Review last lecture

- Characteristic features of a biological system
  - 2 scales: Micro (molecular, genotype) vs Macro (phenotype) - their correspondence. Environmental effect
  - Reductionist corollary of Central Dogma: Biology is underwritten by genes.
  - Genes, a proper functional (coding transcriptome) subset of DNA sequence
  - Reduce to studying DNA sequences – sequence genomics, 1st major entry of computation/math into biology.
- High-throughput gene identification and expression assay (basic principle)
  - Genomic & cDNA libraries (sequencing)
  - SAGE (sequencing) & microarrays (nucleotide complementarity) – short representative oligomers.
  - Gain: Massive parallelity – speed & efficiency
  - Loss: “Noise” & “false positives” up (theme problems)
  - Neutral: Massive datasets.

Outline

- Review last lecture
- Assumptions & questions in high throughput transcriptome studies
  - “Granularity” of questions
  - Paradigm shift in conceptualizing biological problems / systems
  - Prototypical experiment designs
  - 2 theme problems in these studies: noise & high false positive rate.
- Analysis and modeling of transcriptome data
  - Typical work flow (experiment); meta steps (analysis)
  - Mathematical formulation of problem
  - “Correcting” noise and measurement variation / bias
  - Uncovering geometric regularities and variance structures in data
  - Likelihood of regularities, variance structures arising by chance
  - Squaring math results with a priori biological knowledge. Figure of merit: biological vs in silico
- Basic References

Transcriptome studies: Questions

- Granularity of questions, 3 molecular scales
  - Single: Identify single molecules associated with a biological phenomenon
  - Network: Identify molecular networks/interactions associated with a bio phenomenon
  - System: Characterize transcriptomic profile/state of a biological system
- Paradigm shift in conceptualizing biological problems/systems
  - Classical biology: Whole = Sum of its parts
  - Systems biology: Whole \( \geq \) Sum of its parts
- Prototypical experiment designs
  - 2-group comparisons
  - Sequential profiling - parametrized by a continuous scalar variable
  - Hybrid of the above.
**Transcriptome studies: Questions**

- Common questions (different molecular granularities):
  - Given transcriptome profiles of N samples from K clinically distinct diseases. Can we find the minimal gene set distinguishing these diseases with a reasonable specificity/sensitivity? If such gene sets exist, how to determine if they’re descriptive vs. generative for the diseases?
  - Is there a transcriptomic signature in tumors that correlates with some prognostic feature (e.g., survival, relapse) of stage 2 lung adenocarcinoma patients who had tumors removed? (prospective study) Descriptive vs. generative?
  - Are the set of genes up-regulated in hepatic cells subject to a drug D significantly enriched for specific ontologic attributes (e.g., signaling pathways, biochemical processes)? Descriptive vs. generative?

---

**Transcriptome studies: Paradigm shift example 1**

- Combinatorial features. Say we measure 2 genes G1, G2 in 30 patients with disease X, and 30 control subjects O. Neither G1 nor G2 alone discriminate X from O. But (the sign of) G1 – G2 does! G1 – G2 (PC2) is the disease discriminant.

---

**Transcriptome studies: Questions**

- How have high throughput gene expression profiling technologies change the way we think about biological systems/problems?
  - Classical biology: Whole = Sum of its parts
    - Microarrays as a large-scale implementation of northern blots or PCR
  - Systems biology: Whole ⊒ Sum of its parts
    - Practical to think about, model relationship between multiple features.
    - 2 views of one dataset (details later)
      - Genes in Sample space
      - Samples in Gene space

---

**Transcriptome studies: Paradigm shift example 1**

- Where PCA will fail to discriminate X from O. Say we measure 2 genes G1, G2 in 50 patients with disease X, and 50 control subjects O. The principal components PC1 and PC2 line up with the maximal sample variance directions - none of which coincide with the true disease status dichotomy!
Transcriptome studies: Prototypical experiment designs

- Prototypical experiment designs
  - Conceptually similar to commonplace scientific experiment designs, only much more features/variables.
  - 2-group comparisons: disease vs. control, treated vs. non-treated
  - Sequential profiling – parametrized by a continuous scalar variable: time, drug dosage, chemical gradient.
  - Hybrid of 2-group and sequential profiling

Typical data analysis meta steps

Typical data analysis: roadmap

- Mathematical formulation of the biological problem
  - Data representation. Map into a “metric” (or normed linear) space.
  - ”Correcting” noise and systematic measurement variation / bias
    - “Pre-processing”. Normalization. Controls / Replicates.
  - Uncovering geometric regularities and dominant variance structures in dataset
    - “Supervised” vs “unsupervised” analyses, e.g., clustering, machine learning
    - Likelihood of geometric regularities/math results arising by “chance”. False positives (technical / statistical)
      - Modeling ”chance”. Statistics
      - Correspondence between (molecular) regularities and phenotype?
  - Squaring math results with a priori biological knowledge. Figure of merit
    - Statistics
    - False positives (biological). Correlated vs causal (generative)
      - “Integrative genomics” – investigate “similar” system for common themes.
      - “Reverse engineering”
**Data analysis: Starting data set/matrix**

- Almost always transcriptome data analysis, modeling starts off with a genes x samples matrix.
- 2 view of 1 dataset:
  - Genes in sample space
  - Samples in gene space

<table>
<thead>
<tr>
<th></th>
<th>Exp 1</th>
<th>Exp 2</th>
<th>Exp 3</th>
<th>Exp M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>G_{1,1}</td>
<td>G_{1,2}</td>
<td>G_{1,3}</td>
<td>...</td>
</tr>
<tr>
<td>Q 2</td>
<td>G_{2,1}</td>
<td>G_{2,2}</td>
<td>G_{2,3}</td>
<td>...</td>
</tr>
<tr>
<td>Q 3</td>
<td>G_{3,1}</td>
<td>G_{3,2}</td>
<td>G_{3,3}</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>Q N</td>
<td>G_{N,1}</td>
<td>G_{N,2}</td>
<td>G_{N,3}</td>
<td>...</td>
</tr>
</tbody>
</table>

Q I = Gene i
Exp j = Experiment / Sample j

**Data analysis: Math formulation, Data Representation**

- Data representation (DR). First map data into a “metric” space, more generally a normed linear space
  - To determine whether 2 objects are “similar”. Notion of similarity is embodied in the metric (more generally, a dis/similarity measure)
  - Example: 2 different similarity measures are the Euclidean distance (intuitive geometric distance, a true metric), and Pearson linear correlation (not a true metric). Physically, Euclidean distance = difference in displacement, Correlation = difference in velocity

**In Correlation space**
- Magenta & Blue are more similar than Magenta & Green

**In Euclidean space**
- Magenta & Blue are less similar than Magenta & Green

**Data analysis: Noise**

- Model “noise” or systematic measurement biases / variations
  - What is Noise? Deviations from axioms / assumptions about “replicate” states. This deviation may be expressed / reflected in the (numerical) data. Clearly, if detection limit is gross the expression of noise is minimized.
  - Example of logical axiom: Replicate measurements of a system-state should be similar in given metric space.
  - How to correct for noise? Normalization
    - Normalization is a math transformation to minimize noise, while preserving gene expression variation resulting from biologically relevant transcriptome activity.
    - Which transformation? Depends upon reference logical / scientific axiom violated
    - Normalization example: Equalize the mean transcriptome levels across samples.
  - Replicates are critical to characterize noise
Data analysis: Noise, replicates

- Different concepts a Replicate
  - Scatter plots of reported transcriptome levels between replicates

Data analysis: DR & geometric regularities

- Given a transcriptomic data set, we can view the data as
  - Genes in Sample space
  - Samples in Gene space
- Question: Might there be geometric regularities and dominant variance structures in the data?
  - Identify variationally meaningful data structure in feature set
  - Do coherent geometric regularities/variance structures exist?
  - “Supervised” and “unsupervised” math techniques. E.g., clustering, machine learning
- Unsupervised = sample labels are not used by method. Supervised = sample labels are inputs into method.
- Many math methods exist, most ported from physical and engineering science. Which is “best”? 2 rules of thumb
  - Scientific question should guide choice of method. Not other way around
  - Upon deciding on a method, run method on simulated data. Figure of merit

Data analysis: DR & geometric regularities example 1a

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).
  - Genes in Sample space I. Euclidean space.

Data analysis: DR & geometric regularities example 2a-c

- Fourier representation of biological systems with periodic behavior
**Data analysis: DR & regularities example 1b**

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).
  - Genes in Sample space II. Correlation space

**Data analysis: DR & regularities example 1c**

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).
  - Samples in Gene space I. Euclidean space

**Data analysis: DR & regularities example 1d**

- Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).
  - Samples in Gene space II. Correlation space

**Data analysis: DR & regularities example 2a**

- Fourier decomposition. Sum of 3 time sinosoids in frequency space. No noise
Data analysis: DR & regularities example 2b

- Fourier decomposition. Sum of 3 sinusoids in freq space. With small deterministic (periodic) perturbations / noise.

Data analysis: How likely are regularities due to chance?

- Squaring math results with chance
  - Modeling “chance” in the system. Statistics
  - False positives due to:
    - Technical: Noise, Multiple testing
    - Inherent biology: Secondary effects, pleiotropy
  - Assumptions about null distribution (fancy term for “chance”)
  - Permutation testing:
    - Permute data. Run similar analyses to extract geometric regularities/variance structures and their statistic.
    - Get distribution for statistic of regularities in permuted data.
    - Examine statistic from unperturbed data relative to this distribution of statistics from permuted data.

Data analysis: DR & regularities example 2c

- Fourier decomposition. Sum of 3 sinusoids in freq space. With stochastic noise.

Data analysis: Does model mirror physical system, reality?

- Squaring math results with a priori biological knowledge. Figure of merit: biological vs in silico
  - Biological validation: Experiments guided by new hypotheses.
  - In silico validation: “Integrative genomics” - investigate “similar” system for common themes.
  - Coherent / dominant mathematical structures that are identified in dataset ideally have a bio-physical (non technical) correlate.
  - Many models for 1 dataset. Which best mirrors bio-physical situation?
  - 1 physical system --> 1 data set --> >1 possible models --> 1 physical system?
    - How to pick “best” model?
    - Well-definedness
    - Reality checks.
User-friendly references

- I2B2 website, general talks on bleeding edge integrative genomics (streaming video)  
  http://www.i2b2.org/events/index.html

- Refs that demonstrate applied math ethos in physical science & biology.
  
    http://www.dartmouth.edu/~matc/MathDrama/reading/Wigner.html

    http://www.turingarchive.org/browse.php/B/

Not easy

- There is no sense in being precise when you don't even know what you are talking about.
  
  If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is. John von Neumann, 1903-57