Extracting Medical Knowledge from High Throughput Genomics

Ted Goldstein

ted@soe.ucsc.edu
1. What is the problem and Why should you care?

2. What do Computer Scientists need to know about Biology?

3. What do Biomedical Scientists need to know about Big Data?

4. How do we transform Genomic Data into Knowledge?

5. How do we transform Knowledge into Medical Intervention?
1. What is the problem and Why should you care?
What is the problem?

US Center for Disease Control Mortality Index

2010

<table>
<thead>
<tr>
<th>Condition</th>
<th>Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart disease</td>
<td>616067</td>
</tr>
<tr>
<td>Cancer</td>
<td>562875</td>
</tr>
<tr>
<td>Stroke</td>
<td>135952</td>
</tr>
<tr>
<td>Chronic respiratory</td>
<td>127924</td>
</tr>
<tr>
<td>Unintentional injuries</td>
<td>123706</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>74632</td>
</tr>
<tr>
<td>Diabetes</td>
<td>71382</td>
</tr>
<tr>
<td>Influenza &amp; Pneumonia</td>
<td>52717</td>
</tr>
</tbody>
</table>

http://www.cdc.gov/nchs/fastats/deaths.htm
## 2011 Estimated Cancer Rate (AACR)

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Incidence</th>
<th>Death</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Esophagus</td>
<td>16,980</td>
<td>14,710</td>
<td>87%</td>
</tr>
<tr>
<td>Pancreas</td>
<td>44,030</td>
<td>37,660</td>
<td>86%</td>
</tr>
<tr>
<td>Liver &amp; Bile duct</td>
<td>26,190</td>
<td>19,590</td>
<td>75%</td>
</tr>
<tr>
<td>Lung &amp; Bronchus</td>
<td>221,130</td>
<td>156,940</td>
<td>71%</td>
</tr>
<tr>
<td>Ovary</td>
<td>21,990</td>
<td>15,460</td>
<td>70%</td>
</tr>
<tr>
<td>Nervous System</td>
<td>22,340</td>
<td>13,110</td>
<td>59%</td>
</tr>
<tr>
<td>Myeloma</td>
<td>20,520</td>
<td>10,610</td>
<td>52%</td>
</tr>
<tr>
<td>Leukemia</td>
<td>44,600</td>
<td>21,780</td>
<td>49%</td>
</tr>
<tr>
<td>Colorectal</td>
<td>101,340</td>
<td>49,380</td>
<td>49%</td>
</tr>
<tr>
<td>Stomach</td>
<td>21,520</td>
<td>10,340</td>
<td>48%</td>
</tr>
<tr>
<td>Cervix Uterus</td>
<td>12,710</td>
<td>4,290</td>
<td>34%</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
<td>66,360</td>
<td>19,320</td>
<td>29%</td>
</tr>
<tr>
<td>Urinary Bladder</td>
<td>69,250</td>
<td>14,990</td>
<td>22%</td>
</tr>
<tr>
<td>Kidney &amp; Renal Pelvis</td>
<td>60,920</td>
<td>13,120</td>
<td>22%</td>
</tr>
<tr>
<td>Oral</td>
<td>39,400</td>
<td>7,900</td>
<td>20%</td>
</tr>
<tr>
<td>Breast</td>
<td>232,620</td>
<td>39,970</td>
<td>17%</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
<td>8,830</td>
<td>1,300</td>
<td>15%</td>
</tr>
<tr>
<td>Prostate</td>
<td>240,890</td>
<td>33,720</td>
<td>14%</td>
</tr>
<tr>
<td>Melanoma of the skin</td>
<td>70,230</td>
<td>8,790</td>
<td>13%</td>
</tr>
<tr>
<td><strong>All Malignant Cancers</strong></td>
<td><strong>1,596,670</strong></td>
<td><strong>571,950</strong></td>
<td><strong>36%</strong></td>
</tr>
</tbody>
</table>
Interlocking Problems

- surgery
- chemotherapy
- radiation
- clinical trials
- experimental methods
- alternative therapies
2. What do Computer Scientists need to know about Cancer Biology?
Central Dogma of Biology
Information is Represented as Molecules

- Adenine
- Thymine
- Guanine
- Cytosine
- Sugar phosphate backbone

U.S. National Library of Medicine
A multistep model for the development of colorectal cancer

predispositions to cancer that run in some families

*APC* is mutated in about 60% of colorectal cancers.

one mutant *BRCA1* allele (a tumor suppressor gene) has a 60% probability of developing breast cancer before age 50

**Cancer Biology for Computer Scientists**

Biology, Seventh Edition

*Neil Campbell and Jane Reece*

Peerson 2007
Sometimes it's like debugging, a single mutation, like BRAF V600 in metastatic melanoma, targeted drugs can give spectacular results.
It's not like debugging, the bugs adapt

Before initiation of vemurafenib

15 weeks on vemurafenib

23 weeks after therapy

Roche
Mechanisms of Drug Resistance
Mutations Drive Cancer
A mixture of clones within a cancer tissue sample results from somatic selection of subclones.

Initiating ‘driver’ event

‘driver’ events

Last clonal ‘driver’ events

‘passenger’ events

Pathways not Genes

Genomes are the key to the future of cancer treatment
3. What do Biomedical Scientists need to know about Big Data?
The Cancer Genome Atlas: 10,000 tumors from 20 adult cancers

TCGA CENTERS
BC Cancer Research Center
Fred Hutchinson Cancer Research Center
Complete Genomics Inc.
Pacific NW National Laboratory
University of Southern California
Oregon Health & Science University
Institute for Systems Biology
University of California, Santa Cruz

TCGA CENTERS
Boise State University

TCGA CENTERS
Brigham & Women’s Hospital and Harvard Medical School
Broad Institute
John Hopkins University
Memorial Sloan-Kettering Cancer Center

TCGA CENTERS
Nationwide Children’s Hospital

TCGA CENTERS
Vanderbilt University
Washington University Genome Institute

TCGA CENTERS
University of North Carolina

TCGA Centers:
- Biospecimen Core Resource
- Genome Characterization Centers (GCCs)
- Genome Sequencing Centers (GSCs)
- Proteome Characterization Centers (PCCs)
- Data Coordination Center (DCC)
- Genome Data Analysis Centers (GDACs)
The Dawn of Personal Genomics

- The sequencing cost for the first human genome was about $300M.
- The cost is now headed toward $1000 per genome, dramatically outpacing the Moore’s Law rate for the decreasing cost of computer processing capacity.
Cancer Genomics Hub (CGHub)

- Designed for 25,000 cases with average of 200 gigabytes per case
- 5 petabytes ($5 \times 10^{15}$) total, scalable to 20 petabytes
- General Parallel File System, Dual RAID 6 subsystems, Redundant I/O paths
- Currently holds 21,000 files from ~5,000 cases, ~60 gigabytes/case
TCGA System Overview

1. Participant
   - BioSpecimens & Clinical Pathology Data in caBIG Compliant Format
   - Tissue Source Sites (TSSs)

2. Biospecimen Management
   - BioSpecimens & Anonymized Metadata
   - Secure Participant Data
   - Biospecimen Core Resource (BCR) x 2

3. Genome Characterization Tools
   - Send aliquots (DNA/RNA) + UUIDs
   - Genome Characterization Centers (GCCs) x 5

4. Genome Sequencing Tools
   - Send aliquots (DNA) + UUIDs
   - Genome Sequencing Centers (GSCs) x 9

5. Data Coordinating Centers (DCCs)
   - Create New Sample (Specimen metadata w/ UUID)
   - Store SDRF Metadata for UUIDs
   - Get Sample UUIDs
   - Push Sequence Data Catalog

6. TCGA Data Portal Web Service
   - National Institutes of Health (NIH)
   - SRA Sequence Read Archive (Replaced by CGHub)
   - NCBI User Authentication
   - NCBI Sequence Accession

7. CGHub Application Processors
   - Anmai
   - CGHub Admin
   - Sequence Data Management
   - CGHub Storage Controllers
     - Sequence Data Repository

8. CGHub Storage Controllers
   - Storage Repository Data Center
     - Operated by UCSC in San Diego, CA
   - CGHub Systems
   - Human UI

Legend
- Control & Metadata Path
- High-Speed File Transfer
- Sample / Sequence Data Flow
- Server Process
- Workstation
- Web Application
- Other TCGA Systems
CGHub Topology
High Throughput Genomic Technologies

- **Gene expression profiling**
  - Monitoring expression levels for thousands of genes simultaneously used to:
    - **Array CGH (Comparative genomic hybridization)**
      - Assessing genome content in different cells or closely related organisms.
    - **SNP array (single nucleotide polymorphism)**
      - Identifying single nucleotide polymorphism among alleles within or between populations.
    - **ChIP-on-chip (Chromatin immunoprecipitation)**
      - Determining protein binding site occupancy throughout the genome.

- **DNA Sequencing**
  - Base Pair Sequencing to find mutations and alterations
  - Conversion from RNA to DNA for Gene Expressions
  - Cytosine conversion to determine
Old Technology:
Biological Samples in 2D Arrays on Membranes or Glass Slides
Microarrays: Universal Biochemistry Platforms

- Peptides
- Proteins
- DNA
- Lipids
- Carbohydrates
- Small molecules
New Technology
Whole Genome Shotgun Sequencing

3.2 billion base pairs

thousands of segments of 200,000 base pairs each

millions of reads of 500 base pairs

3 billion base pairs

Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

Shotgun sequence

Assembly
4. How do we transform Genomic Data into Knowledge?
Detecting duplications

Small Tandem Duplication

chr2 : 169,227,553

chr2 : 169,210,871

Tandem Duplication Size = 16,682 bp

Zack Sanborn
Pathways
Super Pathway
The Data Deluge

Structural Variations

Copy Num Alterations

Exome Sequences

Methylation

2^n combos
Integration Approach: Detailed models of gene expression and interaction
Integration Approach: Detailed models of expression and interaction

Two Parts:

1. Gene Level Model (central dogma)

2. Interaction Model (regulation)

Josh Stuart Group
PARDIGM Gene Model to Integrate Data

1. Central Dogma-Like Gene Model of Activity

2. Interactions that connect to specific points in gene regulation and signal cascade map

Vaske et al. 2010. *Bioinformatics*
Integrated Pathway Analysis for Cancer

Cohort → Multimodal Data → Pathway Model of Cancer → Inferred Activities

- CNV
- mRNA
- meth

Josh Stuart Group
Data Analytic Strategy

• Turn masses of data into easily “compared signatures”
• Draw inferences from similar and distinct signatures
• Cluster signatures to find patterns
• Prevent bias and multiple test false discovery
Basic Statistic is 2 group \( t \)-test

- At a \( p < 0.05 \), there are 300 genes up and 200 genes down regulated
  - 95% confidence that the means of these genes in the two groups is different
- At a \( p < 0.05 \), \( x \) genes up and \( y \) genes down with a fold change of at least 3.0
OCCAM Pipeline
Digital Pathway Signature Correlation

- Differential Pathway Signature Correlation (DiPSC) calculates the correlation between our signatures computed using our PARADIGM pathway methods.
- Simple Vector Correlation

\[
\text{corr}(x, y) = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_xs_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2},
\]

- Cluster using Euclidean distance

\[
\|a - b\|_2 = \sqrt{\sum_{i}(a_i - b_i)^2}
\]
Pathway signatures reveal connections between mutations and clinical outcomes.

Key cancer pathway components altered in luminal breast tumours.

doi:10.1038/nature11143
Controlling for Multiple Hypothesis

- In lieu of analytical statistics we use brute force (Monte Carlo) computations to generate huge numbers of synthetic or fake samples

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Permutation</strong></td>
<td>Samples are drawn at random from original pool without replacement</td>
</tr>
<tr>
<td><strong>Bootstrap</strong></td>
<td>Samples are drawn with replacement</td>
</tr>
<tr>
<td><strong>Jack Knife</strong></td>
<td>Samples consist of original pool with some withheld</td>
</tr>
</tbody>
</table>
Controlling for Multiple Hypothesis

- Bootstrap resample cohorts to create $k=1000$ sub-cohorts
- Provides accurate estimates of correlation $\rho$, and confidence intervals

$$R_{s,t} = \frac{k}{\sum_{i=1}^{k} \text{corr}(p_s, p_t)} \quad p_s, \sim \text{draw}(0.5), \quad p_t, \sim \text{draw}(0.5)$$
5. How do we transform Knowledge into Medical Intervention?
**Hypothesis:** Tumor addicted to PIK3CA pathway

**Previous:**
- **2010** Patient enrolled in AZ197 clinical trial targeting VEGF; dropped after radiographic evidence that tumor was non-responsive
- **2001** Patient was treated with lumpectomy

**MUTATION SUMMARY**
- HER2: negative
- ER: negative
- PR: negative
- BRAF (V600): negative
- NRAS: negative
- KRAS: negative
- PIK3CA: exon 9: positive, exon 20: negative

**Imaging**
- [Imaging results]

**Mt. Joseph Hospital**
- Patient in waiting room
- B. Bryers, M.D.
- (Oct. 10, 12:45 PST)
  - PIK3CA mutation may indicate use of AKT 1-2 inhibitor+ more
Tumor addicted to PIK3CA pathway

**Re 239471 (Oct. 10, 12:25 PST)**

PIK3CA mutation may indicate use of AKT 1-2 inhibitor as referenced heatmap shows patient current pathways correlate (r=0.75, p-Value 0.03) with 14 other patients with PIK3CA mutations who have been successfully treated with **more**
Laura Choi, Ph.D.

**Academic Scientist’s view**

**Patient 239471**
41 year old ductal carcinoma

Barry Catz, M.D.
Oncologist, MJH

Lauren Bryer, M.D.
MDCC

**Academic researcher:**
Focus is on scientific questions, and developing platforms for diagnosis and therapy

**COHORT ACTIVITY**
- SU2C BRCA
- TCGA COLOREAD
- TCGA OVARIAN
- MDA GBM

**APPS**
- OncoPrint
- Paradigm Pathway Analysis

**TCGA BRCA "Tier 1" genes** including PTEN, CDKN2A mutation status, expression and CNV for days_to_tumor recurrence, days_to_tumor progression, + more

**Virtual Study**
Cohort Complete
(Oct. 10, 12:45 PST)
22 new patients have donated their data and the cohort is now ready for analysis + more

© Copyright 2011. All rights reserved.

PATIENT | DOCTOR | RESEARCHER | BIOLOGIST
Favorites | Locator | Apps | Dialog Alerts | About | Contact

STAND UP TO CANCER
Conclusions

Make personalized medicine a reality by using social networking technology to tie together the patient with their care givers

**Patients** gain is visibility into the process and possibly an improved outcome by making their data available and getting many minds working on their problems

**Physicians / Community Oncologists**, It closes the gap to apply genomic knowledge to their care and be driven by expert curated guidelines; gains access to the most up to date clinical guidelines

**Clinical Researcher Physicians / Academic Researchers** gain access to larger cohorts and allows them to consult to community

**Bioinformatic App Developers** get a market for their development
Remarks for XLDB

• Software
  • Basic UNIX command line with some web tools
  • Compute bound because of null permutations
  • No Database: all data is stored in flat files (CSVs/TSVs)
  • Typical computations are done on 1000 node compute clusters with 64 Gb memory
  • R, Perl, Python, Java, C++,
  • Mix open source and proprietary

• Metadata and version management is a nightmare
  • Not taking advantage of readonly nature of the data
  • No reuse of previously done large operations
  • No way to store provenance of the computation
Requirements for an XLDB for Bioinformatic Medical Systems

Both an OLTP and OLAP system

Bulk of Data is a hyper-cube, but keyed by a multi-graph

Ability to slice and dice the data across many dimensions

Workflow and conclusions must be traceable, re-playable, auditable and sharable

Needs to closely tie to publishing and new forms of reputation systems