A systems biology analysis of Alzheimer's disease identifies links to cardiovascular diseases

Weixiong Zhang

Washington University in St. Louis
Department of Computer Science
Department of Genetics

http://www.cse.wustl.edu/~zhang

Joint work with Monika Ray, Jianhua Ruan and Guandong Wang
What I am going to do

- Quick introduction to gene expression
- Brief discussion of Alzheimer’s disease
- Results on AD
  - Our approach to AD
  - Our results
- Two techniques used in our study
  - Cis-element (motif) discovery
  - Network community discovery
Central dogma of molecular biology
Complex transcriptional regulation

Mutations at DNA level – e.g., single nucleotide polymorphisms (SNPs) – can affect gene expression.
Transcriptional regulatory networks
Transcriptional regulatory logic

- Which?
- When?
- Where?
- How?

Gene

Promoter

RNA-Pol

A and not B

A and B

A or B

Not (A and B)

PNAS 2003;100(9):5136-41
Microarray based gene profiling

- High-throughput – measure numerous genes simultaneously
- “Snapshot” of the cell status
- Identify genes that change expression across conditions
**Gene expression profiling**

<table>
<thead>
<tr>
<th>No. of transcripts / probes</th>
<th>Subject1 (control)</th>
<th>Subject2 (control)</th>
<th>Subject3 (control)</th>
<th>Subject4 (affected)</th>
<th>Subject5 (affected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>mRNA expression level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example**

<table>
<thead>
<tr>
<th>genename</th>
<th>control 1</th>
<th>affecte d1</th>
<th>control 2</th>
<th>control 3</th>
<th>affecte d2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_000026</td>
<td>6.355</td>
<td>4.588</td>
<td>6.564</td>
<td>4.608</td>
<td>6.718</td>
</tr>
<tr>
<td>NM_000075</td>
<td>3.062</td>
<td>3.085</td>
<td>4.566</td>
<td>5.297</td>
<td>3.151</td>
</tr>
<tr>
<td>NM_000141</td>
<td>2.849</td>
<td>2.901</td>
<td>2.761</td>
<td>2.907</td>
<td>2.704</td>
</tr>
<tr>
<td>NM_000143</td>
<td>3.482</td>
<td>3.506</td>
<td>3.134</td>
<td>2.949</td>
<td>4.127</td>
</tr>
<tr>
<td>NM_000148</td>
<td>3.099</td>
<td>2.809</td>
<td>3.498</td>
<td>3.049</td>
<td>3.423</td>
</tr>
<tr>
<td>NM_000177</td>
<td>2.629</td>
<td>2.669</td>
<td>2.606</td>
<td>2.687</td>
<td>4.045</td>
</tr>
<tr>
<td>NM_000178</td>
<td>7.072</td>
<td>6.407</td>
<td>8.21</td>
<td>7.23</td>
<td>7.503</td>
</tr>
<tr>
<td>NM_000202</td>
<td>6.472</td>
<td>5.559</td>
<td>6.107</td>
<td>5.648</td>
<td>6.481</td>
</tr>
<tr>
<td>NM_000210</td>
<td>2.919</td>
<td>3.083</td>
<td>2.82</td>
<td>3.03</td>
<td>2.79</td>
</tr>
<tr>
<td>NM_000240</td>
<td>4.441</td>
<td>4.959</td>
<td>3.994</td>
<td>4.641</td>
<td>4.102</td>
</tr>
<tr>
<td>NM_000289</td>
<td>8.41</td>
<td>5.577</td>
<td>8.974</td>
<td>8.213</td>
<td>8.945</td>
</tr>
<tr>
<td>NM_000305</td>
<td>8.029</td>
<td>8.015</td>
<td>5.9</td>
<td>6.817</td>
<td>7.187</td>
</tr>
</tbody>
</table>
Schedule for the AD part

- What is Alzheimer’s disease (AD)?
  - Different ways to characterize it
- Why is the common practice not effective?
  - Motivation and our general approach
- Study design
  - Data/sample used
  - Methods we developed and tools we used
- The results
  - A module of disease-related genes
  - Hubs, etc
- Conclusion
What is Alzheimer’s disease?

- Clinical Picture
  - Gradual onset and progression of memory impairment combined with deficits in executive functioning, language, visuospatial abilities, personality, behavior and self-care.
How common is Alzheimer’s disease?

- ~4 million people with AD in the U.S. alone
- 100,000 deaths/yr attributed to AD
- Risk of developing AD increases with age
  - 5% of population over 65yrs develop AD
  - 20% of those over 85yrs develop AD
- Social impact
  - Devastating to patients and families
  - High health care cost
What is Alzheimer’s disease? – (2)

- **Physiology**
  - Brain volume shrinkage
  - Neuron death

- **Neurophathology**
  - Extracellular Aβ-peptide plaques
  - Intracellular neurofibrillary tangle (NFT) of tau

What is Alzheimer’s disease? – (3)

- Molecular and genetic picture

Schedule

- What is Alzheimer’s disease (AD)?
  - Different ways to characterize it

- Why is the common practice not effective?
  - Motivation and our general approach

- Study design
  - Data/sample used
  - Methods we developed and tools we used

- The results
  - A module of disease-related genes
  - Hubs, etc

- conclusion
Current practice for finding risk factors

- Linkage analysis and association study
  - Correlate genetic markers with disease phenotypes
  - Effective on autosomal dominant diseases
    - E.g., familial AD – mutations in APP, PS1 and PS2
  - But not effective for polygenic diseases
    - Each factor contributes a small, nearly undetectable, amount of effect
    - A large number of genes and single nucleotide polymorphisms (SNPs) have been indicated
    - APOE4 is the only known genetic risk factor
      - 50% AD cases do not carry APOE4 allele!
Hypothesis and motivation of our work

- Onset of disease is a gradual process
  - It involves gene expression changes
  - Genotype variations lead to gene expression alteration
    - Genetic genomics
  - Gene expression changes occur before phenotype
- Gene expression changes as step stones
  - Bridge the gap between genotypes and phenotypes


Our objectives

- Analyze gene expression in AD brain to understand the genetics of the disease
- Genome-level view of common, ageing-related human diseases
What is Alzheimer’s disease (AD)?
- Different ways to characterize it

Why is the common practice not effective?
- Motivation and our general approach

Study design
- Data/sample used
- Methods we developed and tools we used

The results
- A module of disease-related genes
- Hubs, etc

Conclusion
AD samples and data we used

- **Brain samples**
  - Entorhinal cortex of post mortem samples
  - Single neurons via laser captured microdissection
  - 13 normal controls
    - Braak stages 0-II; average age: 80.1 years
  - 20 AD affected
    - Braak stages III-IV (incipient AD); average age: 84.7 years

- **Affymetrics microarray gene chip**
  - Data were normalized by the gcRMA method
  - Probe sets were mapped to gene names using DAVID

- **Differentially expressed (DE) genes**
  - From significance analysis of microarrays (SAM)
  - 1663 DE genes at a false discovery rate (FDR) of 0.1%
Selecting differentially expressed genes

- by significance analysis of microarrays (SAM)
  - 1663 DE genes at false discovery rate (FDR) 0.1%
- DE genes are involved in biological processes known to be affected in AD were identified.
  - Immune response, inflammatory response, cell development and differentiation are upregulated
  - Processes related to actin and glucose metabolism are down regulated
  - All these processes have been confirmed in other studies
Quality check on the data/genes

- Principal component analysis using the DE genes
Approach to microarray data analysis

Microarray expression data preprocessing

Dimensionality reduction by selection of differentially expressed genes

Network of co-expressed genes

Functional modules of highly co-expressed genes

Biological significance of the results

Objective of this Research
Approach to microarray data analysis


Building & analyzing Co-Exp. network

- **Build a network of co-expressed genes**
  - A node is a gene
  - An edges represents the expression similarities between two genes

- **Community/module finding**
  - Optimization a modularity function by Newman
  - Iterative spectral clustering method

- **Network result**
  - Total 6 Co-Exp modules over 1663 DE genes
  - 2 modules corresponding to upregulated genes
  - 4 modules corresponding to downregulated genes
Overall Co-Exp network
Adjacency matrix of the network
Pearson Correlation coefficients
# Top biological processes in modules

<table>
<thead>
<tr>
<th>Module</th>
<th>Activity</th>
<th>Ease score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 1</td>
<td>Protein biosynthesis</td>
<td>7.14E-06</td>
</tr>
<tr>
<td></td>
<td>Cell development</td>
<td>2.37E-05</td>
</tr>
<tr>
<td></td>
<td>Cell differentiation</td>
<td>4.88E-05</td>
</tr>
<tr>
<td></td>
<td>Macromolecule biosynthesis</td>
<td>8.56E-05</td>
</tr>
<tr>
<td></td>
<td>Cellular nerve ensheathment</td>
<td>1.11E-04</td>
</tr>
<tr>
<td></td>
<td>Neuron development</td>
<td>2.22E-04</td>
</tr>
<tr>
<td></td>
<td>Regulation of action potential</td>
<td>4.37E-04</td>
</tr>
<tr>
<td>Module 2</td>
<td>Response to other organism</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>Immune response</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Defense response</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>Response to stress</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Protein kinase cascade</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Integrin-mediated signalling pathway</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Myeloid cell differentiation</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td>JAK-STAT cascade</td>
<td>0.042</td>
</tr>
</tbody>
</table>
## Significant KEGG pathways in modules

<table>
<thead>
<tr>
<th>Module</th>
<th>KEGG pathway</th>
<th>Ease score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 1</td>
<td>Ribosome</td>
<td>8.16E-07</td>
</tr>
<tr>
<td></td>
<td>Translation</td>
<td>3.41E-14</td>
</tr>
<tr>
<td>Module 2</td>
<td>Phospholipid degradation</td>
<td>0.013</td>
</tr>
<tr>
<td>Module 3</td>
<td>Signal transduction</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Phosphatidylinositol signaling system</td>
<td>0.005</td>
</tr>
<tr>
<td>Module 4</td>
<td>Neuron development</td>
<td>2.22E-04</td>
</tr>
<tr>
<td>Module 6</td>
<td>Nucleotide metabolism</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Statistically significant ($p < 0.05$) KEGG pathways present in the modules of the co-expression network.
Module related to human diseases

- EASE (using the Genetic Association Database1) was used to identify genes that are associated with human diseases/disorder in each module

- Module 1 contains **18** genes linked to cardiovascular diseases (CVD), diabetes, stroke and AD
  - Referred to as “disease associated module”.

- Modules 2-6 did not have a significant enrichment of genes associated with human diseases/disorders
### 18 disease-related genes in module 1

- Relatively well characterized in human diseases

<table>
<thead>
<tr>
<th>Disease/condition</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurodegeneration</td>
<td>VWF, A2M, APOE, FTL, PON2, COMT, MAP4, TF, SERPINA3, ATP1A2, AGT</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>A2M, APOE, PON2, SERPINA3</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>A2M, APOE, SERPINA3, PON2</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>VWF, A2M, APOE, PON2, COMT, WNK1, CBS, SERPINA3, TIMP1</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>APOE, PON2, COMT, SERPINA3</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>VWF, A2M, APOE, PCBD2, HLA-DQB1 (HLA-DQB2), TIMP3, SLC2A1, AGT</td>
</tr>
</tbody>
</table>

Functional annotation clustering of genes in module 1 based on their association to human conditions/diseases.
Further evidence – *cis*-elements (motifs)

- WordSpy motif-finding method on genes in module 1
- 89 motifs were specifically enriched in module 1
  - $p < 0.001$
  - target genes were co-expressed with an average correlation coefficient $> 0.4$.
- 36 motifs match 26 known transcription factor binding sites (TFBS) in JASPAR (match score $\geq 0.8$)
- Implication:
  - Found a set of co-expressed and co-regulated genes
  - Common genes and common regulators
  - Add significance to the hypothesis that many common biochemical pathways are affected and regulated in AD and CVD
Many motifs match the binding sites of TFs associated with CVD, diabetes, stroke and AD

- 139 genes in module 1 contain motifs matched the TFBS of the known TFs associated with these diseases
- e.g., *Arnt-Ahr* dimer TF activates genes responding to hypoxia and hypoglycemia, which are known to play pathophysiological roles in diabetes and AD

<table>
<thead>
<tr>
<th>Transcription factors</th>
<th>Number of target genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI4</td>
<td>9</td>
</tr>
<tr>
<td>Arnt-Ahr</td>
<td>93</td>
</tr>
<tr>
<td>ARRI0</td>
<td>6</td>
</tr>
<tr>
<td>Broad-complex 3</td>
<td>10</td>
</tr>
<tr>
<td>CEBP</td>
<td>20</td>
</tr>
<tr>
<td>Gfi</td>
<td>8</td>
</tr>
<tr>
<td>HAND1-TCF3</td>
<td>279</td>
</tr>
<tr>
<td>Mycn</td>
<td>11</td>
</tr>
<tr>
<td>Myf</td>
<td>8</td>
</tr>
<tr>
<td>Prx2/PRDX2</td>
<td>17</td>
</tr>
<tr>
<td>RELA, REL</td>
<td>10</td>
</tr>
<tr>
<td>RUNX1</td>
<td>4</td>
</tr>
<tr>
<td>Snail</td>
<td>49</td>
</tr>
<tr>
<td>SPI</td>
<td>47</td>
</tr>
<tr>
<td>TBP</td>
<td>6</td>
</tr>
<tr>
<td>E74A</td>
<td>16</td>
</tr>
<tr>
<td>ELK1</td>
<td>16</td>
</tr>
<tr>
<td>SPIB</td>
<td>16</td>
</tr>
<tr>
<td>Hunchback</td>
<td>6</td>
</tr>
<tr>
<td>MAX</td>
<td>11</td>
</tr>
<tr>
<td>USFI</td>
<td>11</td>
</tr>
<tr>
<td>ZNF42 5-13</td>
<td>27</td>
</tr>
<tr>
<td>NFIL3</td>
<td>5</td>
</tr>
<tr>
<td>Agamous</td>
<td>8</td>
</tr>
<tr>
<td>GAMYB</td>
<td>6</td>
</tr>
</tbody>
</table>
Hub genes – why AD-related genes are important in the Co-Exp network?

- Hub are highly-connected nodes/genes in the network
  - Connectivities > 2 times standard deviation above the average
  - Most influential nodes in the network
  - Possibly play important roles in the biological processes

- Totally 107 hub genes, 22 of them are in module 1

<table>
<thead>
<tr>
<th>Module</th>
<th>Number of hubs</th>
<th>Range of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 1</td>
<td>22</td>
<td>42-63</td>
</tr>
<tr>
<td>Module 2</td>
<td>17</td>
<td>41-56</td>
</tr>
<tr>
<td>Module 3</td>
<td>15</td>
<td>40-68</td>
</tr>
<tr>
<td>Module 4</td>
<td>14</td>
<td>40-65</td>
</tr>
<tr>
<td>Module 5</td>
<td>20</td>
<td>40-73</td>
</tr>
<tr>
<td>Module 6</td>
<td>19</td>
<td>40-81</td>
</tr>
</tbody>
</table>

Number of hub genes and their range of connections/links in each module.
Hub genes – why AD-related genes are important in the Co-Exp network? – cont.

- 3 disease-related genes are hubs
- Most disease-related genes are connected to hubs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of links</th>
<th>Number of hub genes it is connected to</th>
</tr>
</thead>
<tbody>
<tr>
<td>VWF</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>A2M</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>APOE</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>FTL</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>PON2</td>
<td>51</td>
<td>8</td>
</tr>
<tr>
<td>COMT</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>MAP4</td>
<td>63</td>
<td>5</td>
</tr>
<tr>
<td>TF</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>SERPINA3</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>ATP1A2</td>
<td>45</td>
<td>7</td>
</tr>
<tr>
<td>AGT</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>TIMP1</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>WNK1</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>CBS</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>PCBD2</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>HLA-DQB1</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>HLA-DQA1</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>SLC2A1</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>TIMP3</td>
<td>14</td>
<td>0</td>
</tr>
</tbody>
</table>
How are they organized?
Disease-related module and genes – why important

- Accident for the disease-related genes to aggregate in one co-expressed module?
  - whole gene set analysis resulted in a similar modular structure
  - Common cis-elements mean common regulation (although under different physiological conditions)

- Another piece of evidence linking AD and CVD?
  - Many individual genes are known to be underlying multiple diseases – e.g., APOE
  - AD has been hypothesized as another form of diabetes

- Indication
  - There may be shared regulatory networks or signalling pathways shared by these complex human diseases
  - Using systems biology approach to work on networks/pathways instead of individual genes to find genetic risk factors
Conclusion for the AD part

- The first systems-biology (computational) study linking AD and CVD at a genomic level
  - Identified a module that contained many genes known to play prominent roles in CVD and AD
  - Identified several cis-regulatory elements, some of which mapped to the binding sites of known TFs involved in neurodegenerative and CVD as well as diabetes and stroke

- Implication
  - The presence of co-expressed genes and cis-elements, which are both related to CVD and AD in a single module, provides strong, genomic level evidence to the hypotheses connecting these two conditions
  - May affect future research on brain-blood barrier related to many common human diseases
What I am going to do

- Quick introduction to gene expression
- Brief discussion of Alzheimer’s disease
- Results on AD
  - Our approach to AD
  - Our results
- Two techniques used in our study
  - Cis-element (motif) discovery
  - Network community discovery
Genome-wide motif finding: the problem

- **Given:** a set of promoter sequences of selected genes
- **Find:** transcription factor (TF)-binding sites (motifs)
Reading a novel – your help, please

- Read novel *Moby Dick* (112K, first 10 chapter) embedded in a cover-text (a 156K long random string).
  - http://uqbar.rocketfeller.edu/~siggia/projects/mobydick/

• Steganography – hiding messages in a stego- or cover-text
  - Bin Laden has been using this technology
Weixiong Zhang

Steganalysis: cracking stegoscripts

Chapter 1 Loomings
Call me Ishmael. Some years ago - never mind how long precisely - having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation.

Chapter asct ing call ishmael some years bag xhjt never mind how long precisely having little ru no soya money in purse wgjnyyv xwf and nothing particular zaoq to interest time mon shore i thought would about vja little and see bb the water ici part of the world to way have of ing off the ddyya and to gulating the irc...tion

chapterptgpqdrftezptqtasctmvivwpecjnsisrmbtqlmlfvetloomingscallmeerishmaesomey lqyearstvhnbagoaxhjtjcokhvneverpmqpmindhowzrbdlzjllonggbhqipreciselysunpvskepfd jktcgarwtnxybgcvdjfbnohavinglilliezorunozsoyapmoneyyvugsgtsqintmyteixpurseiwfmjw gjnynyveqxwftlammbxksrbkyandnthingcgparticularwtzaqsjtnmtoqsnwvxfiupinterestzt imebymonlnshoreeggdithoughtyxfxmhqixceojjzhwouldsailsailpcaboutudxsbsnewtpggvjaasxms vlittleplvycdaowgwlbizjlnzyxandzolwcudthjdosbopxkfdosxardgcseebbthefzrskdhma wateryjikzicimypartmofprtheluworldvtoamfutitazpisagewayrybkioshavebojwphiixofpr malungipjdrivingpkuoikrwxoffodhicbnimtheixyucpdzacspleenqbpcremhwvdyyainwandad abkpvgzmtoregulatinggeetheslcirculationvsuc
Steganalysis: cracking genome?

- Objective: discover conserved functional elements

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>...cactcaca...</td>
</tr>
<tr>
<td>Mouse</td>
<td>...cg...tcacctgc...</td>
</tr>
<tr>
<td>Rat</td>
<td>...ttcaca...</td>
</tr>
<tr>
<td>Chicken</td>
<td>...ttggcagtca...</td>
</tr>
<tr>
<td>Dog</td>
<td>...tttccag...</td>
</tr>
</tbody>
</table>

**Conserved functional elements:**
- Human: CACTG
- Mouse: CACGT
- Rat: CACGT
- Chicken: CACGT
- Dog: CACGT

**Conserved non-functional elements:**
- Human: Ggcg
- Mouse: Gctg
- Rat: gttt
- Chicken: ggg
- Dog: ggcg
WordSpy – an steganographic approach

Main ideas

- A dictionary and a grammar can be used to generate a stegoscript as well as decipher a stegoscript
- Learn a dictionary and grammar while deciphering a stegoscript
  - Start with single base words, iteratively discover longer words and grammar based on shorter words – through word counting
  - In this process, treat the script as if it was generated by the current dictionary (of shorter words) and grammar – guided by a statistical model
    - \{D_1, G_1\} \Rightarrow \{D_2, G_2\} \Rightarrow \{D_3, G_3\} \Rightarrow \ldots.
  - Organize conserved motifs and background words in separate dictionaries
Stegoscripts and Statistical Model

- Stegoscripts are generated by a stochastic regular grammar model (hidden Markov model)
  - Covertext submodel (background words)
  - Secret message submodel (motif words)
- Learn a dictionary and a grammar model from the genome sequences
  - $S$ is generated by $M$
  - $M = \arg \max_M P(S \mid M')$
Results on budding yeast

- Promoters of ~800 cell-cycle genes of budding yeast (*S. cerevisiae*)
  - Full of low complexity repeats
- Comparing WordSpy with MobyDick

<table>
<thead>
<tr>
<th>Known motifs</th>
<th>Known TFs</th>
<th>WordSpy</th>
<th>Z-score</th>
<th>Z-score rank</th>
<th>G-score rank</th>
<th>MobyDick</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGCTGG(CCAGCA)</td>
<td>Ace2, Swi5 [22]</td>
<td>TGCTGG</td>
<td>5.5</td>
<td>106/147</td>
<td>22</td>
<td>TGCTGCTGGA</td>
</tr>
<tr>
<td>RRRCCAGCR(YGCTGGYY)</td>
<td>Ace2, Swi5 [17]</td>
<td>GCTGG</td>
<td>5.3</td>
<td>17/30</td>
<td>10</td>
<td>TGCTGCTGGA</td>
</tr>
<tr>
<td>ACGCGT(ACGCCT)</td>
<td>Swi6, Mbp1 [17, 22]</td>
<td>ACGCGT</td>
<td>17.4</td>
<td>19/147</td>
<td>1</td>
<td>AACGCGT</td>
</tr>
<tr>
<td>CACGAAA(TTTCGTG)</td>
<td>Swi4, Swi6 [17, 22]</td>
<td>CACGAAA</td>
<td>5.4</td>
<td>246/ 419</td>
<td>47</td>
<td>GTCACGAAA</td>
</tr>
<tr>
<td>CGCGAAA(TTTCGCG)</td>
<td>Swi4, Swi6 [22]</td>
<td>CGCGAAA</td>
<td>16.1</td>
<td>24/419</td>
<td>8</td>
<td>CGCGAAA</td>
</tr>
<tr>
<td>ATAAACAA(TTGTATAT)</td>
<td>Fkh1, Fkh2 [22]</td>
<td>ATAAACAA</td>
<td>8.8</td>
<td>193/1015</td>
<td>44</td>
<td>TTGTATAT</td>
</tr>
<tr>
<td>GTAAACAA(TTGTATTAC)</td>
<td>Fkh1, Fkh2 [22]</td>
<td>GTAAACAA</td>
<td>8.5</td>
<td>202/1015</td>
<td>21</td>
<td>n/a</td>
</tr>
<tr>
<td>GTAAACA(TGTATTAC)</td>
<td>Fkh1, Fkh2, Ndd1 [23]</td>
<td>GTAAACA</td>
<td>7.6</td>
<td>128/419</td>
<td>21</td>
<td>GTAAACA</td>
</tr>
<tr>
<td>TTTCTAA(TTACGAAA)</td>
<td>MCM1 [23]</td>
<td>TTTCTAA</td>
<td>6.4</td>
<td>359/1015</td>
<td>28</td>
<td>n/a</td>
</tr>
<tr>
<td>TCACGTG(CACGTGA)</td>
<td>Met4, Met28, Cbf1 [24]</td>
<td>TCACGTG</td>
<td>5.0</td>
<td>288/419</td>
<td>93</td>
<td>n/a</td>
</tr>
<tr>
<td>TGAAACA(TGTTTCA)</td>
<td>Ste12 [25]</td>
<td>TGAAACA</td>
<td>5.5</td>
<td>489/1015</td>
<td>55</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Table 2: Discovered known motifs in the ~800 promoters of yeast cell-cycle genes. The first two columns list the known motifs (and their reverse complimentary) and their potential TFs. The next four columns report the results from WordSpy, followed by the last column for MobyDick. The Z-score rank is based on motif Z-score, where the first number is the ranking and the second is the total number of discovered motifs of the same length. The G-score rank is the ranking among the motifs of the same length.
Comparison on benchmarks

  - sTP: number of known sites overlapped with predicted sites;
  - sFP: number of predicted sites not overlapped with known sites.
WordSpy – properties

- A new perspective on an old, challenging problem
- Combining word counting and statistical modeling
- Learn a background model on the fly
- Genome wide motif finding
- Discriminative motif finding


What I am going to do

- Quick introduction to gene expression
- Brief discussion of Alzheimer’s disease
- Results on AD
  - Our approach to AD
  - Our results
- Two techniques used in our study
  - Cis-element (motif) discovery
  - Network community discovery
Networks are everywhere

- Social networks
- Scientific societies
- Internet/cyber space
- Power grids
- Distribution chains
- Biological networks
Biological networks

- **Vertex:** molecule
  - Gene, protein, cis-element

- **Edge:** relationship
  - Regulation
  - Interaction
  - Association

- Similar to real-world networks in many aspects
  - Small-world
  - Scale-free

It is the interactions of the genes!

(Jeong et al., 2001)
Co-exp. networks from microarray data

- Rows are genes
- Columns are samples
- Colors represent gene activities (expression levels)

**Red**: high activity
**Green**: low activity
Many clustering algorithms available
- K-means
- Hierarchical
- Self organizing maps
Gene co-expression networks

- Genes \( i \) and \( j \) connected if their expression patterns are similar
  - Traditional: correlation coefficient > cutoff
  - \( K \) nearest neighbors (Ruan & Zhang, RECOMB Satellite Conf. Systems Biology, 2006)
Community discovery

- Advantages:
  - Network topology considered
  - Number of clusters determined by the algorithm
  - Easy to integrate other types of data
    - e.g. cis-elements or TFs
    - As additional nodes and edges
Community discovery problem

- Divide a network into relatively densely connected sub-networks
Modularity function ($Q$)

- Measure strength of community structures

$$Q = \sum_{i=1}^{k} \left( \frac{e_{ii}}{M} \right) - \left( \frac{a_i}{M} \right)^2$$

-1 < $Q$ < 1
$Q = 0$ if $k = 1$

Number of communities

$$Q = \frac{68+66}{140} - \left( \frac{71}{140} \right)^2 - \left( \frac{69}{140} \right)^2 = 0.96 - 0.26 - 0.24 = 0.46$$

Observed fraction of edges falling in community $i$
Expected fraction of edges falling in community $i$

$e_{11} + e_{12} = a_1$, $M = a_1 + a_2$
Goal: find the partition that has the highest Q value

But: optimizing Q is NP-hard (Brandes et al., 2006)
The algorithm

1. Construct network H
2. for \( k = K_{\text{min}} \) to \( K_{\text{max}} \)
   a. Treat H as an affinity matrix, and apply a standard k-way spectral clustering
   b. Denote the clusters as \( P^k \)
   c. \( Q_k = Q(P^k, A) \)
3. \( K^* = \arg\max_k Q_k \)
4. \( P^* = P^{k^*} \)

Algorithm NJW (Ng, Jordan, Weiss, 2001)

a. Normalize affinity matrix
b. Compute k largest eigenvectors
c. Stacking the eigenvectors by columns
d. Normalize each row to have unit length
e. Apply k-means to cluster the rows
# Real-world networks

SA: Simulated annealing, Guimera & Amaral, Nature 2005

<table>
<thead>
<tr>
<th></th>
<th>#Vertices</th>
<th>#Edges</th>
<th>Newman</th>
<th>SA</th>
<th>Qcut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>67</td>
<td>142</td>
<td>0.573</td>
<td>0.608</td>
<td>0.587</td>
</tr>
<tr>
<td>Neuron</td>
<td>297</td>
<td>2359</td>
<td>0.396</td>
<td>0.408</td>
<td>0.398</td>
</tr>
<tr>
<td>Ecoli Reg</td>
<td>418</td>
<td>519</td>
<td>0.766</td>
<td>0.752</td>
<td>0.776</td>
</tr>
<tr>
<td>Circuit</td>
<td>512</td>
<td>819</td>
<td>0.804</td>
<td>0.670</td>
<td>0.815</td>
</tr>
<tr>
<td>Yeast Reg</td>
<td>688</td>
<td>1079</td>
<td>0.759</td>
<td>0.740</td>
<td>0.766</td>
</tr>
<tr>
<td>Ecoli PPI</td>
<td>1440</td>
<td>5871</td>
<td>0.367</td>
<td>0.387</td>
<td>0.387</td>
</tr>
<tr>
<td>Internet</td>
<td>3015</td>
<td>5156</td>
<td>0.611</td>
<td>0.624</td>
<td>0.632</td>
</tr>
<tr>
<td>Physicists</td>
<td>27519</td>
<td>116181</td>
<td>--</td>
<td>--</td>
<td>0.744</td>
</tr>
</tbody>
</table>
## Running time (seconds)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#vertices</th>
<th>#Edges</th>
<th>Newman</th>
<th>SA</th>
<th>Qcut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td>67</td>
<td>142</td>
<td>0.0</td>
<td>5.4</td>
<td>2.0</td>
</tr>
<tr>
<td>Neuron</td>
<td>297</td>
<td>2359</td>
<td>0.4</td>
<td>139</td>
<td>1.9</td>
</tr>
<tr>
<td>Ecoli Reg</td>
<td>418</td>
<td>519</td>
<td>0.7</td>
<td>147</td>
<td>12.7</td>
</tr>
<tr>
<td>Circuit</td>
<td>512</td>
<td>819</td>
<td>1.8</td>
<td>143</td>
<td>6.1</td>
</tr>
<tr>
<td>Yeast Reg</td>
<td>688</td>
<td>1079</td>
<td>3.0</td>
<td>1350</td>
<td>13.4</td>
</tr>
<tr>
<td>Ecoli PPI</td>
<td>1440</td>
<td>5871</td>
<td>33.2</td>
<td>5868</td>
<td>41.5</td>
</tr>
<tr>
<td>Internet</td>
<td>3015</td>
<td>5156</td>
<td>253.7</td>
<td>11040</td>
<td>43.0</td>
</tr>
<tr>
<td>Physic peace</td>
<td>27519</td>
<td>116181</td>
<td>--</td>
<td>--</td>
<td>2852</td>
</tr>
</tbody>
</table>

**Note:** The values in the Newman column are in seconds.
Communities found by HQcut

- Small ribosomal subunit (90%)
- RNA poly II mediator (83%)
- Proteasome core (90%)
- Exosome (94%)
- gamma-tubulin (77%)
- respiratory chain complex IV (82%)
Summary on community analysis

- **Community discovery**
  - An efficient algorithm for finding intrinsic community structures by optimizing Q

- **Biological networks**
  - Gene co-expression networks
  - Protein-protein interaction networks

- **Many applications**
  - Alzheimer’s disease
  - Notch signaling pathways
  - Sepsis (infection)
Wrap up - Take home message

- Methods integrating experiments and computational analysis can be powerful for understanding complex diseases.

- There are a lot of interesting computational problems in biomedical areas that computer scientists can study for life.
Acknowledgement

- **Students and collaborators**
  - Monika Ray – AD microarray data analysis
  - Jianhua Ruan – network module identification
  - Guandong Wang – cis-element (motif) finding

- **Funding sources**
  - Two NSF grants
  - A grant from the Alzheimer’s Association
  - Partially by an NIH center grant
  - Several grants from Monsanto Corp.
Research on small noncoding RNAs

- **Experiment oriented** – 3 publications, 3 under review
  - **Goal**: find new genes and characterize their functions
  - **High-throughput sequencing (454, Solexa)**
    - Species – rice, Arabidopsis, Medicago, Cassava, Castor bean (and silkworm and pig)
    - Abiotic and biotic stresses – e.g., control, drought and salinity
    - Small RNA species – miRNAs, nat-siRNAs, tasiRNAs, …

- **Computation oriented** – 2 publications
  - **Goal**: identify stress inducible miRNA genes – functional annotation methods
    - Combining transcriptome analysis
    - Abiotic stress induced miRNAs in plants
      - Cold, drought, UV-B, …
Related publications

- **On Alzheimer’s disease**

- **On cis-element (motif) finding**

- **On network community analysis**

http://www.cse.wustl.edu/~zhang