Cancer Is Not a Single Disease

- Carcinoma
  - Malignant tumor derived from epithelial cells
- Sarcoma
  - Malignant tumor derived from connective tissue (mesenchymal cells)
- Lymphoma and Leukemia
  - Malignant hematopoietic (blood) cells
- Germ cell tumor
  - Malignant totipotent (zygote) cells
- Blastosma
  - Malignant precursor (blast) cells

The American “War on Cancer”

- When did it begin?
  - 1971
- Who started it?
  - Richard Nixon, State of the Union Address
- Concepts for implementation
  - Coordinated effort (Sidney Farber, Boston MD)
  - Loose coordination (Major hospitals)
- How did we proceed?
  - Fragmented efforts
- National Cancer Institute’s 2009 budget?
  - $4.97B, +3% over 2008
  - 5,179 grants
  - 16% of funds to intramural research

Are We Winning?

- ACS estimates of 2009 cancer deaths
  - Men: 292,540
  - Women: 269,800

- ACS estimates of 2009 cancer deaths

- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009
- 2010

- 50,000
- 100,000
- 150,000
- 200,000
- 250,000
- 300,000
- 350,000
- 400,000

- 1930
- 1940
- 1950
- 1960
- 1970
- 1980
- 1990
- 2000

- Men
- Women

- American Cancer Society
- Number of Cancer Deaths

- Beginning of the "War"

- Fortune
  - Mar 22, 2004

- New York Times
  - Oct 29, 2010

Cancer Trends, 1998-2007
NCI Surveillance Epidemiology and End Results

Cancer Correlations

What’s Needed?

- Cancer (“Arms of a crab”) identified by Hippocrates, c. 400BC
- Advances in Science*: Peaceful interludes punctuated by intellectually violent revolutions
- Propositions are true or false within the context of a paradigm**
- Paradigm Shifts
  - Old world view is inadequate
  - Need to create new perspectives
- Conquering cancer requires new paradigms
  - Dynamics of metabolism, mutation, and metastasis
  - Systems approach

* Kuhn, The Structure of Scientific Revolutions, 1962
** Gödel, Incompleteness Theorems, 1931
**Time and Spatial Scales of Processes Related to Cancer**

- **Protein Production**
  - Seconds-minutes; nm
- **Cell Cycle**
  - Hours-months-lifetime; mm
- **Invasive (Primary) Cancer**
  - Years to detection; cm
- **Metastasis (Secondary Cancer)**
  - Years to development; cm
  - However, rapid growth upon onset

---

**The Central Dogma: Core Process of Protein Production**

- **DNA**
  - Transcription
- **mRNA**
  - Translation
- **Protein**
  - 4 Nucleotides (A, T, C, G)
  - 4 Nucleotides (U, A, C, G)
  - 20 Amino Acids (A, C, D, E, ...)

- **Gene**: a sequence of DNA nucleotides (nt) or base pairs
- **Transcript**: cell-specific messenger RNA (mRNA) sequence containing the gene’s code
  - Information coded in 64 nt triplets (codons)
  - with open reading frames defined by Start and Stop codons
- **Protein**: amino acid sequence coded by transcript

---

**System View of The Core Process**

- **Traditional view**: Primary emphasis on products

  ![Traditional View Diagram](image)

- **System view**: Primary emphasis on process

  ![System View Diagram](image)

**Protein Production is More Complex than the Central Dogma Implies**

- ![Protein Production Diagram](image)
Variability in the Protein Process

- Many RNA sequences could code for one protein
  - Redundancy: \((64 - 3) / 20 \approx 3\) codons per amino acid
- Many proteins can originate from one DNA gene
  - Millions of proteins, but only \(\sim 22,000\) genes
  - Alternative splicing
  - RNA editing
  - Gene silencing
  - Other epigenetic effects
- Point and gross mutations alter DNA/RNA sequence and coding frames
  - Replacement
  - Insertion and deletion
- Reverse transcription: DNA from RNA due to retroviruses (v-oncogenes)

The process is stochastic, not deterministic

Example: p53 Pathways

- p53 is an important tumor suppressor gene (Levine et al., 1979, Vogelstein, 1989)
- Mutations and epigenetic effects that inhibit p53 are involved in over half of all cancers

Levine & Oren, 2009

Can a Biologist Fix a Radio? - Or, What I Learned While Studying Apoptosis*

Yuri Lazebnik, Professor, Cold Spring Harbor Laboratory, Cancer Cell, 2, 179-182, Sept 2002

Qualitative Descriptions of Pathways are Imprecise and Inadequate

- Plus many other details:
  - P53 transcription factors
  - pre-mRNA processing proteins
  - post-translational regulation of Mdm2
  - Downstream effects of TP53
  - TP53 interaction with co-activators
  - siRNA effects
  - mutated derivatives
  - DNA
  - RNA
  - stressors that induce gene expression

- Mdm2: murine (mouse) double minute oncogene
- Hdm2: Human homologue
Pathways Must Be Converted to “Circuits”

Hanahan and Weinberg, 2000

Math Model of a Dynamic Process

- Differential or difference equation
- Dynamic Process
  - $x(t) = x(t)$
  - $u = input$
  - $w = exogenous disturbance$
  - $p = parameter$
  - $k = time or event index$
- Observation Process
  - $y(t) = y(t)$
  - $z = measurement$
  - $n = measurement error$

- Stability and transient behavior of a dynamic process
  - determined primarily by structure and parameters
  - secondarily by inputs

Why We Need Dynamic Models

- Estimating measured and hidden variables
  - $\hat{y}(t) = h(x(t), u(t))$
  - $n(t)$

- Estimating parameters of the process
  - $\hat{p}(t)$

- Providing optimal therapy

Structure of a Dynamic System

- Dynamic Process Vectors: Current state depends on prior state
  - $x = dynamic state$
  - $u = input (e.g., treatment)$
  - $w = exogenous disturbance (e.g., infection, radiation)$
  - $p = parameter$

- Observation Process Vectors: Measurement may contain error or be incomplete
  - $y = output (error-free)$
  - $z = measurement$
  - $n = measurement error$
The Cell Cycle

**System View of the Cell Cycle**

- Growth/Inhibitory Factors
  - Nutrients, Environmental Factors
  - Metabolic Ribosomal Proteins
  - DNA Repair
  - Cell Division
- **G2 Checkpoint**
  - Arrest (senescence/apoptosis)
- **G2 Post-DNA Synthesis**
  - Nutrients, Environmental Factors
- **DNA Synthesis**
- **G2 Pre-DNA Synthesis**
  - Metabolic Ribosomal Proteins
  - DNA Repair
- **G2 Quiescent Phase**
  - Exit Cell Cycle (senescence)

**Chronological Progression of Epithelial Cancer**

- 4-6 rate-limiting or mutation events to yield a tumor cell
  - Genetically altered cell (1 cell)
  - Hyperplasia (~ million cells) (~\(2^{20}\))
  - Dysplasia (~ billion cells) (~\(2^{30}\))
  - In situ cancer (~ 20 billion cells) (~\(2^{34}\))
  - Invasive cancer
  - Metastasis

- Heterogeneous zones of tumor: cancer, stromal, immune cells
- Involvement of stem cells
- Tumors as organs/complex tissues
- Normal cell populations are well-regulated (birth/death ≈ 1)
- Cancer cell populations are growth-oriented (birth/death > 1)

* But not necessarily causal

**mRNA Expression During Cell Cycle is Varied**

- **mRNA for Export Proteins**
- **mRNA for Proteins Involved in Metabolism**
- **mRNA for Ribosomal Proteins and DNA**
- **mRNA for Proteins Involved in Duplication of DNA**

**Stage I**

**Stage II**

*But not necessarily causal*
Progression from Invasive Cancer to Metastasis

- **Traditional view**: Metastasis is a late-stage process, selective but not genetically distinct
- **Alternative view**: Early-stage cells may possess metastatic phenotype
- ~90% of cancer deaths due to metastasis

Coupled Structural Models of Cancer Dynamics

Classical View of Metastasis

- Neo-vascularization around primary tumor
- Lymphatic angiogenesis
- Change in cell adhesion
- Change in anchorage-dependence
- Generation of proteins that enhance motility
- Induction of leaky membranes
- Invasion of host stroma
- Survival of cancer cells in circulation
- Arrest in small blood vessels
- Extravasation at secondary site
- Proliferation of cancer cells

Fidler, 2003

Lung and Liver Metastases

- Lung
  - MRI-CT-Bioluminescence
- Liver
  - Autopsy
Cancer Stem Cells?

3 Possible Paths
- Self-renewal
- Stem-cell niche
- Differentiation

Cell Fusion
- Normal and Abnormal Transdifferentiation
- Immune response is a dynamic process triggered by external/abnormal factors
- Initial infection
- Spread of pathogen
- Immune response

Involvement of the Immune System
- Significance of mutation in overcoming body’s defenses
- Cancers may result from
  - Viral infection
  - Immune deficiency
- Evidence for immunosurveillance/editing
- Macrophages and APC as fellow travelers
- Potential for treatment by immunotherapy

Two Hypotheses of Metastasis
- Late Dissemination of Metastasis
  - Parallel Progression of Primary Tumor and Metastasis

Therapy Ultimately Must Control or Prevent Metastasis
- Limited benign growth
- Disseminated malignancies
- Thrombi that target cancer stem cells
- Cancer stem cell
- Tumorigenesis
- Tumour growth
- Metastasis
Cancer Therapy

- **Primary Cancer**
  - Surgery
  - Radiotherapy
  - Chemotherapy
  - Biochemotherapy
  - Immunotherapy
- **Localized treatment**
  - Focus on malignancy
- **Systemic treatment**
  - Expose all cells to therapy, including normal cells
  - Inhibit cell division
  - Block cell receptors

- **Secondary Cancer (Metastasis)**
  - Chemotherapy
  - Radiotherapy
  - Biochemotherapy
  - Immunotherapy

Cancer Detection

- **Visual examination**
  - General medical exam by MD
  - History
  - Biopsy
  - Microscope inspection of sample tissue by pathologist
  - BioMEMS enhancement
  - Colonoscopy/Endoscopy
- **Physical examination**
  - Palpation
- **Biochemical examination**
  - Blood (PSA, electrophoresis, Western blot...)
  - Stool and urine

- **Molecular profiling (pathology) of sample tissue**
  - DNA (e.g., PCR, Southern blot)
  - RNA (e.g., microarray, Northern blot)
  - Single-Nucleotide Polymorphisms (SNPs)

- **Internal medical imaging**
  - Ultrasound (diagnostic sonography)
  - Magnetic resonance imaging (MRI)
  - X-ray, Computer-aided tomography (CAT)
  - Positron emission tomography (PET)

Microscopic Tissue Samples for Detecting Cancer*

- **Sampling**
  - of a pattern
  - of an evolutionary trend
  - of a random ensemble

Virtually All Biological Measurements are Snapshots

- **a. Normal colonic mucosa**
- **b. Well-differentiated tubular adenocarcinoma**
- **c. Poorly-differentiated adenocarcinoma**
- **d. Mucinous adenocarcinoma**

- Experienced pathologists agree on visual microscopic sample classification ~60% of the time

**DNA Microarray**

- Search for expressed genes (i.e., RNA transcripts)
- ~22,200 wild-type transcripts represented on Affymetrix HU133A
- Photolithography deposits known 25-mer sequences (oligonucleotides) at known locations (features) on chip
- ~500,000 features / chip
- 10-20 features describe each transcript’s “probe set”
- ~10^6 nt strands for each feature

**DNA/SNP Microarray Application**

- cDNA produced from sample RNA, labeled with fluorescent tag, and hybridized to the array
- Array is washed, stained, scanned, and quantified

**Goals of DNA/RNA Microarray Analysis**

- Identify genes that are central to biological function
- Find correlations between RNA expression and disease states
- Identify targets for gene therapy
- Screen targets for further analysis
- Define the genetic signatures of cancer

**Paradigm for Microarray Analysis:**

**Wild-type mRNA Transcript Expression Level Infers Function**

- Up-regulation of mRNA transcripts in **tumor cells**
  - Causal input
  - Defensive response
  - Bystander effect
  - Tissue effect
  - Artifact
- Down-regulation of mRNA transcripts in **tumor cells**
  - Present, but **mutated** (and, therefore, not detected)
  - Eliminated or suppressed by tumor growth
  - Tissue effect
  - Artifact
Curse of dimensionality (too many genes = “large p problem”)  
Curse of sparcity (too few samples = “small n problem”)  

Objectives for Classifying Data  
- Class comparison  
  - Identify feature sets (e.g., RNA expression profiles) for predefined classes  
- Class prediction  
  - Develop algorithms that predict class membership for a novel expression profile  
- Class discovery  
  - Identify significant new classes, sub-classes, or features  

Typical Pairs of Colon Cancer Microarray Expression Levels  
- Samples not well differentiated in individual transcript clusters (overlapping, non-separable sets)  

Supervised and Unsupervised Learning of Discriminants  
- Learning depends on “closeness” of related features  
- Previously unknown correlations or features are detected  
- Classification occurs after clustering via exogenous knowledge  
- Same answer (= clusters) given for all questions  
- Learning depends on prior definition and knowledge of class  
- Complex correlation between features revealed  
- Classification is inherent in learning  
- Different answers given for different questions
Characteristics of Classification Features

- Strong feature
  - Individual feature provides good classification
  - Minimal overlap of feature values in each class
  - Significant difference in class mean values, $m_A$ and $m_B$
  - Low variance in each class ($\sigma_A^2$, $\sigma_B^2$) is desirable

Example of Transcript-by-Transcript Colon Tumor/Normal Classification by $t$ Test
(data from Alon, Notterman et al, 1999)

- In example, 1,151 probe sets* are over/under-expressed in tumor/normal comparison, $p \leq 0.003$
- Genetically dissimilar samples are apparent
- $t$ test eliminates bias errors

* Each probe set represents one mRNA transcript

Correlation Matrices

- Transcript correlation
  - Identifies possible components of biological circuits

- Sample tissue correlation
  - Confirms or questions the classification of samples

Ensemble Mean Values

- Treat each probe set ($i^{th}$ row) as a redundant, corrupted measurement of the same tumor/normal indicator, $y$
  - $z_{ij}$: Transcript expression for $i^{th}$ transcript and $j^{th}$ sample
  - $\hat{z}_{ij}$: Error in transcript expression for $i^{th}$ transcript and $j^{th}$ sample
  \[
  z_{ij} = k_i y + e_{ij}, \quad i = 1, m, \quad j = 1, n
  \]

- Compute $j^{th}$ column averages, $\hat{z}_j$, for each sample sub-group, [$A(\cdot)$, $B(\cdot)$, $A(-\cdot)$, $B(-\cdot)$]
  \[
  \hat{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij} = \frac{1}{n} \sum_{i=1}^{n} \left[ k_i y + e_{ij} \right] \quad n \to \infty \rightarrow \left( \hat{k}_j \right)
  \]

- Averaging attenuates the effect of individual random (zero mean) errors
- Scale factor, $k$, can be eliminated by normalization
Class Prediction and Evaluation Using Ensemble Mean Values

- Single-feature prediction
  - \([A(\text{+}), B(\text{+})]\): Cancer-positive transcripts: 2 errors
  - \([A(-), B(-)]\): Cancer-negative transcripts: 1 error
  - \([A(\text{+}), A(-)]\) vs. \([B(\text{+}), B(-)]\)
    Two-feature prediction (cross-plot): 1 error
- Mislabling of samples in lab?

Tumor and normal tissue are heterogeneous.

Clustered of Sample Averages for Primary Colon Cancer vs. Normal Mucosa
[NCI Program Project Grant (PPG), 144-sample set]
- 144 samples, 3,437 probe sets analyzed
- 47 primary colon cancer
- 22 normal mucosa
- Affymetrix HGU-133A GeneChip
- All transcripts "Present" in all samples

Clustered of Sample Averages for Primary Polyp vs. Normal Mucosa
- 21 primary polyp
- 22 normal mucosa

# Probe sets
Up: 1014
Down: 219
Constant: 40
Clustering of Sample Averages for Primary Polyp vs. Primary Colon Cancer

260-Sample Colon Cancer Data Set
- Primary colon cancer (130)
- Primary polyp (31)
- Normal mucosa (24)
- Liver metastasis (33)
- Normal liver (12)
- Lung metastasis (20)
- Normal lung (5)
- Microadenoma (4)
- High grade dysplasia (1)

Artificial Neural Networks
- Sigmoidal neural network with supervised learning

9-Class Colon Cancer Neural Network
- Neural network
  - 18 ensemble-average inputs
  - various # of sigmoidal neurons
  - 9 linear neurons
  - 9 outputs
  - 12-48 transcripts in each ensemble average
9-Class Transcript Signatures by t Test
(260-Sample PPG Colon Cancer Set; 12 transcripts in each sub-class)
### Classification of Primary Colon Cancer According to Gender

- **Male** (61)
- **Female** (44)

<table>
<thead>
<tr>
<th>Transcripts in Training Set</th>
<th>Leave-One-Out</th>
<th>Neurons</th>
<th>Trials</th>
<th>Correct</th>
<th>% Correct</th>
<th>Neurons</th>
<th>Trials</th>
<th>Correct</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>210</td>
<td>168</td>
<td>80%</td>
<td>164</td>
<td>78%</td>
<td>12</td>
<td>210</td>
<td>178</td>
<td>85%</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>210</td>
<td>178</td>
<td>176</td>
<td>84%</td>
<td>12</td>
<td>210</td>
<td>178</td>
<td>85%</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>210</td>
<td>178</td>
<td>176</td>
<td>84%</td>
<td>24</td>
<td>210</td>
<td>184</td>
<td>88%</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>210</td>
<td>184</td>
<td>174</td>
<td>83%</td>
<td>24</td>
<td>210</td>
<td>184</td>
<td>88%</td>
</tr>
</tbody>
</table>

Overlap and Correlation of Primary Colon Cancer by Gender

- Male (yellow dot), female (green square)
- Significant overlap in ensemble averages
- Female (upper left), male (lower right)
- Sample distinction not strong
- Probe distinction is strong

### Classification of Primary Colon Cancer According to Age at Diagnosis

- 50 annotated samples

<table>
<thead>
<tr>
<th>Transcripts in Training Set</th>
<th>Leave-One-Out</th>
<th>Neurons</th>
<th>Trials</th>
<th>Correct</th>
<th>% Correct</th>
<th>Neurons</th>
<th>Trials</th>
<th>Correct</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminant = 50 yr</td>
<td></td>
<td>24</td>
<td>2</td>
<td>200</td>
<td>184</td>
<td>92%</td>
<td>180</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>4</td>
<td>400</td>
<td>368</td>
<td>92%</td>
<td>354</td>
<td>89%</td>
<td></td>
</tr>
<tr>
<td>Discriminant = 60 yr</td>
<td></td>
<td>24</td>
<td>2</td>
<td>400</td>
<td>372</td>
<td>93%</td>
<td>362</td>
<td>91%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>48</td>
<td>2</td>
<td>400</td>
<td>360</td>
<td>90%</td>
<td>360</td>
<td>90%</td>
<td></td>
</tr>
<tr>
<td>Discriminant = 70 yr</td>
<td></td>
<td>24</td>
<td>2</td>
<td>400</td>
<td>312</td>
<td>78%</td>
<td>305</td>
<td>76%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>2</td>
<td>400</td>
<td>336</td>
<td>84%</td>
<td>312</td>
<td>78%</td>
<td></td>
</tr>
</tbody>
</table>

Overlap and Correlation of Primary Colon Cancer According to Age at Diagnosis

- ≤ 50 yr (green)
- ≤ 60 yr (green)
- ≤ 70 yr (green)

- Implication that genetic distinction occurs in the 60s
Genetic Distinctions By Location of Primary Tumor

Conclusions

- Heterogeneity of cancer data
- Large \( p \), small \( n \) problem
  - Solution: Analyze ensemble rather than individual features of data set
- \textit{mRNA} microarray analysis
  - Does not identify mutant genes
    - However, absence of wild-type infers mutation
  - Does identify transcripts and proteins that are likely to be significant in "cancer circuits"
  - Infers presence of
    - Malignant cells
    - Fellow traveler cells
  - Ensemble Mean (\( t \) test)
  - Ensemble Covariance (correlation matrices)
  - Classification using Computational Neural Networks

Future Work

- Develop dynamic models (circuitry) for different cancer types
- Need for temporal, multi-scale data
- Identify "cancer sleeper cells"
- Clarify mechanisms for metastasis
- Improve \textit{mRNA} classification accuracy
- Formulate robust, optimal policies for primary and secondary cancer therapies
Classification of Data

- Data set characterized by two features

Data Clusters

- How many clusters?
Discriminants of Data

- Where are the boundaries between sets?

The Data Set Revealed

The discriminant is the Delaware River

Towns and Crossroads of Pennsylvania and New Jersey

Another Classification Example

- Party for returning graduate alumni/ae
- From this picture, who are the:
  - Grad alums
  - Current students
  - Spouses
  - Children
  - Visitors from abroad
  - Hosts
  - Oldest alums
  - Youngest alums
  - Party crashers?
- What features are used to classify?
Attributes That Could be Used to Define a Class

- Clinical data
  - Primary sample class (normal; tumor stage, location, ...)
  - Metastatic tumor class (organ; stage, location ...)
  - Secondary correlates (antigens, antibodies, health measures)
  - Time to recurrence / of survival

- Genetic data
  - Tumor suppressor/enhancer gene mutation (DNA or RNA)
  - Expression level of other RNA sequences
  - Translocation, chromosomal abnormalities
  - Chemical alteration of DNA or RNA

- Protein data
  - Tumor suppressor/enhancer protein mutation
  - Expression level of other cell-cycle proteins
  - Chemical alteration of proteins

Two-Sample *Student t Test* for Transcript Selection

- \( t \) compares mean values of two data sets
  - \( |t| \) is reduced by uncertainty in the data sets (\( \sigma \))
  - \( |t| \) is increased by number of points in the data sets (\( n \))

\[
t = \frac{(m_A - m_B)}{\sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}}}
\]

- \( m \) = mean value of data set
- \( \sigma \) = standard deviation of data set
- \( n \) = number of points in data set

- \(|t| > 3\), \( m_A \neq m_B \) with \( \geq 99.7\% \) confidence
  - [error \( p \leq 0.003 \) (Gaussian)] \( [N > 25] \)

Effect of Feature Set Dimension

- Best line or curve may classify with significant error
- Best plane or surface classifies with equal or less error

\( t \) test applied to mean tumor/normal expression levels of individual transcripts

... and What if the Data are Distorted?

- Grad alums
- Current students
- Spouses
- Children
- Visitors from abroad
- Hosts
- Oldest alums
- Youngest alums
- Party crashers
SNP Array

- Single nucleotide polymorphism: point mismatch between 2 alleles
  - ~10M identified in human genome
  - Evolutionary conservation within populations
- Array technology is similar to the DNA microarray
- Search for loss of heterozygosity and copy-number variation in DNA

High-Throughput Sequencing

- Sample DNA or RNA fragmented
- Fragments assembled and identified by various methods, e.g.,
  - Clonal amplification
  - Cyclic array sequencing

Chromosome “Hot” and “Cold” Spots Inferred by RNA Arrays

Chromosome “Hot” and “Cold” Spots Identified by SNP Arrays
Primary Colon Tumors

- Ascending
- Cecum
- Transverse
- Descending
- Sigmoid