Intelligent Systems and Computational Biology

Richard H. Lathrop
Dept. of Computer Science

rickl@uci.edu
Donald Bren Hall 4224
949-824-4021
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

**Goal of talk:** The power of information science to influence molecular science and technology
Goal of talk: The power of information science to influence molecular science and technology

“Computers are to Biology as Mathematics is to Physics.”

--- Harold Morowitz
(spiritual father of BioMatrix, and Intelligent Systems for Molecular Biology Conference)
Biology has become Data Rich

- Massively Parallel Data Generation
  - Genome-scale sequencing
  - High-throughput drug screening
  - Micro-array “gene chips”
  - Combinatorial chemical synthesis
  - “Shotgun” mutagenesis
  - Directed protein evolution
  - Two-hybrid protocols for protein interaction
  - Half a million biomedical articles per year
“Data Rich”
Genomic sequence data
“Data Rich”
Protein 3D structure data

Protein Databank
Content Growth
“Data Rich”
Biomedical literature
“Data Rich”
10-100K data points per gene chip
Characteristics of Biomedical Data

- **Noise!!**
  - => need robust analysis methods
- **Little or no theory.**
  - => need statistics, probability
- **Multiple scales, tightly linked.**
  - => need cross-scale data integration
- **Specialized ("boutique") databases**
  - => need heterogeneous data integration
Intelligent Systems are well suited to biology and medicine

- Robust in the face of inherent complexity
- Extract trends and regularities from data
- Provide models for complex processes
- Cope with uncertainty and ambiguity
- Content-based retrieval from literature
- Ontologies for heterogeneous databases
- Machine learning and data mining

- Intelligent systems handle complexity with grace
Intelligent Systems and Computational Biology

**Artificial Intelligence for Biology and Medicine**
- Biology is data-rich and knowledge-hungry
- AI is well suited to biomedical problems

**Examples**
- Machine learning -- drug discovery
- Rule-based systems -- drug-resistant HIV
- Heuristic search -- protein structure prediction
- Constraints -- design of large synthetic genes

**Current Project**
- Machine learning and p53 cancer rescue mutants

**Goal of talk:** The power of information science to influence molecular science and technology
Drug Discovery Background

- Cost and time, per drug, start to finish
  - US$500 million to US$1 billion
  - 6 to 12 years
  - High failure rate

- “Blockbuster” drugs
  - Revenues > US$1 billion/year
  - Profits > US$1 million/day
  - Blood pressure, ulcers, ....

- A risky business....
Positive Examples | Negative Examples

O

O

O

O

O

O

O

O

O

O

O

O

O

O
Digital 3D Shape Representation
The Power of a Good Representation
Learning the “Multiple Instance” Problem

“Solving the multiple instance problem with axis-parallel rectangles”
Dietterich, Lathrop, Lozano-Perez, Artificial Intelligence 89(1997) 31-71
New Start-up Company:
Arris Pharmaceutical Corp.

- Started a Venture Pharmaceutical Company
  - Apply machine learning to drug discovery
- Became a Publicly Traded Company (1993)
  - Value of $60 Million, 59 Full-time Employees
- Merged with Sequana Pharmaceuticals (1998)
  - Merger result became Axys Pharmaceuticals
- Finally bought by Celera Genomics (2001)
  - Purchase price $188 Million
  - Became Celera Therapeutics
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - **Rule-based systems** – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

- **Goal of talk**: The power of information science to influence molecular science and technology
Knowledge-based avoidance of drug-resistant HIV mutants

- Physician’s advisor for drug-resistant HIV
- Rules link HIV mutations & drug resistance
- Rules extracted from literature manually
- Patient’s HIV is sequenced
- Rules identify patient-specific resistance
- Rank approved combination treatments
Input/Output Behavior

**INPUT = HIV SEQUENCES FROM PATIENT:**
- 5 HIV clones (clone = RT + PRO)
- = 5 RT + 5 PRO (RT = 1,299; PRO = 297)
- = 7,980 letters of HIV genome

**OUTPUT = RECOMMENDED TREATMENTS:**
- 12 = 11 approved drugs + 1 humanitarian use
- Some drugs should not be used together
- 407 possible approved treatments
Example Patient Sequence of HIV Reverse Transcriptase (RT)

CCA GTA AAA TTA AAG CCA GGA ATG GAT GGC CCA AAA GTT AAA CAA TGG CCA CCC ATT AGC CCT ATT GAG ACT GTA TTG ACA GAA GAA AAA ATA AAA GCA TTA GTA GAA ATT TGT ACA GAG ATG GAA AAG GAA GGG *AA ATT TCA AAA ATT GGG CCT GAA AAT CCA TAC AAT ACT CCA GTA TTT GCC ATA AAG AAA AAA GAC AGT ACT AAA TGG AGA AAA TTA GTA GAT TTC AGA GAA CTT AAT AAG AGA ACT CAA GAC TTC TGG GAA GTT CAA TTA GGA ATA CCA CAT CCC GCA GGG TAA AAA AAG AAA AAA TCA GTA ACA GTA CTG GAT GTG GGT GAT GCA TAT TTT TCA GTT CCC TTA GAT GAA GAC TTC AGG AAG TAT ACT GCA TTT ACC ATA CCT AGT ATA AAC AAT GAG ACA CCA GGG ATT AGA TAT CAG TAC AAT GTG CTT CCA [CAG] GGA TGG AAA GGA TCA CCA GCA ATA TTC CAA AGT AGC ATG ACA AAA ATC TTA GAG CCT TTT AGA AAA AAT GTA CCA GAC ATA TTG AGT ATT TCT TAC TTT GAA TGG AAT TGG GTT AAG AAT CAG ACA GAA AAA CAT CAG AAA GAA CCT CCA TCC TTT AGG ATG GGT TAT GAA CTC CCT GAT AAA TGG ACA GTA CAG CCT ATA GTG CTG CCA GAA AAA GAC AGC TGG ACT GTG AAT GAC ATA CAG AAG TTA GTG GGG AAA TTG AAT TGG GCA AGT CAG ATT TAC CCA GGG ATT AAA GTA AGG CAA TTA TGT AAA CTC CTT AGA GGA ACC AAA GCA CTA ACA GAA GTA ATA CCA CTA ACA GAA GCA GCA GAG TCA GAA AAT GTA CCA CAA GAG ATT CTA TAA GAA CAA GTA CAT GGA GTG TAT TAT GAC CCA TCA AAA GAC TTA ATA GCA GAA ATA CAG AAG CAG GGG CAA GGC CAA TGG ACA TAT CAA ATT TAT CAA GAG CCA TTT AAA AAT CTG AAA ACA GGA AAA TAT GCA AGA ATG AGG GTG GCC CAC ACT AAT GAT GTA AAA CAA ATA ACA GAG GCA GTG CAA AAA ATA ACC ACA GAA AGC GTA ATG TTA TGG TGA AAG ACT CCT AAA TTT AAA CTG CCC ATA CAA AAG GAA ACA TGG GAA ACA TGG TGG ACA GAG TAT TGG CAA GCC ACC TGG ATT CCT GAG TGG GAG TTT GTT AAT ACC CCT CCC ATA GTG AAA TTA TGG TAC CAG TTA GAG AAA GAA CCC

The bracketed codon [CAG] causes strong resistance to AZT.
Rules represent knowledge about HIV drug resistance

IF <antecedent> THEN <consequent>
[weight] (reference).

IF RT codon 151 is ATG
THEN do not use AZT, ddI, d4T, or ddc.
[weight=1.0] (Iversen et al. 1996)

The weight is the degree of resistance, NOT a confidence or probability.
“Knowledge-based avoidance of drug-resistant HIV mutants”

Lathrop, Steffen, Raphael, Deeds-Rubin, Pazzani

Innovative Applications of Artificial Intelligence Conf.
Madison, WI, USA, 1998
“Knowledge-based avoidance of drug-resistant HIV mutants”

Lathrop, Steffen, Raphael, Deeds-Rubin, Pazzani, Cimoch, See, Tilles

AI Magazine 20(1999)13-25
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - **Heuristic search -- protein structure prediction**
  - Constraints – design of large synthetic genes

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

- **Goal of talk:** The power of information science to influence molecular science and technology
Protein structure prediction

TQVAKKLVTCLPYPANGSILHGMLEHIADADVWVRYQRMRG
HEVNCICADDAHGTKPAQGIQLPGEOMEMSQEHQTDFA
AGFNSYDYHSTHSENRQLSELYSRKENGFIKNTISQLY
DPEKGMFLPDVFKGTCPCKSPDPQYGDNCEVCAGATYSPTEL
IEPKSVVSGATPVMRDSHEFDFLDPSFSEMLOAATRSGALQEG
VANKMQEFESQLOQWDISRDPFPFGFIEPAPGKYFYLYLD
APIGYMGSKRENLEDKRGDSVSFDEYWKDOSTAELYHFIGKDI
VYFHSLLFWPMLEGSNFRIKPSNLFTVHYVTVVNGAKMSKSRGT
FIKASTWLNHFADSLRYYTAKLSSRIDDIIIDNLLEDVFQVN
ADINVKVNLASRNAGFINKRF DGVLASDELAPQLYKRFTDA
AEVIGEAWRESERFGKAVREIMALA DLANRYVDEQAP WVVAK
QEGRDADLQAI AQWGINLFRVLMTYLKPVLPKLTERAEFLN
TEL TWGIOQPLLGHKVNPFKALYNRIMRQVEA LVEASKEE
VKAAAAPVTGPLADDPODGCHRDRVVDGSK
“Protein Threading”

(A) Two Structurally Similar Proteins and a Core of Four Segments

(C) One Possible Threading with a Novel Sequence

(D) This Set of Possible Threadings Will Be Split Into Three Subsets at Segment I
“Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions”

Lathrop and Smith

Multi-queue Branch-and-Bound

- Originally developed for protein structure prediction by “protein threading”
- Abstracted to a general-purpose method
  - Have also run on SAT, Traveling Salesman
  - Small molecule conformation search, DNA motif discovery, DNA nanotechnology, synthetic gene design and assembly

- Difficult applications inspire new techniques. New techniques enable other applications.
Optimal Exponent < Proof Exponent

Small organic molecule conformation search

Protein-DNA binding motif search

X axis = log10(search space size), Y axis = log10(time in seconds)
“x” = time to best result, “o” = time to optimal, “*” = search abandoned
“A multi-queue branch-and-bound algorithm for anytime optimal search with biological applications”

Lathrop, Sazhin, Sun, Steffen, Irani

Genome Informatics
12(2001)73-82

GIW’2001, Tokyo, Japan
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

- **Goal of talk**: The power of information science to influence molecular science and technology
Current Application: Synthetic gene design

- Constraint-based search through sequence space
- Constraints guarantee correct self-assembly and other desirable properties
- Underlying domain-specific computation done on a 64-node cluster of 3GHz Xeon dual processors
Assembly of Integrase Gene (1640 bp)
Problem: Melting Temperatures of correct and incorrect hybridization assemblies overlap

- CODA

**solid:** correct overlap of oligos
**dashdot:** correct overlap of intermediate fragments
**dashed:** incorrect overlap of small oligos
**dotted:** incorrect overlap of intermediate fragments

Result: Intermediate assemblies contain multiple fragments and error products (smears above)
Solution: Fragments designed with non-overlapping melting temps

solid: correct overlap of oligos
dashdot: correct overlap of intermediate fragments
dashed: incorrect overlap of small oligos
dotted: incorrect overlap of intermediate fragments

+ CODA

Result: Single products assembled correctly.
In vivo Expression of Gag protein in E. coli

Confirmed by Western blot and sequencing
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

- **Goal of talk:** The power of information science to influence molecular science and technology
p53 is a central tumor suppressor protein
“The guardian of the genome”

Controls apoptosis (programmed cell death) and cell cycle arrest
Monitors cellular distress

The most-mutated gene in human cancers
All cancers must disable the p53 apoptosis pathway.

p53 core domain bound to DNA

Image generated with UCSF Chimera

Cho, Y., Gorina, S., Jeffrey, P.D., Pavletich, N.P.
Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations.
*Science* v265 pp.346-355, 1994
Consequences of p53 mutations

~250,000 US deaths/year

Loss of DNA contact  Disruption of local structure  Denaturation of entire core domain

Over 1/3 of all human cancers express full-length p53 with only one a.a. change

A Long-held Goal of Anti-cancer Therapy

Restore integrity of p53 core domain

Restore apoptosis (cell death) in tumor cells
Suppressor Mutations

Several second-site mutations restore functionality to some p53 cancer mutants \textit{in vivo}.
PREDICT

Cancer Mutant

Rescued Cancer Mutant

R158L

S240R
Machine Learning: Predict phenotype from genotype

A functional census of p53 “cancer rescue” mutants

=> Catalog *all* Actives
Class Labels: Active/+ or Inactive/-

p53 Transcription Assay

Initial: Yeast Growth Selection, Sequencing

**ACTIVE (+)**

Will grow.

- Wild-type
- P152L(N)
- P152L_q100i(S)
- P152L_y103c(S)
- P152L_k101n(W)
- P152L_s106p(S)
- P152L_q100s(S)
- P152L_q100t(S)

(S) = Strong
(W) = Weak
(N) = Negative

**INACTIVE (-)**

Will not grow.

- Human p53 consensus
- URA-

Confirm: Human 1299 Cell-based Luciferase

First measurement
Firefly luciferase p53 dependent

Second measurement
Renilla luciferase p53 independent

Baroni, T.E., et al., 2004
Danziger, S.D., et al., 2009
Baronio, R., et al., 2010
Multi-dimensional Features

2D Surface Map

3D Structural Changes

4D Unfolding Trajectory

Active Machine Learning for Biological Discovery

Find New Cancer Rescue Mutants

Knowledge

Theory

Experiment
Known Mutants: 16,722
Known Actives: 143

Assuming up to 5 mutations in 200 residues
How Many Mutants are There?: $\sim 10^{11}$

Spiral Galaxy M101
http://hubblesite.org/
$\sim 10^9$ stars.
Computational Active Learning

Pick the Best (= Most Informative) Unknown Examples to Label

Train the Classifier

Choose Examples to Label

Add New Examples To Training Set

Known
Example 1
Example 2
Example 3
...
Example N

Training Set

Unknown
Example N+1
Example N+2
Example N+3
...
Example M

Classifier

Example 2

Example 3
Example of Active Learning:
Minimum Marginal Hyperplane

Should unknown Mutant 1 or Mutant 2 be added to the training set?

Select Mutant 2
Example of Active Learning: Maximum Curiosity

Should Mutant 1 or Mutant 2 be added to the training set?

Training Set

<table>
<thead>
<tr>
<th>Training Set + Mutant 1 (Active)</th>
<th>Cross-validator</th>
<th>.0411</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set + Mutant 1 (Inactive)</td>
<td>Cross-validator</td>
<td>-.6014</td>
</tr>
<tr>
<td>Training Set + Mutant 2 (Active)</td>
<td>Cross-validator</td>
<td>.0309</td>
</tr>
<tr>
<td>Training Set + Mutant 2 (Inactive)</td>
<td>Cross-validator</td>
<td>.0276</td>
</tr>
</tbody>
</table>

Select Mutant 1
Which is the Best Active Learning Method?

**TYPE I:** Select mutants that most improve the classifier if correctly predicted.
- Maximum Curiosity
- Composite Classifier
- Improved Composite Classifier

**TYPE II:** Select mutants that most improve the classifier.
- Additive Curiosity
- Additive Bayesian Surprise

**TYPE III:** Common methods taken from the literature.
- Minimum Marginal Hyperplane
- Maximum Entropy

**TYPE IV:** Variations on methods from the literature.
- Maximum Marginal Hyperplane (negative control)
- Minimum Entropy (negative control)
- Entropic Tradeoff

**TYPE C:** Controls
- Non-iterated Prediction
- Predict All Inactive
- Random (30 trials)
The Active Learning Tradeoff: How Fast Does It Learn?

Forward Prediction Accuracy (25 Predicts 236)

Danziger, et al., ISMB 2007 (also Bioinformatics 2007)
Blind *in vivo* Trial of MIP On Three p53 Regions

- **MIP Positive Region:**
  - Predicted to be *Informative & Many Positive Mutants*

- **MIP Negative Region:**
  - Predicted to be *Informative & Few Positive Mutants*

- **Expert Region:**
  - Selected by a *Human Cancer Biologist*

- **No previous single-a.a. rescue mutants in any region**

Danziger, et al. (2009)
Visualization of Selected Regions

- **Positive Region:**
  Predicted Active
  96-105 (Green)

- **Negative Region:**
  Predicted Inactive
  223-232 (Red)

- **Expert Region:**
  Predicted Active
  114-123 (Blue)

Danziger, et al. (2009)

<table>
<thead>
<tr>
<th></th>
<th>MIP Positive (96-105)</th>
<th>MIP Negative (223-232)</th>
<th>Expert (114-123)</th>
</tr>
</thead>
<tbody>
<tr>
<td># Strong Rescue</td>
<td>8</td>
<td>0 (p &lt; 0.008)</td>
<td>6 (not significant)</td>
</tr>
<tr>
<td># Weak Rescue</td>
<td>3</td>
<td>2 (not significant)</td>
<td>7 (not significant)</td>
</tr>
<tr>
<td>Total # Rescue</td>
<td>11</td>
<td>2 (p &lt; 0.022)</td>
<td>13 (not significant)</td>
</tr>
</tbody>
</table>

P-Values are two-tailed, comparing Positive to Negative and Expert regions. Danziger, et al. (2009)

**No significant differences between the MIP Positive and Expert regions.**

**Both were statistically significantly better than the MIP Negative region.**

**The Positive region rescued for the first time the cancer mutant P152L.**

**No previous single-a.a. rescue mutants in any region.**
Mutations Rescue Cancerous p53

Wild Type
Active p53

Cancer Mutation
Inactive p53

Cancer+Rescue Mutations
Active p53

Ultimate Goal

Cancer Mutation
Inactive p53

Anti-Cancer Drug

Active p53
The long road to a future anti-cancer drug

Peter Kaiser
Rommie Amaro
Dick Chamberlin
Melanie Cocco
Hudel Luecke
Wes Hatfield
Chris Wassman
Roberta Baronio
Ozlem Demir
Faezeh Salehi
Edwin Vargas
Da-Wei Lin
Intelligent Systems and Computational Biology

- Artificial Intelligence for Biology and Medicine
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- Examples
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes
  - DNA nanotechnology and space-filling DNA tetrahedra

- Current Project
  - Machine learning and p53 cancer rescue mutants

Goal of talk: The power of information science to influence molecular science and technology
The spectrum of \textit{p53} mutations

IARC \textit{TP53} Mutation Database (R7, Sep. 2002):
- 17,689 somatic mutations
- 12,631 (71\%) missense mutations affecting codons 100-300
- 974 different amino acid changes in core domain
Rescued Cancer Mutant

Mutant Cancer Mutant Suppressor Mutant

Wildtype p53

Molecular Models

\{ 4D trajectory data \\
3D structure data \\
2D surface data \\
1D sequence data \} Machine Learning
Intelligent Systems and Computational Biology

- **Artificial Intelligence for Biology and Medicine**
  - Biology is data-rich and knowledge-hungry
  - AI is well suited to biomedical problems

- **Examples**
  - Machine learning -- drug discovery
  - Rule-based systems – drug-resistant HIV
  - Heuristic search -- protein structure prediction
  - Constraints – design of large synthetic genes
  - DNA nanotechnology and space-filling DNA tetrahedra

- **Current Project**
  - Machine learning and p53 cancer rescue mutants

- **Goal of talk:** The power of information science to influence molecular science and technology
3D DNA Nanostructures

Christopher D. Wassman
UC Irvine
Dept. of Computer Science
Why DNA Nanotechnology

- DNA has an well understood 3D structure
- DNA is easily synthesized and manipulated
- DNA Feature Sizes:
  - 3.6 nm per helical rise,
  - 2 nm helical width
- Intel Feature Sizes:
  - Current chips, 45nm feature size
  - Research chips, 32nm feature size (Sept, 2008)
- Bio-Nanotechnology is a emerging field
  - Lots to do, and lots of fun to be had!
Tiling 3-Space

- A familiar concept
  - Building blocks
  - Cubes fill space
  - Cylinders do not

- Other building blocks are possible
  - We will focus on tetrahedral building blocks, constructed by “folding DNA”
Irregular Tetrahedra…

Can Tile 3-Space Completely!
Full Tetrahedron
A Closer Look
Atomic Force Microscopy (AFM)
Experimental AFM Image
Simulated AFM Image

The image shows a simulated AFM (Atomic Force Microscope) image with a color scale ranging from 0 to 35 nanometers. The image features three distinct points on the X-axis (nanometers) and Y-axis (nanometers).
Experimental AFM Image