Bioinformatics 2

From genomics & proteomics to biological 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 separate DNA molecules).
- Can be separated based on charge or mass.
- 2D Page separates a protein extract in two dimensions.
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)

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

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.

Armstrong, 2007

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.

Armstrong, 2007

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

Armstrong, 2007

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.

Armstrong, 2007

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

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

<table>
<thead>
<tr>
<th>k</th>
<th>fraction</th>
<th>%lethal</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;6</td>
<td>93%</td>
<td>21%</td>
</tr>
<tr>
<td>&gt;15</td>
<td>0.7%</td>
<td>62%</td>
</tr>
</tbody>
</table>

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.

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


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

- Producing a tree can predict functional clustered arrangements.

Reconstructing the cognitive proteome

J Douglas Armstrong
Edinburgh Centre for Bioinformatics
University of Edinburgh

Cluster analysis on the network

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
**Synapse proteomes**

<table>
<thead>
<tr>
<th>Study</th>
<th>#PSD proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collins et al</td>
<td>620</td>
</tr>
<tr>
<td>Yoshimura et al</td>
<td>441</td>
</tr>
<tr>
<td>Jordan et al</td>
<td>401</td>
</tr>
<tr>
<td>Peng et al</td>
<td>328</td>
</tr>
<tr>
<td>Li et al</td>
<td>151</td>
</tr>
<tr>
<td>Saloh et al</td>
<td>46</td>
</tr>
<tr>
<td>Wallkonis et al</td>
<td>29</td>
</tr>
<tr>
<td>Total PSD proteins</td>
<td>1124</td>
</tr>
<tr>
<td>Literature</td>
<td>119</td>
</tr>
<tr>
<td>NRC/MASC</td>
<td>186</td>
</tr>
<tr>
<td>Consensus PSD</td>
<td>466 (2 or more studies)</td>
</tr>
</tbody>
</table>

**Consensus PSD**

- **Post Synaptic proteome**: 1168 proteins
- **NRC / MASC**: 186 proteins

**Protein complexes at the Synapse**

1124 proteins

**NRC / MASC**

- **2-3 MDa**: 186 proteins

**The synaptic proteome is enriched for proteins containing signalling related domains**

<table>
<thead>
<tr>
<th>Domain</th>
<th>% MASC</th>
<th>% Mouse</th>
<th>ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein kinase</td>
<td>11.8</td>
<td>3.75</td>
<td>3.16</td>
</tr>
<tr>
<td>Ser/Thr protein kinase</td>
<td>10.1</td>
<td>1.69</td>
<td>6.05</td>
</tr>
<tr>
<td>GTPases</td>
<td>8.06</td>
<td>1.51</td>
<td>5.33</td>
</tr>
<tr>
<td>Pleckstrin-like</td>
<td>5.91</td>
<td>1.35</td>
<td>4.47</td>
</tr>
<tr>
<td>PDZ/DHR/CLGF</td>
<td>5.71</td>
<td>0.74</td>
<td>8.04</td>
</tr>
<tr>
<td>Small GTP-binding domain</td>
<td>5.38</td>
<td>1.49</td>
<td>3.62</td>
</tr>
<tr>
<td>Pleckstrin homology</td>
<td>4.84</td>
<td>1.08</td>
<td>4.49</td>
</tr>
<tr>
<td>Calcium-binding EF-hand</td>
<td>4.84</td>
<td>1.65</td>
<td>2.93</td>
</tr>
<tr>
<td>C2</td>
<td>4.30</td>
<td>0.82</td>
<td>5.26</td>
</tr>
<tr>
<td>IQ calmodulin-binding region</td>
<td>3.76</td>
<td>0.31</td>
<td>12.0</td>
</tr>
</tbody>
</table>

**Non-Sequence Annotation**

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

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

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Molecules</th>
<th>Molecules bound to NR2A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Mental retardation</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>LTP</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>Rodent spatial learning</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>Rodent fear conditioning</td>
<td>25</td>
<td>1</td>
</tr>
</tbody>
</table>

(186)

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?

aim: rebuild the network from the proteomics list

Netpro (commercial)
- 56 proteins, 94 interactions
- 40% agreement in predictions
- $50,000

Netpro (commercial)
- BIND/MINT etc
- 22 proteins, 16 interactions
- $200

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
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)

Simulated disruption vs. mutations

Linear correlation between simulation and in vivo assay

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
Average 3.4 links between any two pairs of proteins in the entire network

The interconnected network provides significant robustness

Clusters of proteins are linked to known function

Community structure 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
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

regulation/dynamics

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

phospho-regulation in NRC/MASC

Comparative genomics of postsynaptic proteome:

- 570 genes: 186, NRC/MASC; 570 PSD
- 19 species

- number of synapse orthologues

Eukaryotes

- Menaaus
- Chordata

Domain number

Domain type
Phylogeny of modules in NRC/MASC

Vertebrate NR2 cytoplasmic C-terminal motifs absent in invertebrates

<table>
<thead>
<tr>
<th>Name</th>
<th>motif (ENSP0000027853)</th>
<th>Start</th>
</tr>
</thead>
<tbody>
<tr>
<td>PXSP motif</td>
<td>PRP</td>
<td>1118</td>
</tr>
<tr>
<td>Src substrate</td>
<td>WWAKKMYKMQKVD</td>
<td>1232</td>
</tr>
<tr>
<td>Acetylase</td>
<td>A</td>
<td>1276</td>
</tr>
<tr>
<td>ESSP motif</td>
<td>POSP</td>
<td>1292</td>
</tr>
<tr>
<td>Cdc2 kinase</td>
<td>RQHSYD</td>
<td>1300</td>
</tr>
<tr>
<td>PKC α</td>
<td>S</td>
<td>1303</td>
</tr>
<tr>
<td>PKC α</td>
<td>S</td>
<td>1323</td>
</tr>
<tr>
<td>p85 P13K binding</td>
<td>YXXM</td>
<td>1336</td>
</tr>
<tr>
<td>Cdk site</td>
<td>Y</td>
<td>1336</td>
</tr>
<tr>
<td>AP-2 binding</td>
<td>YERK</td>
<td>1371</td>
</tr>
<tr>
<td>Src site</td>
<td>Y</td>
<td>1372</td>
</tr>
<tr>
<td>ESSP motif</td>
<td>ESDV</td>
<td>1481</td>
</tr>
</tbody>
</table>

Signaling complexity increased
Evolution of brain complexity across multiple levels

NRCMASC modules phylogeny

Recent innovations

Ancient conserved

what does the complex look like across the brain?

Western Blot

Immunohistochemistry

Microarray

In situ hybridisation

Combinations of expression profiles defines brain regions

Western data – 65 proteins, 4 regions

microarray data – 148 mRNAs, 22 regions

Recently evolved proteins show higher regional variation in expression level

Least variable genes mostly ancient origin.
Most variable genes mostly recent/chordate origin
Armstrong, 2007

Expression variance verses module

molecular computation in fly neurons

<table>
<thead>
<tr>
<th>behav</th>
<th>lethal</th>
</tr>
</thead>
<tbody>
<tr>
<td>67%</td>
<td>65%</td>
</tr>
<tr>
<td>40%</td>
<td>75%</td>
</tr>
<tr>
<td>50%</td>
<td>100%</td>
</tr>
</tbody>
</table>

90 orthologues in fly
1,671 total alleles

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

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, Ian Deary

Mental Retardation

Extending the original model

First working model of synaptic function

Spatial / temporal complexity

Wet lab

Informatics

G2C

Learning and memory
Psychiatric Studies
Volumetric analysis
Connectivity mapping
LTP
Neuroculture
Calcium imaging
Immunohistochemistry
Proteomics
Gene Knock-outs
Microarrays
Gene mapping/sequencing

Behaviour
Brain
Network models
Gene expression databases
LTP database
Protein network modelling
Mutational and SNP analysis
Protein and Gene Databases

Behavioural databases
Computer vision tracking
Machine learning behaviour

Network models
Gene expression databases
LTP database
Protein network modelling
Mutational and SNP analysis
Protein and Gene Databases