Recent approaches to understanding global gene expression

Rohan Williams
John Curtin School of Medical Research
Australian National University
22nd Canberra International Physics Summer School
18 December 2008

“correlation pleiades” by Paul V. Terentjev, Biometrika, 1931

What I’ll talk about today

• Genes
  – “the central dogma”, expression, variation
• Impact of genome sequencing
• Some current problems:
  – Understanding regulation of gene expression
  – Data-driven models of “molecular machines”
What are genes?

- Fundamental unit of heredity
- **Sequential information** encoded as a biopolymer called DNA
- Simplified representation…
  ....ACGTTGACATAGC....
  ....TGCAACTGTATCG....

What is gene “expression”?  

- “Central dogma of molecular biology”  
  Frances Crick (1958)
- “Information **transfer** between **information carrying biopolymers** in living **cells**”
- Involves **3 types** of biopolymers  
  – DNA, RNA, proteins
What is gene “expression”? 

• “Central dogma of molecular biology” 
  Frances Crick (1958)

<table>
<thead>
<tr>
<th>General</th>
<th>Special</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA → DNA</td>
<td>RNA → DNA</td>
<td>Protein → DNA</td>
</tr>
<tr>
<td>Replication</td>
<td>Reverse transcription</td>
<td></td>
</tr>
<tr>
<td>DNA → RNA</td>
<td>RNA → RNA</td>
<td>Protein → RNA</td>
</tr>
<tr>
<td>Transcription</td>
<td>Viral copying</td>
<td></td>
</tr>
<tr>
<td>RNA → protein</td>
<td>DNA → protein</td>
<td>Protein → protein</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Gene expression is controlled by multiple complex processes

Slide: Peter Little
Genetic variation

- We have ~20,000 genes
- Have mostly *same genes*, but they come in different “*flavours*”: known as *mutations*, or more generally, as *variants*
- *Different variants* > different products
- *Different gene-products* (proteins) will have *different functional properties*, or may influence *broader cellular function in different ways*


International HapMap Project: www.hapmap.org

1000 Genomes: www.1000genomes.org/page.php?page=home

Genome Sequencing

http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html
**Genome Sequencing**

Hierarchical shotgun sequencing

Genomic DNA → BAC library → Organized mapped large clone contigs → BAC to be sequenced → Shotgun clones → Shotgun sequence → Assembly

Images sourced from:
http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html

**Expression microarrays**

(Brown lab Stanford, mid-90s+)

“spotted” cDNA e.g. Sod1 (superoxide dismutase 1)

Images sourced from:
http://www.ifr.ac.uk/safety/microarrays/
Brown et al. World Journal of Surgical Oncology 2003 1:21
‘Omics I

“Secretome”
Population of gene products secreted from the cell

“Interactome”
List of interactions between all macromolecules in a cell

“Proteome”
Population of proteins in the cell, weighted by their expression levels

“Transcriptome”
The population of mRNA transcripts in the cell, weighted by their expression levels

“Genome”
The full complement of genetic information both coding and non coding in the organism

‘Omics II
Can we put it all together?

‘Omics III
Exploration vs confirmation?*

Figure taken from: Vidal, A Biological Atlas of Functional Maps, Cell, 2001
*Jaeger and Halliday, On confirmatory versus exploratory research, Herpetologia, 1998

What can I do with all this data?

Control mRNA  Test mRNA
[replicated] microarray experiment  Statistical analysis

You find that 500 transcripts show evidence of being differentially expressed between control and test conditions

What are these 500 genes (gene-products) ‘doing’ in the biological system under study?
Constructing models of ‘molecular machines’

Gene Expression Data

(experimental conditions, time points, strains, etc)

<table>
<thead>
<tr>
<th>Genes</th>
<th>array1</th>
<th>array2</th>
<th>array3</th>
<th>array4</th>
<th>array5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.30</td>
<td>0.80</td>
<td>1.51</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>-0.10</td>
<td>0.49</td>
<td>0.24</td>
<td>0.06</td>
<td>0.46</td>
</tr>
<tr>
<td>3</td>
<td>0.15</td>
<td>0.74</td>
<td>0.04</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>-0.45</td>
<td>-1.03</td>
<td>-0.79</td>
<td>-0.56</td>
<td>-0.32</td>
</tr>
<tr>
<td>5</td>
<td>-0.06</td>
<td>1.06</td>
<td>1.35</td>
<td>1.09</td>
<td>-1.09</td>
</tr>
</tbody>
</table>

Gene expression level of gene \(i\) on slide \(j\)

\[ M = \log_2 \left( \frac{\text{Red intensity}}{\text{Green intensity}} \right) \]

Control Test

\(-\text{mRNA levels} \uparrow \)
Gene Expression Data

How correlated are the expression changes of genes 1 & 4 across early embryogenesis? Use Pearson’s correlation co-efficient (PCC)

Gene-gene correlation matrix

<table>
<thead>
<tr>
<th>Genes</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>-0.2</td>
</tr>
<tr>
<td>3</td>
<td>-0.6</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Gene-gene network
Constructing models of ‘molecular machines’


Expression profiles generated using microarrays

Taken from Gunsalus 2005

Combine 3 data types together

PPI
Expression
Phenotype
Constructing models of ‘molecular machines’

Combine all 3 types of evidence together to form putative network

Taken from Gunsalus 2005

---

Early *C. elegans* embryogenesis

Taken from Gunsalus 2005
But isn’t this just a problem for computer geeks?

Science, 1 July 2005 (125 Anniversary Special Issue)
“…explores 125 big questions that face scientific inquiry over the next quarter-century….

How Will Big Pictures Emerge From a Sea of Biological Data?

Biology is rich in description data—data gathering enables all the more important means of resolving complex problems, such as DNA sequencing, molecular biology, and automated gene expression studies. They are filling new databases to the brim. Many scientists think that their approaches to disclose have gone digital, and in a result, observations are more precise and more plentiful. A central question now confronting essentially all fields of biology is whether scientists may deduce from the vast trove of molecular data how systems and complex interactions work. All the information needs to be linked, organized, manipulated, and most importantly—interpreted or a way that makes sense.

As the field gets a big boost from the completion of the human genome sequence, the producer of a sequence, no matter how important, the goal is now to turn that data into functional patterns. This revolution in biology has been hailed as an important milestone, yet most researchers are finding the challenges of building comprehensive models of complex biological systems are unprecedented.

The data can be used to unlock the secrets of cellular processes, such as how cells communicate, how genes are regulated, and how diseases are caused. However, the sheer volume of data is overwhelming, and researchers need to develop new tools and methods to analyze and interpret this information.

Control of gene expression

- Three spatial scales of control:
  - Local ⇒ transcription factors
  - Extended ⇒ chromatin state
  - Large ⇒ nuclear organisation
“Local” control elements

Activator proteins
“transcription factors”

Post-transcriptional control
MicroRNAs, alternative splicing, alternative polyadenylation, RNA-binding proteins, etc.

miRNA RNA binding proteins

Transcriptional control
Transcription factors, chromatin state, combinatorial control, co-factors, alternative promoters, etc.

Specific short binding sequences “regulatory motifs”

Coordinated control of gene expression

Komili and Silver, Nature Reviews Genetics, 2008
**Ciona muscle development:**
An example of coordinated control of gene expression

Motifs are clustered but no additional patterning of order, spacing or orientation that might inform us about organisational principles of regulatory modules

Johnson et al., Genome Res., 2005
Brown et al., Science, 2007

---

**Gene expression programs**

Co-functionality

- GO
- Complexes
- Metabolites

---

Co-expression

- mRNA levels
- Protein levels

---

Co-regulation

- ChIP
- Chromatin-state
- DNA methylation
- 3C+

---

Co-control

- Motifs
- Sequences

"Shared controls" approach

Does the potential for co-regulation inform us about biological function of groups of genes?
Measuring shared control

<table>
<thead>
<tr>
<th>Gene pairs</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1BG, ABL2</td>
<td>0/12 = 0.0</td>
</tr>
<tr>
<td>A1BG, A2M</td>
<td>1/10 = 0.1</td>
</tr>
<tr>
<td>PLA2G5, SLC22A6</td>
<td>4/5 = 0.8</td>
</tr>
</tbody>
</table>

Share no motifs
Share few motifs
Share most motifs

Finding groups of gene with potential for shared control

Data
11,653 genes; 837 motifs

Hierarchical clustering
(0.895 threshold, complete-linkage)
2577 clusters
2478 non-singletons

Cluster validity*
Resample 50%-genes x 50 repeats
Measure cluster recovery (gene membership)
Select those with >60% recovery

334 “stable” clusters

Cluster properties

Relative contributions of transcriptional v. post-transcriptional controls
How can we identify *modules of genes* co-ordinately controlled by this machinery?

*Control of gene expression: Transcriptional and post-transcriptional*

“Extended” control elements: *chromatin*
Chromatin modifications.

The histone (modification) “code”
“Large-scale” control elements: nuclear organisation

Fraser and Bickmore, Nature, 2007
Much information, not enough theory?


Acknowledgements

• Prof. Peter Little (NUS)
• Dr Chris Cotsapas (Harvard)
• Mr. Mark Cowley (Garvan)
• Dr. Eva Chan (CSIRO)
• Prof. David Nott (NUS)
• Prof. Marc Wilkins (UNSW)
• Mr. Florian Breitweiser (UNSW)
• Mr. Michael Liu (UNSW)
• Mr Jeremy Pulvers (MPI: MCBG)
• Prof. Simon Easteal (JCSMR)
• Dr. Gavin Huttley (JCSMR)
• Dr. Steve Ohms (JCSMR/BRF)
• Dr. Conrad Burden (JCSMR/CMA)
• Prof. Frances Shannon (JCSMR)
• Dr. Kristine Hardy (JCSMR)
• Mr. Won-Min Song (RSPhysSE)
• Prof. Tiziana Di Matteo (RSPhysSE)
• Prof. Tomaso Aste (RSPhysSE)
• Dr. Yvonne Pittelkow (CMA)

Molecular Systems Biology Group
John Curtin School of Medical Research
Australian National University

Vicky Cho  Ros Attenborough
Hugh French  James McCracken
Oscar Luo  Cath Lawrence

Recruits and visitors very welcome….
Rohan.Williams@anu.edu.au

Funding: NHMRC Australia (R.W), JCSMR (O.L.)