Gene expression bioinformatics
Part 2: Data representation and modeling

September 27, 2017

Alvin T. Kho
Boston Children's Hospital

alvin_kho@hms.harvard.edu
Outline

- Review key points from last lecture.
  - Gene ↔ DNA identification opens door to computation.
  - Central dogma.

- Assumptions & questions in high throughput transcriptome studies – “Big Data”
  - “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
Review last lecture

- Characterizing a biological system
  - Organizational scales and constituents: Interaction of Micro (molecular, genotype) + Macro (phenotype) + Environment
  - Central Dogma reductionism: Key role of genes in biology.
  - Gene as double helix DNA sequence:
    - Bridge between biology and chemistry/physics.
    - 1st significant entry of computation into biology.

- High-throughput gene identification & quantification:
  - Identification: Genomic & cDNA libraries. Sequencing.
  - Quantification: Short (unique) representative sequences for a gene:
    - Sequencing: SAGE, (next generation seq generalization)
    - Nucleotide complementarity: microarrays
  - Key data features: Massively parallel. “Noise” & “false positives” (theme problems)
  - Genomics: Holistic study of structure & function of the genome.
Transcriptome studies: Questions

- Granularity of questions, 3 molecular scales
  - **Single**: Identify single molecules associated with a biological system state
  - **Network**: Identify molecular networks/interactions associated with a biological system state.
  - **System**: Whole transcriptome profile of a biological system state
- Technology/Big data changes how we conceptualize biological problems
  - **Classical biology**: Whole = Sum of its parts
  - **Systems biology**: Whole ≥ 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:
  - Given molecular profiles of samples from 3 clinically distinct inflammatory bowel diseases, identify the minimal gene set distinguishing these diseases. Is gene set descriptive/correlative vs. causative?
  - Is there a molecular signature in tumors resected from stage 1 lung adenocarcinoma patients that correlate with prognosis (e.g., survival, response to drug X)? (Prospective study) Descriptive vs. generative?
  - Are genes up-regulated in hepatic cells subject to a growth factor X significantly enriched for specific ontologic attributes (e.g., signaling pathways, molecular function)? Descriptive vs. generative?
  - **Theme:** Solve
How has high throughput gene expression profiling technology changed the way we conceptualize biological problems?

**Classical biology**: Whole = Sum of its parts
- Microarray = many independent northern blots or PCR

**Systems biology**: Whole ≥ Sum of its parts
- Model relationships across multiple features & scales.
- 2 views of one dataset (details later)
  - Genes in Sample space
  - Samples in Gene space

---

Transcriptome studies: Questions

Data matrix

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Patient ID</th>
<th>G1</th>
<th>G2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>101</td>
<td>8.439</td>
<td>8.432</td>
</tr>
<tr>
<td>O</td>
<td>102</td>
<td>9.139</td>
<td>8.711</td>
</tr>
<tr>
<td>X</td>
<td>103</td>
<td>7.314</td>
<td>7.461</td>
</tr>
<tr>
<td>O</td>
<td>104</td>
<td>6.709</td>
<td>6.568</td>
</tr>
<tr>
<td>O</td>
<td>105</td>
<td>10.361</td>
<td>10.426</td>
</tr>
<tr>
<td>O</td>
<td>106</td>
<td>8.026</td>
<td>7.974</td>
</tr>
<tr>
<td>O</td>
<td>107</td>
<td>10.588</td>
<td>10.576</td>
</tr>
<tr>
<td>X</td>
<td>108</td>
<td>9.992</td>
<td>10.119</td>
</tr>
<tr>
<td>O</td>
<td>109</td>
<td>14.617</td>
<td>14.680</td>
</tr>
<tr>
<td>X</td>
<td>110</td>
<td>11.923</td>
<td>11.896</td>
</tr>
<tr>
<td>X</td>
<td>111</td>
<td>10.821</td>
<td>11.021</td>
</tr>
<tr>
<td>X</td>
<td>112</td>
<td>8.990</td>
<td>8.941</td>
</tr>
<tr>
<td>O</td>
<td>113</td>
<td>8.528</td>
<td>8.564</td>
</tr>
<tr>
<td>O</td>
<td>114</td>
<td>9.405</td>
<td>9.660</td>
</tr>
<tr>
<td>X</td>
<td>115</td>
<td>8.544</td>
<td>8.597</td>
</tr>
<tr>
<td>O</td>
<td>116</td>
<td>10.718</td>
<td>10.521</td>
</tr>
<tr>
<td>X</td>
<td>117</td>
<td>7.487</td>
<td>7.452</td>
</tr>
<tr>
<td>O</td>
<td>118</td>
<td>11.166</td>
<td>11.192</td>
</tr>
<tr>
<td>X</td>
<td>119</td>
<td>10.581</td>
<td>10.669</td>
</tr>
<tr>
<td>O</td>
<td>120</td>
<td>12.686</td>
<td>12.672</td>
</tr>
<tr>
<td>X</td>
<td>121</td>
<td>14.035</td>
<td>14.029</td>
</tr>
<tr>
<td>O</td>
<td>122</td>
<td>7.738</td>
<td>7.283</td>
</tr>
<tr>
<td>O</td>
<td>123</td>
<td>8.523</td>
<td>8.653</td>
</tr>
<tr>
<td>X</td>
<td>124</td>
<td>8.442</td>
<td>8.343</td>
</tr>
<tr>
<td>X</td>
<td>125</td>
<td>9.006</td>
<td>9.185</td>
</tr>
<tr>
<td>O</td>
<td>126</td>
<td>7.783</td>
<td>7.922</td>
</tr>
<tr>
<td>O</td>
<td>127</td>
<td>7.383</td>
<td>7.383</td>
</tr>
<tr>
<td>X</td>
<td>128</td>
<td>9.126</td>
<td>8.938</td>
</tr>
<tr>
<td>O</td>
<td>129</td>
<td>12.198</td>
<td>12.298</td>
</tr>
<tr>
<td>X</td>
<td>130</td>
<td>11.389</td>
<td>11.415</td>
</tr>
</tbody>
</table>

...
Transcriptome studies: Whole >Σ , Example 1

- Combinatorial features. Say we assay G1, G2 in 30 diseased patients X, and 30 controls O. Neither G1 nor G2 alone discriminate X O. But (the sign of) G1 – G2 does! G1 – G2 (PC2) = disease discriminant.

* Principal component analysis
  Singular value decomposition of 2x2 covariance matrix
Example PCA fails to discriminate X O. Say we assay G1, G2 in 50 patients with disease X, and 50 control subjects O. The principal components PC1 & PC2 line up with maximal sample variance directions – none correlate with true disease status!

Transcriptome studies: Whole >Σ, Example 1

* Principal component analysis
Singular value decomposition of 2x2 covariance matrix
Transcriptome studies: Prototypical experiment designs

- Prototypical experiment designs
  - **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
  - Key: Assay many features/variables per sample.

![Graph showing expression levels of gene G in disease and control groups]

- Each colored line = 1 gene
Typical work flow in a transcriptomic study

Biological question

Biological system → Appropriate control, replicate, experiment design → Extract RNA (make cDNA) → Chip hybridization and scanning OR Sequencing → Data analysis dataset matrix → Biological validation

Our focus
expanded next...
Typical data analysis meta steps

Map data into metric/measure space, model appropriate to biological question

Math formulation
Data representation

(Biological proxy/representation)
Transcriptome, Proteome

Biological system-state

Prediction. Inferential statistic. Correlation vs causality “Integrative genomics” (investigate similar system for common themes)

Do regularities/variances have biological correlates?

False positives – biological

Uncover regularities / dominant variance structures in data

False positives – technical, statistical

Likelihood of geometric regularities/variance structures arising by chance alone False positives

Normalization
Replicates

Correct for noise, variation arising not from relevant transcriptome program

Noise – technical & biological variation

Un/supervised math techniques. E.g., clustering, networks, graphs, myriad computational techniques guided by overriding scientific question!

Chance modeled by null hypothesis
Statistics
Permutation analyses

Gene    P1-1  P3-1  P5-1  P7-1  P10-1
Csrp2   -2.4  74.6  25.5  -30.7  14.6
Mxd3    126.6 180.5 417.4  339.2  227.2
Mxi1    2877.2 1535 2195.6  3681.3  3407.1
Zfp422  458.5  353.3  581.5  520  348
Nm1c1   4130.3 2984.2 3145.5  3895  2134.3
E2f1    1244  1761.5  1503.6  1434.9  487.7
Atoh1   94.9  181.9  268.6  184.5  198
Hmgb2   9737.9 12542.9 14502.8 12797.7 8950.6
Pax2    379.3  584.9  554  438.8  473.9
Tcfap2a 109.8  152.9  349.9  223.2  169.1
Tcfap2b 4544.6 5299.6 3418.1 3429.5 1579.4

Gene    P1-1  P3-1  P5-1  P7-1  P10-1
Csrp2   -2.4  74.6  25.5  -30.7  14.6
Mxd3    126.6 180.5 417.4  339.2  227.2
Mxi1    2877.2 1535 2195.6  3681.3  3407.1
Zfp422  458.5  353.3  581.5  520  348
Nm1c1   4130.3 2984.2 3145.5  3895  2134.3
E2f1    1244  1761.5  1503.6  1434.9  487.7
Atoh1   94.9  181.9  268.6  184.5  198
Hmgb2   9737.9 12542.9 14502.8 12797.7 8950.6
Pax2    379.3  584.9  554  438.8  473.9
Tcfap2a 109.8  152.9  349.9  223.2  169.1
Tcfap2b 4544.6 5299.6 3418.1 3429.5 1579.4

Math formulation
Data representation

(Biological proxy/representation)
Transcriptome, Proteome
Typical data analysis: Roadmap

- Mathematical formulation of the biological problem
  - Data representation. Map into a “metric” space.
- “Correcting” noise and systematic measurement variation / bias
  - 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
- Squaring math results with *a priori* biological knowledge. Figure of merit
  - Statistics
- Correspondence between (molecular) regularity & biology (phenotype)?
- False positives (biological). Correlative vs. causative.
  - “Integrative genomics” – investigate “similar” system for common themes.
    - Meta analysis
  - “Reverse engineering”.
Data analysis: Starting data set/matrix

- Almost always transcriptome data analysis, modeling starts off with a genes x samples matrix.

- 2 different views of 1 dataset:
  - Genes in sample space
  - Samples in gene space

**MACROSCOPIC**

<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: Math formulation, DR

- Different DR emphasize different inherent patterns/regularities in complex datasets (if they exist!). Such regularities may have biologic correlates.
  - What are regularities? Features that pass particular statistical criteria.
  - Toy example, genes mapped into a 2-d sample space.
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 assumed reference mathematical / scientific axiom(s).
    - Normalization example: Equalize the mean transcriptome levels across samples.

- Replicates are critical to characterize noise
Data analysis: Noise, replicates

- Different concepts of 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

- Again, any genes $\times$ samples data matrix can be viewed as,
  - **Genes in Sample space**
  - **Samples in Gene space**
  - Typically for transcriptome data, $\#$ Genes $>> \#$ Samples
  - These spaces may have different similarity measures

- DR example 1a-d in the following 4 slides: 6,000 distinct RNA-gene levels measure in developing mouse cerebellum day 1-60.
  - Visualizing Gene-wise: Profiles (“Speed”, correlation) #2 versus Absolute Intensity (euclidean) #1
  - Visualizing Dev Stage-wise: Profiles (“Speed”, correlation) #2 versus Absolute Intensity (euclidean) #1
  - DR tool: Principal Component Analysis (PCA), an affine change-of-coordinates. Similarity in the euclidean sense.
  - Central Limit Theorem (CLT) normalization / “standardization” --> Mean 0, Variance 1 if Profile “Speed” matters rather than Absolute intensity. Later.

- DR example 2a-c: Fourier representation of biological systems with periodic behavior
**Data analysis: DR & regularities example 1**

- Example: Mouse whole cerebella transcriptome @ day P1, 3, 5, 7, 10, 15, 21, 30, 50, 60 – in duplicates. 10K genes measured

<table>
<thead>
<tr>
<th>Probe Set ID</th>
<th>Entrez ID</th>
<th>Gene Symbol</th>
<th>Day 1 a</th>
<th>Day 1 b</th>
<th>Day 3 a</th>
<th>Day 3 b</th>
</tr>
</thead>
<tbody>
<tr>
<td>162448_f_at</td>
<td>80888</td>
<td>Hspb8</td>
<td>8.439</td>
<td>8.432</td>
<td>8.258</td>
<td>8.346</td>
</tr>
<tr>
<td>162449_f_at</td>
<td>11536</td>
<td>Gpr182</td>
<td>9.139</td>
<td>8.711</td>
<td>8.459</td>
<td>8.725</td>
</tr>
<tr>
<td>162450_f_at</td>
<td>16005</td>
<td>Igfals</td>
<td>7.314</td>
<td>7.461</td>
<td>7.222</td>
<td>7.406</td>
</tr>
<tr>
<td>162451_r_at</td>
<td>57443</td>
<td>Fbxr3</td>
<td>6.709</td>
<td>6.568</td>
<td>6.750</td>
<td>6.644</td>
</tr>
<tr>
<td>162452_at</td>
<td>18545</td>
<td>Pcp2</td>
<td>10.361</td>
<td>10.426</td>
<td>10.374</td>
<td>10.409</td>
</tr>
<tr>
<td>162454_f_at</td>
<td>22024</td>
<td>Crip2</td>
<td>8.026</td>
<td>7.974</td>
<td>7.859</td>
<td>8.029</td>
</tr>
<tr>
<td>162455_f_at</td>
<td>68796</td>
<td>Tmem214</td>
<td>10.588</td>
<td>10.576</td>
<td>10.540</td>
<td>10.583</td>
</tr>
<tr>
<td>162458_i_at</td>
<td>108037</td>
<td>Shmt2</td>
<td>11.923</td>
<td>11.898</td>
<td>11.807</td>
<td>11.903</td>
</tr>
<tr>
<td>162459_f_at</td>
<td>12833</td>
<td>Col6a1</td>
<td>10.821</td>
<td>11.021</td>
<td>11.145</td>
<td>10.950</td>
</tr>
<tr>
<td>162460_f_at</td>
<td>18518</td>
<td>Igbp1</td>
<td>8.990</td>
<td>8.941</td>
<td>8.869</td>
<td>8.928</td>
</tr>
<tr>
<td>162461_f_at</td>
<td>12332</td>
<td>Capg</td>
<td>8.528</td>
<td>8.564</td>
<td>8.680</td>
<td>8.562</td>
</tr>
<tr>
<td>162463_at</td>
<td>21985</td>
<td>Tdp52</td>
<td>8.544</td>
<td>8.597</td>
<td>8.516</td>
<td>8.584</td>
</tr>
<tr>
<td>162464_f_at</td>
<td>19064</td>
<td>Ppy</td>
<td>10.718</td>
<td>10.521</td>
<td>10.476</td>
<td>10.549</td>
</tr>
<tr>
<td>162465_i_at</td>
<td>234825</td>
<td>Klhd4</td>
<td>7.487</td>
<td>7.452</td>
<td>7.551</td>
<td>7.531</td>
</tr>
<tr>
<td>162466_at</td>
<td>629182</td>
<td>Anapc13</td>
<td>11.166</td>
<td>11.192</td>
<td>11.030</td>
<td>11.091</td>
</tr>
<tr>
<td>162467_f_at</td>
<td>101513</td>
<td>2700078K21Ri</td>
<td>10.581</td>
<td>10.669</td>
<td>10.641</td>
<td>10.739</td>
</tr>
<tr>
<td>162468_at</td>
<td>11853</td>
<td>Rhoc</td>
<td>12.668</td>
<td>12.672</td>
<td>12.694</td>
<td>12.732</td>
</tr>
<tr>
<td>162469_f_at</td>
<td>66445</td>
<td>Cyc1</td>
<td>14.035</td>
<td>14.029</td>
<td>13.977</td>
<td>14.014</td>
</tr>
<tr>
<td>162471_f_at</td>
<td>68523</td>
<td>Fam96b</td>
<td>7.738</td>
<td>7.283</td>
<td>7.503</td>
<td>7.509</td>
</tr>
<tr>
<td>162472_f_at</td>
<td>14459</td>
<td>Gast</td>
<td>8.523</td>
<td>8.653</td>
<td>8.719</td>
<td>8.687</td>
</tr>
<tr>
<td>162473_r_at</td>
<td>68198</td>
<td>Ndufb2</td>
<td>8.442</td>
<td>8.343</td>
<td>8.018</td>
<td>8.140</td>
</tr>
<tr>
<td>162477_r_at</td>
<td>20813</td>
<td>Srp14</td>
<td>7.783</td>
<td>7.922</td>
<td>7.936</td>
<td>7.896</td>
</tr>
<tr>
<td>162478_r_at</td>
<td>15929</td>
<td>Idh3g</td>
<td>7.383</td>
<td>7.383</td>
<td>7.458</td>
<td>7.485</td>
</tr>
<tr>
<td>162480_f_at</td>
<td>13590</td>
<td>Lefty1</td>
<td>9.126</td>
<td>8.938</td>
<td>8.894</td>
<td>8.896</td>
</tr>
<tr>
<td>162481_f_at</td>
<td>109689</td>
<td>Arnb1</td>
<td>12.198</td>
<td>12.298</td>
<td>12.159</td>
<td>12.134</td>
</tr>
<tr>
<td>162482_at</td>
<td>22032</td>
<td>Traf4</td>
<td>11.389</td>
<td>11.415</td>
<td>11.324</td>
<td>11.390</td>
</tr>
</tbody>
</table>

**Microarray / Chip**

**Data matrix**

**Heatmap, genes/rows standardized**
Data analysis: DR & regularities example 1a

- Example: Mouse cerebellar development 10K genes at 9 time stages (duplicate).

- Genes in Sample space I. Euclidean space.

Each dot is a gene

PCA representation

Regularities here?

Each line is a gene

Cereb postnatal days

Signal

0 5 10 15

1 3 5 7 9 11 13 15 17 19

PC1 58.64%

PC2 24.26%
Data analysis: DR & regularities example 1b

- Example: Mouse cerebellar development 10K genes at 9 time stages (duplicate).
  - Genes in Sample space II. Correlation space.

Each line is a gene

Each dot is a gene

PCA representation

Regularities here?
Data analysis: DR & regularities example 1c

- Example: Mouse cerebellar development 10K genes at 9 time stages (duplicate).

  - Samples in Gene space I. Euclidean space

Do configurations say anything bio-meaningful?
Data analysis: DR & regularities example 1d

- Example: Mouse cerebellar development 10K genes at 9 time stages (duplicate).

  Samples in Gene space II. Correlation space

Each dot is a sample
Data analysis: DR & regularities example 2a

- Fourier decomposition. Sum of 3 time sinosoids in frequency space. No noise

Y-axis depends on discretization of time

Boundary condition artifact!

X axis depends on inherent frequencies

Application in sequence analysis: \{A,T,C,G\} -> \{0,1,2,3\} -> Fourier
Data analysis: DR & regularities example 2b

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

Y-axis depends on discretization of time

Small periodic perturbations

X axis depends on inherent frequencies
Data analysis: DR & regularities example 2c

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

Y-axis depends on discretization of time

Not robust with stochastic noise

X axis depends on inherent frequencies
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.

Original, un-permuted data matrix

1000 permutation cycles of data matrix
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 $\rightarrow$ 1 data set $\rightarrow$ >1 possible models $\rightarrow$ 1 physical system?
  - How to pick “best” model?
  - Well-definedness
  - Reality checks.
Revisit: Typical data analysis meta steps

Map data into metric/measure space, model appropriate to biological question

\[
\begin{array}{c|c|c|c|c|c}
\text{Gene} & P1-1 & P3-1 & P5-1 & P7-1 & P10-1 \\
\hline
\text{Csrp2} & -2.4 & 74.6 & 25.5 & -30.7 & 14.6 \\
\text{Mxd3} & 126.6 & 190.5 & 417.4 & 399.2 & 227.2 \\
\text{Mxi1} & 287.2 & 1535 & 2195.6 & 3681.3 & 3407.1 \\
\text{Zfp422} & 458.5 & 353.3 & 581.5 & 520 & 348 \\
\text{Nmyc1} & 4130.3 & 2894.2 & 3145.5 & 3886 & 2134.3 \\
\text{E2f1} & 1244 & 1761.5 & 1503.6 & 1434.9 & 487.7 \\
\text{Atoh1} & 94.9 & 181.9 & 268.6 & 184.5 & 198 \\
\text{Hmgb2} & 9737.9 & 12542.9 & 14502.8 & 12797.7 & 8950.6 \\
\text{Pax2} & 379.3 & 1529.9 & 349.9 & 223.2 & 169.1 \\
\text{Tcfap2a} & 109.8 & 152.9 & 349.9 & 223.2 & 169.1 \\
\text{Tcfap2b} & 4544.6 & 5299.6 & 3418.1 & 3429.5 & 1579.4 \\
\end{array}
\]

Un/Supervised math techniques. E.g., clustering, networks, graphs, myriad computational techniques guided by overriding scientific question!

Normalization Replicates

Correct for noise, variation arising not from relevant transcriptome program

Do regularities/variances have biological correlates?

False positives – technical, statistical

False positives – biological

Uncover regularities / dominant variance structures in data

Likelihood of geometric regularities/variance structures arising by chance alone

False positives

Noise – technical & biological variation

Math formulation Data representation

Biological system-state

(Biological proxy/representation) Transcriptome, Proteome

Prediction. Inferential statistic. Correlation vs causality

“Integrative genomics” (investigate similar system for common themes)
Example study: Human lung control vs adenocarcinoma

<table>
<thead>
<tr>
<th>Gene</th>
<th>Fold T/N</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSRP2</td>
<td>-0.228</td>
<td>0.150</td>
</tr>
<tr>
<td>MXD3</td>
<td>1.268</td>
<td>0.010</td>
</tr>
<tr>
<td>MXI1</td>
<td>-0.190</td>
<td>0.500</td>
</tr>
<tr>
<td>ZFP422</td>
<td>0.200</td>
<td>0.300</td>
</tr>
<tr>
<td>NMYC1</td>
<td>0.346</td>
<td>0.450</td>
</tr>
<tr>
<td>E2F1</td>
<td>0.312</td>
<td>0.210</td>
</tr>
<tr>
<td>ATOH1</td>
<td>0.537</td>
<td>0.050</td>
</tr>
<tr>
<td>HMGB2</td>
<td>-0.560</td>
<td>0.050</td>
</tr>
<tr>
<td>PAX2</td>
<td>-0.397</td>
<td>0.700</td>
</tr>
<tr>
<td>TCFAP2</td>
<td>0.214</td>
<td>0.770</td>
</tr>
<tr>
<td>TCFAP3</td>
<td>-0.447</td>
<td>0.080</td>
</tr>
</tbody>
</table>

Is candidate gene set enriched for specific ontologic attributes?

Integrative genomics To identify common regularities at genetic or ontologic levels

Principal component analysis (PCA) to see global sample variations

GSE10072 human AD vs N lung

Permutation test for false discovery rate

Permutation testing

Ontologic enrichment?

Permutation testing
## Genome Wide Association Study

- Typical matrix representation of big data. G, T, S positive integers, where G & T >> S.

<table>
<thead>
<tr>
<th>Examples</th>
<th>Features</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
<th>Subject 4</th>
<th>Subject 5</th>
<th>...</th>
<th>Subject S</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11655198</td>
<td>Genome Location 1</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>Genome Location 2</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>Genome Location 3</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>Genome Location 4</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genome Location G</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>asthma, no asthma</td>
<td>Phenotypic Trait 1 (binary, N=2)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>race</td>
<td>Phenotypic Trait 2 (categorical, N=3)</td>
<td>P3</td>
<td>P1</td>
<td>P1</td>
<td>P2</td>
<td>P2</td>
<td>...</td>
<td>P1</td>
</tr>
<tr>
<td>height</td>
<td>Phenotypic Trait P (continuous real number)</td>
<td>158.1</td>
<td>180.0</td>
<td>173.7</td>
<td>147.6</td>
<td>149.9</td>
<td></td>
<td>182.5</td>
</tr>
<tr>
<td></td>
<td>Expression Gene 1</td>
<td>1.619</td>
<td>0</td>
<td>1.565</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.491</td>
</tr>
<tr>
<td></td>
<td>Expression Gene 2</td>
<td>0.834</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Typical analyses

- Typical analyses, relationships between pairs:
  - Asthma / No asthma vs Genome Location 2
  - Asthma / No asthma vs Gene 3 expression
  - Height vs Gene 173 expression

<table>
<thead>
<tr>
<th></th>
<th>Asthma</th>
<th>No asthma</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome Location 2: A</td>
<td>3324</td>
<td>2104</td>
<td>5428</td>
</tr>
<tr>
<td>Genome Location 2: G</td>
<td>2676</td>
<td>1896</td>
<td>4572</td>
</tr>
<tr>
<td>Total</td>
<td>6000</td>
<td>4000</td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio: 1.119
95% confidence interval: 1.033 - 1.213
p value: 0.006
Typical analyses

- Typical analyses, relationships between everything:
  - before O & after X steroid tx in blood

  *line connects same subject*

---

Logsdon BA, Hoffman GE, Mezey JG
BMC Bioinformatics, 2012
PMCID: PMC3338387
If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.

The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann, 1903-57