Proteins can be run through a poly acrylamide gel (similar to that used to seqparate DNA molecules).
- Can be separated based on charge or mass.
- 2D Page separates a protein extract in two dimensions.



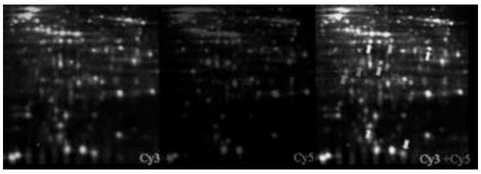
2D Page



DiGE

- We want to compare two protein extracts in the way we can compare two mRNA extracts from two paired samples
- Differential Gel Electrophoresis
- Take two protein extracts, label one green and one red (Cy3 and Cy5)

DiGE



- The ratio of green:red shows the ratio of the protein across the samples.

Identifying a protein 'blob'

- Unlike DNA microarrays, we do not normally know the identify of each 'spot' or blob on a protein gel.
- We do know two things about the proteins that comprise a blob:
  - mass
  - charge

Identifying a protein 'blob'

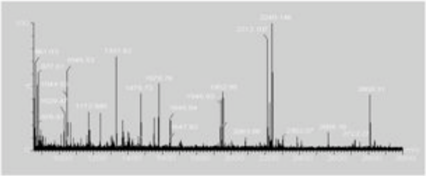
- Mass and Charge are themselves insufficient for positive identification.
- Recover from selected blobs the protein (this can be automated)
- Trypsin digest the proteins extracted from the blob (chops into small pieces)

Identifying a protein ‘blob’

- Take the small pieces and run through a mass spectrometer. This gives an accurate measurement of the weight of each.
- The total weight and mass of trypsin digested fragments is often enough to identify a protein.
- The mass spec is known as a MALDI-TOFF

Armstrong, 2010

Identifying a protein ‘blob’



MALDI-TOFF output from myosin
Good for rapid identification of single proteins.
Does not work well with protein mixtures.

Armstrong, 2010

Identifying a protein ‘blob’

- When MALDI derived information is insufficient. Need peptide sequence:
- Q-TOF allows short fragments of peptide sequences to be obtained.
- We now have a total mass for the protein, an exact mass for each trypsin fragment and some partial amino acid sequence for these fragments.

Armstrong, 2010

Protein Interactions

- Individual Proteins form functional complexes
- These complexes are semi-redundant
- The individual proteins are sparsely connected
- The networks can be represented and analysed as an undirected graph

Armstrong, 2010

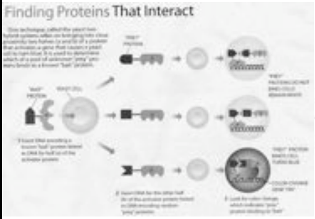
How to build a protein network

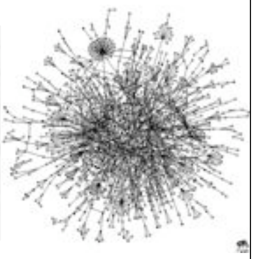
- Biological sample – how to you isolate your complex?
- What is in your complex?
- How is it connected?
  - Databases and Literature Mining
  - Yeast two hybrid screening & other cellular interaction assays
  - Mass-spec analysis
- Building and analysing the network
- An example

Armstrong, 2010

Yeast protein network

Nodes: proteins
Links: physical interactions (binding)



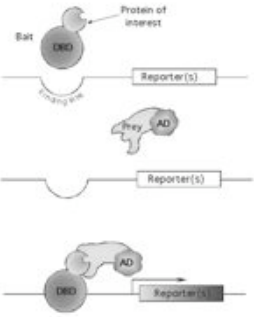


P. Uetz, *et al.* *Nature* **403**, 623-7 (2000). Armstrong, 2010

Slide from http://www.nd.edu/~networks/

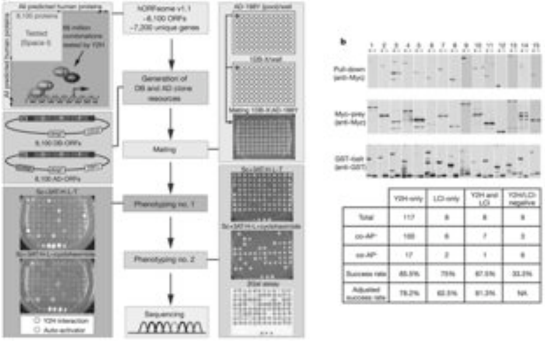
Yeast two hybrid

- Use two mating strains of yeast
- In one strain fuse one set of genes to a transcription factor DNA binding site
- In the other strain fuse the other set of genes to a transcriptional activating domain
- Where the two proteins bind, you get a functional transcription factor.



Data obtained

- Depending on sample, you get a profile of potential protein-protein interactions that can be used to predict functional protein complexes.
- False positives are frequent.
- Can be confirmed by affinity purification etc.

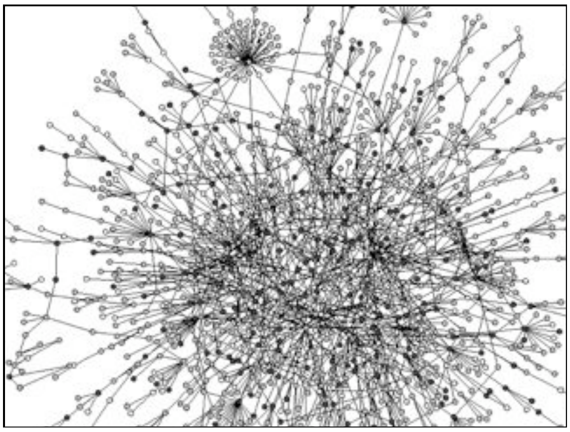


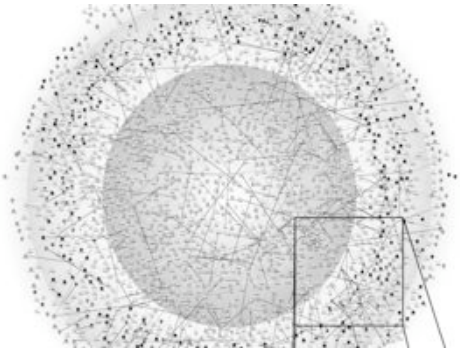
	Y2H-only	LCI-only	Y2H and LCI	Y2H/LCI-negative
Total	117	8	8	9
co-AP+	100	6	7	3
co-AP−	17	2	1	6
Success rate	85.5%	75%	87.5%	33.3%
Adjusted success rate	78.2%	62.5%	81.3%	NA

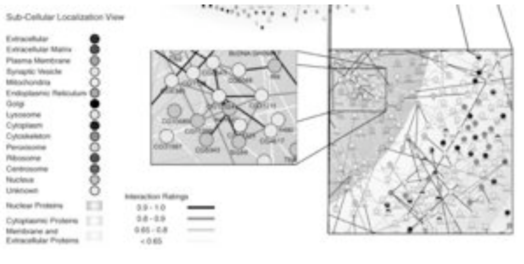
Interaction mapping schema from Rual et al 2005

Protein Networks

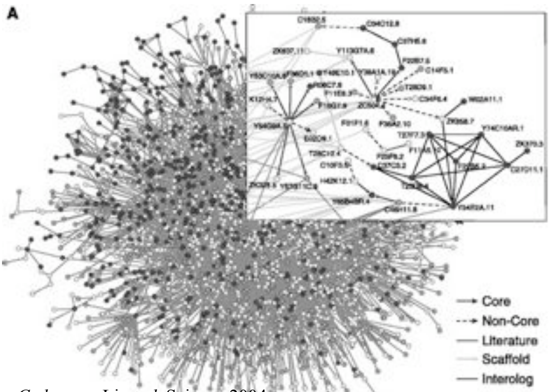
- Networks derived from high throughput yeast 2 hybrid techniques
  - yeast
  - *Drosophila melanogaster*
  - *C.elegans*
- Predictive value of reconstructed networks



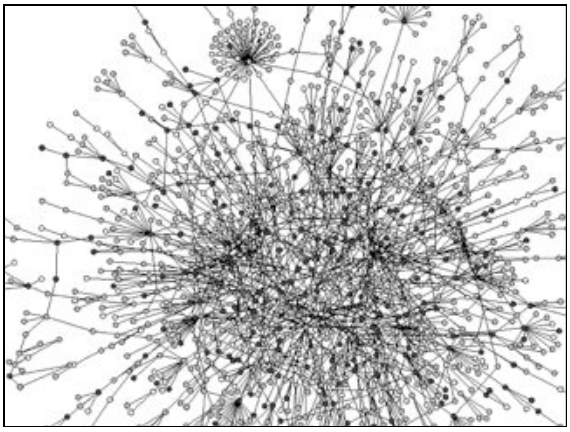




Giot *et al*, Science 2003



C.elegans: Li *et al,* Science 2004



Predictive value of networks

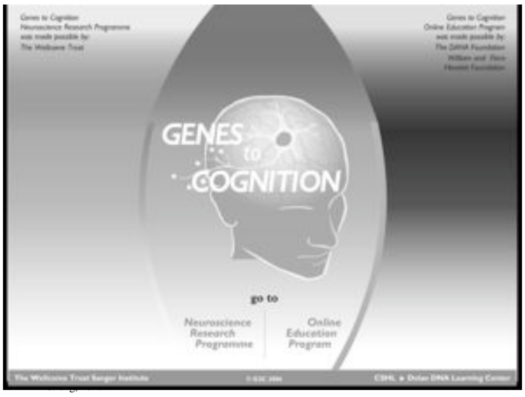
Jeong et al., (2001) Lethality and Centrality in protein networks. Nature 411 p41

- In the yeast genome, the essential vs. unessential genes are known.
- Rank the most connected genes
- Compare known lethal genes with rank order

k	fraction	%lethal
<6	93%	21%
>15	0.7%	62%

A walk-through example…

See linked papers on for further methodological details



Genes to Cognition
Neuroscience Research Programme
was made possible by:
The Wellcome Trust

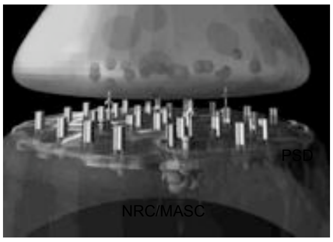
Genes to Cognition
Online Education Program
was made possible by:
The DANA Foundation
William and Flora
Hewlett Foundation

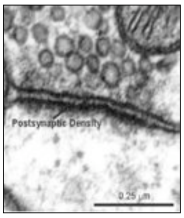
GENES to COGNITION

go to

Neuroscience Research Programme | Online Education Program

The Wellcome Trust Sanger Institute © G2C 2006 CSHL ► Dolan DNA Learning Center

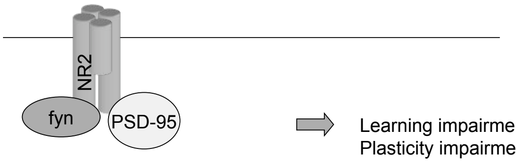




Armstrong, 2010

Genetic evidence for postsynaptic complexes

NMDA receptor



Learning impairments
Plasticity impairments

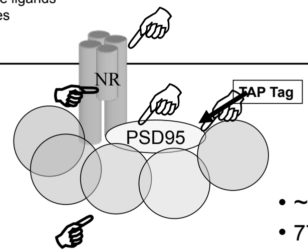
Grant, et al.	Science, 258, 1903-10. 1992
Migaud et al,	Nature , 396; 433-439. 1998
Sprengel et al.	Cell 92, 279-89. 1998

Armstrong, 2010

Proteomic characterisation of NRC / MASC

(MAGUK Associated Signaling Complex)

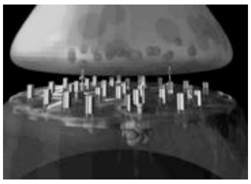
- glutamate ligands
- antibodies
- peptides
- TAP Tag



- ~2 MDa
- 77 proteins (2000)
- 186 (2005)

Husi et al.	Nature Neuroscience, 3, 661-669. 2000.
Husi & Grant.	J. Neurochem, 77, 281-291. 2001
Collins et al,	J. Neurochem. 2005

Armstrong, 2010



Post Synaptic Density	1124
ER:microsomes	491
Splicesome	311
NRC/MASC	186
Nucleolus	147
Peroxisomes	181
Mitochondria	179
Phagosomes	140
Golgi	81
Choroplasts	81
Lysosomes	27
Exosomes	21

Armstrong, 2010

Grant. (2010) Biochemical Society Transactions. 34, 59-63. 2010

Literature Mining

- 680 proteins identified from protein preps
- Many already known to interact with each other
- Also interact with other known proteins
  - Immunoprecipitation is not sensitive (only finds abundant proteins)
- Literature searching has identified a group of around 4200 proteins
  - Currently we have extensive interaction data on 1700

Armstrong, 2010

Annotating the DB

- How do we find existing interactions?
  - **Search PubMed with keyword and synonym combinations**
  - Download abstracts
  - Sub-select and rank-order using regex’s
  - Fast web interface displays the most ‘productive’ abstracts for each potential interaction

Keyword and synonym problem

- PSD-95:
  - DLG4,PSD-95,PSD95,Sap90,Tip-15,Tip15, Post Synatpic Density Protein - 95kD, PSD 95, Discs, large homolog 4, Presynaptic density protein 95
- NR2a:
  - Glutamate [NMDA] receptor subunit epsilon 1 precursor (N-methyl D-aspartate receptor subtype 2A) (NR2A) (NMDAR2A) (hNR2A) NR2a
- Protein interactions:
  - interacts with, binds to, does not bind to….

.+\sand\s.+\sinteract

(1..N characters) (space) and (1..N characters) interact

.+\s((is)|(was))\sbound\sto\s.+\s

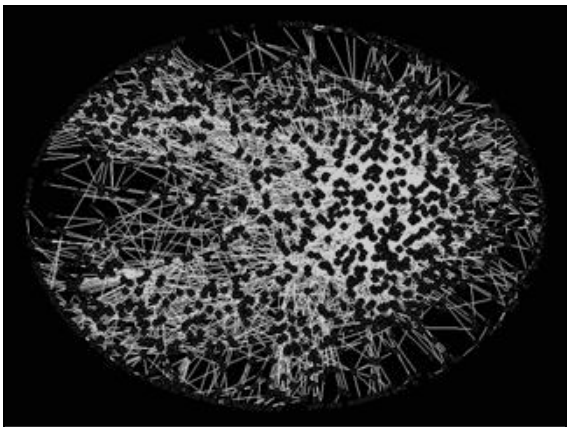
(1..N characters) (space) (is or was) (space) bound (space) to (1..N characters) (space)

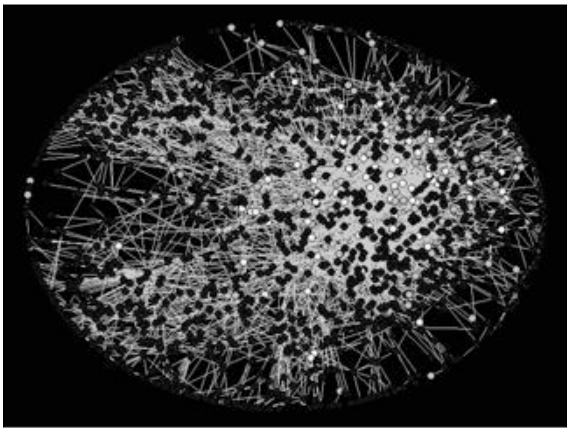
.+\sbinding\sof\s.+\s((and)|(to))\s.+

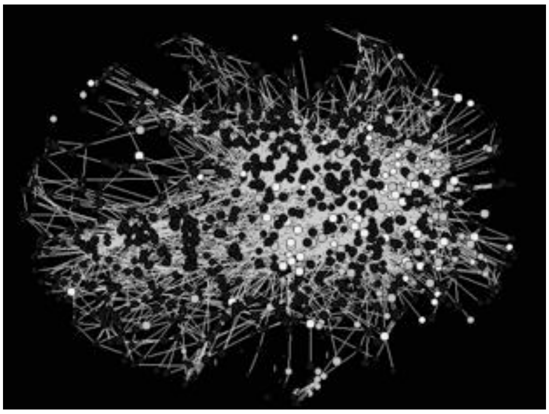
(1..N characters) (space) binding (space) of (and or to) (space) (1..N characters)

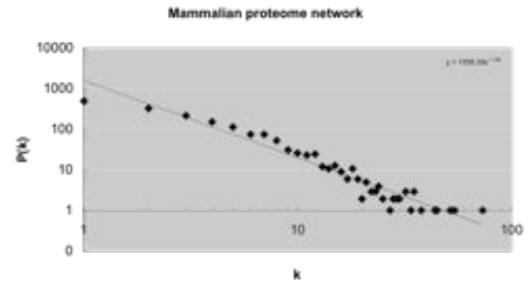
Annotating the DB

- How do we find existing interactions?
  - Search PubMed with keyword and synonym combinations
  - Download abstracts
  - Sub-select and rank-order using regex’s
  - Fast web interface displays the most ‘productive’ abstracts for each potential interaction
  - *Learn from good vs. bad abstracts*



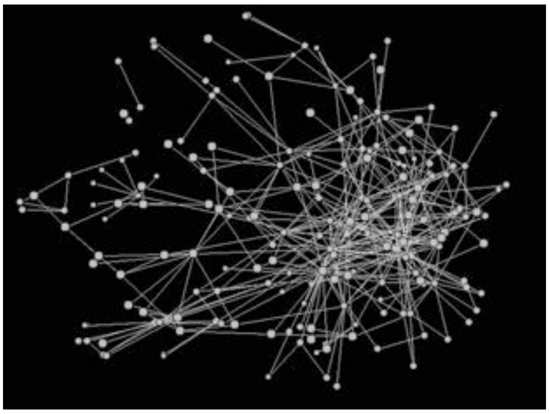


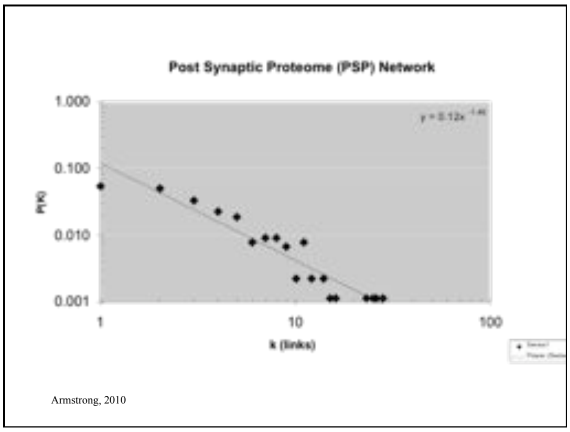




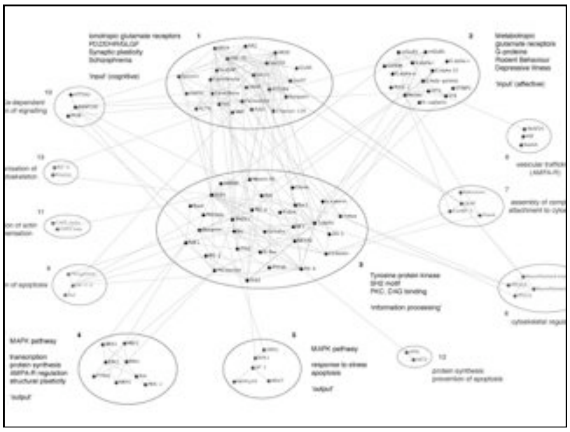
1729 proteins, 7739 links

Armstrong, 2010





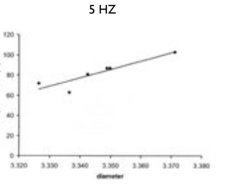
Armstrong, 2010



Simulated disruption vs. mutations

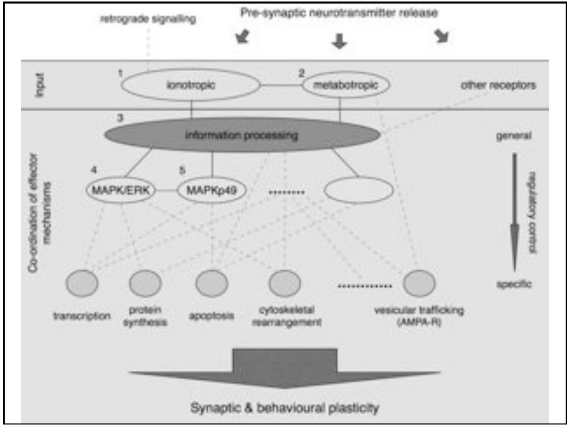
Linear correlation between simulation and *in vivo* assay

Details: Mutations in MEK1, SynGAP, NR2AC, PKA, PI3-kinase, PSD-95 were all analysed in a single laboratory (TJ O'Dell, UCSD) under controlled conditions and LTP disruption measured. (p<0.05)



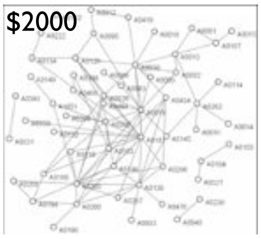
H. Husi J. Choudhary L. Yu M. Cumiskey W. Blackstock T.J. O'Dell P.M. Visscher J.D. Armstrong S.G.N. Grant, unpublished

Armstrong, 2010

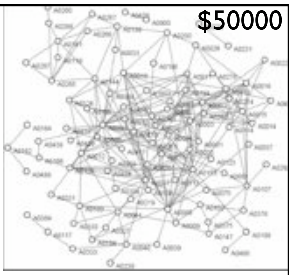


datasources

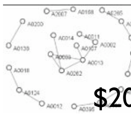
(jan 2005)



$2000



$50000



$200

Netpro (commercial)
56 proteins, 94 interactions
40% agreement in predictions

Armstrong, 2010

BIND/MINT etc
22 proteins
16 interactions

Synapse proteome summary

- Protein parts list from proteomics
- Literature searching produced a network
- Network is essentially scale free
- Hubs more important in cognitive processes
- Network clusters show functional subdivision
- Overall architecture resembles bow-tie model
- Expensive…

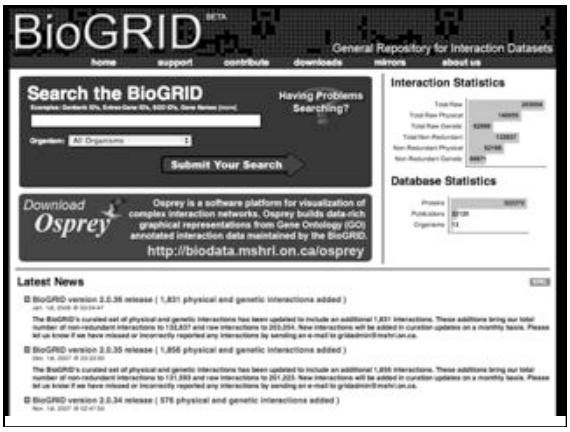
Armstrong, 2010

Protein (and gene) interaction databases

BioGRID- A Database of Genetic and Physical Interactions
DIP - Database of Interacting Proteins
MINT - A Molecular Interactions Database
IntAct - EMBL-EBI Protein Interaction
MIPS - Comprehensive Yeast Protein-Protein interactions
Yeast Protein Interactions - Yeast two-hybrid results from Fields' group
PathCalling- A yeast protein interaction database by Curagen
SPiD - Bacillus subtilis Protein Interaction Database
AllFuse - Functional Associations of Proteins in Complete Genomes
BRITE - Biomolecular Relations in Information Transmission and Expression
ProMesh - A Protein-Protein Interaction Database
The PIM Database - by Hybrigenics
Mouse Protein-Protein interactions
Human herpesvirus 1 Protein-Protein interactions
Human Protein Reference Database
BOND - The Biomolecular Object Network Databank. Former BIND
MDSP - Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectromet
Protcom - Database of protein-protein complexes enriched with the domain-domain structures
Proteins that interact with GroEL and factors that affect their release
DPIDB - DNA-Protein Interaction Database
YPD™ - Yeast Proteome Database by Incyte

Source with links: http://proteome.wayne.edu/PIDBL.html

Armstrong, 2010



Osprey is a software platform for visualization of complex interaction networks. Osprey builds data-rich graphical representations from Gene Ontology (GO) annotated interaction data maintained by the BioGRID.
http://biodata.mshri.on.ca/osprey

BioGRID version 2.0.36 release (1,831 physical and genetic interactions added)

The BioGRID's curated set of physical and genetic interactions has been updated to include an additional 1,831 interactions. These additions bring our total number of non-redundant interactions to 133,837 and raw interactions to 203,054. New interactions will be added in curation updates on a monthly basis. Please let us know if we have missed or incorrectly reported any interactions by sending an e-mail to gridadmin@mshri.on.ca.

BioGRID version 2.0.35 release (1,856 physical and genetic interactions added)

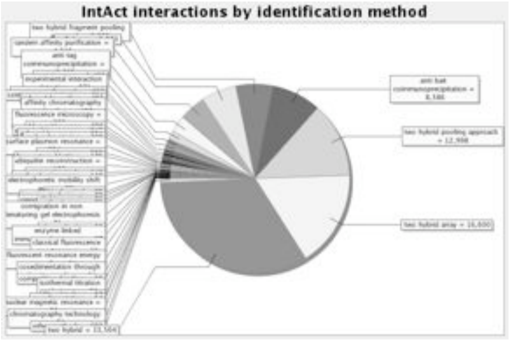
The BioGRID's curated set of physical and genetic interactions has been updated to include an additional 1,856 interactions. These additions bring our total number of non-redundant interactions to 131,593 and raw interactions to 201,223. New interactions will be added in curation updates on a monthly basis. Please let us know if we have missed or incorrectly reported any interactions by sending an e-mail to gridadmin@mshri.on.ca.

BioGRID version 2.0.34 release (576 physical and genetic interactions added)

IntAct : www.ebi.ac.uk/intact

Armstrong, 2010

IntAct : www.ebi.ac.uk/intact

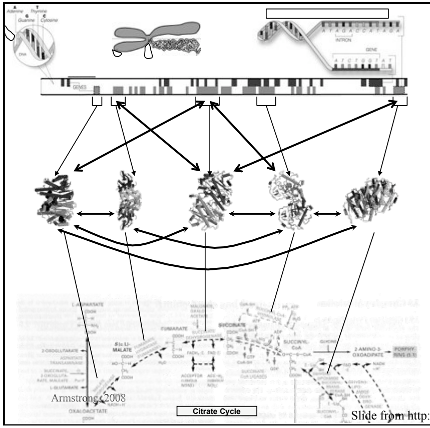


Armstrong, 2010

comparing two approaches

- Pocklington et al 2006
  - Emphasis on QC and literature mining
  - Focussed on subset of molecules
- Rual et al 2005
  - Emphasis on un-biased measurements
  - Focussed on proteome wide models
- Both then look at disease/network correlations

Armstrong, 2010



GENOME

protein-gene interactions

PROTEOME

protein-protein interactions

METABOLISM

Bio-chemical reactions

Armstrong, 2008

Slide from http://www.nd.edu/~networks/