Skinner Lab Meeting

March 29, 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

Might be accomplished with simple aggregation of differentially expressed gene (DEG) lists across studies, but:

- this is vulnerable to the varying methods of assessing significance of differential expression across studies
- this does not incorporate data from genes that don't make a given author's cut to be in the DEG list, even though the study may have collected data on those genes
- some DEG lists reported from one study have little or no overlap with another study, making comparisons impossible

- Meta-analysis allows for a more systematic use of all data across studies
    - can obtain data from central repositories such as GEO, in addition to individual studies
    - vulnerable to variation across studies with regard to different microarray platforms and normalization methods
        - but this still allows more data to pass through, and with less bias, than the above approach

Comparative microarray analysis

Can be thought of as a microarray meta-analysis:

- Or, a statistical comparison of gene expression across multiple different studies
- A meta-analysis “pools” data across many studies, increasing sample size, which can:
    - increase statistical power
    - resolve discrepancies between individual studies
    - refine estimates of effect size
    - answer new questions not asked in the original studies
- Each microarray study has multiple samples divided into two groups:
    - with or without disease (Alzheimer's, Parkinson's, prion disease, etc.)
    - for this meta-analysis: +/- disease ~ +/- amyloid formation
        - note that the original studies intended to study a single specific disease, and we're extrapolating to the amyloid formation common to many diseases
            - this may mean wide variation in effect sizes across diseases/models

Basic & Clinical Biostatistics. B Dawson-Saunders & RG Trapp. Lange, 1994

Microarray meta-analysis

Many techniques have been developed to perform microarray meta-analysis:

- Many of them are available as R packages
- The different techniques can be broken down, based on which metric is combined across the multiiple studies:
  - ranks
  - effect sizes
  - p-values - “Fisher's technique”
- Campan and Yang compared these approaches on common datasets and found:
  - rank technique performed poorly
  - effect size techniques had variable performance, depending on the specific statistical model employed
  - p-value technique performed well

Campain and Yang BMC Bioinformatics 2010, 11:408

Microarray meta-analysis

Meta-analysis by combining p-values:

- Combine p-values from individual studies to estimate an overall p-value for each gene across all studies
- p-values come from performing t-tests to compare normalized gene expression values between the disease group and the control group
- Technique originally developed by Fisher in the 1930s
    - one of the more simple techniques
- Not examining effect sizes may make this approach more robust
    - avoids direct data comparison
        - avoids cross-platform and normalization differences
    - good for combining the very disparate data we are after
- Once this technique tells us which genes hold true across multiple studies, we can go back and look at effect direction and size

Ramasamy A et al. PLoS Medicine 2008, 5:e184.

Rhodes DR at el. Cancer Research 2002,62:4427-4433.

Hu P et al. Cancer Informatics 2006: 2 289–300.

Microarray meta-analysis – in practice

Before even getting to the analysis, must contend with difficult issue of massaging data so that the R meta-analysis package will accept them

- Different microarray platforms provide unique links from their probe identifiers to different gene identifiers:
    - Affymetrix links to a combination of GenBank accession number and RefSeq
    - Illumina links to RefSeq only
    - The link between GenBank and Refseq, or any other identifier, such as EntrezID, is not unique:
        - it's many-to-many
        - this requires modifying the expression datasets to incorporate all combinations
        - ideally, would do this in a way to maximize discovery, so that:

            Links with many probes to one gene => select lowest p-value

            Links with one probe to many genes => expand dataset so there's a record for each gene

Microarray meta-analysis – in practice

Data obtained via GEOquery in R, and by ftp from NCBI

Data massaged in both R and PHP/MySQL

Analysis performed using R package metaMA (part of MAMA suite of microarray meta-analysis packages)

So far have worked with 5 studies of human brains with and without Alzheimer's and Parkinson's

Marot G et al. Bioinformatics 2009;25:2692–2699.

Microarray meta-analysis – in practice

Example of how it would work:

- study1: Alzheimer's disease and the normal aged brain – GSE5281

  Affymetrix platform – GPL570

  161 samples/arrays

- study2: Transcriptional analysis of multiple brain regions in Parkinson's disease – GSE20295

  Affymetrix platform – GPL96

  93 samples/arrays

- study3: Genetic control of human brain transcript expression in Alzheimer’s Disease – GSE15222

  Illumina platform – GPL2700

  363 samples/arrays

Microarray meta-analysis – in practice

Example of how it would work:

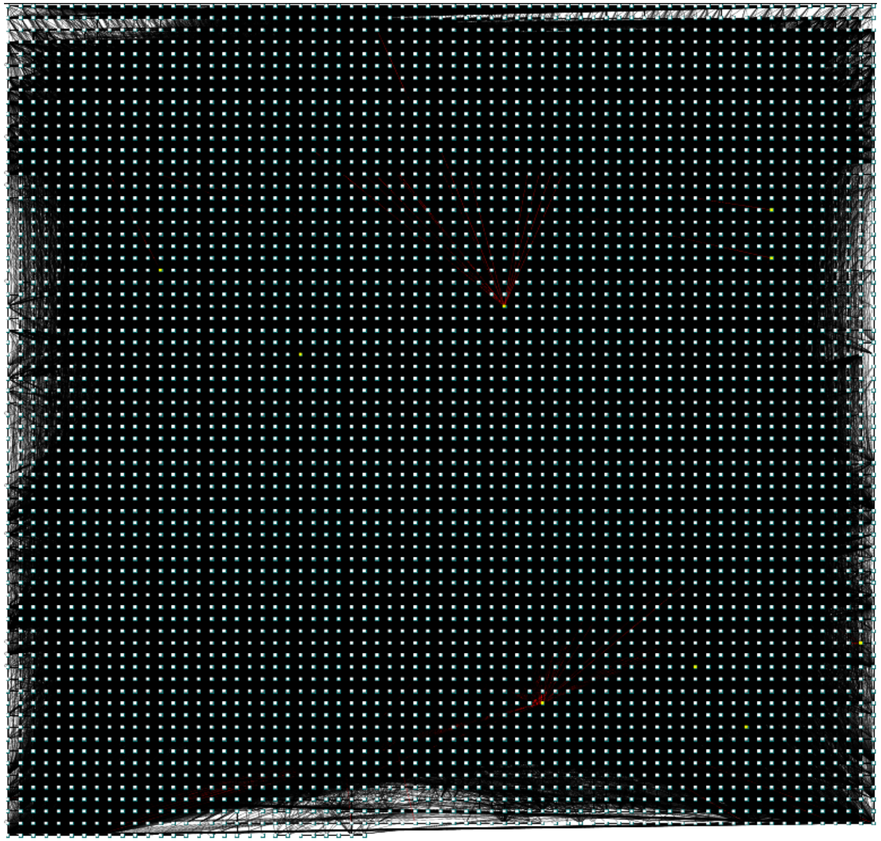
- study1: Alzheimer's disease and the normal aged brain – GSE5281
  Affymetrix platform – GPL570
  161 samples/arrays => for 23 AD, 13 control patients
- study2: Transcriptional analysis of multiple brain regions in Parkinson's disease – GSE20295
  Affymetrix platform – GPL96
  93 samples/arrays => for 15 PD, 15 control patients
- study3: Genetic control of human brain transcript expression in Alzheimer’s Disease – GSE15222
  Illumina platform – GPL2700
  363 samples/arrays => for 176 PD, 187 control patients

Output (Note that I did not group by patient – this is just to show workflow):

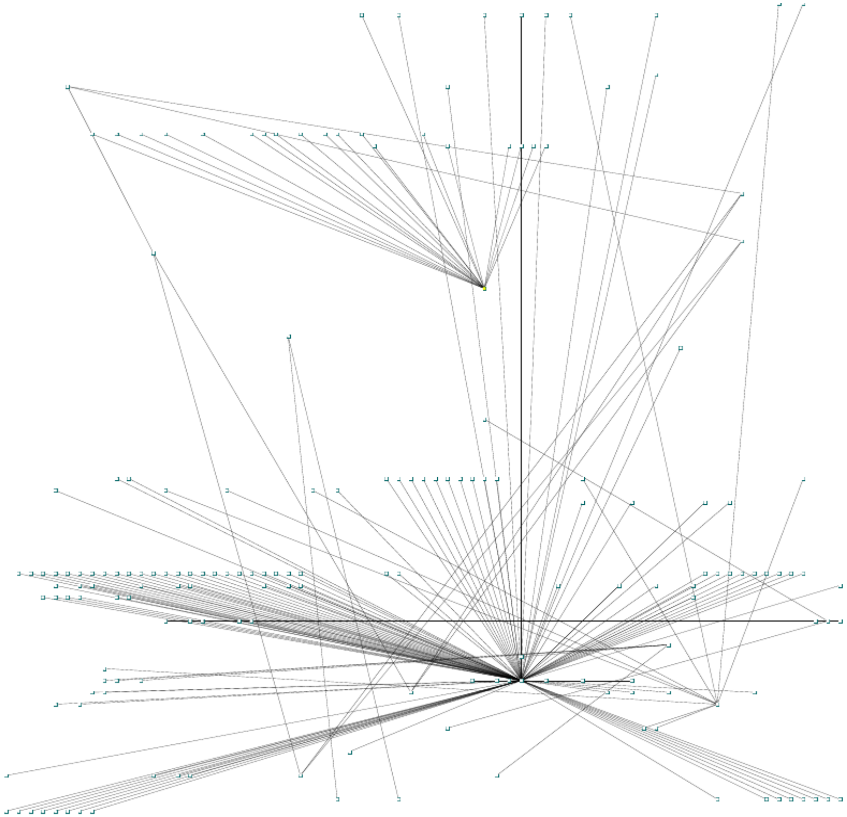
	#DEGs	
study1	1565	
study2	0	=========> (because not grouping by patient...the paper did find DEGs*)
study3	1405	
AllIndStudies	2562	
Meta-analysis	945	=========> DEGs not seen in individual studies: 28 “discoveries”

*Zhang Y et al. American Journal of Medical Genetics Part B (Neuropsychiatric Genetics) 2005; 137B:5 –16.

Pulling the “discovery” genes of interest, and their nearest interacting proteins, from a protein-protein interaction network in Cytoscape:



>>



Next steps

Assimilate more microarray data into analyses

- human, mouse and yeast for now
- from both GEO and manually individual studies

Resolve patient grouping and many-to-many issues, likely in PHP

- consider dividing data into brain regions as an option in analyses

Stats review with Rendahl

- consider other meta-analysis techniques

Develop/refine meta-analysis output displays in Cytoscape