Molecular Machine Learning: A Personal Introduction

Byoung-Tak Zhang

Adapted from “Biological Machine Learning: From Biology to Machine Learning and Back” presented November 9, 2002 as Invited Talk at Korea Information Science Society SIG CVPR

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

http://bi.snu.ac.kr/
http://cbit.snu.ac.kr/
From BT to IT and Back: Biological Machine Learning

- Bioinformatics
- Neural Computation
- Evolutionary Computation
- Molecular Computation
- Bio-Inspired Machine Learning
- Biotechnical Machine Learning
- Biocomputing

S/W

BT

H/W

S/W

IT

H/W

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Outline

- Biological Machine Learning: BIT
- ML Inspired by Biology: NN & GA
  - ML for Biology: Bioinformatics
- ML Using Biology: Molecular ML
  - Conclusions and Outlook
ML Inspired by Biology:

Neural Networks
Genetic Algorithms
The Brain vs. Computer

1. \(10^{11}\) neurons with \(10^{14}\) synapses
2. Speed: \(10^{-3}\) sec
3. Distributed processing
4. Nonlinear processing
5. Parallel processing

1. A single processor with complex circuits
2. Speed: \(10^{-9}\) sec
3. Central processing
4. Arithmetic operation (linearity)
5. Sequential processing
From Biological Neuron to Artificial Neuron

- Cell Body (Soma)
- Dendrites
- Presynaptic Terminal
- Node of Ranvier
- Axon
- Myelin
- Nucleus

Activation function: \( \varphi(\cdot) \)

Input signals:
- \( x_1 \)
- \( x_2 \)
- \( \ldots \)
- \( x_p \)

Synaptic weights:
- \( w_{k1} \)
- \( w_{k2} \)
- \( w_{kF} \)

Threshold: \( \theta_k \)

Output: \( y_k \)
Multilayer Perceptron (MLP)

- **Input Layer**: Scaling Function
- **Hidden Layer**: Activation Function
- **Output Layer**: Activation Function

**Error Backpropagation**

\[ w_i \leftarrow w_i + \Delta w_i, \quad \Delta w_i = -\eta \frac{\partial E}{\partial w_i} \]

**Output Comparison**

\[ E_d(\hat{w}) = \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - o_k)^2 \]

**Information Propagation**

\[ o = f(x) \]
Learning as Error Minimization

Gradient

\[ \nabla E[\vec{w}] = \left[ \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \right] \]

Training rule:

\[ \Delta \vec{w} = -\eta \nabla E[\vec{w}] \]

i.e.,

\[ \Delta w_i = -\eta \frac{\partial E}{\partial w_i} \]
Gradient Descent

\[ w_i \leftarrow w_i + \Delta w_i \quad \text{,} \quad \Delta w_i = -\eta \frac{\partial E}{\partial w_i} \]

\[ \frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \left( \frac{1}{2} \sum_d (t_d - o_d)^2 \right) \]
\[ = \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (t_d - o_d)^2 \]
\[ = \frac{1}{2} \sum_d 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) \]
\[ = \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \overline{w} \cdot \overline{x}_d) \]
\[ \frac{\partial E}{\partial w_i} = \sum_d (t_d - o_d) (-x_{i,d}) \]

\[ \Delta w_i = \eta \sum_{d \in D} (t_d - o_d) x_{i,d} \]
Application Example: Autonomous Land Vehicle (ALV)

- NN learns to steer an autonomous vehicle.
- 960 input units, 4 hidden units, 30 output units
- Driving at speeds up to 70 miles per hour
Charles Darwin (1859)

“Owing to this struggle for life, any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species, in its infinitely complex relations to other organic beings and to external nature, will tend to the preservation of that individual, and will generally be inherited by its offspring.”
Evolutionary Computation

- What is the Evolutionary Computation?
  - Stochastic search (or problem solving) techniques that mimic the metaphor of natural biological evolution.
- Metaphor

<table>
<thead>
<tr>
<th>EVOLUTION</th>
<th>PROBLEM SOLVING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Candidate Solution</td>
</tr>
<tr>
<td>Fitness</td>
<td>Quality</td>
</tr>
<tr>
<td>Environment</td>
<td>Problem</td>
</tr>
</tbody>
</table>
General Structure of GAs

1. Encoding
2. Selection
3. Evaluation
4. Crossover
5. Mutation

New population

Solutions

Fitness computation

Chromosomes
Major Evolutionary Algorithms

- Genetic representation of candidate solutions
- Genetic operators
- Selection scheme
- Problem domain

Hybrids: BGA
Selection Strategies

- **Proportionate Selection**
  - Reproduce offspring in proportion to fitness $f_i$.

- **Ranking Selection**
  - Select individuals according to $\text{rank}(f_i)$.

- **Tournament Selection**
  - Choose $q$ individuals at random, the best of which survives.

- **Other Ways**

\[ p_s(a_i^t) = \frac{f(a_i^t)}{\sum_{j=1}^{\lambda} f(a_j^t)} \]

\[ p_s(a_i^t) = \begin{cases} 
\frac{1}{\mu}, & 1 \leq i \leq \mu \\
\frac{i}{\lambda}, & \mu < i \leq \lambda \\
0, & \lambda < i \leq \mu 
\end{cases} \]
ES: Recombination Operator

\[ r_{x} r_{\sigma} r_{\alpha} \], where \( r_{x}, r_{\sigma}, r_{\alpha} \in \{-, d, D, i, I, g, G\} \), e.g. \( r_{dII} \)

\[
x'_{i} = \begin{cases} 
  x_{S,i} & \text{no recombination} \\
  x_{S,i} \text{ or } x_{T,i} & \text{discrete} \\
  x_{S,i} \text{ or } x_{T,i} & \text{panmictic discrete} \\
  x_{S,i} + (x_{T,i} - x_{S,i}) / 2 & \text{intermediate} \\
  x_{S,i} + (x_{T,i} - x_{S,i}) / 2 & \text{panmictic intermediate} \\
  x_{S,i} + \chi \cdot (x_{T,i} - x_{S,i}) & \text{generalized intermediate} \\
  x_{S,i} + \chi_{i} \cdot (x_{T,i} - x_{S,i}) & \text{panmictic generalized intermediate} 
\end{cases}
\]

- \( r'_{-} \)
- \( r'_{d} \)
- \( r'_{D} \)
- \( r'_{i} \)
- \( r'_{i} \)
- \( r'_{g} \)
- \( r'_{G} \)
ES: Mutation Operator

- $m_{\{\tau, \tau', \beta\}} : I^\lambda \rightarrow I^\lambda$ is an asexual operator.
  - $n_\sigma = n$, $n_\alpha = n(n-1)/2$
    - $\sigma'_i = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$
    - $\alpha'_j = \alpha_j + \beta \cdot N_j(0,1)$
    - $\bar{x}' = \bar{x} + \tilde{N}(\bar{0}, \mathbf{C}(\bar{\sigma}', \bar{\alpha}'))$
  - $1 < n_\sigma < n$, $n_\alpha = 0$
    - $\sigma'_i = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$
    - $x'_i = x_i + \sigma'_i \cdot N_i(0,1)$
  - $n_\sigma = 1$, $n_\alpha = 0$
    - $\sigma' = \sigma \cdot \exp(\tau_0 \cdot N(0,1))$
    - $x'_i = x_i + \sigma' \cdot N_i(0,1)$

$t \propto \left(\frac{2}{\sqrt{n}}\right)^{-1}$
$t' \propto \left(\frac{2}{n}\right)^{-1}$
$\beta \approx 0.0873$

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Geometric Analogy - Fitness Landscape
ML for Bioinformatics
Computers Meet Biosciences

Bioinformation Technology (BIT)
Areas and Workflow of Bioinformatics

Microarray (Biochip)

Structural Genomics  Functional Genomics  Proteomics  Pharmacogenomics

Infrastructure of Bioinformatics
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
Machine Learning Techniques for Bioinformatics

- **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
Gene Finding

Upstream

Open Reading Frame

Downstream

mRNA

DNA

Exon 1
Exon 2
Exon 3
Exon 4

Intron 1
Intron 2
Intron 3

5’

Promoter
TATA

Splice site
GGTGAG

Splice site
CAG

Pyrimidine
tract

polyA signal

Translation
Initiation
ATG

Branchpoint
CTGAC

Stop codon
TAG/TGA/TAA

3’
<table>
<thead>
<tr>
<th>Name</th>
<th>Methods</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER</td>
<td>Discriminant Analysis</td>
<td>Human, Arabidopsis</td>
</tr>
<tr>
<td>GENSCAN (seems the most accurate)</td>
<td>Semi Markov Model</td>
<td>vertebrate, caenorhabditis, arabidopsis, maize</td>
</tr>
<tr>
<td>GRAIL</td>
<td>Neural Network</td>
<td>human, mouse, arabidopsis, drosophila, E.coli</td>
</tr>
<tr>
<td>GenLang</td>
<td>Definite Clause Grammer</td>
<td>Vertebrate, Drosophila, Dicot</td>
</tr>
<tr>
<td>GenView</td>
<td>Linear combination</td>
<td>Human, Mouse, Diptera</td>
</tr>
<tr>
<td>GeneFinder(FGENEH, etc.)</td>
<td>LDA</td>
<td>Human, E.coli, Drosophila, Plant, Nematode, Yeast</td>
</tr>
<tr>
<td>GeneID</td>
<td>Perceptron, rules</td>
<td>Vertebrate</td>
</tr>
<tr>
<td>GeneMark</td>
<td>5th–Markov</td>
<td>Almost all model organism</td>
</tr>
<tr>
<td>GeneParser</td>
<td>Neural networks</td>
<td>Human</td>
</tr>
<tr>
<td>Genie</td>
<td>GHMM</td>
<td>Human (vertebrate)</td>
</tr>
<tr>
<td>Glimmer</td>
<td>Interpolated Markov models (IMMs)</td>
<td>microbial</td>
</tr>
<tr>
<td>MORGAN</td>
<td>Decision Tree</td>
<td>vertebrate</td>
</tr>
<tr>
<td>MZEF</td>
<td>Quadratic Discriminant Analysis</td>
<td>Human, mouse, Arabidopsis, Pombe</td>
</tr>
<tr>
<td>NetPlantGene</td>
<td>Combined Neural Networks</td>
<td>A. thaliana</td>
</tr>
<tr>
<td>OC1</td>
<td>Decision tree</td>
<td>Human</td>
</tr>
<tr>
<td>PROCRUSTES</td>
<td>Spliced alignment</td>
<td>vertebrate</td>
</tr>
<tr>
<td>Sorfind</td>
<td>Rule base</td>
<td>Human</td>
</tr>
<tr>
<td>VEIL</td>
<td>HMM</td>
<td>vertebrate</td>
</tr>
<tr>
<td>Hogeheghe</td>
<td>Wonderful method</td>
<td>extraterrestrial</td>
</tr>
</tbody>
</table>
Learning Scheme

Training set
AATGCCTACCT
CATACGACCAC
AACGAATGAAT
ATGATGT………

Test set
TCGACTACGAG
CCTCATCGACG
AACGAATGAAT
ATGATGT………

Prediction Method

Learning (Model Construction)

Output
input
output

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Fig. 1. Schematic of the neural network for evaluating internal protein coding exons in GRAIL.
**Neural Networks in GRAIL**

**Known Sequence**
CATATCAAGAATTGAAGCGTGAGT
CCTGACTTGAGAGCTGTAGATGACGT
GCTTATATGTC

**Coding potential value**

<table>
<thead>
<tr>
<th></th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>...</th>
<th></th>
<th>Xₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.8</td>
<td>0.1</td>
<td>0.3</td>
<td>...</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>0.6</td>
<td>0.1</td>
<td>...</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9</td>
<td>0.3</td>
<td>0.1</td>
<td>...</td>
<td>0.8</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**Donor**

<table>
<thead>
<tr>
<th></th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>...</th>
<th>tₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Exon**

**Length**

**Intron vocabulary**

**GC Composition**

**Preprocessing**

**Unknown Sequence**
ATGACGTACGATCCTCGTGACGGTGA
CGTGAGCTGTCCCGTGACGGTGTA
ATTTAGCGTGCA

**Input Layer**

**Weights**

\[ w_i \leftarrow w_i + \Delta w_i \]
\[ \Delta w_i = -\eta \frac{\partial E}{\partial w_i} \]

**Hidden Layer**

**Output Layer**

\[ E_d(\bar{w}) \equiv \frac{1}{2} \sum_{k \in \text{outputs}} (t_k - o_k)^2 \]
Protein Structure Prediction

- Amino acid sequences of protein determine its 3D conformation

MNIHRSTPITIARYGRSRNKT
QDFEELSSIRSAEPSQSFSNPL
GSPSPPETPNLSHCVSCIGKY
LLLEPLEGHDVFRAVHLHSG
EELVCKVFDISCYQESLAPCF

Sequence → Structure → Function

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Motif, Domain, Family

**Motif** (protein sequence pattern): is recognizable combinations of $\alpha$ helices and $\beta$ strands that appear in a number of proteins.

**Domain** consists of combinations of motifs, the size of domains varies from about 25 to 30 amino acid residues to about 300, with an average of about 100.

**Protein family** consist of members which has

1) Same function; and
2) Clear evolutionary relationship; and
3) Patterns of conservation, some positions are more conserved than the others, and some regions seem to tolerate insertions and deletions more than other regions, the similarity usually $> 25\%$.
Neural Network for Structure Prediction

...I L...
...I K...
...L I...
...H M...

Helix  Strand  Coil
0.8  0.2  0.1

25  0.5  0  0.25  0  0  0  0.25  0.25  0.25  0.25
H I K L M  H I K L M
Sequence-to-Structure Network Architecture

Each input unit consists of 21 frequencies

Frequency of amino acid ‘E’ in multiple alignments of some protein family

Current input window
Overall Architecture

First level: sequence-to-structure

Second level: structure-to-structure

Third level: jury decision

The input is based on a profile made from amino acid occurrences in columns of a multiple alignment of sequences with high similarity of the query sequence.
cDNA Microarray

- **cDNA clones** (probes)
- **PCR product amplification purification**
- **Printing**
- **Microarray**
- **Hybridize target to microarray**
- **Overlay images and normalize**
- **Analysis**

**Excitation**
- Laser 1
- Laser 2

**Scanning**
- Emission

**0.1nl/spot**

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Image Analysis

- Scanned images → probe intensities → numerical values for higher-level analysis

Array target segmentation

Background intensity extraction

Target detection

Ratio analysis

Target intensity extraction

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Gene Expression Data Mining

Data preprocessing:
- Normalization
- Discretization
- Gene selection

Learning:
- Greedy search
- EM algorithm

Classification
Clustering
Analysis of gene regulations
Cancer Classification with DNA Microarrays

1. Filter genes (88 x 2306)
2. Reduce dimensionality PCA (88 x 10)
3. Random partition 63 training experiments into 3 groups
4. Select validation group
5. Train
6. Re-select (x 3)
7. Repartition (x 1250)
8. Rank genes (sensitivity measurement)
9. Minimize number of genes
Gene Expression Profiling

[DNA microarray dataset]
Cell Cycle-regulated Genes in *S. cerevisiae* (Yeast)

- Identify **cell cycle-regulated genes** by cluster analysis.
  - 104 genes are already known to be cell-cycle regulated.
  - Known genes are clustered into 6 clusters.
- Cluster 104 known genes and other genes together.
- The same cluster ➔ similar functional categories.

[Fig.] 104 known gene expression levels according to the cell cycle (*row*: time step, *column*: gene).
Probabilistic Clustering Using Latent Variables

\[ g_i: \text{ith gene} \]
\[ z_k: \text{kth cluster} \]
\[ t_j: \text{jth time step} \]
\[ p(g_i|z_k): \text{generating probability of ith gene given kth cluster} \]
\[ v_k = p(t|z_k): \text{prototype of kth cluster} \]

\[
p(g_i \in z_k) = p(z_k | g_i) = \frac{p(g_i | z_k)p(z_k)}{p(g_i)}
\]

\[
f(g, t, z) = \sum_i \sum_j \sum_k g_{ij} \sum_k \log(p(z_k)p(g_i | z_k)p(t_j | z_k)) : (* \text{ objective function (maximized by EM)})
\]

\[
\text{similarity } (x_i, v_k) = \sum_j x_{ij} v_{kj}
\]
Experimental Results:
Prototype Expression Levels of Found Clusters

- The genes in the same cluster show similar expression patterns during the cell cycle.
- The genes with similar expression patterns are likely to have correlated functions.

[Fig.] Prototype expression levels of genes found to be cell cycle-regulated (4 clusters).
NCI Drug Discovery Program

The NCI Cancer Drug Discovery - Development Pipeline

- ~500,000 cmpds
- >70,000 cmpds
- Inventory
- 60-Cell line screen
- In vivo studies
- Clinical Trials

Database A

(Activity Patterns)

>70K cmpds

60 cell lines (>4,000,000 numbers)

Database T

Gene expression

T_r (9,703 genes)

T_i (41 individually assessed targets)

NCI 60 cell lines data set
Bayesian Networks

- Represent the joint probability distribution among random variables efficiently using the concept of conditional independence.

An edge denotes the possibility of the causal relationship between nodes.

- $A$, $C$ and $D$ are independent given $B$.
- $C$ asserts dependency between $A$ and $B$.
- $A$, $B$ and $E$ are independent given $C$.

\[
P(A, B, C, D, E) = P(A)P(B \mid A)P(C \mid A, B)P(D \mid A, B, C)P(E \mid A, B, C, D) \text{ (by chain rule)}
\]
\[
= P(A)P(B)P(C \mid A, B)P(D \mid B)P(E \mid C) \text{ (by the example Bayes net)}
\]
Bayesian Network Learning

- **Dependence analysis** [Margaritis ’00]
  - Mutual information and $\chi^2$ test

- **Score-based search**

\[
p(D, S) = p(S)p(D \mid S)
\]
\[
= p(S) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}
\]

- **NP-hard problem**
- **Greedy search**
- **Heuristics to find good massive network structures quickly** (local to global search algorithm)
A Small Bayesian Network for Classification of Cancer

- The Bayesian network was learned by full search using BD (Bayesian Dirichlet) score with uninformative prior [Heckerman ’95] from the DNA microarray data for cancer classification (http://waldo.wi.mit.edu/MPR/).

[Table] Comparison of the classification performance with other methods [Hwang ’00].

<table>
<thead>
<tr>
<th></th>
<th>Training error</th>
<th>Test error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bayes nets</strong></td>
<td>0/38</td>
<td>2/34</td>
</tr>
<tr>
<td><strong>Neural trees</strong></td>
<td>0/38</td>
<td>1/34</td>
</tr>
<tr>
<td><strong>RBF networks</strong></td>
<td>0/38</td>
<td>1.3/34</td>
</tr>
</tbody>
</table>
Experimental Results: Drug-Drug Dependency

- Drug-drug activity correlations

- Three drugs “Aphidicolin-glycinate”, “Floxuridine”, and “Cytarabine” directly depend on each other.

- “Cyclocytidine” directly depends on “Cytarabine” and vice versa.

Confirmation
Experimental Results: Gene-Drug Dependency

- Gene expression-drug activity correlations

- The negative correlation between “ASNS” and “L-asparagine” is mediated by two other genes.

- The relationships revealed by the Bayesian network is putative and should be verified by biological experiments.

<Part of the learned Bayesian network structure>
ML Using Biology: Biocomputing
Bioinformation Technology (BIT)

Bioinformatics (in silico Biology)

Biocomputing (e.g. DNA Computing)
Biology as Inspiration: History

- **Neural Networks**
  - McCulloch & Pitts (1943)
  - Rosenblatt (1958)

- **Genetic Algorithms**
  - Fogel (1960s)
  - Rechenberg (1960s)
  - Holland (1975)

- **Artificial Life**
  - Langton (1988)

- **DNA Computing**
  - Adleman (1994)
Biology and Artificial Intelligence (AI)

Symbolic AI
- Rule-Based Systems

Connectionist AI
- Neural Networks

Evolutionary AI
- Genetic Algorithms

Molecular AI: DNA Computing
DNA Computing: Biology as Technology

011001101010001

ATGCTCGAAGCT

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Liquid-Phase 3D Biochemical Reaction as Computing

\[ \neg R \lor S \lor T \lor \neg Q \]

\[ Q \lor \neg T \lor \neg S \]

\[ \neg T \lor S \]

\[ T \lor P \]

\[ P \]

\[ R \lor \neg P \lor \neg Q \]

\[ \neg R \]

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Computing: Hybridization & Ligation

Hybridization

Ligation

Dehybridization

AGCTTACCTTAGGCT
AGCTTAGGATGGCATGGAATCCGATGCATGGC
AATCCGATGCATGGC
CGTACCTTTAGGCT
Solution Detection: Bead Separation

Magnetic Beads

Magnet

Complementary
DNA Computing: Bio-Lab Procedure

Node 0: ACG  Node 3: TAA
Node 1: CGA  Node 4: ATG
Node 2: GCA  Node 5: TGC
Node 6: CGT

Encoding

ATG
ACG  GCA  CGA
CGT  TGC  TAA

Ligation

... TAAACG ...
ATGTGCTAACGAACG
ACGGCAGCATAAAATGTGCACGCGT
... TAAACGGCAACG ...
ACGGGAGCATAAAATGTGCCGTACGCGAGCATAAATGTGCCGT
... CGACGTAGCCGT ...
ACGCGTAGCCGT

Gel Electrophoresis

Solution

ACGCGAGCATAAAATGTGCCGT
ACGGCAGCATAAAATGTGCACGCGT
ACGCGAGCATAAAATGTGCCGTACGCGAGCATAAATGTGCCGT

Decoding

Node 0: ACG  Node 3: TAA
Node 1: CGA  Node 4: ATG
Node 2: GCA  Node 5: TGC
Node 6: CGT

Affinity Column

... ACGCGTAGCCGT ...
ACGCGAGCATAAAATGTGCACGCGT
ACGCGAGCATAAAATGTGCCGTACGCGAGCATAAATGTGCCGT
... ACGCGT ...
ACGCGT

PCR (Polymerase Chain Reaction)

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Why DNA Computing?

- $6.022 \times 10^{23}$ molecules / mole
- **Immense, brute force search** of all possibilities
  - Desktop: $10^9$ operations / sec
  - Supercomputer: $10^{12}$ operations / sec
  - $1 \ \mu$mol of DNA: $10^{26}$ reactions
- **Favorable energetics**: Gibb’s free energy
  \[ \Delta G = -8 \text{kcal mol}^{-1} \]
- $1 \ J$ for $2 \times 10^{19}$ operations
- **Storage capacity**: 1 bit per cubic nanometer
## DNA Computers vs. Conventional Computers

<table>
<thead>
<tr>
<th>DNA-based computers</th>
<th>Microchip-based computers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Slow</strong> at individual operations</td>
<td><strong>Fast</strong> at individual operations</td>
</tr>
<tr>
<td>Can do <strong>billions of operations</strong> simultaneously</td>
<td>Can do substantially <strong>fewer operations</strong> simultaneously</td>
</tr>
<tr>
<td>Can provide <strong>huge memory</strong> in small space</td>
<td><strong>Smaller memory</strong></td>
</tr>
<tr>
<td>Setting up a problem may involve considerable <strong>preparations</strong></td>
<td>Setting up only requires <strong>keyboard input</strong></td>
</tr>
<tr>
<td>DNA is <strong>sensitive</strong> to chemical deterioration</td>
<td><strong>Electronic data are vulnerable</strong> but can be backed up easily</td>
</tr>
</tbody>
</table>

---

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
DNA Computing for AI and CogSci

- **Memory**
  - Associative memory with enormous density
  - $6 \times 10^{23}$ molecules / mole
  - 3 grams of water contains $10^{22}$ molecules

- **Inference**
  - Massive parallel reactive process
  - $10^{19}$ operations for Joule

- **Learning**
  - Generalization through biochemical reaction
  - 3-dimensional diffusion in solution
  - Global parameterization, e.g. temperature
Molecular Machine Learning

Molecular Learning Machines

Molecular Computing

Machine Learning

Bio-Memory

Bio-Processor

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Example: Concept Learning for a Janitor Robot

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>department</td>
<td>cs, ee</td>
</tr>
<tr>
<td>status</td>
<td>faculty, staff</td>
</tr>
<tr>
<td>floor</td>
<td>four, five</td>
</tr>
</tbody>
</table>

**Concepts**

- $h_1$: <faculty>
- $h_2$: <cs, faculty>
- $h_3$: <cs, faculty, four>

More general: $h_1 > g h_2 > g h_3$

**Examples:**

- $x_1$: <cs, faculty, four>+
- $x_2$: <ee, faculty, five>-
Bio-Lab Procedure

1. Sequence Design and Synthesis

2. Hybridization

3. Ligation

Initial Version Space

4. Learning (Affinity Separation)

5. Classification

---

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
<th>Tm (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs</td>
<td>5'- CTCGG TCGAA TTACC TCTAA ATAGC CTCGA -3'</td>
<td>65.2</td>
</tr>
<tr>
<td></td>
<td>3'- GAGGC AGCTT GATCG AGATT -5'</td>
<td>48.9</td>
</tr>
<tr>
<td>ee</td>
<td>5'- ACAAC GCCTC AGNAC CAAGC ATAGC CTCGA -3'</td>
<td>69.3</td>
</tr>
<tr>
<td></td>
<td>3'- TTTCC CGGAG TCTTG GTTAC -5'</td>
<td>53.7</td>
</tr>
<tr>
<td>staff</td>
<td>5'- ATGAT GAGTC ATCTG TCGCA ATAGC CTCGA -3'</td>
<td>66.9</td>
</tr>
<tr>
<td></td>
<td>3'- TAGTA CATCG TCGAC AGCCG -5'</td>
<td>50.0</td>
</tr>
<tr>
<td>faculty</td>
<td>5'- CGTGA TTGGG TTGGC GCTAA CAAGC GGGGC -3'</td>
<td>77.1</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>70.3</td>
</tr>
<tr>
<td>staff</td>
<td>5'- CGTGA TTGGG TTGGC GCTAA CAAGC GGGGC -3'</td>
<td>73.7</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>66.9</td>
</tr>
<tr>
<td>four</td>
<td>5'- ATCGG GTATG GGTGA GCTTT CAAGC GGGGC -3'</td>
<td>74.0</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>66.9</td>
</tr>
<tr>
<td>five</td>
<td>5'- CGTGA TTGGG TTGGC GCTAA CAAGC GGGGC -3'</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>76.5</td>
</tr>
<tr>
<td>five</td>
<td>5'- CGTGA TTGGG TTGGC GCTAA CAAGC GGGGC -3'</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>79</td>
</tr>
<tr>
<td>seven</td>
<td>5'- GAGTG TCAGG ATCTG AAGCT -3'</td>
<td>48.5</td>
</tr>
<tr>
<td></td>
<td>3'- GTAGG GAGCT TCAGT CAAGC ACTCG GCTCT - 5'</td>
<td>75.2</td>
</tr>
</tbody>
</table>
1. Generating the Concept Space

- One DNA molecule for one hypothesis
- Sticky ends and “don’t care symbols”
2. Given a Positive Example:

<cs, faculty, four>+

Given a positive example, select all hypotheses that are consistent with the example (and don’t cares)

A ↔ A ∩ B

Affinity separation by magnetic beads
3. Given a Negative Example:

Given a negative example, **filter out all hypotheses that are consistent with the example**

![Diagram](attachment:image.png)

Affinity separation by magnetic beads
4. Given an Unknown Example:

\[ \text{<cs, staff, four>?} \]

- Compute \( Y = A \cap B \) and \( N = A - B \)
- Answer \textit{yes} if \( |Y| > |N| \), answer \textit{no}, otherwise.

\[ Y \leftarrow A \cap B \]
\[ N \leftarrow A - B \]

This example is negative!
Experimental Results (1)

- **Generation of the initial concept space**
  - Sequence design: H-measure, similarity, and Tm
  - Attribute: 20 mer
  - Sticky end: 10 mer
  - Total: 80 mer (3 attributes per hypothesis)

**Gel Electrophoretogram for hybridization & ligation**

- 100 bp
- 75 bp
- 50 bp
- 25 bp

(c) 2002 SNU Biointelligence Lab, http://bi.snu.ac.kr/
Experimental Results (2)

- Learning from the first (positive) example
  - \( <cs, \text{faculty}, \text{four}> + \)
  - \( <cs, \text{faculty}, \text{four}>, <cs, \text{faculty}, ?>, <cs, ?, \text{four}>, <?, \text{faculty}, \text{four}>, <cs, ?, ?>, <?, \text{faculty}, ?>, <?, ?, \text{four}>, <?, ?, ?> \)

M : 25bp marker (KDR)
- lane 1: cs / 4
- lane 2: cs / ?_{floor}
- lane 3: ?_{dept} / 4
- lane 4: ?_{dept} / ?_{floor}
- lane 5: ee / 5
- lane 6: No Primers
- lane 7: ligation mixture
- lane 8: sup. after cs/ ?_{dept}
- lane 9: sup. after faculty / ?_{stat}
- lane 10: sup. after 4/ ?_{floor}
Experimental Results (3)

- Learning from the second (negative) example
  - \(<\text{cs, staff, five}> -\)
    - \(<\text{cs, faculty, four}>, <\text{cs, faculty, } >, <\text{cs, }?, \text{four}>
    - \(<?, \text{faculty, four}>, <?, \text{faculty, } >, <?, ?, \text{four}>

- Answering to the query (unknown example)
  - \(<\text{cs, staff, four}> ?\)
    - (+): \(<\text{cs, }?, \text{four}>, <?, ?, \text{four}>
    - (-): \(<\text{cs, faculty, four}>, <\text{cs, faculty, } >, <?, \text{faculty, four}>, <?, \text{faculty, } >\)
    - Should be classified as “negative”
    - Result
      - \((+): (-) = 0.762 : 1.134\)
Experimental Results (4)

(+): <cs, ?, four>, <?, ?, four>

(-): <cs, faculty, four>, <cs, faculty, ?>, <?, faculty, four>, <?, faculty, ?>

M: 25 bp marker (KDR)
lane p1: cs / 4
lane p2: ?dept / 4
lane p3: ee / 5
lane p4: No Primers
lane n1: cs / 4
lane n2: cs / ?floor
lane n3: ?dept / 4
lane n4: ?dept / ?floor
lane n5: No Primers

Figure 3. PCR product of majority voting confirmed by 3% agarose gel electrophoresis
Conclusions and Outlook

- Bio-inspired ML: Machine learning models have been inspired by biological systems.
  - Neural networks from central nervous systems
  - Genetic algorithms from natural selection

- ML for Biology: Bio-inspired machine learning techniques are applied back to study biology.
  - Neural networks for protein structure analysis
  - Genetic algorithms for sequence alignment

- Biotechnical ML: The resulting advancement in biotechnology will further develop the machine learning technology, not just in models but in substrates as well.
  - DNA-based inductive machine learning
  - Learning to diagnose diseases from natural DNA
Acknowledgements

Collaborating labs

서울대 바이오지능 & 인공지능 연구실
서울대 세포 및 미생물공학 연구실
한양대 프로테오믹스 연구실
㈜바이오니아 & ㈜바이오인포메틱스

Supported by

과기부 국가지정연구실사업
산자부 차세대신기술연구개발사업

More information at

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