Bioinformatics of high-throughput bio-molecular measurement technologies. Part 2

September 23, 2008

Alvin T. Kho
Children’s Hospital Boston

alvin_kho@hms.harvard.edu
Lecture outline

- Review key points of last Thursday's lecture
- Assumptions & questions in microarray studies
  - “Granularity” of questions
  - Paradigm shift in thinking about biological systems, questions
  - Prototypical experiment designs
- Typical work flow in microarray studies
- Analysis and modeling of transcriptome data (rules of thumb),
  - 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 \textit{a priori} biological knowledge. Figure of merit: biological vs \textit{in silico}
- User-friendly References
Review of last lecture

- Characterizing a biological system
  - Features (molecular, genotype), their interactions + environment to engender an observable event (phenotype). Multi-scalar in space & time.

- Genes as key players in biological events (genetic basis):
  - heritable traits -> modulate macroscopic aspects of a bio system?
  - genes <-> DNA. Central Dogma: DNA -> RNA -> Protein
  - convergent phenotypes: different genes -> 1 phenotypic endpoint
  - sequence genomics: 1st major confluence of high thru-put bio & computation, math modeling

- High-throughput gene expression assays (basic principle)
  - SAGE (sequencing) & microarrays (nucleotide complementarity)
  - Gain: Massive parallelity, high throughput
  - Loss: Specificity down, “noise” up

- Theme problems for massive datasets
  - High false positive rate (type 1 error)
  - Modeling
Microarray studies: Definitions & assumptions

• Definition: Functional genomics
  • De-constructing the genome to assign biological function: Genotype -> Biological Function -> Phenotype

• Transcriptomics: As above, but restricted to expressed subset.

• Implicit assumption
  • Penomenological event of a biological system engages or is modulated by its transcriptome.
Microarray 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:** Transcriptomic profile / state of a biological system
- Paradigm shift in thinking about a biological system / question
  - **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 continuously-varying scalar variable
  - Hybrid of 2-group and sequential profiling
Some questions that can be practically investigated with transcriptome profiling:

- 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 that correlates with prognosis of stage 2 lung adenocarcinoma patients? Descriptive vs. generative?

- Are the set of genes up-regulated in hepatic cells subject to a drug D significantly enriched for specific functional / ontologic attributes? Descriptive vs. generative?
Microarray studies: Questions

• How have high throughput gene expression profiling technologies change the way we think about and model biological systems/problems? ... paradigm shift.

• **Classical biology**: Whole = Sum of its parts
  - Microarrays as a large-scale application of northern blots or PCR

• **Systems biology**: Whole ≥ Sum of its parts
  - Practical to think about, model, test combinatorial / multi-factorial features.
  - 2 views of a common dataset (more later)
    - Genes in Sample space
    - Samples in Gene space
Microarray studies: Paradigm shift toy 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.

* Principal component analysis
Singular value decomposition of 2x2 covariance matrix
Microarray studies: Paradigm shift toy example 1

- Where PCA will fail to discriminate \( X \) from \( O \). Say we measure 2 genes \( G_1, G_2 \) 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!

*Principal component analysis
Singular value decomposition of 2x2 covariance matrix*
Microarray 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 continuously-varying scalar variable: time course, dosage-varying study
  - Hybrid of 2-group and sequential profiling
Work flow in microarray experiments

• Appropriate experimental design: Replicates – biological and measurement / technical. Controls.

• RNA target / probe preparation: Extract mRNA. Convert to single strand cDNA (typically). Label with fluorescence.

• Probe hybridization. Fluorescence scan.

• (Fluorescence) Image analysis.

• (Post image) Data analysis and modeling to generate more focused hypotheses.

• Biological validation
Work flow in microarray experiments

Biological question

Biological system

Appropriate tissue, condition, experiment design

Extract RNA (make cDNA)

Chip hybridization and scanning

Image analysis

Data analysis

Our focus
expanded next ...

Biological validation
Microarray data analysis meta steps

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

Math formulation

Data representation

Correct for noise, variation arising not from bio-relevant transcriptome program

Normalization Replicates

Uncover regularities / variance structures in data

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

Prediction. Inferential statistic. Hamilton’s principle – minimizing an energy functional

Correlation vs causality

Likelihood of regularities / variance structures arising by chance alone

Chance modeled by null hypothesis
Statistics
Permutation analyses

Do regularities/variances mirror biological parameters, system?

The Big Picture
Data analysis: Transcriptome data

- Mathematical formulation of the biological / physical problem
  - Mapping physical problem into a normed linear (Banach) or metric space. Data representation. Modeling.

- “Correcting” noise and systematic measurement variation / bias
  - “Pre-processing”. Normalization. Replicate measurements.

- Uncovering geometric regularities and dominant variance structures intrinsic to data
  - “Supervised” and “unsupervised” math techniques. E.g., clustering, machine learning

- Likelihood of geometric regularities/math results arising by “chance”
  - Modeling “chance”. Statistics

- Correspondence between regularities and phenotype?

- Squaring math results with a priori biological knowledge. Figure of merit
  - Statistics

- “Reverse engineering”. Correlation (descriptive) vs. causality (generative)
  - Graph and network theory.
Data analysis: The beginning

- Almost always microarray data analysis, modeling starts off with a genes × samples matrix post image analysis ...

\[ \begin{array}{cccc}
\text{Exp 1} & \text{Exp 2} & \text{Exp 3} & \text{Exp M} \\
Q_1 & G_{1,1} & G_{1,2} & G_{1,3} & \ldots & G_{1,M} \\
Q_2 & G_{2,1} & G_{2,2} & G_{2,3} & \ldots & G_{2,M} \\
Q_3 & G_{3,1} & G_{3,2} & G_{3,3} & \ldots & G_{3,M} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Q_N & G_{N,1} & G_{N,2} & G_{N,3} & \ldots & G_{N,M} \\
\end{array} \]

\[ Q_i = \text{Gene } i \]
\[ \text{Exp } j = \text{Experiment / Sample } j \]

This is “Data”
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 do we mean by regularities? Toy example illustrating regularities at different scales:

Data embedded in some metric space representation

Bio significance?
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 of a Replicate
  - Scatter plots of reported transcriptome levels between replicates
Data analysis: DR & geometric regularities

Given a (normalized) 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 subsets (structure) from the mixture of all features
- 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 necessary input 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 × 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 behaviour
Data analysis: DR & regularities example 1a

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

Each dot is a gene

PC1 58.64%

PC2 24.26%

PCA representation

Regularities here?

Genes in Sample space I. Euclidean space

Example: Mouse cerebellar development 6K genes at 9 time stages (duplicate).

Genes in Sample space I. Euclidean space

Each dot is a gene

PCA representation

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

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

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

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

Each dot is a sample

Do configurations say anything bio meaningful?
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\} \rightarrow \{0, 1, 2, 3\} \rightarrow \text{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
  - X-axis depends on inherent frequencies
  - Small periodic perturbation here
Data analysis: DR & regularities example 2c

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

**In Time space**

- Y-axis depends on discretization of time
- Not robust with stochastic noise

**In Frequency space**

- 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

- Assumptions about null hypothetic 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: Does model mirror physical system, reality?

- Squaring math results with *a priori* biological knowledge. Figure of merit: biological vs *in silico*
  - Biological: Experiments guided by new hypotheses.
  - In silico: Statistics
  - In silico: Integrative genomics
- Coherent / dominant mathematical structures that are uncovered via math from data ideally have a physical/biological (non technical) correlate.
- Many analytic methods and attendant models for 1 dataset. Which best mirrors physical system?
- 1 physical system --> 1 data set --> >1 possible models --> 1 physical system?
  - How to pick?
  - Well-definedness
  - Reality checks. How likely is this data and methods
Microarray data analysis meta steps

- Map data into metric/measure space, model appropriate to biological question
- Data representation
- Math formulation
- Correct for noise, variation arising not from bio-relevant transcriptome program
- Replicates
- Normalization
- Un/supervised math techniques. E.g., clustering, networks, graphs, myriad computational techniques guided by scientific question
- Uncover regularities/variance structures in data
- Likelihood of regularities/variance structures arising by chance alone
- Chance modeled by null hypothesis
- Statistics
- Permutation analyses
- Do regularities/variances mirror biological parameters, system?
- Prediction. Inferential statistic. Hamilton's principle – minimizing an energy functional
- Correlation vs causality

The Big Picture

<table>
<thead>
<tr>
<th>Gene</th>
<th>P1-1</th>
<th>Pr-1</th>
<th>P0-1</th>
<th>Pr-1</th>
<th>P1-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Csrpt</td>
<td>-3.6</td>
<td>16.3</td>
<td>20.0</td>
<td>-3.7</td>
<td>16.7</td>
</tr>
<tr>
<td>Mxdn</td>
<td>172.7</td>
<td>144.0</td>
<td>18.7</td>
<td>372.7</td>
<td>257.3</td>
</tr>
<tr>
<td>Mxi1</td>
<td>###</td>
<td>107.0</td>
<td>###</td>
<td>###</td>
<td>###</td>
</tr>
<tr>
<td>Zfp651</td>
<td>408.0</td>
<td>304.0</td>
<td>401.0</td>
<td>47.0</td>
<td>41.6</td>
</tr>
<tr>
<td>Nmci1</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>489.0</td>
<td>###</td>
</tr>
<tr>
<td>Erf1</td>
<td>144.4</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>444.4</td>
</tr>
<tr>
<td>Atoh1</td>
<td>96.9</td>
<td>181.9</td>
<td>184.7</td>
<td>184.0</td>
<td>196</td>
</tr>
<tr>
<td>Hmgb2</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>###</td>
</tr>
<tr>
<td>Paxr</td>
<td>779.2</td>
<td>481.9</td>
<td>80.0</td>
<td>572.8</td>
<td>479.3</td>
</tr>
<tr>
<td>Tcfap2</td>
<td>1.98</td>
<td>103.9</td>
<td>249.9</td>
<td>143.9</td>
<td>119.1</td>
</tr>
<tr>
<td>Tcfap2</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>###</td>
<td>###</td>
</tr>
</tbody>
</table>
User-friendly references

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

- Refs below not microarray specific. Capture applied math ethos in science, biology. Think of standard analytic approaches in engineering/physical sci that can now be practically applied to biological domain.
  
    http://www.dartmouth.edu/~matc/MathDrama/reading/Wigner.html

    http://www.turingarchive.org/browse.php/B/22

- *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