Skinner Lab Meeting

May 24, 2011

Amyloid bioinformatics project

Adrian Marinovich, MD, MPH

Introduction

Questions:

How are amyloid diseases the same, and how are they different?

More specifically:

1) What molecular pathways are common to many or all amyloid diseases?

2) What molecular pathways are unique to certain amyloid diseases (such as prion disease or Alzheimer's disease?)

Rationale:

A bioinformatics approach allowing comparisons of gene expression, gene annotation, and protein-protein and protein-DNA interaction data may answer these questions and provide leads for biomarker discovery

Aims and Methods

1) Comparative microarray analysis:

Create a unified amyloid bioinformatics database using gene transcription data

Outcome: two target lists of genes of interest – by their expression across many amyloid diseases, or by expression unique to individual amyloid diseases

2) Ontology annotation and data integration:

Create a semi-automated text mining and gene annotation tool using known ontologies, as well as with associated gene-disease, protein-DNA and protein-protein interaction data

Outcome: function annotation and clustering of genes in the target lists above

3) Data visualization and network inference:

Create a data visualization and network inference tool, using as input the unified amyloid database and the ontology annotation tool described above

Outcome: concise data visualization and network inference to prioritize target lists of genes and gene products for further analysis as biomarkers

Aims and Methods

1) Comparative microarray analysis:

Create a unified amyloid bioinformatics database using gene transcription data

Outcome: two target lists of genes of interest – by their expression across many amyloid diseases, or by expression unique to individual amyloid diseases

2) Ontology annotation and data integration:

Create a semi-automated text mining and gene annotation tool using known ontologies, as well as with associated gene-disease, protein-DNA and protein-protein interaction data

Outcome: function annotation and clustering of genes in the target lists above

3) Data visualization and network inference:

Create a data visualization and network inference tool, using as input the unified amyloid database and the ontology annotation tool described above

Outcome: concise data visualization and network inference to prioritize target lists of genes and gene products for further analysis as biomarkers

Comparative microarray analysis

Meta-analysis

- systematic use of all data across studies
- starting with central repository of array data in NCBI's GEO

Did rough 'proof of concept' using human array data

- now have turned to mouse data (more abundant)

Much variability in data among studies re: probe ID linkage to gene, gene ID, missing values, phenotype coding, etc.

- Detailed review of each mouse in each individual study required to determine non-overlap across platforms, correct coding, appropriate groupings
- Some studies broken into sub-studies & analyzed separately from each other
- One study had no normalized data posted in GEO (just had .CEL files)

Comparative microarray analysis, cont'd

Made RefSeq/Unigene combination ID ("RefUni") to ensure maximum capture of genes

- All arrays eventually locked onto common RefUni template (about 44,000 mouse 'genes')
- If multiple probes per gene, then expression values averaged and one average value associated with that gene
    - May also try using highest expression value, or biggest difference in values between groups
- If same probe for multiple genes, then that probe's values repeated across all relevant genes

Created data repository using PHP/MySQL

- Brought Skinner lab data into db

Final subsetting of studies, and analysis, performed in R

Modified R microarray meta-analysis package (MetaMA) to allow for analysis of log-ratios as well as straight expression values

- uses p values (not effect sizes)

Mouse microarray meta-analysis

10 studies broken down into 17 substudies:

(2 other studies removed due to restrictive glycosylation array, and missing normalized expression values)

Sub-study	n	model	time point	other	status
mexprs1840pad	16	scrapie	21 d.p.i.		
mexprs1840pbe	20	scrapie	100 d.p.i.		
mexprs1840pcf	18	scrapie	endpoint		
mexprs10310_4134pa	6	scrapie	endpoint		
mexprs10310_6412pabc	25	scrapie	160-460 d.p.i.	C57 mice	pending*
mexprs10310_6412pd	8	scrapie	200 d.p.i.	VM mice	pending*
mexprsski2006pa	6	scrapie	146 d.p.i.		pending*
mexprsski2008pa	5	scrapie	104 d.p.i.		
mexprs23182pa	6	scrapie	126 d.p.i.		
mexprs7207_339pa	6	scrapie	145 d.p.i.		
mexprs9914_1261pa	10	SCA7	5 weeks		
mexprs9914_339pabc	12	SCA1	4-12 weeks		pending*
mexprs2867pbd	11	SCA1	4-11wks	cerebellum	
mexprs2867pac	9	SCA1	4-11wks	forebrain	
mexprs19677_1261pa	8	Huntington's	12 months		
mexprs19677_1261pb	10	Huntington's	24 months		
mexprs14499pab	18	Alzheimer's/APP	6 months		

*data constancy problem - may need larger groupings

Mouse microarray meta-analysis

	DE	IDD	Loss	IDR	IRR
All studies	1	0	930	0.00	99.89
Scrapie only (all dpi)	14	1	725	7.14	98.24
21-104 dpi	0	0	6	NaN	100
126Dpi-endpoint	23	2	712	8.70	97.14
Among all degenerative dz	1	0	204	0.00	99.51
Among SCA	60	31	174	51.67	85.71
Among Huntington's dz	0	0	1	NaN	100
Among scrapie & SCA	11	2	920	18.18	99.03

DE = the number of significant genes in Meta-Analysis

IDD = genes which are significant in Meta-Analysis but not in individual studies

Loss = genes significant in individual data sets but not in Meta-Analysis

IDR & IRR are the percentages of Integration Driven Discoveries and Integration Driven Revisions in identified differentially expressed genes

(Neither Serum amyloid P nor Ataxin (1 or 7) found)

Next steps

Expand search for non-GEO expression data and newer generation expression data

Network genes of interest (Cytoscape and/or Ingenuity)

Look at other subgroups (scrapie strain, brain regions, etc.)

Determine cause of / fix data constancy problem

- May need to further modify R package and/or create larger grouping where possible - consider sign test instead of t test

Contact authors for missing normalized data

Look at yeast