DNA Microarrays
Introduction Part II
Readings Assigned

- Text: Chapters 3 + 5
- Be ready to discuss Spellman Cell Cycle paper next class
mRNA Expression Profiling
Image Analysis & Data Visualization

<table>
<thead>
<tr>
<th>Cy3</th>
<th>Cy5</th>
<th>Cy5/Cy3</th>
<th>log₂(Cy5/Cy3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>10000</td>
<td>50.00</td>
<td>5.64</td>
</tr>
<tr>
<td>4800</td>
<td>4800</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9000</td>
<td>300</td>
<td>0.03</td>
<td>-4.91</td>
</tr>
</tbody>
</table>

Underexpressed

Overexpressed
Genes involved in a specific biological process (i.e. heat shock)

• “Guilt by association” - assumption that genes with same pattern of changes in expression are involved the same pathway

• If we know what some genes in a group are doing, it hints that others with similar expression may be involved in same cellular process
Classification

• In some cases, don’t need to know what any genes in a group are doing to be useful

• Total expression profile can be used as a “fingerprint” identifying cell type
  – example: Tumor classification - predict outcome / prescribe appropriate treatment based on clustering with “known outcome” tumors
Types of Analysis to Form Gene Associations

- Hierarchical clustering
- Partitioning of Data in Groups
  (supervised & unsupervised)
  - Self-organizing maps (SOM)
  - K-means clustering
  - Gene shaving
  - Support vector machines (SVM)
- Modeling
  - Singular value decomposition (SVD) / Principle component analysis (PCA)
- Ranking Tests
### Data flow

**New Data** → **ScanAlyze** → **Database** → **Data Selection** → **Complete Data Table (cdt)** → **Cluster** → **SOM** → **K-means** → **SVD**
Developing New Methods

• How do you know when your method performs better than a previous method?
• A “gold standard” test set for benchmarking array data doesn’t exist
• There is too much biology we don’t know: if a new method classifies a gene in the “wrong” gene group, is it recognizing new biology, or just getting it wrong??
Limitations of Arrays

- Do not necessarily reflect true levels of proteins - protein levels are regulated by translation initiation & degradation as well.
- Generally, do not “prove” new biology - simply suggest genes involved in a process, a hypothesis that will require traditional experimental verification.
- Expensive! $20-$100K to make your own / buy enough to get publishable data.
Practical Problems

• Distinguishing background noise from low level, biologically important signals
• Difficult to independently verify so many data points, so success of array based on “anecdotal evidence”
• Multiple repetitions of array experiments is crucial, not always possible with limited samples
• Not easy to dissect specific from non-specific expression changes for a spec. experiment
  – Example: “common stress response” in yeast
Common Reference Problem

• Difficult to directly compare array experimental results from different labs / technologies
• For spotted arrays, all expression data are “relative” to a reference (ratios), not an absolute quantity
• Affy arrays are all absolute measurements, so must be converted to ratios for comparison to spotted arrays
• References should be consistent to facilitate comparisons, but vary a lot in practice
  – For human arrays, a pool of reference mRNA from 22+ cell lines is made to allow wide comparisons
Integrating Array Data with Sequence Analysis

One example: Promoter motif extraction

1. Cluster / classify genes with common response pattern
2. Align upstream promoter regions (Gibb’s sampler) or count over-represented X-mers
3. Develop profile / motif from set & search genome for new candidates w/ motif
4. Return to array data, look for supporting evidence for new members
5. Carry out experiment to support hypothesis