Machine Learning for Biological Data Mining

Byoung-Tak Zhang

Center for Bioinformation Technology (CBIT) &
Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University

http://cbit.snu.ac.kr/
http://bi.snu.ac.kr/

Outline

• Introduction to Bioinformatics
• Introduction to Machine Learning and Data Mining
• Gene Finding
• Protein Structure Prediction
• Promoter Prediction
• Gene Expression Profiling
• Biological Text Mining
• Conclusion
Bioinformatics

- What is a Bioinformatics?

  Bioinformatics is a new term referring to the discipline that employs computers to store, retrieve, analyze and assist in understanding biological information.

- The application of information technology and computer science to the study of biological systems.
- The analysis of the massive (and constantly increasing) amount of genetic information
- Sophisticated computer technologies to enable discovery in all fields of life sciences.
Bioinformation Technology (BIT)

Bioinformatics (*in silico* Biology)

Biocomputing (*in vivo* Computing)

Reductionistic and Synthetic Approaches in Biology

Biological System

(Organism)

Reductionistic Approach (Experiments) → Synthetic Approach (Bioinformatics)

Building Blocks

(Genes/Molecules)
**Motivations for Bioinformatics**

- Growth in molecular-biology knowledge
  
  ![Bar Chart](chart.png)  
  **EMBL Database Growth**
  
  **total nucleotides (gigabases)**
  
  ![Bar Chart](chart.png)  
  **EMBL Database Growth**
  
  **total record number (millions)**

- Genomics
  - Study of genomes through DNA sequencing
  - Industrial Biology

**The Challenge of the Information Space (1/2)**

<table>
<thead>
<tr>
<th>Data source</th>
<th>Data size</th>
<th>Bioinformatics topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw DNA sequences</td>
<td>8.2 million sequences (9.5 billion bases)</td>
<td>Separating coding and non-coding regions, Identification of introns and exons, Gene product prediction, Forensic analysis</td>
</tr>
<tr>
<td>Protein sequences</td>
<td>300,000 sequences (~300 amino acids each)</td>
<td>Sequence comparison algorithms, Multiple sequence alignments algorithms, Identification of conserved sequence motifs</td>
</tr>
<tr>
<td>Macromolecular structure</td>
<td>13,000 structures (~1,000 atomic coordinates each)</td>
<td>Secondary, tertiary structure prediction, 3D structural alignment algorithms, Protein geometry measurements, Surface and volume shape calculations, Intermolecular interactions, Molecular simulations (forces-field calculations, molecular movements, docking predictions)</td>
</tr>
</tbody>
</table>

(Source: Yearbook of Medical Informatics 2001)
## The Challenge of the Information Space (2/2)

<table>
<thead>
<tr>
<th>Data source</th>
<th>Data size</th>
<th>Bioinformatics topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomes</td>
<td>40 complete genomes (1.6 million – 3 billion bases each)</td>
<td>Characterisation of repeats, Structural assignments to genes, Phylogenetic analysis, Genomic-scale censuses (characterisation of protein content, metabolic pathways), Linkage analysis relating specific genes to disease</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Largest: ~20 time point measurements for ~6,000 genes</td>
<td>Correlating expression patterns, Mapping expression data to sequences, structural and biochemical data</td>
</tr>
<tr>
<td>Other data</td>
<td>11 million citations</td>
<td>Digital libraries for automated bibliographical searches, Knowledge databases of data from literature</td>
</tr>
<tr>
<td>Literature</td>
<td></td>
<td>Pathway simulations</td>
</tr>
<tr>
<td>Metabolic pathways</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Source: Yearbook of Medical Informatics 2001)

---

## Areas and Workflow of Bioinformatics

Microarray (Biochip)
- AGCTAGTCATCA
- TGATCTCAACTAG
- CTACGATATTAG
- AGTGTAGTACGTA
- TCAGTCAACTAG
- CTGACCTGAGAGT

- Structural Genomics
- Functional Genomics
- Proteomics
- Pharmaco-genomics

IT Infrastructure
- Bio-information Software
- Bio-information Service
Topics in Bioinformatics

Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding

Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling

Expression analysis
- Gene expression analysis
- Gene clustering

Pathway analysis
- Metabolic pathway
- Regulatory networks

Multidisciplinary Field

- It is a multidisciplinary field requiring in-depth knowledge of:

  - Math & Stat
  - Genome Science
  - Molecular Biology
  - Algorithms
  - Pharmacology
  - Machine Learning
  - Biophysics
  - Software Programming
  - Hardware Engineering
  - Hardware Engineering
Bioinformatics as Information Technology

Industry Challenges

- Key challenges to industry
  - To capture relevant data
  - To have a platform that enables data integration and mining

- Key challenges on the technology
  - To create data mining tools that are capable of retrieving the desired information from the massive databases with the highest “signal-go-noise” ratio possible
  - To do so in a cross-platform manner
Bioinformatics Market Size

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E-based business-to-business market</td>
<td>$800 million</td>
<td>$100.0 billion</td>
</tr>
<tr>
<td>Business-to-business biomedical information market</td>
<td>$300 million</td>
<td>$1.0 billion</td>
</tr>
<tr>
<td>Pharmacogenomics data gathering and analysis alliances</td>
<td>$1.0 billion</td>
<td>$3.5 billion</td>
</tr>
<tr>
<td>Biochip-based data gathering and analysis alliances</td>
<td>$500 million</td>
<td>$4.0 billion</td>
</tr>
</tbody>
</table>

Sources: Cognia (www.cognia.com); Biovista (www.biovista.com)

Business Fields

Software offering
- Data visualization
- Data management
- Gene and protein analysis
- Data filtering and transformation
- Clustering and classification
- Tools supporting laboratory experiment

Data offering
- DNA sequence data
- Gene expression data
- Protein data
- Medical genetics data
- Biological text data

Business structure offering
- Networking and service solution
- Supercomputer
- High performance storage system
### Major Companies and Their Areas of Interest

<table>
<thead>
<tr>
<th>Companies in Bioinformatics</th>
<th>Microarray analysis</th>
<th>EST-Clustering</th>
<th>Protein analysis</th>
<th>SNP Analysis</th>
<th>Passway</th>
<th>Web-potal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celera Paracel Inc.</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compugen</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DoubleTwist</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>eBioinformatics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Informax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Lion Bioscience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Molecular Mining</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosetta Inpharmatics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silicon Genetics</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
### IT Companies Offering Bioinformatics Products

<table>
<thead>
<tr>
<th>Companies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent Technologies</td>
<td>In 1999, Agilent entered into a strategic collaboration with Rosetta Inpharmatics to make and sell gene expression analysis systems, including hardware and software.</td>
</tr>
<tr>
<td>Compaq</td>
<td>Compaq has a major strategic alliance with Celera to provide integrated bioinformatics hardware, software, networking and service solutions.</td>
</tr>
<tr>
<td>IBM</td>
<td>IBM is conducting research into high value-added data mining and protein structure determination methods. IBM offers a variety of enterprise-wide IT solutions for the life sciences market, and recently initiated a collaboration with NetGenics.</td>
</tr>
<tr>
<td>Silicon Graphics</td>
<td>SGI offers visual computing and high-performance computer systems. SGI systems support a wide variety of bioinformatics software applications.</td>
</tr>
<tr>
<td>Sun Microsystems</td>
<td>Sun systems support a wide variety of bioinformatics software applications.</td>
</tr>
</tbody>
</table>

---

### Effect and Applications of Biological Data Mining

Biocomputing  
Increase and Improvement of Farm Products  
Renewable Energy  

**Biological Data Mining**  
store, retrieve, analyze and assist in understanding biological information  

- Diagnosis with Chip  
- SNP (Single Nucleotide Polymorphism)  
- Customized Drug
Current Trend

- Government driven commercialization
- More than 150 companies
- The total market for bioinformatics could $2.0 billion within five years
- Bioinformatics is becoming critical to Life Science R&D.
- Commercial applications of bioinformatics are increasing numerously.

Collaboration for Research and Development

Machine Learning and Data Mining
Why Machine Learning?

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Budding industry

Three niches for machine learning

- Data mining: using historical data to improve decisions
  - Medical records -> medical knowledge
- Software applications we can’t program by hand
  - Autonomous driving
  - Speech recognition
- Self customizing programs
  - Newsreader that learns user interests

Machine Learning

- Supervised Learning
  - Estimate an unknown mapping from known input-output pairs
  - Learn $f_w$ from training set $D=\{(x, y)\}$ s.t.
  - Classification: $y$ is discrete $f_w(x) = y = f(x)$
  - Regression: $y$ is continuous
- Unsupervised Learning
  - Only input values are provided
  - Learn $f_w$ from $D=\{(x)\}$ s.t.
  - Compression $f_w(x) = x$
  - Clustering
Methods in Machine Learning (1/2)

- Symbolic Learning
  - Version Space Learning
  - Case-Based Learning
- Neural Learning
  - Multilayer Perceptrons (MLPs)
  - Self-Organizing Maps (SOMs)
  - Support Vector Machines (SVMs)
- Evolutionary Learning
  - Evolution Strategies
  - Evolutionary Programming
  - Genetic Algorithms
  - Genetic Programming

Methods in Machine Learning (2/2)

- Probabilistic Learning
  - Bayesian Networks (BNs)
  - Helmholtz Machines (HMs)
  - Latent Variable Models (LVMs)
  - Generative Topographic Mapping (GTM)
- Other Machine Learning Methods
  - Decision Trees (DTs)
  - Reinforcement Learning (RL)
  - Boosting Algorithms
  - Mixture of Experts (ME)
  - Independent Component Analysis (ICA)
Machine Learning Techniques for Bio Data Mining

- Sequence Alignment
  - Simulated Annealing
  - Genetic Algorithms
- Structure and Function Prediction
  - Hidden Markov Models
  - Multilayer Perceptrons
  - Decision Trees
- Molecular Clustering and Classification
  - Support Vector Machines
  - Nearest Neighbor Algorithms
- Expression (DNA Chip Data) Analysis
  - Self-Organizing Maps
  - Bayesian Networks

Why Data Mining?

- Traditional analysis
  - One gene in one experiment
  - Small data sets
  - Simple analytical methods

- High-throughput genomic analysis
  - Simultaneous measurements of thousands of gene expression levels
    - massive data sets
  - Statistical methods
  - Machine learning approach
Data Mining

Selection & Sampling
Preprocessing & Cleaning
Transformation & reduction
Data Mining
Interpretation/ Evaluation

Database/data warehouse
Target data
Cleaned data
Transformed data
Patterns/ model

Knowledge
Performance system

Gene Fining
Gene Structure

Gene Finding

- From an unannotated DNA sequence, find putative expressive regions
- Distinguish exon and intron regions (in eukaryote).

- Solutions
  - Compare known genes (alignments)
  - Simply find conserved region
  - Using statistical models like hidden Markov models and others
Gene Finding and Gene Prediction

- Machine Learning for Gene Prediction
  - Hidden Markov model for motif extraction [Yada et al., 1998]
  - EM algorithm for common site identification [Lawrence and Reilly, 1990]
  - Information theory for protein-binding site identification [Stormo and Hartzell, 1989]
  - SOM for human protein sequence mapping [Ferran et al., 1994]
  - NNs for translational initiation site identification [Stormo et al., 1982], for splice junction prediction [Nakata et al., 1985], for promoter site prediction [Horton and Kanehisa, 1992]

- Machine Learning for Gene Finding
  - NNs for protein coding region finding [Uberbacher and Mural, 1991]
  - Markov chain for gene recognition [Borodovsky and McIninch, 1993]

Decision Trees: Architecture

- Nodes: attributes
- Edges: values
- Terminal nodes: class labels
Decision Trees: Learning

- Two-Phase Learner
  - Building Phase
    - Recursively split nodes using the best splitting attribute for the node.
  - Pruning Phase
    - Smaller imperfect DTs generally achieve better accuracy.
    - Prune leaf nodes recursively to prevent over-fitting.

- Learning Algorithms
  - ID3 [Quinlan, 1986]
  - C4.5 [Quinlan, 1993]
  - CART [Breiman et al., 1984]
  - AdaBoost DT
  - Bagging DT

Decision Trees for gene finding

- MORGAN: A decision tree system for gene finding. Coding and non-coding regions finding/exon finding

by Markov Chains

- **d+a**: donor and acceptor site score
- **hex**: in-frame hexamer freq.
- **asym**: Fickett’s position asymmetry statistic

<table>
<thead>
<tr>
<th>d+a&lt;3.4?</th>
<th>d+a&lt;5.3?</th>
<th>hex&lt;-5.6?</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

- d+a<1.3?
  - yes: (6,560)
  - no: (18,160)
- d+a<5.3?
  - yes: (9,49)
  - no: (142,73)

- donor<0.0?
  - yes: (5,21)
  - no: (23,16)
- hex<16.3?
  - yes: (24,13)
  - no: (737,50)

- hex<0.1?
  - yes: (737,50)
  - no: (1,5)
- hex<-5.6?
  - yes: (1,5)
  - no: (737,50)
Leading Gene-Finding Systems

<table>
<thead>
<tr>
<th>Gene finder</th>
<th>Coding bases</th>
<th>Exact exons</th>
<th>ME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sn</td>
<td>Sp</td>
<td>AC</td>
</tr>
<tr>
<td>MORGAN</td>
<td>0.81</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>GENSCAN</td>
<td>0.83</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>VEIL</td>
<td>0.83</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>Genie</td>
<td>0.78</td>
<td>0.84</td>
<td>0.77</td>
</tr>
<tr>
<td>FGENEH</td>
<td>0.77</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>GRAIL 2</td>
<td>0.72</td>
<td>0.87</td>
<td>0.75</td>
</tr>
<tr>
<td>GeneID</td>
<td>0.63</td>
<td>0.81</td>
<td>0.67</td>
</tr>
<tr>
<td>GeneParser2</td>
<td>0.66</td>
<td>0.79</td>
<td>0.67</td>
</tr>
<tr>
<td>GenLang</td>
<td>0.72</td>
<td>0.79</td>
<td>0.69</td>
</tr>
<tr>
<td>SorFind</td>
<td>0.71</td>
<td>0.85</td>
<td>0.73</td>
</tr>
<tr>
<td>Xpound</td>
<td>0.61</td>
<td>0.87</td>
<td>0.68</td>
</tr>
</tbody>
</table>

*AC is the approximate correlation proposed by Burset and Guigo (1996) [30] as a replacement for the correlation coefficient. Sensitivity (Sn) is the fraction of true coding bases that were correctly predicted as coding, and specificity (Sp) is the number of bases predicted to be in coding regions that actually were coding; their average is given in the Avg column. The “exact exon” columns show the corresponding results for prediction of whole exons. ME (missing exons) is the fraction of whole coding exons that are missed completely.

Support Vector Machines: Architecture

<table>
<thead>
<tr>
<th>Type of SVM</th>
<th>Inner product kernel $K(x,x_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial learning machine</td>
<td>$(1 + x^T y)^p$</td>
</tr>
<tr>
<td>Radial-basis function network</td>
<td>$\exp\left(-\frac{1}{2\sigma^2}|x-x_i|^2\right)$</td>
</tr>
<tr>
<td>Two-layer perceptron</td>
<td>$\tanh(\beta_0 x^T x_i + \beta_1)$</td>
</tr>
</tbody>
</table>
Support Vector Machine
Theory & Learning (1/2)

\[ f(x) = w^T x + b \]
\[ h(x) = \text{sgn}(f(x)) \]

Margin \( r = \frac{|w^T x + b|}{||w||} \geq \frac{1}{||w||} \)

SVM's objective: Large margin classifier with constraints

\[ \min_{w,b} \frac{1}{2} ||w||^2 \]
\[ d_i(w^T x + b) \geq 1, \ \forall i \]

Support Vector Machines:
Theory & Learning (1/2)

- SV (Support Vector): data on the margins
- SVMs find optimal hyperplane maximizing margins
  \[ w^T \varphi(x) + b = 0 \]
- Solution: Quadratic Programming

Minimize \( F(\Lambda) = -\Lambda d + \frac{1}{2} \Lambda H\Lambda \)
subject to
\[ \Lambda d = 0 \]
\[ \Lambda \leq C_1 \]
\[ \Lambda \geq 0 \]
\[ H_{ij}=d_i d_j K(x_i, x_j), \ \Lambda = (\lambda_1, \lambda_N) \]
Support Vector Machines for Functional Classification of Genes (1)

- Classifying gene functional class using gene expression data from DNA microarray hybridization experiments
  - Dataset: 2467 genes, 79 experiments (2467x79 matrix)

1. Tricarboxylic-acid pathway
2. Respiration chain complexes
3. Cytoplasmic ribosomal proteins
4. Proteasome
5. Histones
6. Helix-turn-helix

Functional classes defined from MYGD

121 Expression profiles of the cytoplasmic ribosomal proteins. (Similarity can be found!)

Support Vector Machines for Functional Classification of Genes (2)

<table>
<thead>
<tr>
<th>Class</th>
<th>Method</th>
<th>Learned threshold</th>
<th>FP</th>
<th>FN</th>
<th>TP</th>
<th>TN</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tricarboxylic acid</td>
<td>Radial SVM</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2442</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>2439</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>2445</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>4</td>
<td>12</td>
<td>5</td>
<td>2446</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>4</td>
<td>12</td>
<td>5</td>
<td>2446</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>2441</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>7</td>
<td>17</td>
<td>0</td>
<td>2443</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MOC1</td>
<td>3</td>
<td>16</td>
<td>1</td>
<td>2446</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Respiration</td>
<td>Radial SVM</td>
<td>9</td>
<td>6</td>
<td>24</td>
<td>2428</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>21</td>
<td>10</td>
<td>20</td>
<td>2416</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>7</td>
<td>14</td>
<td>16</td>
<td>2430</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>3</td>
<td>15</td>
<td>15</td>
<td>2434</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>22</td>
<td>10</td>
<td>20</td>
<td>2415</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>2427</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>18</td>
<td>17</td>
<td>13</td>
<td>2419</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MOC1</td>
<td>12</td>
<td>26</td>
<td>4</td>
<td>2425</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Ribosome</td>
<td>Radial SVM</td>
<td>9</td>
<td>4</td>
<td>117</td>
<td>2337</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-1 SVM</td>
<td>13</td>
<td>6</td>
<td>115</td>
<td>2333</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-2 SVM</td>
<td>7</td>
<td>10</td>
<td>111</td>
<td>2339</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dot-product-3 SVM</td>
<td>3</td>
<td>18</td>
<td>103</td>
<td>2343</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parzen</td>
<td>6</td>
<td>8</td>
<td>113</td>
<td>2340</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLD</td>
<td>15</td>
<td>5</td>
<td>116</td>
<td>2331</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C4.5</td>
<td>31</td>
<td>21</td>
<td>100</td>
<td>2315</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MOC1</td>
<td>26</td>
<td>26</td>
<td>95</td>
<td>2320</td>
<td>78</td>
<td></td>
</tr>
</tbody>
</table>

Cost = FP + 2FN

FLD: Fisher’s linear discriminant
C4.5 and MOC1: Decision trees
Parzen: Parzen windows (similar nonparametric density estimation technique)

Comparison of error rates for various classification methods on 4 classes
Protein Structure Prediction

Structure Prediction

- DNA, RNA and protein structure prediction
- These structures affect their function, stability, etc.
- Protein structure prediction by calculation of biochemical properties of each amino acid
  - In most cases of protein secondary or tertiary structure prediction, it is a terribly huge computing job (almost impossible)
- Prediction based on known structures
Protein Structure

Primary structure (amino acid sequence) determine its secondary structure (partial structure: \( \alpha \) helix, \( \beta \)-sheet) and tertiary structure

- Protein function: determined by its structure (e.g.: enzyme inactivated by heat - because its structure was changed by heat.)

Structure and Function Prediction

- Protein secondary structure prediction
  - Neural networks [Qian and Sejnowski, 1988] [Rost and Sander, 1993]
  - Hidden Markov model [Asai et.al., 1993]
  - Logic programming [Muggleton et.al., 1992]

- Protein modeling
  - HMM [Krogh et.al., 1994] [Brown et.al., 1993]

- Gene Finding and Gene Prediction (see below)
Hidden Markov Models: Architecture & Learning

- HMM is a probabilistic finite state automata
  - Nodes: states
  - Edges: transition and emission probabilities
  - Learning: Expectation-Maximization (EM) or Gibbs sampling
- HMM for sequence analysis (finding a sequence motif that consists of 5 match states)

Hidden Markov Models for Protein Modeling

- 20 alphabets (20 amino acids)
- $m_0$: start state, $m_5$: end state, $m_k$: match states
- $i_k$: insertion states, $d_k$: deletion states
- $T(s_j|s_j)$: transition probabilities
- $P(x|m_k)$: alphabet generating probabilities ($x$: letter; amino acid)
Nearest Neighbor Algorithms: Algorithms

Training algorithm:
- For each training example \( <x, f(x)> \),
  add the example to the list \( training\_examples \)

Classification algorithms:
- Given a query instance \( x_q \) denote
  the \( k \) instances from \( training\_examples \) that are nearest to \( x_q \)
- Return
  \[
  \hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))
  \]
  where \( \delta(a,b)=1 \) if \( a=b \) and \( \delta(a,b)=0 \)
  where otherwise

Nearest Neighbor Algorithms for 3D Protein Classification

- 3D shape similarity model by shape histograms [Ankerst, 1999]

\[
d^2_A(x, y) = (x - y) \cdot A \cdot (x - y)^T
\]
\[
= \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(x_i - y_i)(x_j - y_j)
\]
\[
a_{ij} = e^{-\sigma \cdot d(i, j)}
\]
\( d(i,j) \): distance of the cells that corresponds to the bins \( i, j \).
The cell distance is calculated from the difference of the shell radii and the angles between the sectors.
Promoter Prediction

Promoter Prediction in the human genome

- Promoter
  : Signals around the Transcriptional Start Site (TSS)
  - Core Promoter (~100bp)
  - Transcription Factor Binding Sites. (~5-20bp)
- CpG Island.
  - Stretches of un-methylated DNA with a higher frequency of CpG dinucleotide.
  - 50% of all genes have CpG islands at the 5’end of transcript
- Promoter Prediction
  - Partitioning a Genome into Genes: Gene Finding
  - Determining the Correct Protein Translation
  - Determining the Expression Context.
- **Basal Transcription and Core Promoter**: binding site for RNA polymerase II and General transcription factors (GTFs)
  - TATA-box: 30bp downstream.
    - Recognized by TATA-binding protein (TBP)
    - Consensus sequence or position weight matrix (PWM)
  - Initiator (Inr): pypyAN[TA]PyPy
  - Downstream promoter element (DPE)
  - GC box and CAAT-box

**Activated transcription and regulatory sequence**
(binding sites for transcription factor)
Promoter Prediction

- Audic/Claverie
  : Markov Models of vertebrate promoter sequence.
- Autogene
  : Clustering Algorithm based on the consensus site occurrence.
- GeneID/Promoter2.0
  : Neural Network and Genetic Algorithm
- NNPP
  : Time delay neural net architecture.
- TSSG/TSSW
  : Linear discriminant function

Promoter 2.0

- Neural Network
  : input a small window of DNA sequence
  : output of other neural networks.
- Genetic algorithm:
  : the weights in the neural networks are optimized to discriminate maximally between promoters and non-promoters.
DNA Microarrays

- Monitor thousands of gene expression levels simultaneously \(\leftrightarrow\) traditional one gene experiments.
- Fabricated by high-speed robotics.
A Comparative Hybridization Experiment

Types of DNA Microarrays

- Oligonucleotide chips
  - An array of oligonucleotide (20 ~ 80-mer oligos) probes is synthesized.

- cDNA microarrays
  - Probe cDNA (500 ~ 5,000 bases long) is immobilized to a solid surface.
Data Preparation for Data Mining

Image analysis

An Example of Data Mining: Clustering
Analysis of DNA Microarray Data

Previous Work

- Characteristics of data
  - Analysis of expression ratio based on each sample
  - Analysis of time-variant data
- Clustering
  - Self-organizing maps [Golub et al., 1999]
  - Singular value decomposition [Orly Alter et al., 2000]
- Classification
  - Support vector machines [Brown et al., 2000]
- Gene identification
  - Information theory [Stefanie et al., 2000]
- Gene modeling
  - Bayesian networks [Friedman et al., 2000]

DNA Microarray Data Mining

- Clustering
  - Find some groups of genes that show the similar pattern in some conditions.
  - PCA (Principle component analysis)
  - SOM (Self-organizing maps)

- Genetic network analysis
  - Determine the regulatory interactions between genes and their derivatives.
  - Linear models
  - Neural networks
  - Probabilistic graphical models
Bayesian Networks & Generative Topographic Mapping

- Bayesian Networks
  A graphical model for probabilistic relationships among a set of variables
- Generative Topographic Mapping
  A graphical model through a nonlinear relationship between the latent variables and observed features.

Why Bayesian Networks?

- The joint probability distribution $\rightarrow$ all the knowledge about the (biological) system.
- Efficient probabilistic inference $\rightarrow$ prediction
- The network structure $\rightarrow$ an insight into the intricate relationships between components (i.e., genes) of the biological system $\rightarrow$ dependency analysis
- Robust to the noise and error in gene expression data
Gene Expression Data Analysis with Bayesian Networks

Data preprocessing:
- Normalization
- Discretization
- Gene selection

Learning:
- Greedy search
- EM algorithm

- Classification
- Clustering
- Analysis of gene regulations

An Introductory Example

- A Bayesian network classifier for acute leukemias [Hwang et al. 2001]

<table>
<thead>
<tr>
<th>Zyxin</th>
<th>Zyxin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>ALL or AML</td>
</tr>
<tr>
<td>LTC4S</td>
<td>Leukotriene C4 synthase (LTC4S) gene</td>
</tr>
<tr>
<td>C-myb</td>
<td>C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds</td>
</tr>
<tr>
<td>MB-1</td>
<td>MB-1 gene</td>
</tr>
</tbody>
</table>

\[ P(X) = \prod_{i=1}^{n} P(X_i | P_{a_i}) \]
Probabilistic Graphical Models

- The joint probability distribution over \( X = \{X_1, X_2, \ldots, X_n\} \)
  - Chain rule: \( P(X) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \)

- Conditional independence \( X \perp Y \mid Z \)
  \[ P(X \mid Y, Z) = P(X \mid Z) \]

- Efficient representation of the joint probability distribution using conditional independencies encoded by the graph (network) structure
  \[ P(X) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} P(X_i \mid V(X_i)) \]
  - \( X_i \perp \{X_1, \ldots, X_{i-1}\} \mid V(X_i) \mid V(X_i) \)

Gene-Drug Dependency Analysis Using Bayesian Networks

- To discover
  - Gene-gene expression dependency
  - Gene expression-drug activity dependency
  - Drug-drug activity dependency
Bayesian Networks: the Qualitative Part

- The directed-acyclic graph (DAG) structure
  - Conditional independencies between variables → dependency analysis

<Conditional independencies>
- Gene B and Gene D are independent given Gene A.
- Gene B asserts dependency between Gene A and Gene E.
- Gene A and Gene C are independent given Gene B.

Bayesian Networks: the Quantitative Part

- The joint probability distribution over all the variables in a Bayesian network.
  \[ P(X) = \prod_{i=1}^{n} P(X_i | \text{Pa}_i) \]

Local probability distribution for \( X_i \)
  \[ \Theta_i = (\theta_{i1}, \ldots, \theta_{i_q}) \sim \text{parameter for } P(X_i | \text{Pa}_i) \]
  \[ P(\Theta_i) = \text{Dir}(\theta_{ij} | \alpha_{ij1}, \ldots, \alpha_{ijn}) \]
  \[ q_i : \# \text{ of configurations for } \text{Pa}_i \]
  \[ r_i : \# \text{ of states for } X_i \]

\[
\]
Analysis Procedure

Gene Expression Data

Discretization and Selection

- Selected genes and the target variable

Bayesian Network Learning

Gene A
Gene B
Gene C
Gene D
Target

Gene A ➔ Gene B ➔ Gene C ➔ Gene D ➔ Target

Gene B
Gene C
Gene D
Target

<Learned Bayesian network>

- Classification
- Dependency analysis

Clustering Using Generative Topographic Mapping

- GTM: a nonlinear, parametric mapping $y(x;W)$ from a latent space to a data space.

Grid

$y(x;W) \rightarrow$

$<Latent space> \rightarrow <Data space>$
CAMDA-2000 Data Sets

- **CAMDA**
  - Critical Assessment of Techniques for Microarray Data Mining
  - Purpose: Evaluate the data-mining techniques available to the microarray community.

- **Data Set 1**
  - Identification of cell cycle-regulated genes
  - Yeast Saccharomyces cerevisiae by microarray hybridization.
  - Gene expression data with 6,278 genes.

- **Data Set 2**
  - Cancer class discovery and prediction by gene expression monitoring.
  - Two types of cancers: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).
  - Gene expression data with 7,129 genes.

---

CAMDA-2000 Data Set 1
Identification of Cell Cycle-regulated Genes of the Yeast by Microarray Hybridization

- **Data given:** gene expression levels of 6,278 genes spanned by time
  - α Factor-based synchronization: every 7 minute from 0 to 119 (18)
  - Cdc15-based synchronization: every 10 minute from 10 to 290 (24)
  - Cdc28-based synchronization: every 10 minute from 0 to 160 (17)
  - Elutriation (size-based synchronization): every 30 minutes from 0 to 390 (14)

- Among 6,278 genes
  - 104 genes are known to be cell-cycle regulated
    - classified into: M/G1 boundary (19), late G1 SCB regulated (14), late G1 MCB regulated (39), S-phase (8), S/G2 phase (9), G2/M phase (15).

- 250 cell cycle-regulated genes
CAMDA-2000 Data Set 1
Characteristics of data (α Factor-based Synchronization)

- M/G1 boundary
- Late G1 SCB regulated
- Late G1 MCB regulated
- S Phase
- S/G2 Phase
- G2/M Phase

CAMDA-2000 Data Set 2
Cancer Class Discovery and Prediction by Gene Expression Monitoring

- Gene expression data for cancer prediction
  - Training data: 38 leukemia samples (27 ALL, 11 AML)
  - Test data: 34 leukemia samples (20 ALL, 14 AML)
  - Datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood.

- Graphical models used:
  - Bayesian networks
  - Non-negative matrix factorization
  - Generative topographic mapping
Applications of GTM for Bio Data Mining (1)

- DNA microarray data provides the whole genomic view in a single chip.

- The intensity and color of each spot encode information on a specific gene from the tested sample.

- The microarray technology is having a significant impact on genomics study, especially on drug discovery and toxicological research.

(Figure from http://www.gene-chips.com/sample1.html)

Applications of GTM for Bio Data Mining (2)

- Select cell cycle-regulated genes out of 6179 yeast genes. (cell cycle-regulated : transcript levels vary periodically within a cell cycle)

- There are 104 known cell cycle-regulated genes of 6 clusters
  - S/G2 phase : 9 (train:5 / test:2)
  - S phase : 8 (Histones) (train:5 / test:3)
  - M/G1 boundary (SWI5 or ECB (MCM1) or STE12/MCM1 dependent) : 19 (train:13 / test:6)
  - G2/M phase: 15 (train: 10 / test:5)
  - Late G1, SCB regulated : 14 (train: 9 / test:5)
  - Late G1, MCB regulated : 39 (train: 25 / test:12)
  (M-G1-S-G2-M)
### The comparison of entropies for each method

<table>
<thead>
<tr>
<th>cluster</th>
<th>size (3)</th>
<th>mean response to Cln3p</th>
<th>mean response to Clb2p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2758</td>
<td>-0.125</td>
<td>-0.130</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>-0.095</td>
<td>-0.094</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>-0.085</td>
<td>-0.070</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>1.142</td>
<td>-0.510</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>-0.406</td>
<td>1.316</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>-0.180</td>
<td>-0.152</td>
</tr>
<tr>
<td>7</td>
<td>132</td>
<td>0.795</td>
<td>-0.552</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>-0.629</td>
<td>-0.180</td>
</tr>
<tr>
<td>9</td>
<td>25</td>
<td>-0.447</td>
<td>0.066</td>
</tr>
<tr>
<td>10</td>
<td>47</td>
<td>-0.194</td>
<td>0.102</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>0.025</td>
<td>-0.104</td>
</tr>
<tr>
<td>12</td>
<td>225</td>
<td>-0.086</td>
<td>-0.120</td>
</tr>
<tr>
<td>13</td>
<td>45</td>
<td>-0.284</td>
<td>0.060</td>
</tr>
<tr>
<td>14</td>
<td>25</td>
<td>0.079</td>
<td>-0.010</td>
</tr>
<tr>
<td>15</td>
<td>53</td>
<td>-0.437</td>
<td>-0.158</td>
</tr>
<tr>
<td>16</td>
<td>53</td>
<td>-0.058</td>
<td>-0.088</td>
</tr>
<tr>
<td>17</td>
<td>23</td>
<td>-0.178</td>
<td>-0.050</td>
</tr>
<tr>
<td>18</td>
<td>46</td>
<td>-0.122</td>
<td>-0.204</td>
</tr>
<tr>
<td>19</td>
<td>86</td>
<td>-0.230</td>
<td>-0.138</td>
</tr>
<tr>
<td>20</td>
<td>76</td>
<td>-0.002</td>
<td>-0.004</td>
</tr>
<tr>
<td>21</td>
<td>28</td>
<td>-0.140</td>
<td>-0.104</td>
</tr>
<tr>
<td>22</td>
<td>35</td>
<td>-0.046</td>
<td>-0.058</td>
</tr>
<tr>
<td>23</td>
<td>34</td>
<td>-0.167</td>
<td>-0.168</td>
</tr>
<tr>
<td>24</td>
<td>117</td>
<td>-0.213</td>
<td>-0.120</td>
</tr>
<tr>
<td>25</td>
<td>52</td>
<td>-0.214</td>
<td>-0.198</td>
</tr>
<tr>
<td>unclassified</td>
<td>204</td>
<td>-0.189</td>
<td>-0.088</td>
</tr>
</tbody>
</table>

### Clusters identified by various methods

<table>
<thead>
<tr>
<th>cluster</th>
<th>GTM1</th>
<th>GTM2</th>
<th>SOM1</th>
<th>SOM2</th>
<th>PCA</th>
<th>Average Linkage</th>
</tr>
</thead>
<tbody>
<tr>
<td>size (3)</td>
<td>71</td>
<td>37</td>
<td>21</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td># known genes</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Cln3p</td>
<td>0.856</td>
<td>1.142</td>
<td>1.34</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Clb2p</td>
<td>-0.415</td>
<td>-0.510</td>
<td>-0.554</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>size (G2/M1)</td>
<td>123</td>
<td>42</td>
<td>-</td>
<td>63</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td># known genes</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Cln3p</td>
<td>-0.446</td>
<td>-0.406</td>
<td>-</td>
<td>-0.452</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Clb2p</td>
<td>0.563</td>
<td>1.316</td>
<td>0.962</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>size (G1)</td>
<td>202</td>
<td>132</td>
<td>62</td>
<td>47</td>
<td>-</td>
<td>147</td>
</tr>
<tr>
<td># known genes</td>
<td>29</td>
<td>25</td>
<td>21</td>
<td>19</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Cln3p</td>
<td>0.684</td>
<td>0.795</td>
<td>-0.890</td>
<td>1.283</td>
<td>0.439</td>
<td>-</td>
</tr>
<tr>
<td>Clb2p</td>
<td>-0.481</td>
<td>-0.552</td>
<td>-0.103</td>
<td>-0.993</td>
<td>-0.361</td>
<td>-</td>
</tr>
</tbody>
</table>

[Clusters identified by various methods]

[The comparison of entropies for each method]
Biological Text Mining

- “Literature mining”
- Problems and Motivation
  - Vast amounts of sequences have accumulated in public databases.
  - The next step in genome analysis
    - Definition of the function of each gene
    - Determination of its role in biological pathways
  - Protein-protein interaction (gene-gene relation)
    - Form the basis of phenomena such as DNA replication and transcription, metabolic pathway, signal pathway, and cell cycle control.
**Biological Text Mining**

- Automated extraction of information on protein-protein interaction
  - Protein-protein interaction data have been collected in several DBs.
  - The information exist in the form of text literature, and the collection of these data takes too much time and labor.
  - Efficient processing of these large data requires an intelligent information extraction method.

- Extract useful information from the biological literature
  - Gene name extraction
  - Gene-gene relation mining
  - Protein-protein interaction mining

- Visualize mined information (with graph etc.)

---

**Overall Architecture**

- Query Processing and Literature Retrieval System
- Information Extractor
  - Gene/protein name identification
  - Gene/protein/drug relation extraction
  - (activation, suppression)
  - Disease-related gene/protein identification

- IR system
  - Thematic Analysis
  - Topical clustering
  - Summarization

- Extraction Rule Learner
  - HMMs, LVM, NLP
  - Rule induction

- Machine Learning

- Literature DB
  - MEDLINE

- Processed 2nd DB
  - Gene Thesaurus

- Medical DBs
  - OMIM, PDB Swiss-Prot
Biological Text Mining

- Applicable ML methods
  - Hidden Markov Models
    - Text entity modeling and extraction.
    - Relation extraction for pathway design.
      - Protein-protein interaction, Gene-gene relation.
    - DNA sequence modeling.
    - DNA splice site modeling and identification, etc.
  - Latent Variable Model
    - Topic word extraction, topical classification, etc.
  - NLP technique
    - Shallow parsing, full sentence parsing, tagging
      - Appropriate for improving extraction precision.

References

- Toshihide Ono et al.

- C. Blaschke et al.

- J. Stapley and G. Benoit
Conclusion

- Challenges of Machine Learning and Data Mining Applied to Biosciences
  - Huge data size
  - Noise and data sparseness
  - Unlabeled and imbalanced data
- Probabilistic graphical models have been applied to gene expression analysis
  - Bayesian networks
  - Generative topographic mapping
- Probabilistic graphical models such as Bayesian networks are useful for gene expression profiling and dependency analysis.
  - Dependency analysis → exploratory research on gene regulatory interactions.
  - Bayesian networks are robust against noise and missing data.