Computing and Genomic Medicine

Isaac S. Kohane

http://www.chip.org

Overview

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

The Long Path from Genotype to Function

DNA \[\rightarrow\] Transcription \[\rightarrow\] Translation

Sequencing and genotyping \[\rightarrow\] Transcriptional Profiling \[\rightarrow\] Proteomics \[\rightarrow\] Structural Biology
Myhrvold’s Madonna Complex
**RNA/DNA expression detection chips**


**The Promise: The New Diagnostics**

Alizadeh et al.

**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
Changes in Use of “Every day” Medications

A common polymorphism associated with antibiotic-induced cardiac arrhythmia

Table S3: EGFR mutation status in gefitinib-treated lung cancer

Growth in Our Knowledge of Variants

Polymorphism Publications

Year (1980-2001)
One-pass genotyping

- Chromosome 21: 21,676,868 bases (67%) of unique sequence were assayed for variation with high-density oligonucleotide arrays
- Synthesized 3.4 x 10^9 oligonucleotides on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation.

Bioinformatics: 1998 Ambitions

- Find the functions of all 30,000 genes using:
  - DNA sequence
  - Genetics maps
  - Physical maps
  - Polymorphisms
  - Structure information
  - Existing biomedical literature
  - Gene transcription patterns
  - Protein translation/activity
- With growing databases containing data, this becomes a problem in the realm of bioinformatics.

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!”
Genomic vs genetic

<table>
<thead>
<tr>
<th>Genetic Medicine</th>
<th>Genomic Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low frequency of ~1000 of usually high penetrance genes</td>
<td>The genetic risk for common diseases will often be due to disease-producing alleles with relatively high frequencies (&gt;1%). All genes may be disease causing.</td>
</tr>
<tr>
<td>1000's of 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>Mostly assessed indirectly &amp; focused on single genes</td>
<td>Most experimental access to the entire genome</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
Determining the aforementioned interaction is hard and runs counter to traditional biostatistical techniques.

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

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

Threshold $r^2$ was 0.8.
- 202 networks
- 834 features out of 11,692 (7.1%)
- 1,222 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 (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-(phenylethyl)-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 tumorigenicity 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

Overview

- The future is now
- 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.
  - \( H^2 = V_G / V_P \), 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
Calories from Fat Among US Adults


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.
- We need to capture the effect of the genotype and environmental effects on the phenotype.
- These two comprise
  - History
    - Physical
    - Laboratory Studies
    - Imaging
The New Medicine
A More Expansive Reductionism

- More to the state description than genome
- Given necessity to capture both environment, genomic state and their interaction.
- Only then we can elucidate \( V_E \) and \( V_G \) and \( V_{EG} \)
- Required for effective new therapies
- Required for deeper understanding of mechanism
- Requires capturing the aforementioned interactions
- The less we capture, the more undetermined the system

Overview

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

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

\[ P_h = \frac{P_h P_{\text{sel}(1)} P_{\text{sel}(1)}}{\sum P_h P_{\text{sel}(1)} P_{\text{sel}(1)}} \]

- Performing \( i \) of \( m \) possible tests.
- If every test has \( r \) possible results, then there will be \( rm^r \) possible test histories after \( k \) 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

The scope of the CPDB Group is as follows:

1. **Cystic Fibrosis**
   - The burden of cystic fibrosis.
   - Genetics.
   - Genetics of diabetes mellitus.
   - Genetics of obesity.
   - Genetics of pyloric stenosis.
   - Genetics of rhizomelic dysplasia.
   - Genetics of short stature.
   - Genetics of very low birth weight.
   - Genetics of Wilson’s disease.

2. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

3. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

4. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

5. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

6. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

7. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

8. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

9. **Cystic Fibrosis and Genetic Disorders**
   - Genetics of cystic fibrosis.
   - Genetics of other genetic disorders.
   - Genetics of other genetic disorders.

10. **Cystic Fibrosis and Genetic Disorders**
    - Genetics of cystic fibrosis.
    - Genetics of other genetic disorders.
    - Genetics of other genetic disorders.

**What kind of testing are doctors ordering?**

- Just looking at Cancer Susceptibility Tests (CST)
  - 1251 physicians (820 primary care)
  - In 12 months, approx 30% ordered or referred (only 7% directly ordered)
- Factors affecting ordering:

Northeast, feeling competent, advertising materials, and most importantly having the patient ask for it.
A simple new test could save your baby’s life. The test describes a newborn screening kit that can detect more disorders than most state screening programs: for your baby and for your peace of mind.18

- An advertisement for diagnostic testing, appearing in a Jewish magazine, once the headline shows a photograph of a weary, dark-haired woman. “If you’re Jewish, it could be Gaucher disease.”14

- An advertisement for carrier testing in a Jewish community newspaper asks its readers, “Are you a carrier?” and lists 8 “Jewish genetic conditions” for which the audience may be at risk.
Current trends in criminal inventory and state screening

The privacy challenge is now

How many variants define a person uniquely?

The men of Wee Waa

Tension between research and privacy

Altman, Science 2004

Insufficient for future genomic research

Needed to find genetic relationships

Insufficient for privacy protection

Trade-offs between SNPs and privacy.
History

- The modern history of medical controlled vocabularies begins with the U.S. Army General Surgeon who petitioned Congress to fund a medical library.
- The position eventually became "The US Surgeon General" and the library the National Library of Medicine
  ✓ http://www.nlm.nih.gov/

History

- Library collection was indexed with Index Medicus (created by NLM) which is published in book form.
- Index Medicus was extended to index medical literature articles.
- Index Medicus was extended further to provide online indexing (1960). This became the Medical Subject Headings (MeSH).
- Exponential growth has continued for a century, well past the capacity of paper records
- Today: 15 million citations for biomedical articles back to the 1950's.

As the molecular aspects of disease became clear

- 15 million citations for biomedical articles back to the 1950's.
- 1998, National Center for Biotechnology Information (NCBI) formed as a complement to NLM.
- Databases and standardized nomenclatures.
- One NCBI database is GenBank, the nucleic acid sequence database that contains sequence information from more than 100,000 different organisms
- Now constructed with the understanding of the fundamentally interwoven character of biological information.
  ✓ Closes we have to a unified thesaurus if not a unified biological theory.
Your assignment

● A 60 year old man has been found face down, and dehydrated. You are one of a team of assistants working with a the genomic diagnostician. He gives you the following sequence.

AATLGQSTELKFLNASSILINSLADVRGWALV

Your assignment is to determine if this sequence might inform her about the cause of the presentation.

● In answering the question:
  ✓ What part of the genome is this?
  ✓ What diseases are associated with it?
  ✓ Is it an unusual variant?
  ✓ In which other organisms is this kind of sequence found?
  ✓ Are there model organisms to study this variant?

● Document how you arrive at these answers.