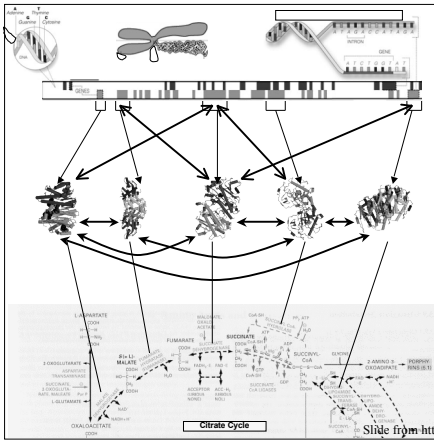
Bioinformatics 2

From genomics & proteomics to biological networks



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

Biological Profiling

- Microarrays
  - cDNA arrays
  - oligonucleotide arrays
  - whole genome arrays
- Proteomics
  - yeast two hybrid
  - PAGE techniques

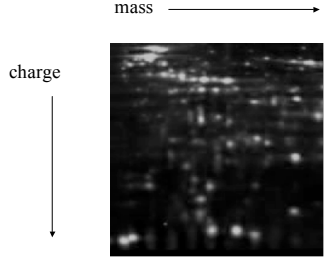
How to build a protein network

- What is there
- High throughput 2D PAGE
- Automatic analysis of 2D Page
- How is it connected
- Yeast two hybrid screening
- Building and analysing the network
- An example

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.

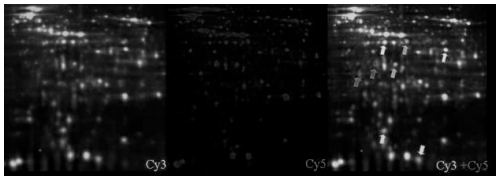
2D Page



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

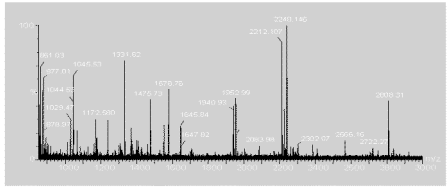
Identifying a protein 'blob'

- 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

Identifying a protein 'blob'



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

Identifying a protein 'blob'

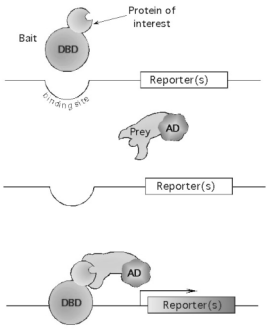
- When MALDI derived information is insufficient. Need peptide sequence:
- Q-TOF allows short fragments of peptide sequences to be obtained.
- We now have a total mass for the protein, an exact mass for each trypsin fragment and some partial amino acid sequence for these fragments.

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.

How to build a protein network

- What is there
- High throughput 2D PAGE
- Automatic analysis of 2D Page
- How is it connected
- Yeast two hybrid screening
- Building and analysing the network
- An example

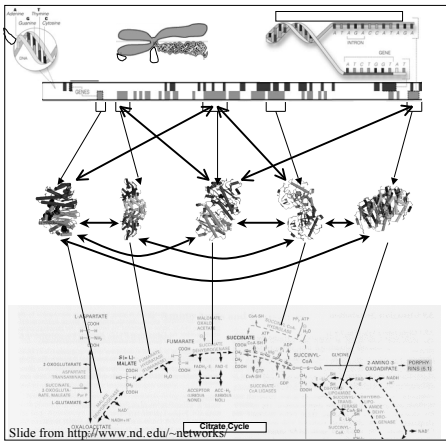


Data obtained

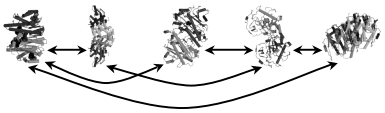
- Depending on sample, you get a profile of potential protein-protein interactions that can be used to predict functional protein complexes.
- False positives are frequent.
- Can be confirmed by affinity purification etc.

How to build a protein network

- What is there
- High throughput 2D PAGE
- Automatic analysis of 2D Page
- How is it connected
- Yeast two hybrid screening
- Building and analysing the network
- An example



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



Armstrong, 2007

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

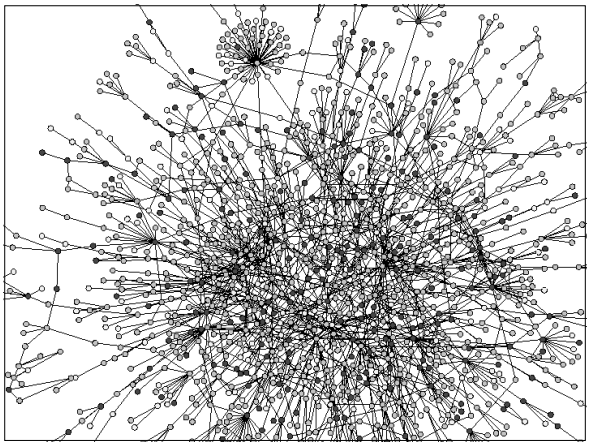
Protein Networks

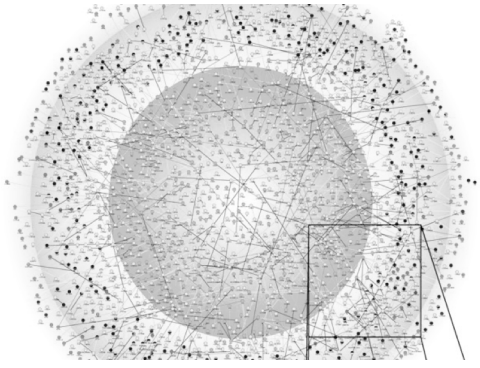
Armstrong, 2007

Protein Networks

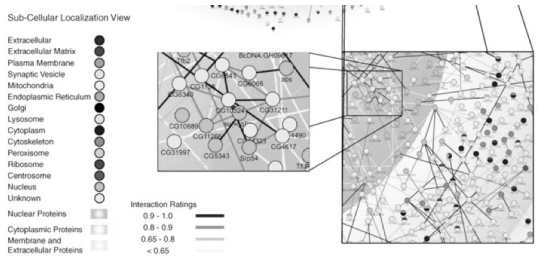
- Networks derived from high throughput yeast 2 hybrid techniques
  - yeast
  - *Drosophila melanogaster*
  - *C.elegans*
- Predictive value of reconstructed networks
- Sub-clusters and sub-architecture
- Comparison with known sub-networks, pathways and protein complexes

Armstrong, 2007



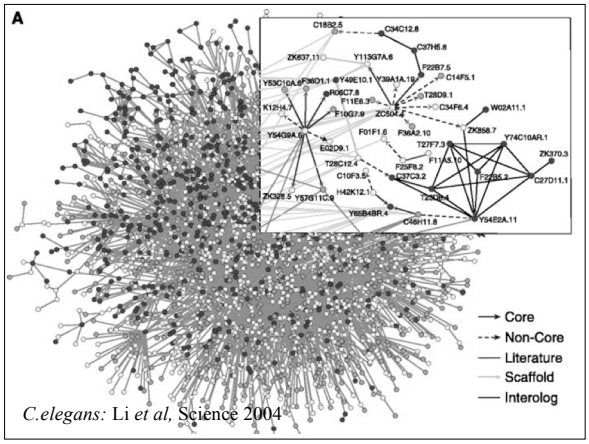


Armstrong, 2007



Giot *et al*, Science 2003

Armstrong, 2007



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

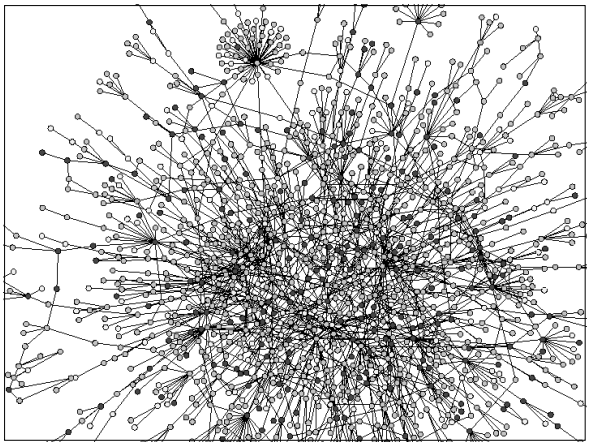
Predictive value of networks

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%

Armstrong, 2007



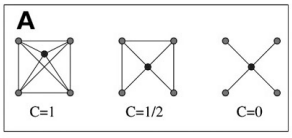
What about known complexes?

- 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.

Armstrong, 2007

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



Armstrong, 2007

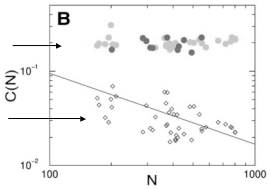
Clustering co-efficients and networks.

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 →

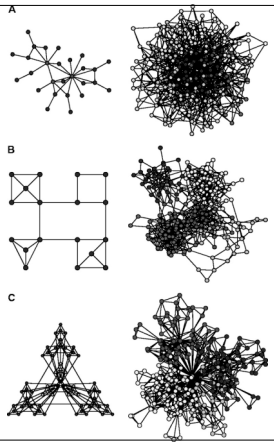
Non modular network →



No modularity
Scale-free

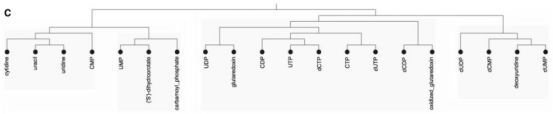
Highly modular
Not scale free

Hierarchical network
Scale-free



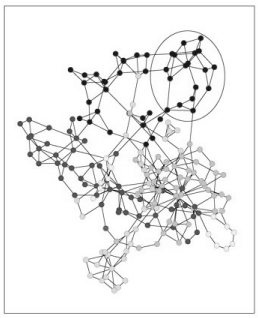
Clustering on C

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



- Producing a tree can predict functional clustered arrangements.

Cluster analysis on the network



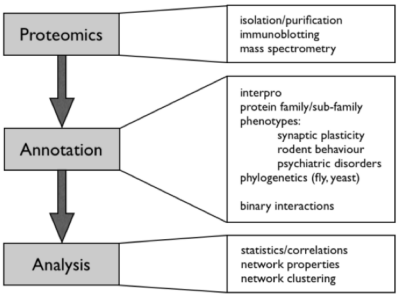
Reconstructing the cognitive proteome

J Douglas Armstrong
Edinburgh Centre for Bioinformatics
University of Edinburgh

Genes 2 Cognition
www.genes2cognition.org

University of Edinburgh
Wellcome Trust Sanger Institute
MRC Human Genetics Unit

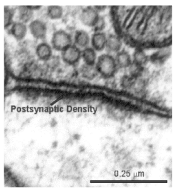
Informatics; Rodent Models (functional genomics, proteomics, gene knock-outs and replacement, behaviour and electrophysiology); Human molecular psychiatry
PI - Seth Grant, 12 co-PIs



Armstrong, 2007

Synapse proteomes

	#PSD proteins
Collins et al	620
Yoshimura et al	441
Jordan et al	401
Peng et al	328
Li et al	151
Satoh et al	46
Walikonis et al	29
Literature	119
Total PSD	1124
Consensus PSD	466 (2 or more studies)

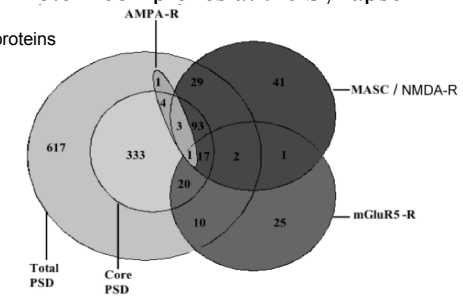


Total PSD proteins	1124
NRC/MASC	186
Post Synaptic proteome	1168

Armstrong, 2007

Protein complexes at the Synapse

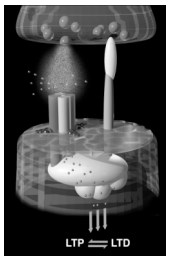
1124 proteins



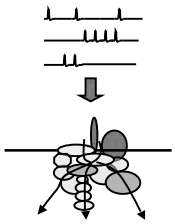
Collins et al, 2005

Armstrong, 2007

NRC / MASC



2-3 MDa
186 proteins



synaptic strength
gene expression
mRNA turnover
Protein turnover

Behaviours
Cognition & plasticity

Migaud et al, Nature, 396; 433-439. 1998;
Husi et al, Nature Neuroscience, 2000

Armstrong, 2007

The synaptic proteome is enriched for proteins containing signalling related domains

	% MASC	% Mouse	ratio
Protein kinase	11.8	3.75	3.16
Ser/Thr protein kinase	10.2	1.69	6.05
SH3	8.06	1.51	5.33
Pleckstrin-like	5.91	1.25	4.72
PDZ/DHR/GLGF	5.91	0.74	8.04
Small GTP-binding domain	5.38	1.49	3.62
Pleckstrin homology	4.84	1.08	4.49
Calcium-binding EF-hand	4.84	1.65	2.93
C2	4.30	0.82	5.26
IQ calmodulin-binding region	3.76	0.31	12.0

Armstrong, 2007

Non-Sequence Annotation

- Clinical:
  - Schizophrenia, Mental Retardation, Bipolar Disorder, Depression
- Model Organisms:
  - Rodent behaviour
  - Rodent electrophysiology: LTP/LTD.
- *Text mining*

Mark Cumiskey, Mike Marshall, Keri Page.

Armstrong, 2007

Annotation of MASC proteins

		expect (random)
Schizophrenia	33	3
Bipolar disorder	12	
Depression	14	
Mental retardation	23	
LTP	44	3
Rodent spatial learning	32	2
Rodent fear conditioning	25	1
	(186)	

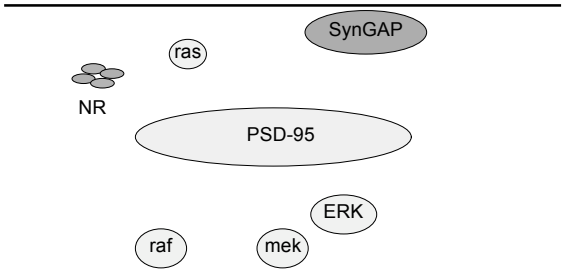
Armstrong, 2007

Protein list

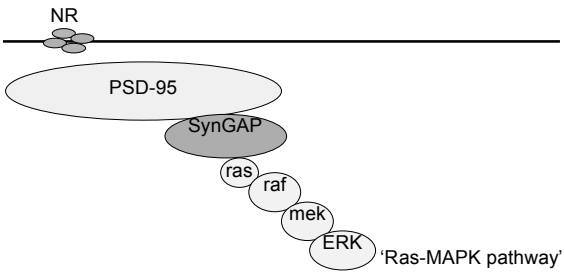
- 186 molecules closely bound to NR2A
- >1000 molecules in PSD
- heavily enriched for signalling proteins
- heavily enriched for proteins linked to human cognition and rodent behaviour
- *what about pathways and structure?*

Armstrong, 2007

aim: rebuild the network from the proteomics list



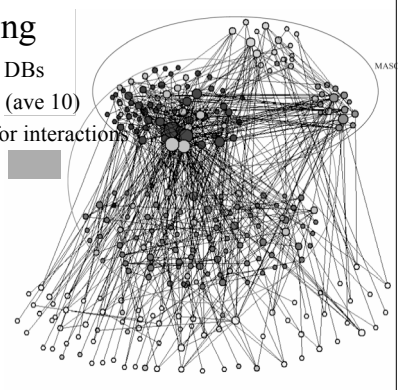
Armstrong, 2007



Armstrong, 2007

Text Mining

Start with existing DBs
Find all synonyms (ave 10)
REGEX patterns for interactions
Manual Curation
Checked twice

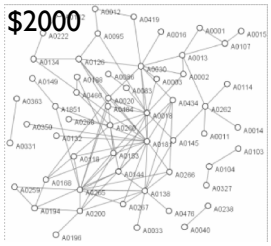


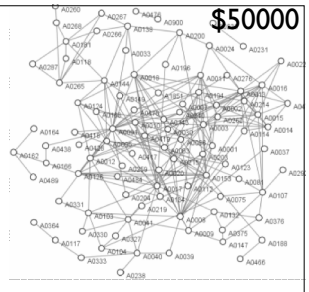
Mark Cumiskey, Keri Page & Mike Marshall
www.ppid.org

Armstrong, 2007

datasources

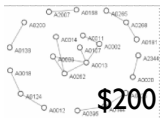
(jan 2005)

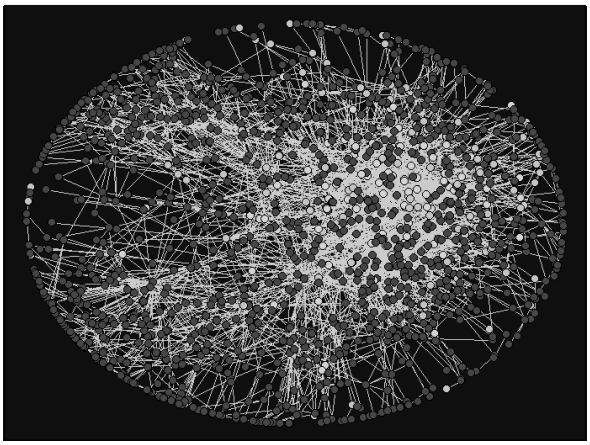




Netpro (commercial)
56 proteins, 94 interactions
40% agreement in predictions

BIND/MINT etc
22 proteins
16 interactions





Predictive value of networks

Jeong et al Nature 2000

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

Synapse proteome predictions

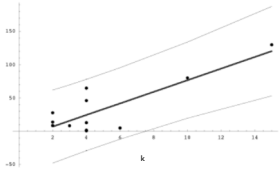
- Biology:
  - LTP - change in neuron response after experience (electrophysiological)
  - Mouse KOs
- Network Analysis
  - vertex degree (number of protein interactions)
  - network diameter (average shortest path after simulated protein deletion)

Armstrong, 2007

Simulated disruption vs. mutations

100Hz LTP data crated from literature.

Linear correlation between simulation and *in vivo* assay. ($p<0.01$)

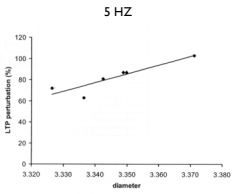


Armstrong, 2007

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, UCLA) under controlled conditions and LTP disruption measured. ($p<0.05$)



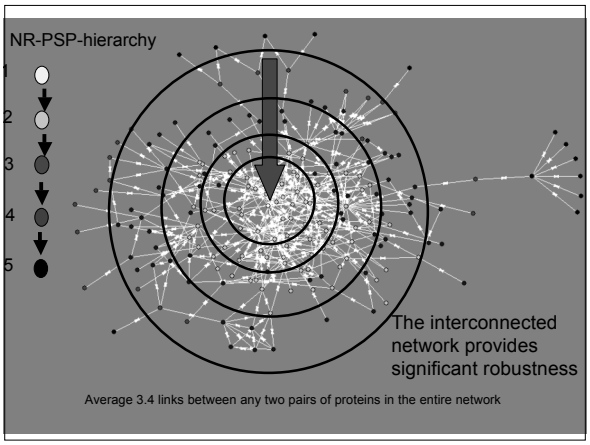
H. Hsiao J. Choudhary L. Yu M. Cumiskey W. Blackstock T.J. O'Dell P.M. Visscher J.D. Armstrong S.G.N.Grant, unpublished

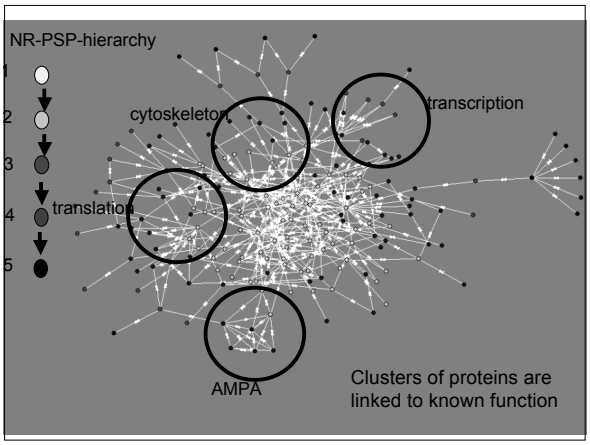
Armstrong, 2007

robust network

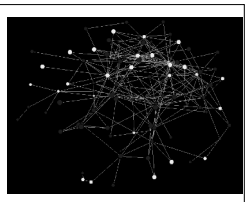
- Biological or Simulated disruption of key molecules the network does not abolish LTP
- Redundancy in signalling pathways
- Need to consider multiple targets/pathways

Armstrong, 2007





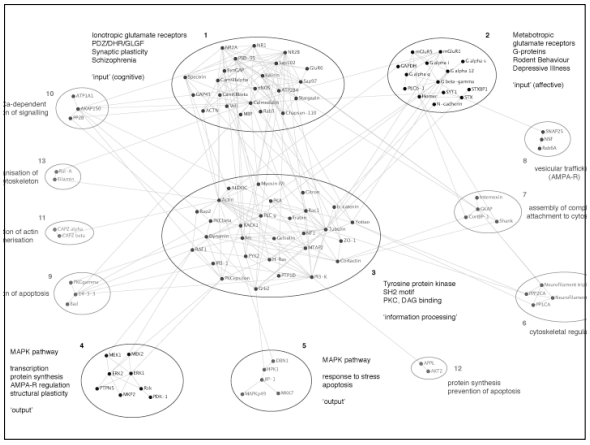
Community stucture based clustering

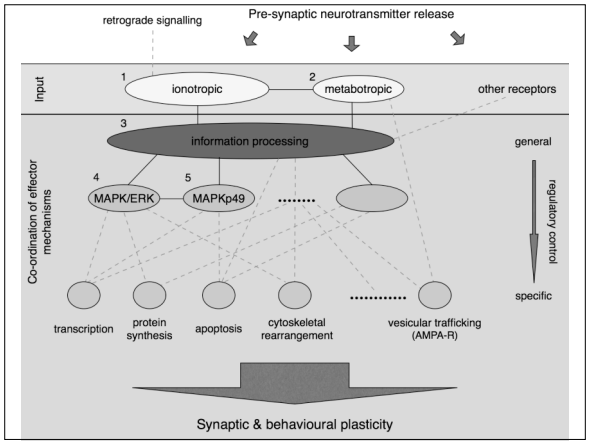


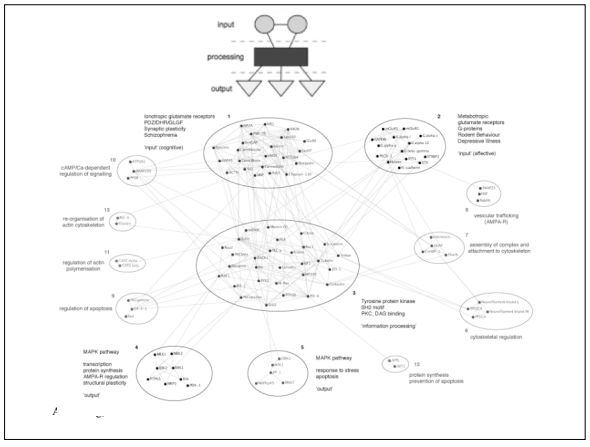
- Choose a start node/protein at random
- Follow a random walk adding 1 to the value of each interaction passed
- Repeat
- Select highest valued interaction and remove
- Continue until network fragments

Newman and Girvan 2003

Armstrong, 2007

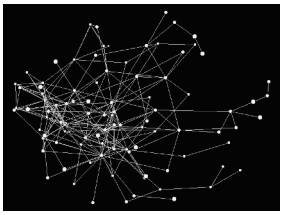






core network properties

architecture relates to function
small world nature gives robustness
underlying modular substructure
modules have specific functionality



what about dynamics?

- regulation within network
- evolution from simple nervous systems
- expression patterns across brain regions

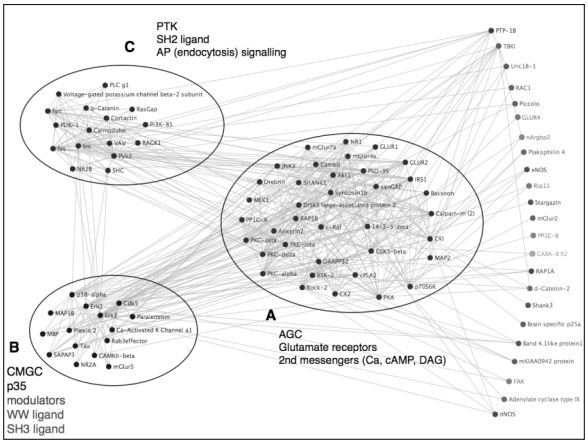
Armstrong, 2007

regulation/dynamics

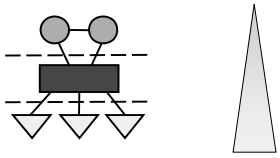
- 25 kinases
- 600 potential phosphorylation sites in PSP
- phospho-peptide array
- existing models of a few kinase pathways

Armstrong, 2007

Marcelo Coba



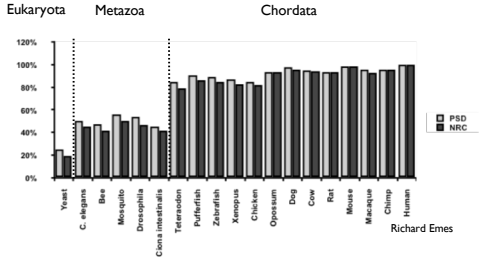
phospho-regulation in NRC/MASC



Increasing phospho regulation

Comparative genomics of postsynaptic proteome:

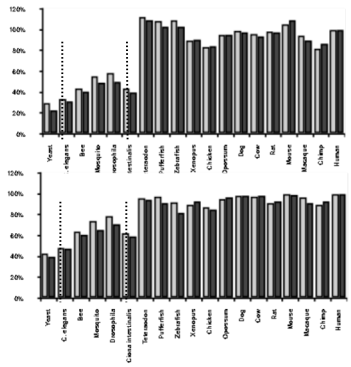
- 570 genes: 186, NRC/MASC; 570 PSD
- 19 species
  - number of synapse orthologues



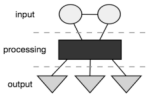
Richard Emes

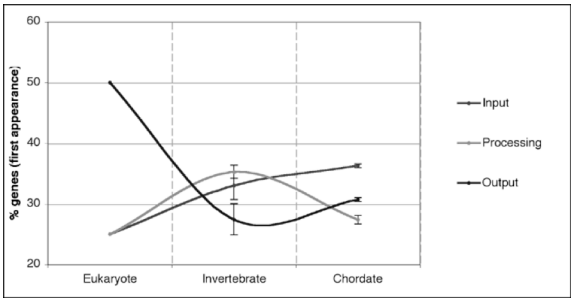
Domain number

Domain type

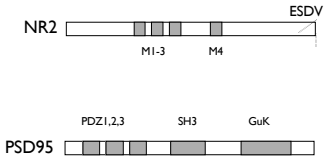


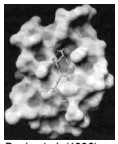
Phylogeny of modules in NRC/MASC



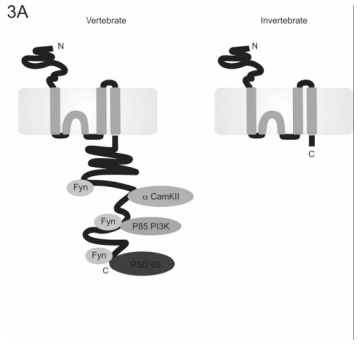


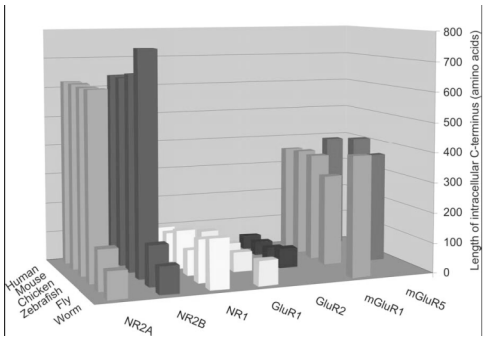
Forming the complex between NMDA receptor and MAGUKs





Doyle et al, (1996)

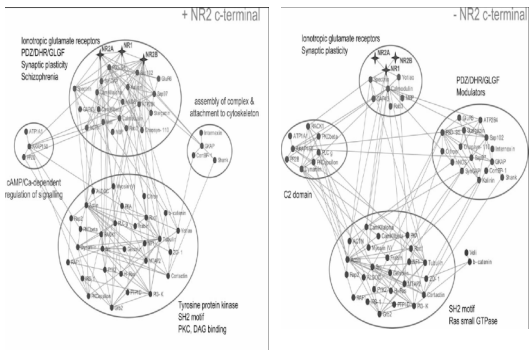




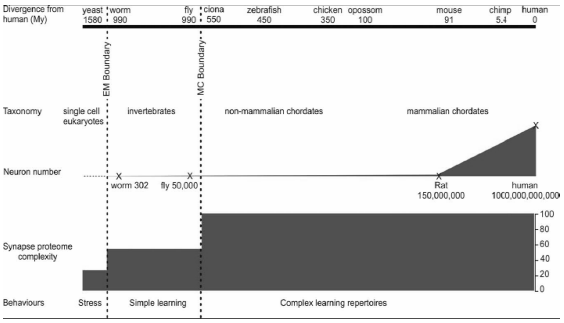
Vertebrate NR2 cytoplasmic C-terminal motifs absent in invertebrates

Name	motif (ENSP00000279593)	Start
PXSP motif	PRSP	1114
not named ?	CxxCxxxxNLYDIxED	1242
fyn site	Y	1246
PXSP motif	PQSP	1282
CanKII binding	RQHSYD	1300
PKC	S	1303
PKC	S	1323
p85 P13K binding	YxxM	1336
fyn site	Y	1336
AP-2 binding	YEKL	1474
fyn site	Y	1474
CK2 site	S	1481
PDZ binding	ESDV	1481

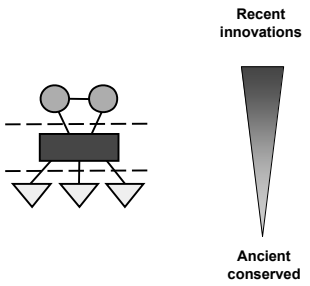
signaling complexity increased



Evolution of brain complexity across multiple levels



NRCMASC modules phylogeny



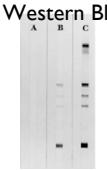
what does the complex look like across the brain?

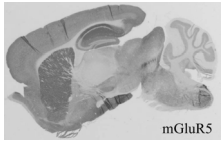
Chris Anderson

Western Blot

Immunohistochemistry

65

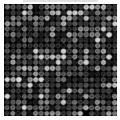


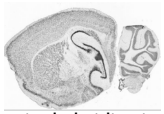


28

protein

148





32

mRNA

Microarray

In situ hybridisation

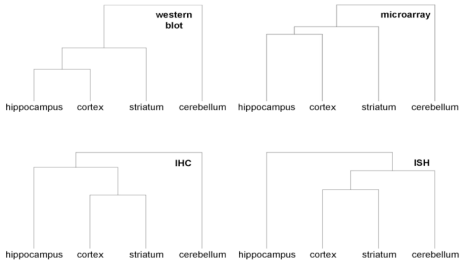
Zapala PNAS 2005

Armstrong, 2007

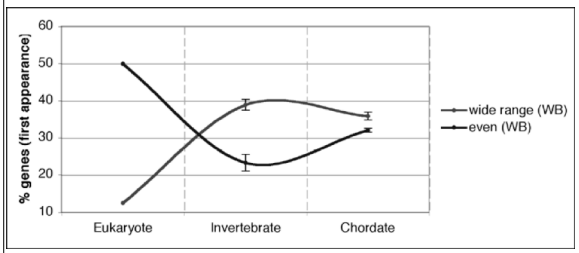
St Jude Brain Gene Expression Map

Quantitative → Qualitative

Combinations of expression profiles defines brain regions

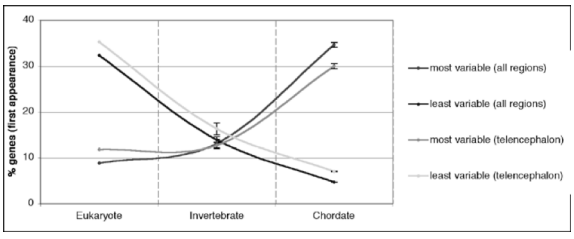


Western data – 65 proteins, 4 regions



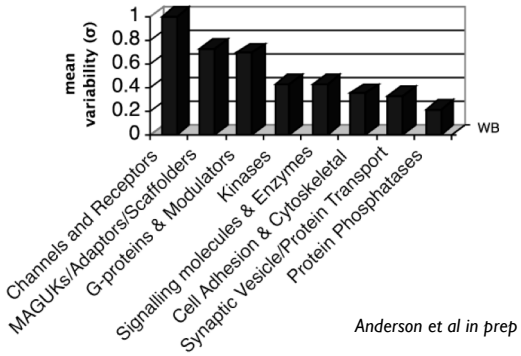
Recently evolved proteins show higher regional variation in expression level

microarray data – 148 mRNAs, 22 regions



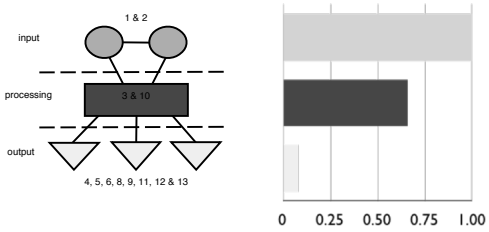
Least variable genes mostly ancient origin.
Most variable genes mostly recent/chordate origin

variability with protein type



Anderson et al in prep

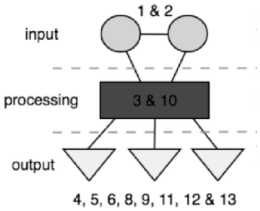
Expression variance verses module



molecular computation in fly neurons

behav	lethal
67%	65%
40%	75%
50%	**100%**

90 orthologues in fly
1471 total alleles



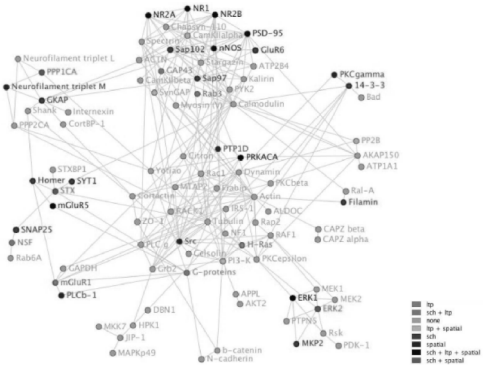
All main clusters and structure represented

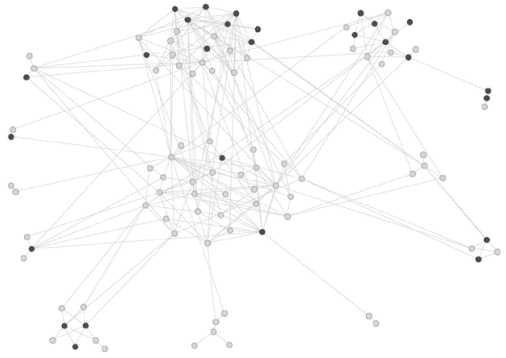
Significant correlation between fly behaviour phenotype and affective (P<0.004) and cognitive (P<0.01) disorders

Armstrong, 2007

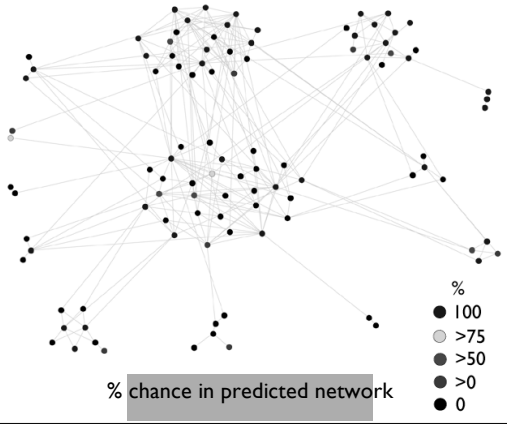
Emes et al

Can we use network structure to predict new disease genes?

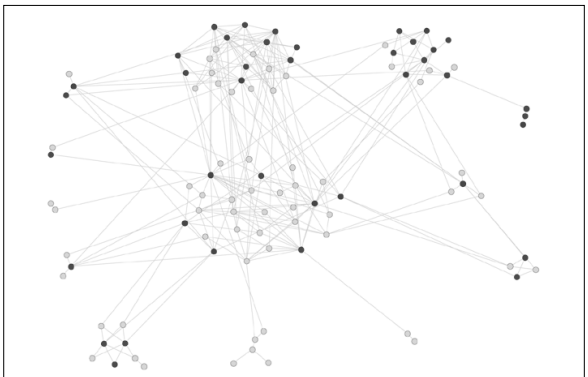




genes linked to schizophrenia (magenta)

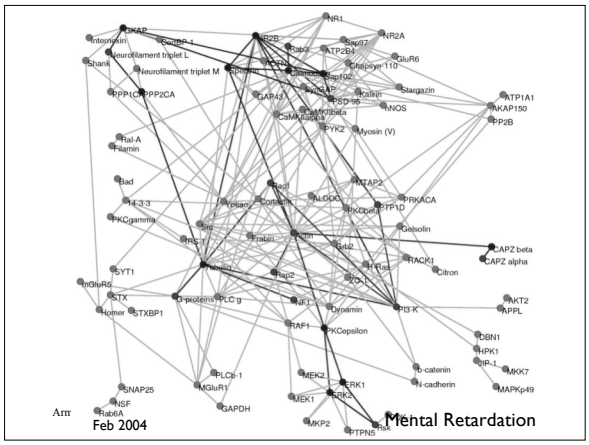


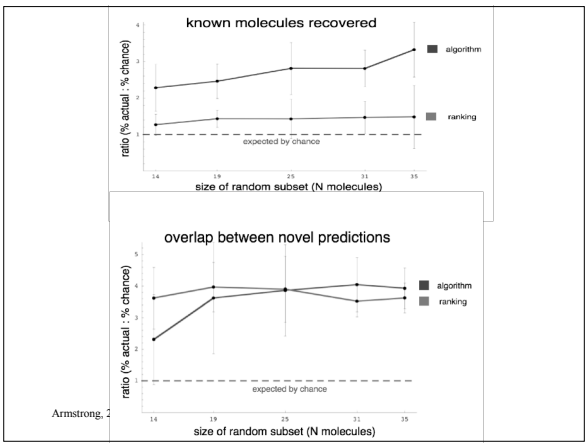
% chance in predicted network



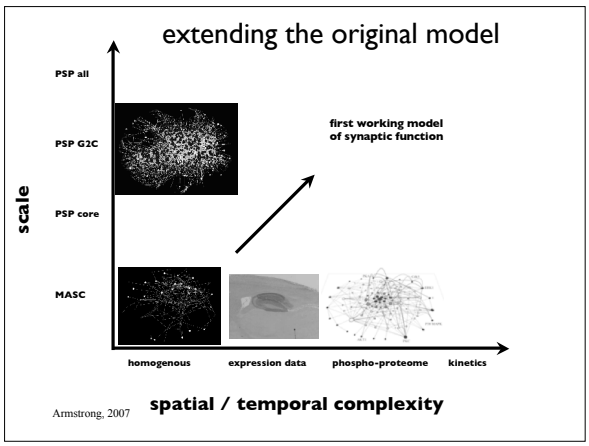
testing by exon sequencing all nodes (>1000 patients)

David Porteous, Douglas Blackwood, Walter Muir, Iain Deary

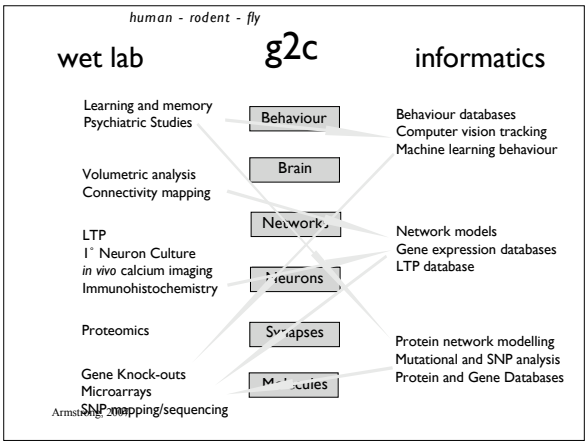




Armstrong, 2



Armstrong, 2007



Armstrong, 2007