Computation in the New Medicine

Isaac S. Kohane

Overview

- The future is now and it computes
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

The Long Path from Genotype to Function

DNA → Transcription → Translation

Sequencing and genotyping
Transcriptional Profiling
Proteomics
Structural Biology

Magnitude of the Task

Nathan Baker UCSD
Magnitude of the Task

\[ \times 1000 \]

**Madonna Complex**

- Some say \( \text{Madonna}_{\text{music}} > \text{Madonna}_{\text{person}} \)
- 4.7 GB vs
- \((3 \times 10^9) \times 2 \text{ (bits/base)} / 8 \text{ (bits/byte)} = 0.75 \text{GB}\)

- Is Madonna, her DNA sequence?
- No, and her current state is captured by...
  - Alternative splicing \((x\ 3/\text{gene})\)
  - Post-translational modification \((x\ 100-1000)\)
  - Location of gene product \((x\ 10^{12})\)
- She’s a little more complicated than her music (are you surprised?)

**RNA/DNA expression detection chips**

The Promise: The New Diagnostics

Retreading the cancer chemotherapeutic protocol
- Cancer and Leukemia Group B (CALGB)
- CALGB has grown into a national network of 29 university medical centers, over 185 community hospitals and more than almost 3000 physicians who collaborate in clinical research studies aimed at reducing the morbidity and mortality from cancer.
- Dozens of new protocols (breast cancer, prostate cancer, renal cancer) that use genome-wide
  - Which genes best predict survival?
  - Which adjuvant improves surgical outcome the best?
  - Can we find expression measure proxies for Stage, Grade and Cell Type.

New Taxonomy of Human Disease
- Clinicians may have moved on from calling ‘fever’ a disease, but they still rely on phenotypic criteria to define most diseases.
- Yet these may obscure the underlying mechanisms and often mask significant heterogeneity.
- Thomas Lewis pointed out in 1944, diagnosis of most human disease provides only “insecure and temporary conceptions.”
- Of the main common diseases, only the infectious diseases have a truly mechanism-based nomenclature.

Changes in Use of “Every day” Medications
- A common polymorphism associated with antibiotic-induced cardiac arrhythmia

Growth in Our Knowledge of Variants
- Polymorphism Publications

"You're not ill yet, Mr. Blandelli. But you've got potential."
Growth in monogenic disease

With growing databases containing 30,000 cases (1000s-1,000,000) of unique sequence were assayed for variation with high-density oligonucleotide arrays. Synthesized 3.4 x 10^17 oligonucleotides on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation.

Bioinformatics: 1998 Aspirations

- Find the functions of all 30,000 genes using:
  - DNA sequences
  - Genomics maps
  - Physical maps
  - Polymorphisms
  - Structure information
  - Existing biomedicai literature
  - Gene transcription patterns
  - Protein translation/variability

- With growing databases containing data, this becomes a problem in the realm of bioinformatics.

Dangers of Dimensionality

- Given 1000 stadia full of people with necklaces with beads of 10 colors
- 1 in 1000 necklace beads are different every baseball season.
- You notice that the third seat on the fifth row of all games has a yellow bead in the middle of the necklace every year in the season opener at the largest NY stadium in the last 26 wins in last 102 years that the NY Yankees have won the World Series.
- How good a bet is that bead for the next time the NY Yankees are in the World Series?
- Extra credit: DIY

Determining the aforementioned interaction is hard and runs counter to traditional biostatistical techniques

- Variables (10^4 - 10^5)
- Cases (10^1 - 10^2)
- Variables (10^5 - 10^7)
- Cases (10^2 - 10^3)

High-dimensionality systems with insufficient data are undetermined
Not tractable by standard biostatistical techniques

The New Pharmacology

- RNA expression in NCI 60 cell lines was determined using Affymetrix Hu6000 arrays
  - 5,223 known genes
  - 1,193 expressed sequence tags
- The RNA expression data set and Anti-cancer susceptibility data set were merged, using the 60 cell lines the two tables had in common
  - 6,000 genes
  - 5,000 anti-cancer agents
Elevated levels of J02923 (lymphocyte cytosolic protein-1, LCP1, L-plastin, pp65) is associated with increased sensitivity to 624044. Agent 624044 is 4-Thiazolidinecarboxylic acid, 3-[6-[2-oxo-2-(phenylthio)ethyl]-3-cyclohexen-1-yl[acetyl]-2 thioxo, methyl ester, [1R- (1a(R*),6a)]- (9CI)).

LCP1 is an actin-binding protein involved in leukocyte adhesion. A role for LCP1 in tumorogenicity has been previously postulated. Low-level expression of LCP1 is thought to occur in most human cancer cell lines. Other thiazolidine carboxylic acid derivatives are known to inhibit tumor cell growth.

Butte et al. PNAS 2000

### Applications

<table>
<thead>
<tr>
<th>Layer</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. Applications</td>
<td>User and application level tasks (e.g. resource sharing)</td>
</tr>
<tr>
<td>6. Presentation</td>
<td>Display, storage form etc.</td>
</tr>
<tr>
<td>5. Session</td>
<td>Establishes session: OS to transport abstraction</td>
</tr>
<tr>
<td>4. Transport</td>
<td>Guarantees end-to-end control within a session (who’s calling, answering etc)</td>
</tr>
<tr>
<td>3. Network</td>
<td>Addressing, routing through nodes</td>
</tr>
<tr>
<td>2. Data link</td>
<td>Error free path between nodes</td>
</tr>
<tr>
<td>1. Physical</td>
<td>Electrical and mechanical</td>
</tr>
</tbody>
</table>

### What are the major lessons you learned in 6.001?

- The future is now and it computes
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

### Overview

Genes and Anti-Cancer Agents
- Threshold r^2 was 0.8
- 202 networks
- 834 features out of 11,692 (7.1%)
- 1,229 links out of 68,345,586 (.0018%)
- Only one link between a gene and anti-cancer agent.

Genes and Anti-Cancer Agents
- Elevated levels of J02923 is associated with increased sensitivity to 624044.
- Agent 624044 is 4-Thiazolidinecarboxylic acid, 3-[6-[2-oxo-2-(phenylthio)ethyl]-3-cyclohexen-1-yl[acetyl]-2 thioxo, methyl ester, [1R- (1a(R*),6a)]- (9CI)).
- LCP1 is an actin-binding protein involved in leukocyte adhesion.
- A role for LCP1 in tumorogenicity has been previously postulated.
- Low-level expression of LCP1 is thought to occur in most human cancer cell lines.
- Other thiazolidine carboxylic acid derivatives are known to inhibit tumor cell growth.

Butte et al. PNAS 2000
An engineer, a physicist, a mathematician, a computer scientist, and a statistician are on a train heading north, and had just crossed the border into Scotland. They look out the window and see a black sheep for the first time.

The engineer exclaims, “Look! Scottish sheep are black!”

The physicist yells, “No, no. Some Scottish sheep are black.”

The mathematician looks irritated and says, “There is at least one field, containing at least one sheep, of which at least one side is black.”

The computer scientist says, “Oh, no, a special case!”

Finally, the statistician says, “It is not statistically significant!”

<table>
<thead>
<tr>
<th>Genetic vs Genomic</th>
<th>Genetic Medicine</th>
<th>Genomic Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural genomics</td>
<td>Low frequency of ~1000 of usually high penetrance genes</td>
<td>The genetic risk for common diseases will then be due to disease-producing alleles with relatively high frequencies (&gt;1%). All genes may be disease causing.</td>
</tr>
<tr>
<td>Genomics</td>
<td>500’s or relatively uncommon diseases (1/300 for most common)</td>
<td>Common disorders due to the interactions of multiple genes and environmental factors</td>
</tr>
<tr>
<td>Map-based gene discovery</td>
<td>Mostly assessed indirectly &amp; focused on single genes</td>
<td>Direct experimental access to the entire genome</td>
</tr>
<tr>
<td>Monogenic disorders</td>
<td>Genetic risk for common diseases</td>
<td>Genomic risk for common diseases</td>
</tr>
<tr>
<td>Specific DNA diagnosis</td>
<td>Pathogenesis (mechanism)</td>
<td>Pathogenesis (mechanism)</td>
</tr>
<tr>
<td>Analysis of one gene</td>
<td>Genes may be disease causing.</td>
<td>Genes may be disease causing.</td>
</tr>
<tr>
<td>Analysis of multiple genes in gene families, pathways, or systems</td>
<td>Etiology (specific mutation)</td>
<td>Etiology (specific mutation)</td>
</tr>
<tr>
<td>Etiology (specific mutation)</td>
<td>Gene action</td>
<td>Gene action</td>
</tr>
<tr>
<td>One species</td>
<td>Pathway analysis</td>
<td>Pathway analysis</td>
</tr>
<tr>
<td>Several species</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Comprehensive Bioinformatics Approach: All Data Are Grist

Reductionist methods that take into account data-type particularities

Interactions between all the “grist” is relevant to the health state

Malignant maths (From The Economist)
Overview

- The future is now and it computes
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

Heritability: the way a population geneticist would think of it.

- Heritability in the Broad Sense (H)
  - This measure of heritability includes all genetic influences on the phenotype, whether due to additive, dominance, or interactive effects.
  - \( H^2 = \frac{V_G}{V_P} \) where \( V_G = V_A + V_D + V_I \)

Obesity

- Don’t some people just eat and not get fat?
- Isn’t it in their genes?

Heritability

All those of you with… leave the room

Obesity

- National Center for Health Statistics:
  - Over 50% of US adults have BMI > 25
  - About 22% of US adults have BMI > 30
- National Health & Nutrition Examination Survey III:
  - 20% of U.S. children overweight
- Behavioral Risk Factor Surveillance System (CDC)
  - Prevalence of obesity up by 50% from 1991 - 1998
Heritability is defined with respect to environment

- How do we define environment?
  - Diet
  - Daily habits
  - Environmental insults
  - Medical care
  - ...

Genotype does not capture the individual patient state

- We need to capture and quantify the environmental influences.
- These two comprise
  - History
    - Physical
  - Laboratory Studies
  - Imaging
  - Diet
  - Daily habits
  - Environmental insults
  - Medical care
  - ...

The New Medicine: A New Expansion

- The future is now and it computes
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation

The less we capture, the more undetermined the system

Pediatric Growth Monitoring

- Data:
  - heights, weights
  - family history
  - bone ages
  - pubertal data, stages
  - Disorders show characteristic patterns on growth chart.

Work-up of Short Stature with Poor Growth

- T4, IGF-1, ESR, CBC, anti-gliaden Ab...
- Insulin Tolerance Test/Glucagon GH Test
  - 6 hours in the hospital
  - IV insulin with symptomatic hypoglycemia
  - Glucagon with nausea
  - $1000-$2000
- Interpretation remains controversial
  - Significant false positive rate: Why?
  - Significant false negative rate: Why?
1. What is the most common chromosomal cause of short stature?
2. 2.5% of idiopathic short stature children (including males) have SHOX mutations
3. Mutants are not growth hormone deficient but…
   They respond to Growth Hormone therapy!

From SMA-12 to SMA-30000

- \( Q_1 \left( < T_{sel}(1), R_{sel}(k, s) > \right) \)
- \( P_{HJ} \left( \sum P_{null} x_{null} \right) \)
- Performing \( m \) possible tests; 
  - we can choose \( P \left( m! / (m-\ell)! \right) \) test sequences
- If every test has \( r \) possible results, then there will be \( r^n \) possible test histories after \( \ell \) tests
- sum over test histories of every length and multiply by the number of hypotheses, \( n \)
- \( m=10 \) hypotheses, \( m=5 \) binary tests ( \( m^2 \))
  - the analysis requires 63,300 conditional probabilities

Re-engineering the knowledge-base

Cochrane Library
1992 Founded for medical Review Group
7000 Collaborators, internationally

- Anaesthesia Group
- Back Group
- Breast Cancer Group
- Colorectal Cancer Group
- Consumers and Communication Group
- Cystic Fibrosis & Genetic Disorders Group
- Dementia and Cognitive Improvement Group
- Depression, Anxiety and Neurosis Group
- Developmental, Psychosocial and Learning Problems Group
- Drugs and Alcohol Group

Overview

- The future is now and it computes
- Genomic vs genetic
- Heredity
- Resequencing of the diagnostic process
- Accelerating consumer activation
Current trends in criminal inventory and state screening

The privacy challenge is now

Course Administrivia

- Both MIT and Harvard Spring Breaks observed
- Problem sets:
  - 2 total
- Final project (up to 2 persons per project)
  - Project selected by March 15th

Course Overview

- Biology Refresher
- Genomic Measurement Techniques
- Functional Genomics and Microarrays
- Limits of the Technologies: Noise
- Information Science at the Center of Genomic Medicine
- Informational Resources
- Modeling, Reverse Engineering
- The Importance of Data Representation
Course Review

- Machine Learning Approach
- Association with Markers
- Case Hx: Complex Trait Analysis
- Complex Traits: What to believe
- Microarray Disease Classification I
- Direct Prediction Outcome/Mortality
- Histopathology Case History
- Microarray Disease Classification II
- Practical Genomic Medicine: Today's Practice
- Individualized Pharmacology

Course Review

- Finding new drugs
- Ethical and Social Considerations
- Commercial and Regulatory Barriers
- Newborn Testing
- The New Microbiology