Bayesian Computing and the Incorporation of Prior Knowledge in Translational-Genomic Modeling

Edward R. Dougherty

Department of Electrical and Computer Engineering
Center for Bioinformatics and Genomic Systems Engineering
Texas A&M University
Medicine at the Right Level

- Source of disease (ex: cancer) – molecular scale
  - Genes and proteins

Organs → Cells → Molecules
Complex Diseases

• Most diseases do not result from a single gene product.

• Complex diseases require complex personalized mathematical analysis.
Patient-Specific Treatment

- Specificity means much higher success rate.

Nano-aspirate/biopsy

Analyze

Compute

Custom drug manufacture

Enhanced disease management

Nano scale chemistry lab

http://gsp.tamu.edu
Translational Genomics

• Genomics is the study of genes as they interact in a system that governs cell behavior.

• Goals of translational genomics:
  – *Screen for key genes and gene families that explain specific cellular phenotypes (disease).*
  – *Use genomic signals to classify disease on a molecular level.*
  – *Mathematically model dynamical system behavior to derive therapeutic strategies to alter undesirable behavior.*
Central Dogma of Molecular Biology

DNA → Transcription → RNA → Translation → Protein
Gene Regulation

Gene regulatory controls

DNA damage

Hypoxia

E1A
Rb
E2F
Myc

p53

MDM2

Gene expression
the process by which gene products (proteins) are made

transcription

translation

protein

http://gsp.tamu.edu
Genomic Classification of Disease

- Abundance of RNA is measured for each gene (gene-expression microarray, RNA-Seq).
- A rule is used to train a classifier from the data.
Classification of Hereditary Breast Cancer

• Classifier discriminates types of breast cancer using two-gene signature.
• If treatment for BRCA1 and BRCA2 differ, then early detection is critical.
Glioma Application

- Data from four types of glioma: OL, GM, AO, AA
- Find small gene sets to separate each type from others.
- Small sample: 25 patients.
3-Gene Glioma Classification

- 3-gene linear discrimination for anaplastic oligodendroglioma from others.
A Huge Challenge

• Janet Woodcock (Director, Center for Drug Evaluation and Research, FDA): [As much as 75 percent of published biomarker associations are not replicable] “This poses a huge challenge for industry in biomarker identification and diagnostics development.”
Small Samples Don’t Work

• There are tens of thousands of genes and a small number of replicates, usually less than 100 – Big data can be very small data.

• If the sample is large (many replicates), then the data can be split into training (classifier design) and testing (error estimation).

• Small data sets cannot be split because there would be insufficient data for both training and testing.

• Vain hope train and test on the same data.
  – *This results in poor error estimation – not reproducible.*
Bayesian Classification

• Integrate prior (existing) biological with new data to design a classifier and estimate the error.
  – *If one had full knowledge of the system, one would derive the optimal classifier need no data.*
  – *Partial knowledge constraints the space of classifiers, thereby allowing more efficient use of the data.*

• Obstacles:
  – *Mathematically much more difficult.*
  – *Computationally much more difficult: involves high-dimensional Markov-chain-Monte-Carlo computational integrations and complex optimizations to incorporate prior knowledge.*
Growth Factor (GF) Signaling Pathways

- Biochemical pathways constrain the feature-label distribution.
- Key problem: Transform pathways into usable prior knowledge.
OBC for Gaussian Model

- Polynomial Optimal Bayesian Classifier (red line)
  - Dotted lines are level curves for the densities corresponding to the average means and covariance matrix.
  - Black solid line is linear classifier corresponding to the optimal classifier for the average mean and covariance matrix.
Control of Gene Networks

• The therapeutic problem is to model a gene regulatory network and then find an optimal treatment strategy.
  – Consider an external control variable and a cost function depending on desired outcome.
  – Minimize the cost function by a sequence of control actions over time – control policy (drugs).
  – Design optimal treatment regime to drive the system away from undesirable states.

• Problem 1: Infer network from data.
• Problem 2: Mathematically derive optimal controller.
A p53 Network

• Consider the DNA double strand break repair pathways involving the tumor suppressor gene p53.
• p53 is a master guardian gene tightly controlling various activities like cell cycle progression, senescence and apoptosis.
• Mutation in p53 is observed in 30% - 50% of common human cancers.
• We consider 4 genes: ATM, p53, Mdm2, Wip1.
ATM-p53-Mdm2-Wip1 Pathways

- dna_dsb refers to DNA damage.
Mutated Mammalian Cell Cycle PBN

- If CycD and Rb are simultaneously down-regulated, then the cell cycles in the absence of any growth factor.
- Intervention tries to stop simultaneous down regulation.

<table>
<thead>
<tr>
<th>Product</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CycD</td>
<td>Input</td>
</tr>
<tr>
<td>Rb</td>
<td>$\overline{\text{CycD}} \land \overline{\text{CycE}} \land \overline{\text{CycA}} \land \text{CycB}$</td>
</tr>
<tr>
<td>E2F</td>
<td>$\overline{\text{Rb}} \land \overline{\text{CycA}} \land \overline{\text{CycB}}$</td>
</tr>
<tr>
<td>CycE</td>
<td>$\overline{\text{E2F}} \land \overline{\text{Rb}}$</td>
</tr>
<tr>
<td>CycA</td>
<td>$(\overline{\text{E2F}} \land \overline{\text{Rb}} \land \overline{\text{Cdc20}} \land \overline{\text{Cdh1} \land \text{Ubc}}) \lor (\text{CycA} \land \overline{\text{Rb}} \land \overline{\text{Cdc20}} \land \overline{\text{Cdh1} \land \text{Ubc}}))$</td>
</tr>
<tr>
<td>Cdc20</td>
<td>CycB</td>
</tr>
<tr>
<td>Cdh1</td>
<td>$(\overline{\text{CycA}} \land \overline{\text{CycB}}) \lor \text{Cdc20}$</td>
</tr>
<tr>
<td>Ubc</td>
<td>$(\overline{\text{Cdh1}}) \lor (\text{Cdh1} \land \text{Ubc} \land (\text{Cdc20} \lor \text{CycA} \lor \text{CycB}))$</td>
</tr>
<tr>
<td>CycB</td>
<td>$(\overline{\text{Cdc20}} \land \overline{\text{Cdh1}})$</td>
</tr>
</tbody>
</table>
WNT5A Network

• Up-regulated WNT5A associated with increased metastasis.
• Cost function penalizes WNT5A being up-regulated.
• Optimal control policy with Pirin as control gene.
Sample Trajectory
Optimal (infinite horizon) control with pirin has shifted the steady-state distribution to states with WNT5A down-regulated: (a) with control; (b) without control.
Bayesian Control

• Network models are uncertain owing to insufficient data and natural regulatory variability among cells.
• Bayesian control: design a control policy that has best average performance across an uncertainty class of networks.
• Computational issues:
  – Assuming a given network, a common design method is dynamic programming, which suffers from the “curse of dimensionality.”
  – Bayesian control is much more computational owing to a huge search space and difficult optimizations – much research is necessary.