Machine Learning and Genetic Microarrays

Jude Shavlik & David Page
University of Wisconsin-Madison

Goals

- Learn about **microarray technology**
- See some ML problem formulations in the computational-biology literature
- A little on experimental pitfalls to avoid
- Overviews of newest "high-throughput" molecular-level data being gathered
- Very little on the details of machine learning algorithms

Outline

- Molecular Biology and Microtechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

Outline

- Molecular Biology and Nanotechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

Combining Three “Hot” Technologies

- Information Technology
- Biotechnology
- “Nanotechnology”

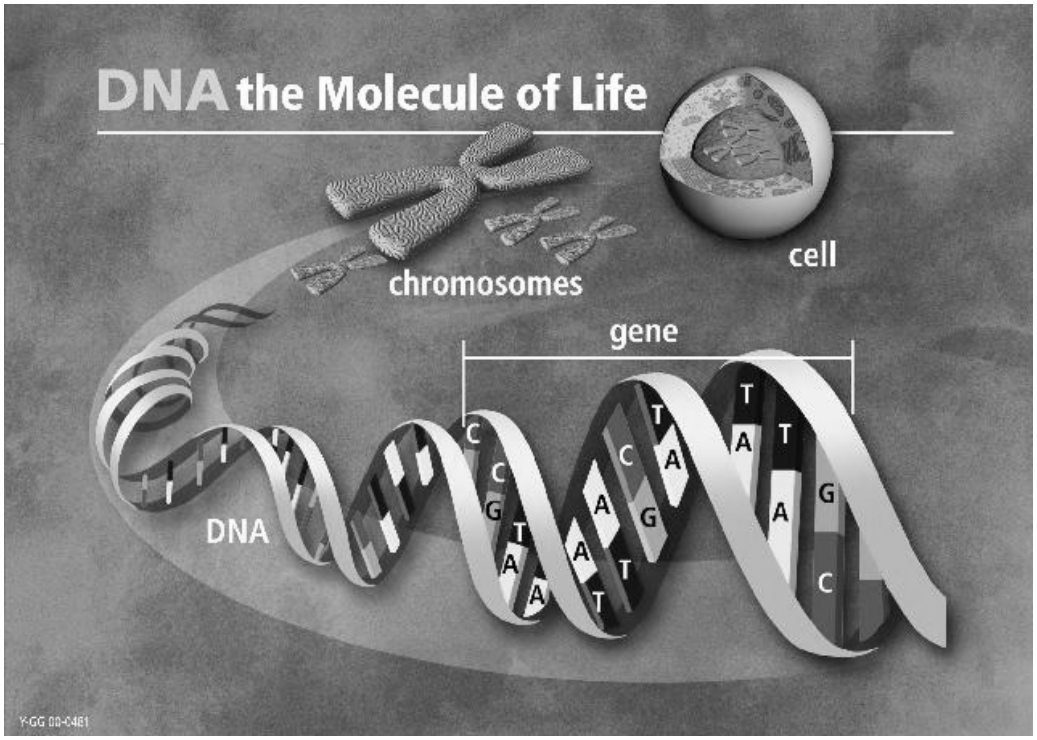
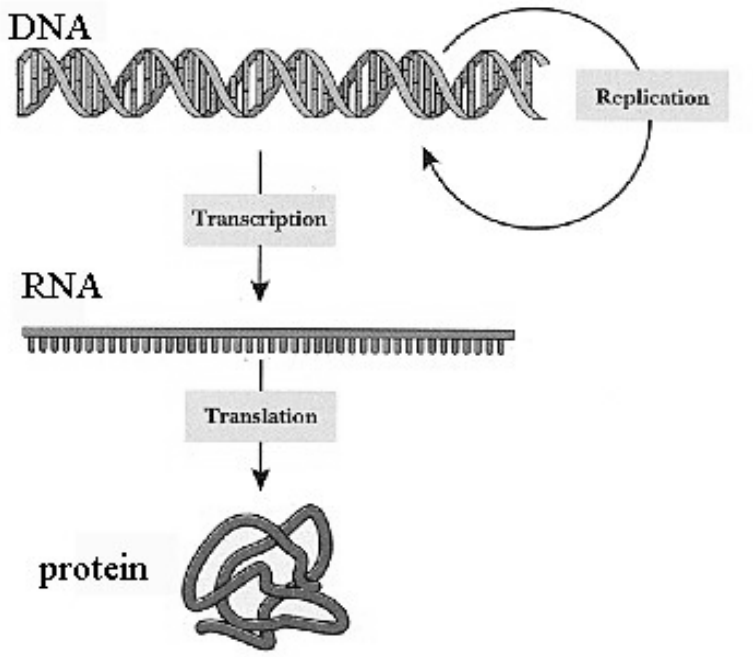


image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

The "Central Dogma" of Mol Bio

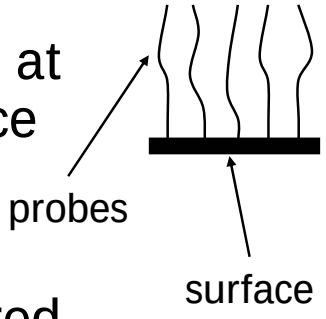


The Big Picture

- We'd like to know which **proteins** are present in a given type of cell, under some conditions, etc
  - eg, cancerous vs. non-cancerous cells
- However, currently can only measure **RNA** well

Microarrays ("Gene Chips")

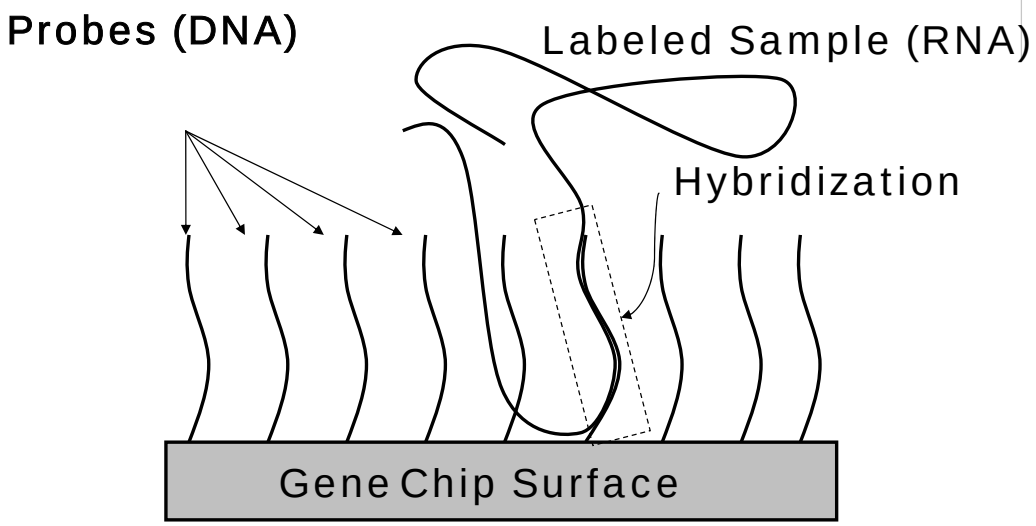
- Specific probes synthesized at known spot on chip's surface
- Probes complementary to RNA of genes to be measured
- Typical gene (1kb+) MUCH longer than typical probe (24 bases)



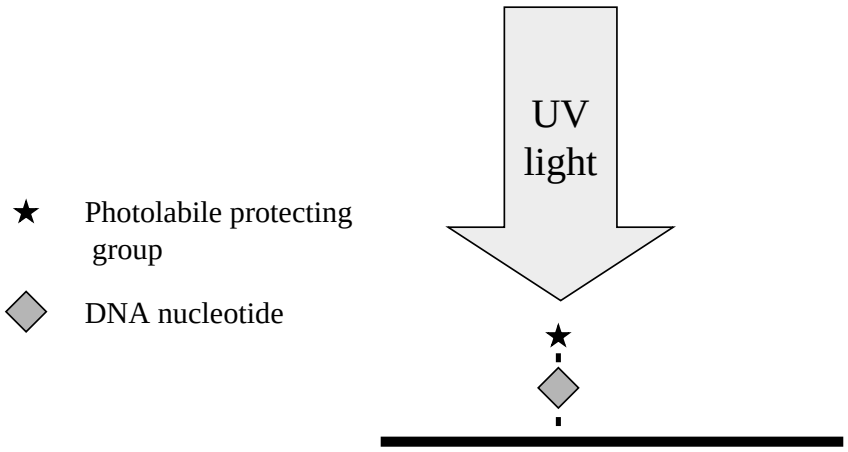
Microarray Technologies

- Alternate technologies exist (eg, "spotted arrays")
- We won't cover them all
- We'll focus on the "market leader" (Affymetrix)

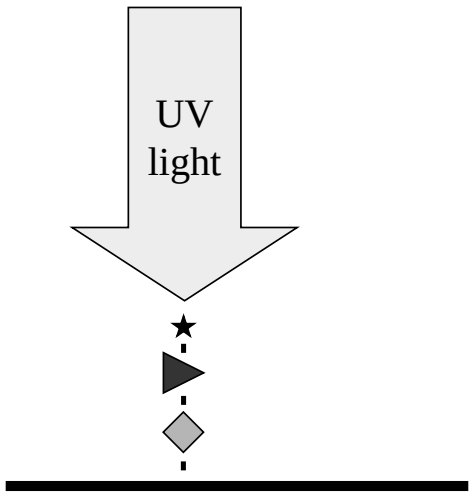
How Microarrays Work



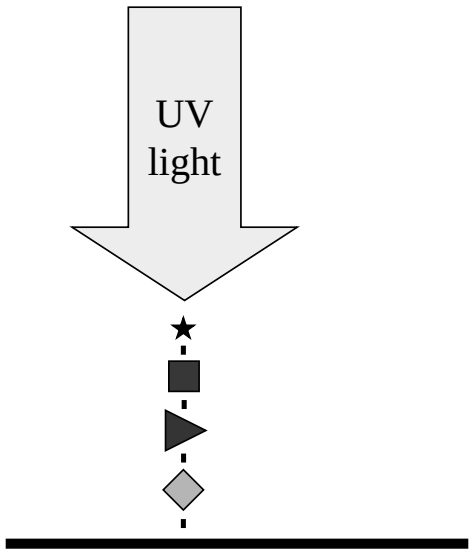
DNA Synthesis can be Controlled by Light



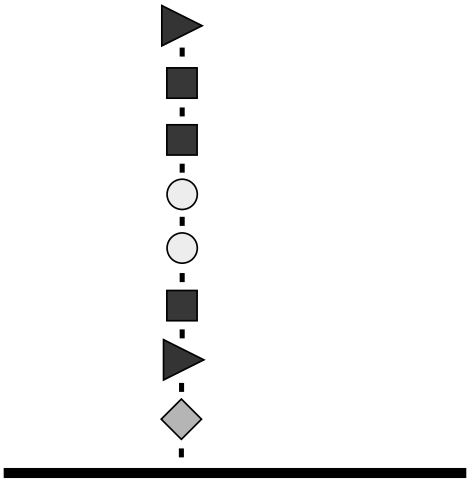
DNA Synthesis can be Controlled by Light (cont.)



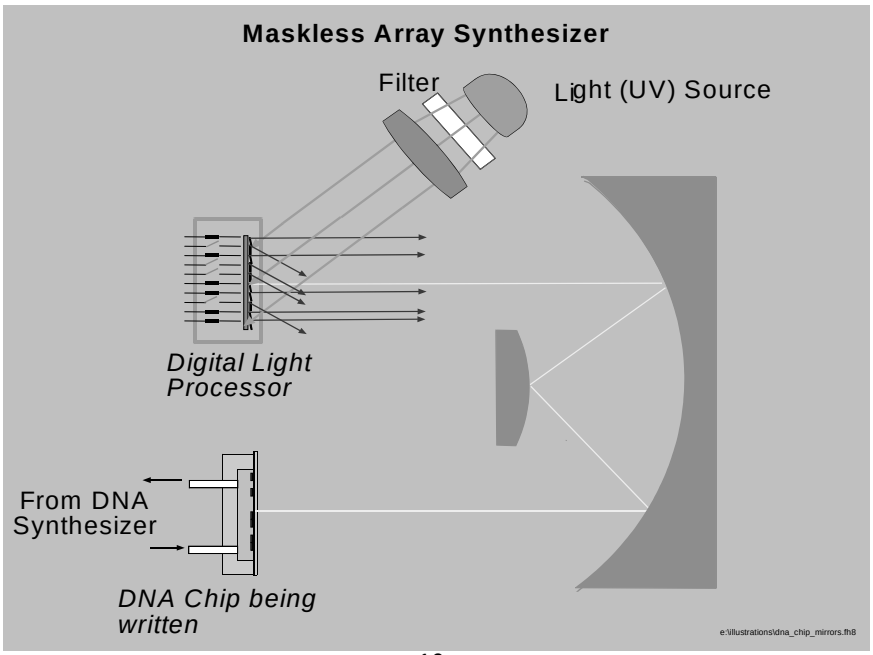
DNA Synthesis can be Controlled by Light (cont.)



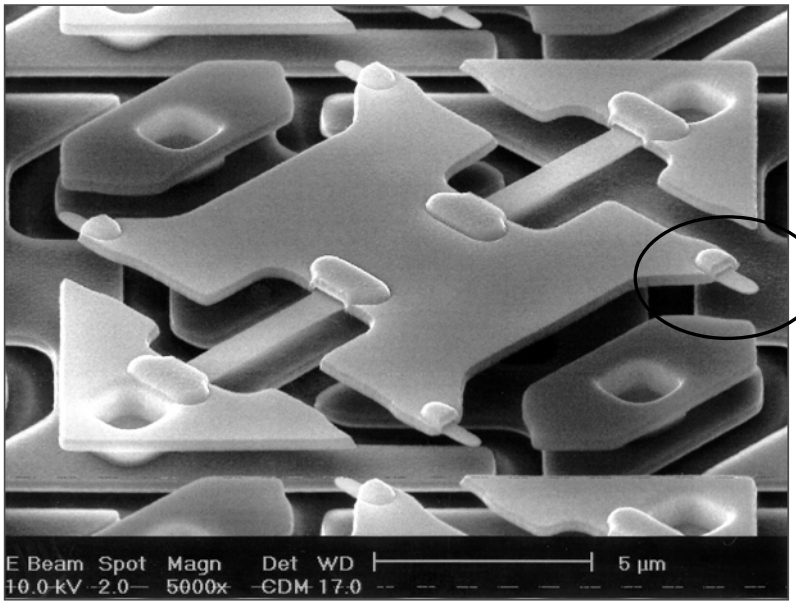
DNA Synthesis can be Controlled by Light (concluded)



Example: NimbleGen, Inc



Micro Mirrors from TI



From Probes Back to Genes

- Need algorithm for converting measured probe intensities into gene-expression levels
- Could simply use average probe value
- More (too?) complicated approaches exist (eg, Affymetrix')

Cleaning Up the Data – "Controlling Variance"

- Often look at *relative* expression levels

  Measurement(cancerCell) / Measurement(normalCell)

- Often correct for small values

  MeasurementToUse(Gene1)
  = ActualMeasurement(Gene1) + Constant

- Often use a mismatch ("near miss") probe

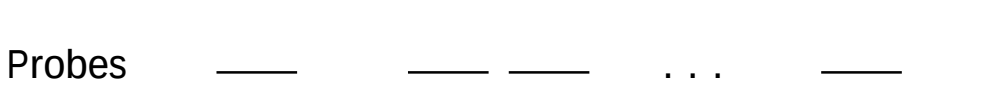
  MeasurementToUse(Probe1)
  = ActualMeasurement(Probe1)
  – ActualMeasurement(MismatchProbe1)

Outline

- Molecular Biology and Microtechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

The Probe-Selection Task

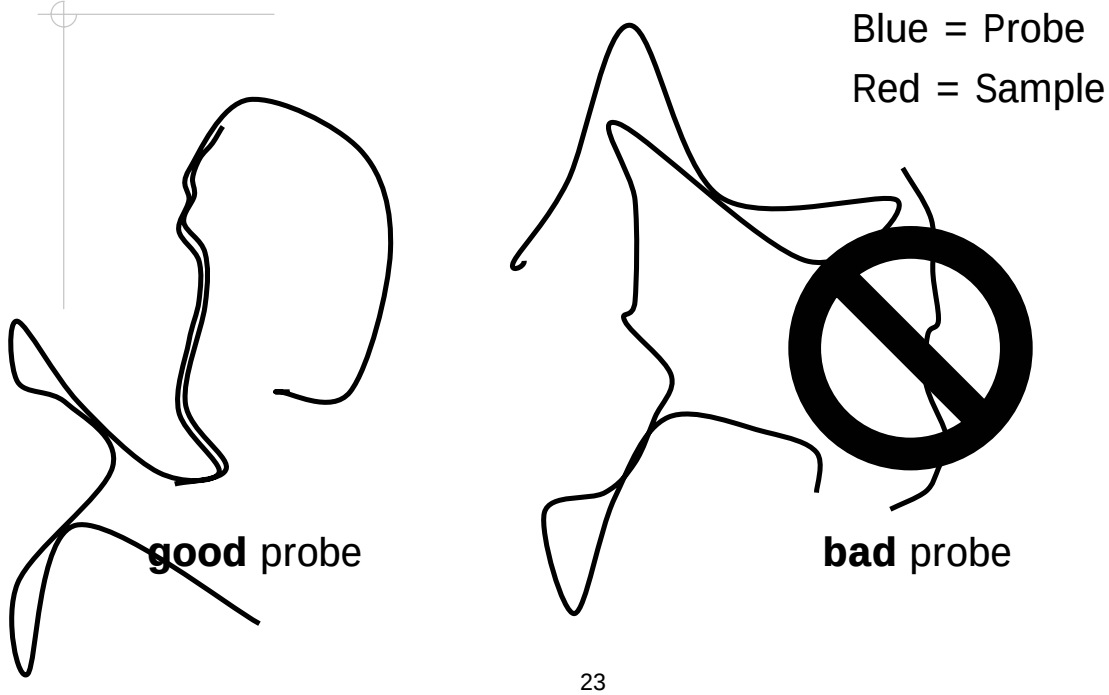
- Need to pick 5-15 probes for each gene we want to monitor
  - Recall, genes about 1000 “bases” long
  - Can only create probes about 24-bases long



Goals

- Probes should bind tightly to target
- Probes should not bind well to other mRNAs ... *cross-hybridization* should be rare

Probes: Good vs. Bad



Supervised Learning Task 1

- Given: probes as examples

  DNA sequence features describe each example; class is **good** or **bad**

  Probe is **good** if it binds tightly to target and **bad** otherwise

- Do: learn model that accurately predicts probe-quality class of new probes

The Data

- Tilings of 8 genes (from *E. coli* & *B. subtilus*)
  - Every possible probe (~10,000 probes)
  - Genes known to be “expressed” in sample

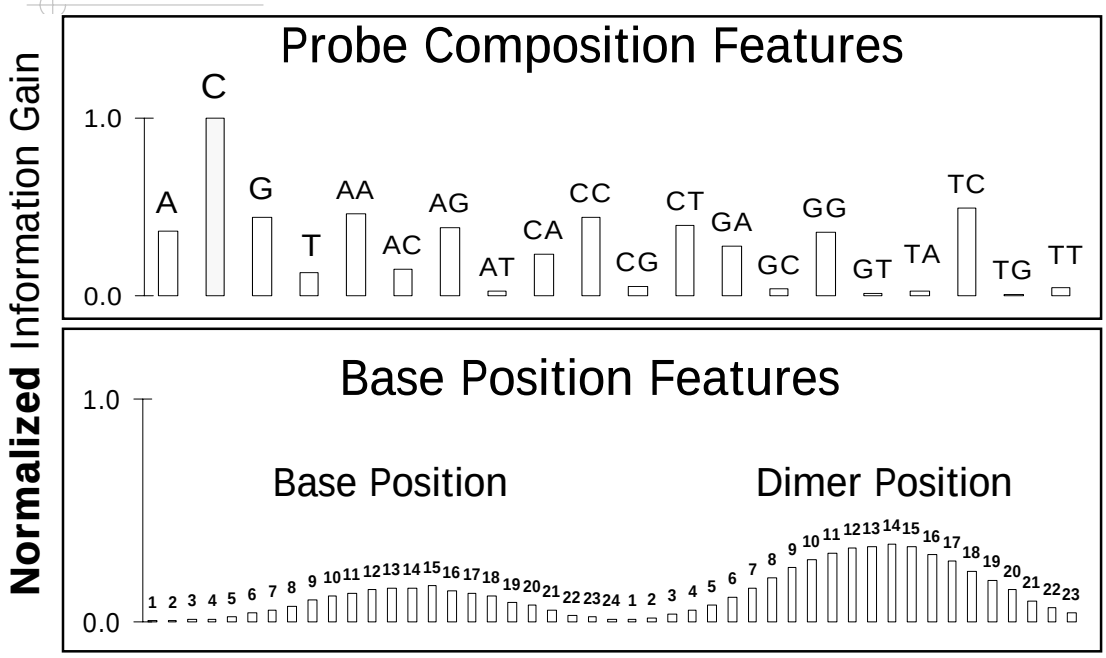
Gene Sequence:	**GTAGCTAGCATTAGCATGGCCAGTCATG…**
Complement:	**CATCGATCGTAATCGTACCGGTCAGTAC…**
Probe 1:	CATCGATCGTAATCGTACCGGTCA
Probe 2:	ATCGATCGTAATCGTACCGGTCAG
Probe 3:	TCGATCGTAATCGTACCGGTCAGT
…	…

Microarray that Created Examples

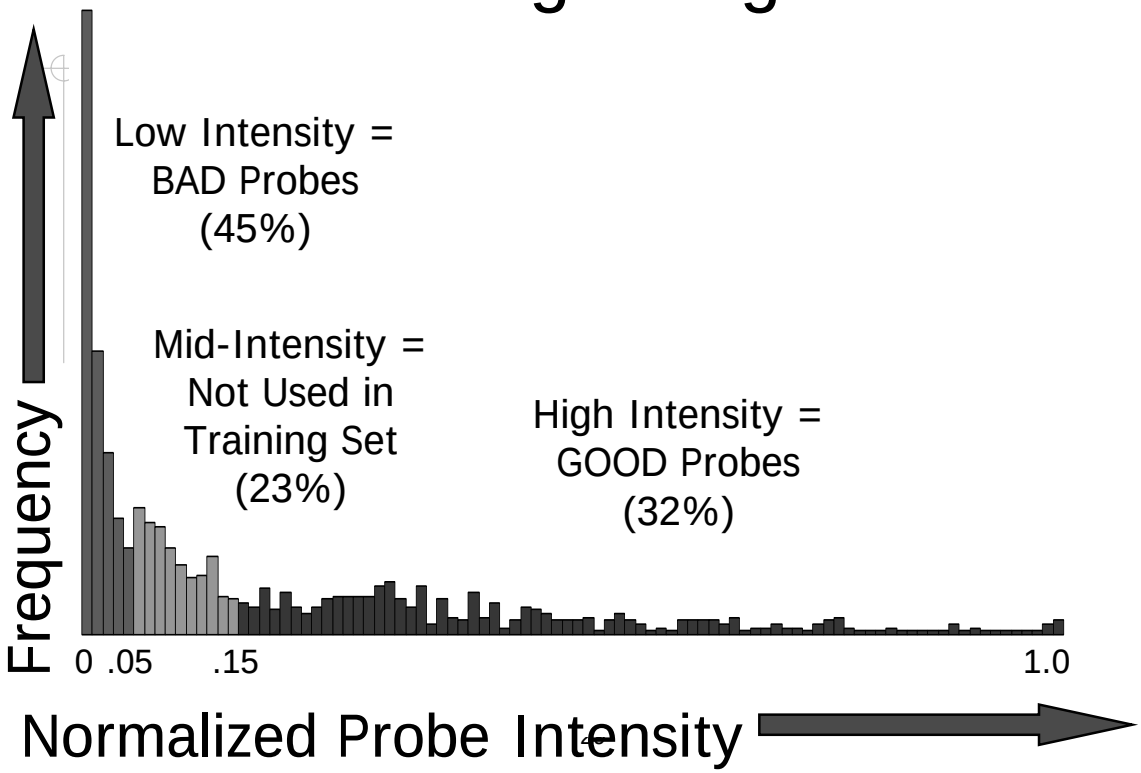
The Features (Tobler et al., ISMB 2002)

Feature Name	Description
fracA, fracC, fracG, fracT	The fraction of A, C, G, or T in the 24-mer
fracAA, fracAC, fracAG, fracAT, fracCA, fracCC, fracCG, fracCT, fracGA, fracGC, fracGG, fracGT, fracTA, fracTC, fracTG, fracTT	The fraction of each of these dimers in the 24-mer
n1, n2, …., n24	The particular nucleotide (A, C, G, or T) at the specified position in the 24-mer
d1, d2, …, d23	The particular dimer (AA, AC,…TT) at the specified position in the 24-mer

Information Gain per Feature



Defining Categories



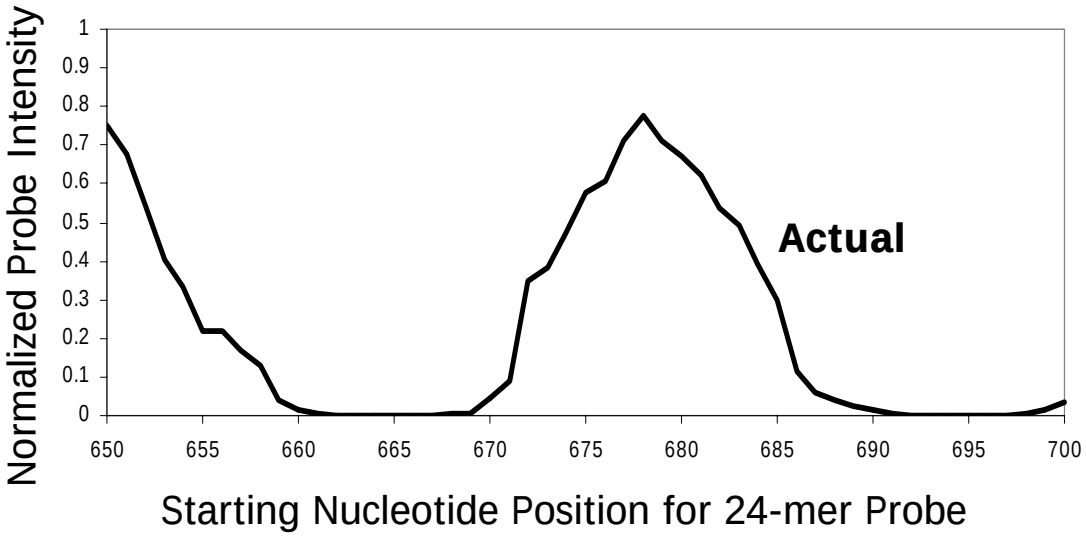
The Machine Learning Techniques

- Naïve Bayes (Mitchell 1997)
- Neural Networks (Rumelhart *et al.* 1995)
- Decision Trees (Quinlan 1996)
- Can interpret predictions of each learner probabilistically

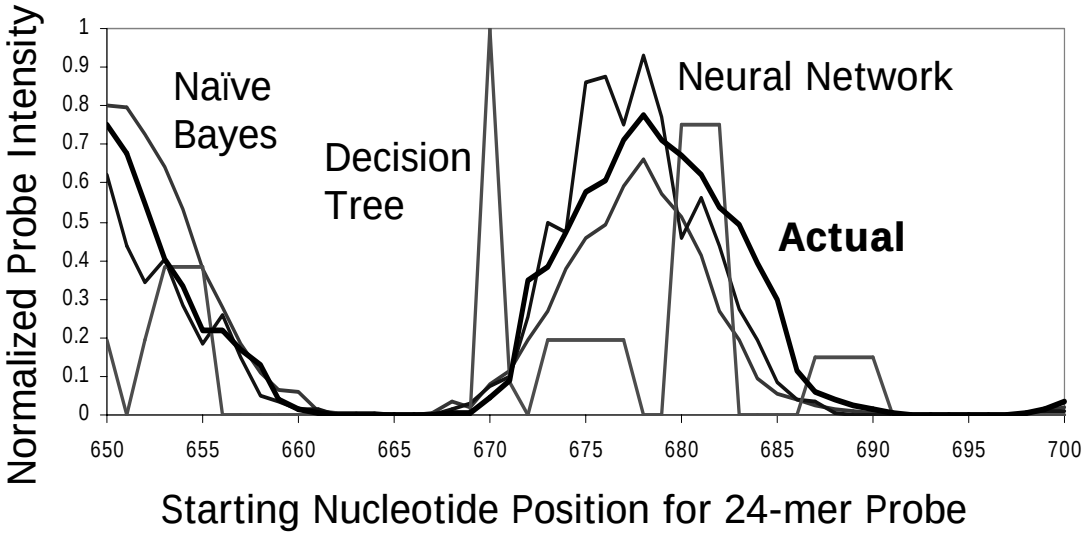
Leave-One-Gene-Out X-Validation

- Leave-one-gene-out testing:
  - For each gene (of the 8)
    - **Train** on all <u>but</u> this gene
    - **Test** on this gene
    - Record result
    - Forget what was learned
  - Average results across 8 test genes
- In mol bio tasks, be carefully how you split into train and test sets!

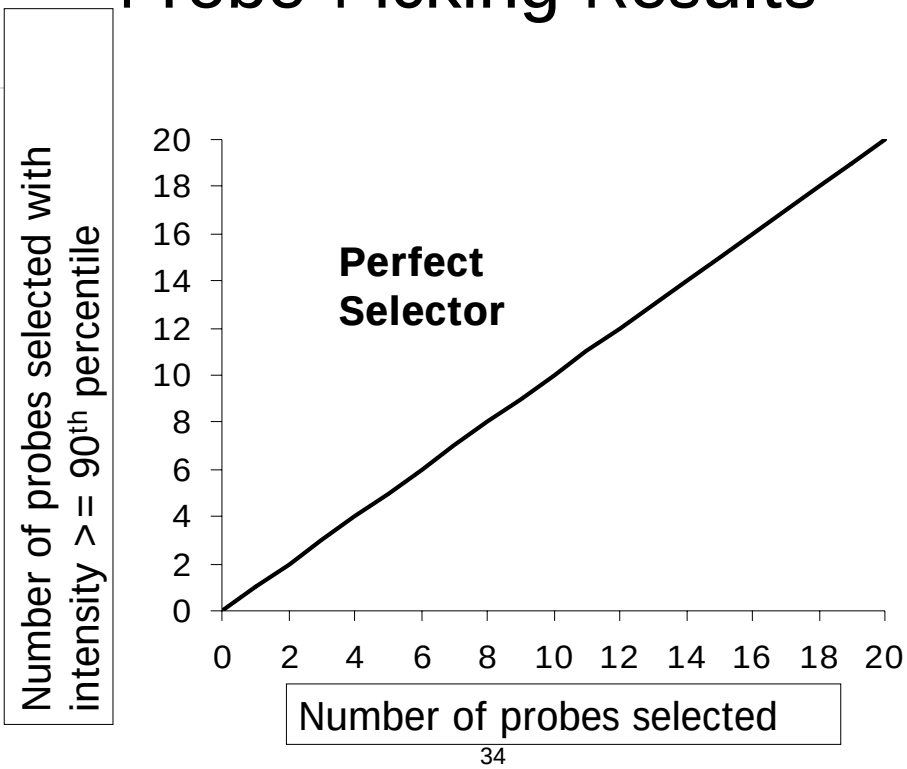
Typical Probe-Intensity Prediction Across Short Region



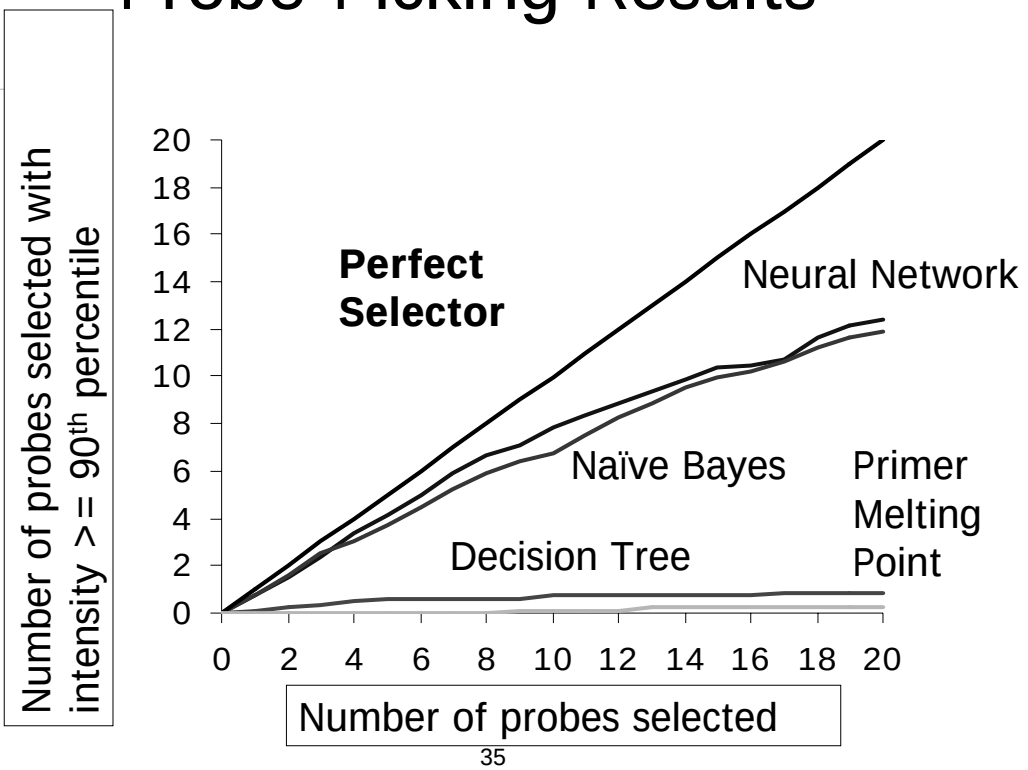
Typical Probe-Intensity Prediction Across Short Region



Probe-Picking Results



Probe-Picking Results



Outline

- Molecular Biology and Microtechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

Two Views of Microarray Data

- Data points are genes
  - Represented by expression levels across different samples (ie, features=samples)
  - **Goal**: categorize new genes
- Data points are samples (eg, patients)
  - Represented by expression levels of different genes (ie, features=genes)
  - **Goal**: categorize new samples

Two Ways to View The Data

Person ▼ Gene ▶	A28202_ac	AB00014_at	AB00015_at	. . .
Person 1	1142.0	321.0	2567.2	. . .
Person 2	586.3	586.1	759.0	. . .
Person 3	105.2	559.3	3210.7	. . .
Person 4	42.8	692.1	812.0	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .

Data Points are Genes

Person ▼ Gene ▶	A28202_ac	AB00014_at	AB00015_at	. . .
Person 1	1142.0	321.0	2567.2	. . .
Person 2	586.3	586.1	759.0	. . .
Person 3	105.2	559.3	3210.7	. . .
Person 4	42.8	692.1	812.0	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .

Data Points are Samples

Person ▼ Gene ▶	A28202_ac	AB00014_at	AB00015_at	. . .
Person 1	1142.0	321.0	2567.2	. . .
Person 2	586.3	586.1	759.0	. . .
Person 3	105.2	559.3	3210.7	. . .
Person 4	42.8	692.1	812.0	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .
. . .	.	.	.	. . .

Supervision: Add Class Values

Person ▼ Gene ▶	A28202_ac	AB00014_at	AB00015_at	. . .	Class
Person 1	1142.0	321.0	2567.2	. . .	normal
Person 2	586.3	586.1	759.0	. . .	cancer
Person 3	105.2	559.3	3210.7	. . .	normal
Person 4	42.8	692.1	812.0	. . .	cancer
. . .	.	.	.	. . .	
. . .	.	.	.	. . .	
. . .	.	.	.	. . .	

Supervised Learning Task 2

- **Given**: a set of microarray experiments, each done with mRNA from a different patient (same cell type from every patient)

  Patient's expression values for each gene constitute the features, and patient's disease constitutes the class

- **Do**: Learn a model that accurately predicts class based on features

Location in Task Space

	Data Points are: Genes	Samples
Clustering		
Supervised Data Mining		Predict the class value for a patient based on the expression levels for his/her genes

Leukemia (Golub *et al.*, 1999)

- Classes

  Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML)

- Approach

  Weighted voting (essentially naïve Bayes)

- Cross-Validated Accuracy

  Of 34 samples, declined to predict 5, correct on other 29

Cancer vs. Normal

- Relatively easy to predict accurately, because so much goes “haywire” in cancer cells
- Primary barrier is noise in the data... impure RNA, cross-hybridization, etc
- Studies include breast, colon, prostate, lymphoma, and multiple myeloma

X-Val Accuracies for **M**ultiple **M**yeloma (74 **MM** vs. 31 Normal)

Trees	98.1
Boosted Trees	99.0
SVMs	100.0
Vote	100.0
Bayes Nets	97.0

Prognosis and Treatment

- Features same as for diagnosis
- Rather than disease state, class value becomes life expectancy with a given treatment (or **positive response** vs. **no response** to given treatment)

Breast Cancer Prognosis
(Van't Veer *et al.*, 2002)

- Classes

  good prognosis (no metastasis within five years of initial diagnosis) vs. **poor** prognosis
- Algorithm

  Ensemble of voters
- Results

  83% cross-validated accuracy on 78 cases

A Lesson

- Previous work selected features to use in ensemble by looking at the entire data set
- Should have repeated feature selection on each cross-val fold
- Authors also chose ensemble size by seeing which size gave highest cross-val result
- Authors corrected this in web supplement; ***accuracy went from 83% to 73%***
- Remember to "tune parameters" separately for each cross-val fold!

Prognosis with Specific Therapy

(Rosenwald *et al.*, 2002)

- Data set contains gene-expression patterns for 160 patients with *diffuse large B-cell lymphoma*, receiving *anthracycline chemotherapy*
- Class label is *five-year survival*
- One test-train split 80/80
- True positive rate: 60%
  False negative rate: 39%

Some Future Directions

- Using gene-chip data to select therapy
  - Predict which therapy gives best prognosis for patient
- Comparing cancer with related benign conditions, rather than with normal
  - Tougher, but may give more insight

Unsupervised Learning Task 1

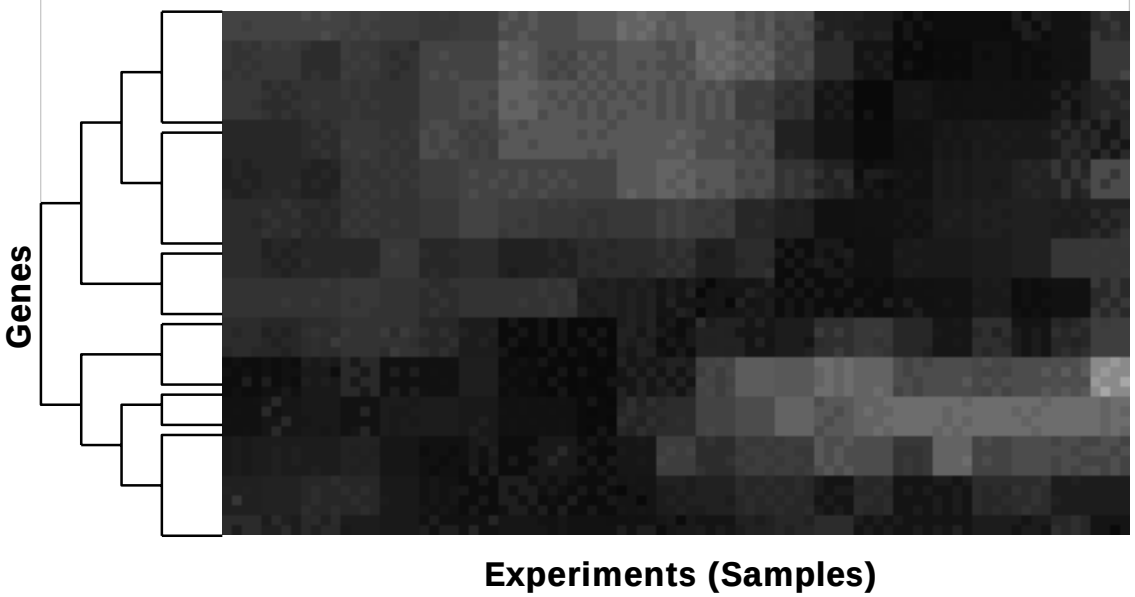
- **Given**: a set of microarray experiments under different conditions
- **Do**: cluster the genes, where a gene described by its expression levels in different experiments

Location in Task Space

	Data Points are: Genes	Samples
Clustering	Group genes into clusters, where all members of a cluster tend to go up or down together	
Supervised Data Mining		

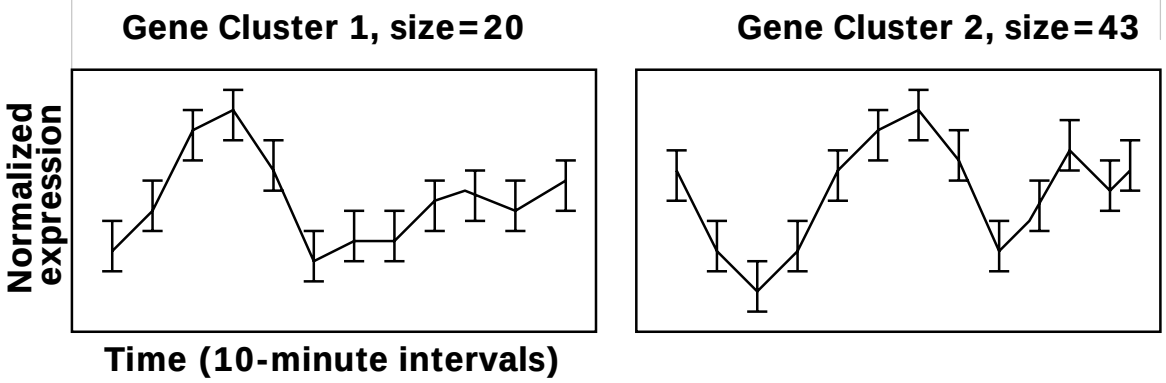
Example

(Green = up-regulated, Red = down-regulated)



Visualizing Gene Clusters

(eg, Sharan and Shamir, 2000)



Unsupervised Learning Task 2

- **Given**: a set of microarray experiments (samples) corresponding to different conditions or patients
- **Do**: cluster the experiments

Location in Task Space

	Data Points are: Genes	Data Points are: Samples
Clustering		Group samples by gene expression profile
Supervised Data Mining		

Examples

- Cluster samples from mice subjected to a variety of toxic compounds (Thomas *et al.*, 2001)
- Cluster samples from cancer patients, potentially to discover different subtypes of a cancer
- Cluster samples taken at different time points

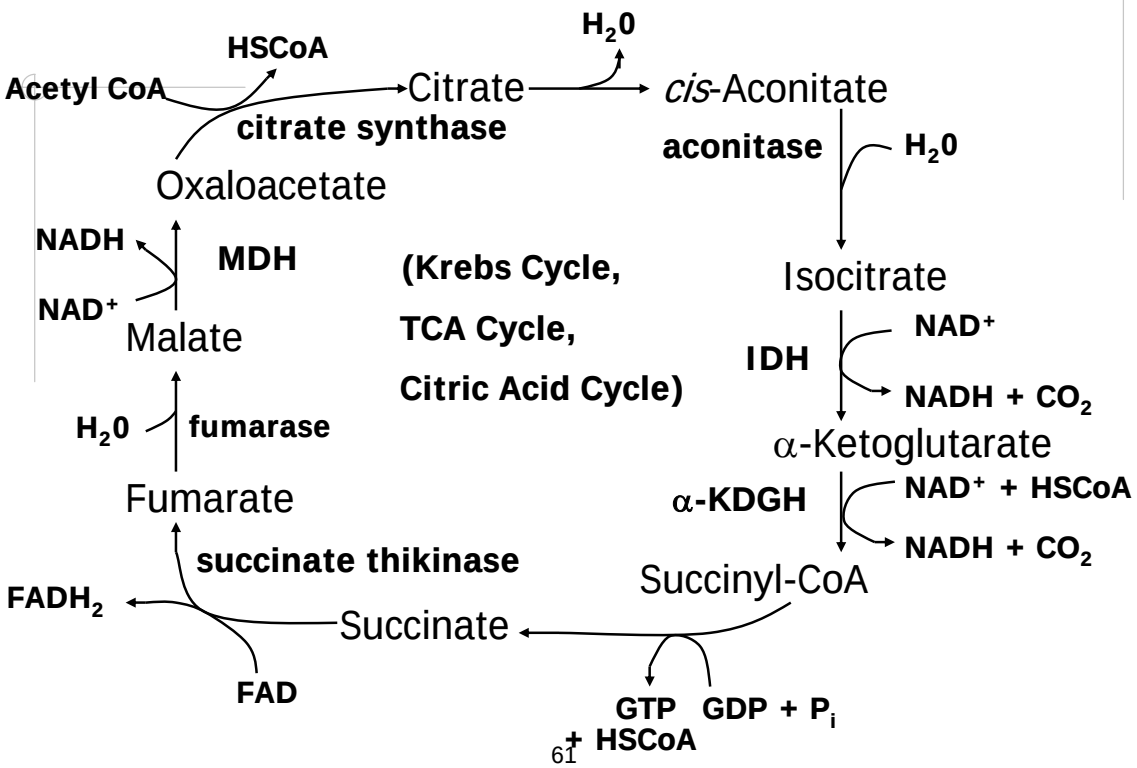
Outline

- Molecular Biology and Microtechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

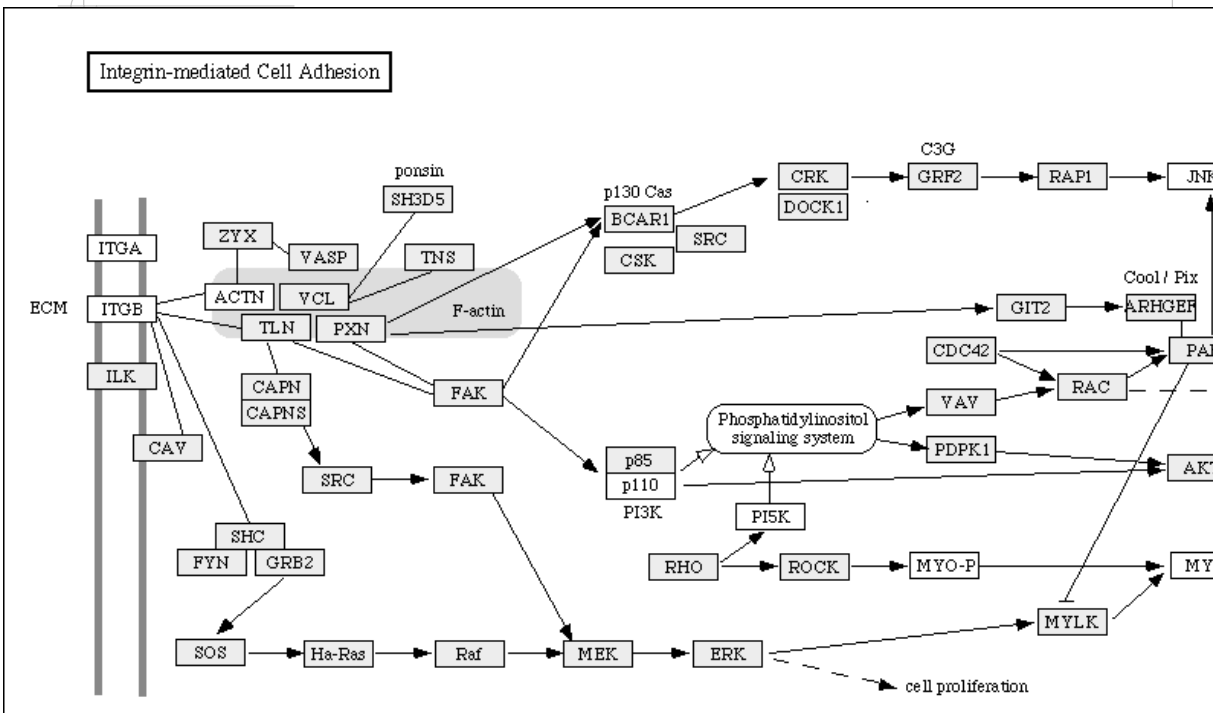
Some Biological Pathways

- Regulatory pathways
  - Nodes are labeled by genes
  - Arcs denote influence on transcription
  - G1 codes for P1, P1 inhibits G2's transcription
- Metabolic pathways
  - Nodes are metabolites, large biomolecules (eg, sugars, lipids, proteins and modified proteins)
  - Arcs from biochemical reaction inputs to outputs
  - Arcs labeled by enzymes (themselves proteins)

Metabolic Pathway Example



Regulatory Pathway (KEGG)



Using Microarray Data Only

- Regulatory pathways
  - Nodes are labeled by genes
  - Arcs denote influence on transcription
  - G1 codes for P1, P1 inhibits G2's transcription
- Metabolic pathways
  - Nodes are metabolites, large biomolecules (eg, sugars, lipids, proteins, and modified proteins)
  - Arcs from biochemical reaction inputs to outputs
  - Arcs labeled by enzymes (themselves proteins)

Supervised Learning Task 3

- **Given**: a set of microarray experiments for same organism under different conditions
- **Do**: Learn graphical model that accurately predicts expression of some genes in terms of others

Some Approaches to Learning Regulatory Networks

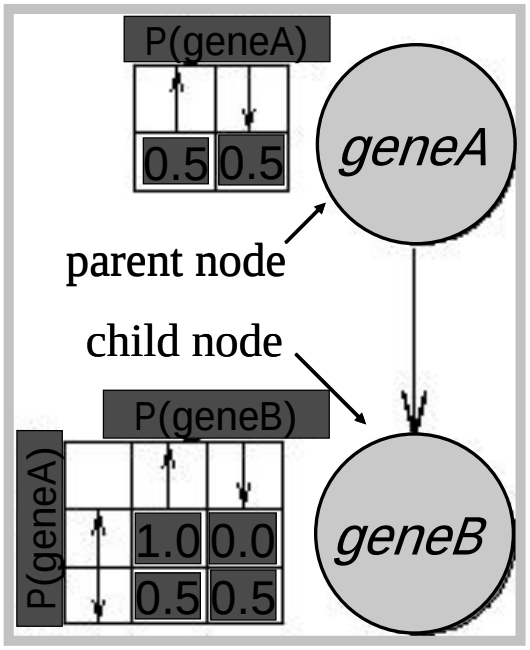
- Bayes Net Learning
  (Friedman & Halpern, 1999)
- Boolean Networks (Akutsu, Kuhara, Maruyama & Miyano, 1998; Ideker, Thorsson & Karp, 2002)
- Related Graphical Approaches
  (Tanay & Shamir, 2001; Chrisman, Langley, Baay & Pohorille, 2003)

Bayesian Network (BN)

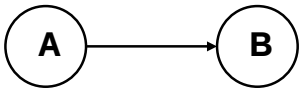
Data

	geneA	*geneB*
Expt1	↑	↑
Expt2	↑	↑
Expt3	↓	↓
Expt4	↓	↑

Note: direction of arrow indicates dependence not causality

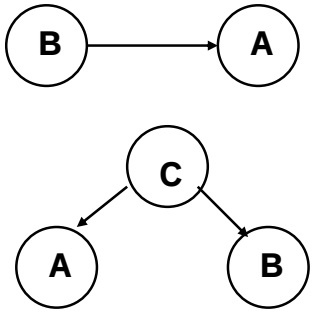


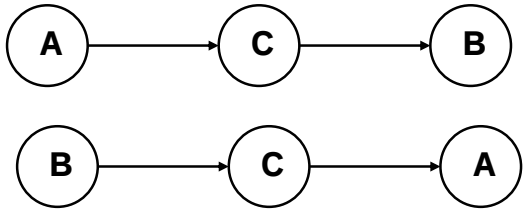
Problem: Not Causality



A is a good predictor of B. But is A <u>regulating</u> B??

Ground truth might be:



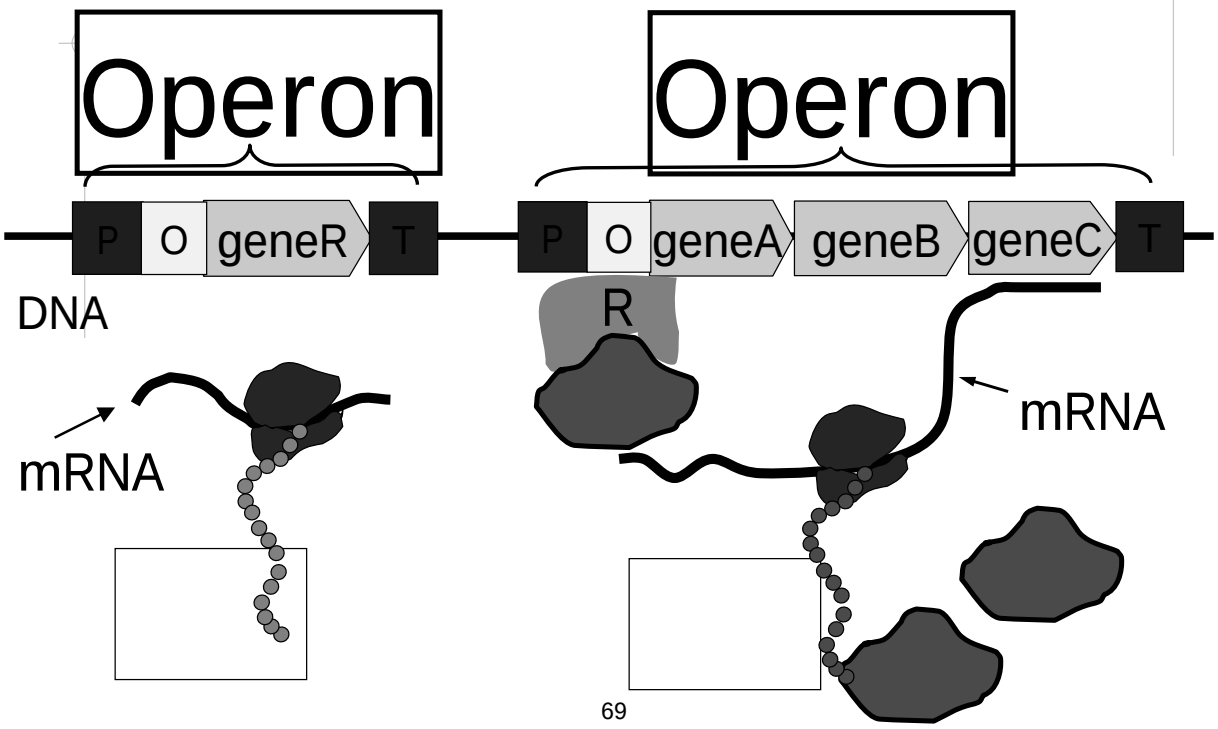


Or a more complicated variant

Approaches to Get Causality

- Use "knock-outs" (Pe'er, Regev, Elidan and Friedman, 2001). But not available in most organisms.
- Use time-series data and Dynamic Bayesian Networks (Ong, Glasner and Page, 2002). But even less data typically.
- Use other data sources, eg sequences upstream of genes, where transcription regulators may bind. (Segal, Barash, Simon, Friedman and Koller, 2002).

Transcription Regulation

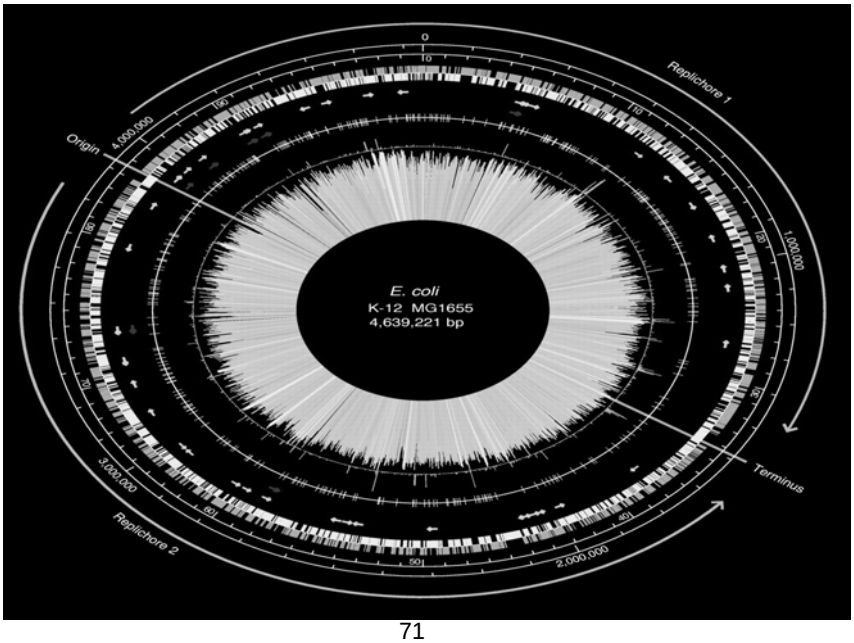


Another Way Around Limitations

- Identify smaller part of the task that is <u>a step toward</u> a full regulatory pathway
  - Part of a pathway
  - Classes or groups of genes
- Example:

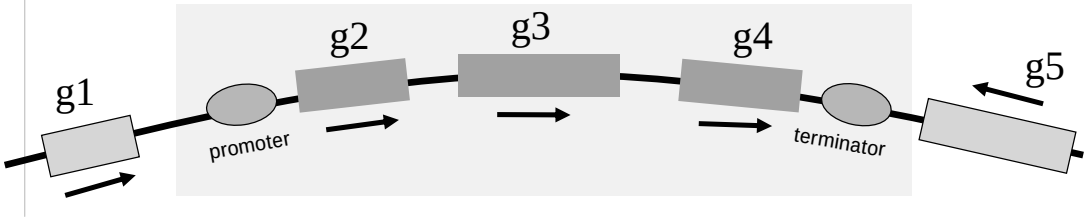
  Predicting the operons in *E. coli*

The *E. Coli* Genome



Finding Operons in *E. coli*

(Craven, Page, Shavlik, Bockhorst and Glasner, 2000)

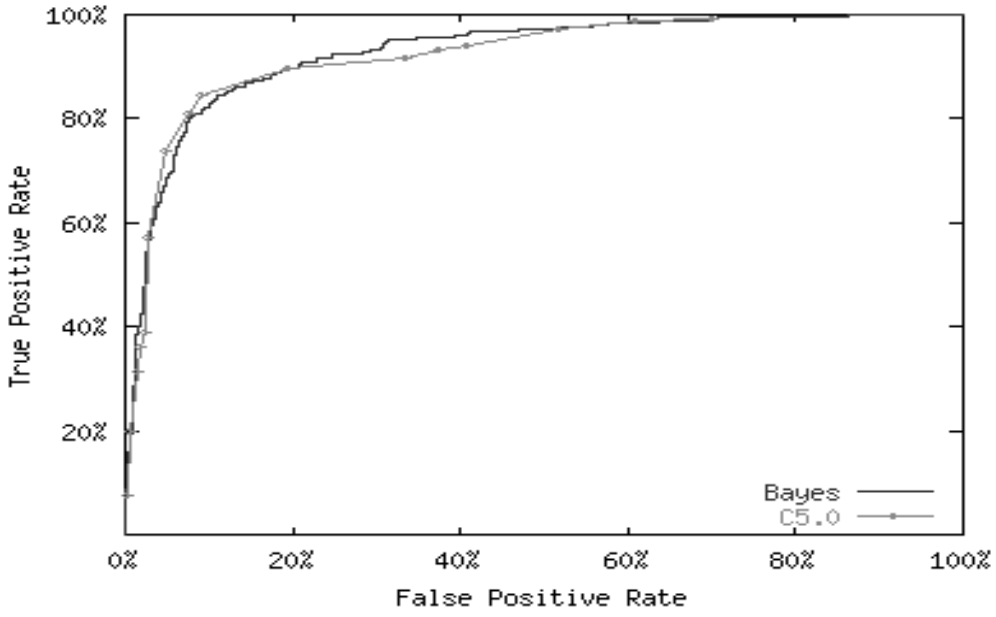


- Given: known operons and other *E. coli* data

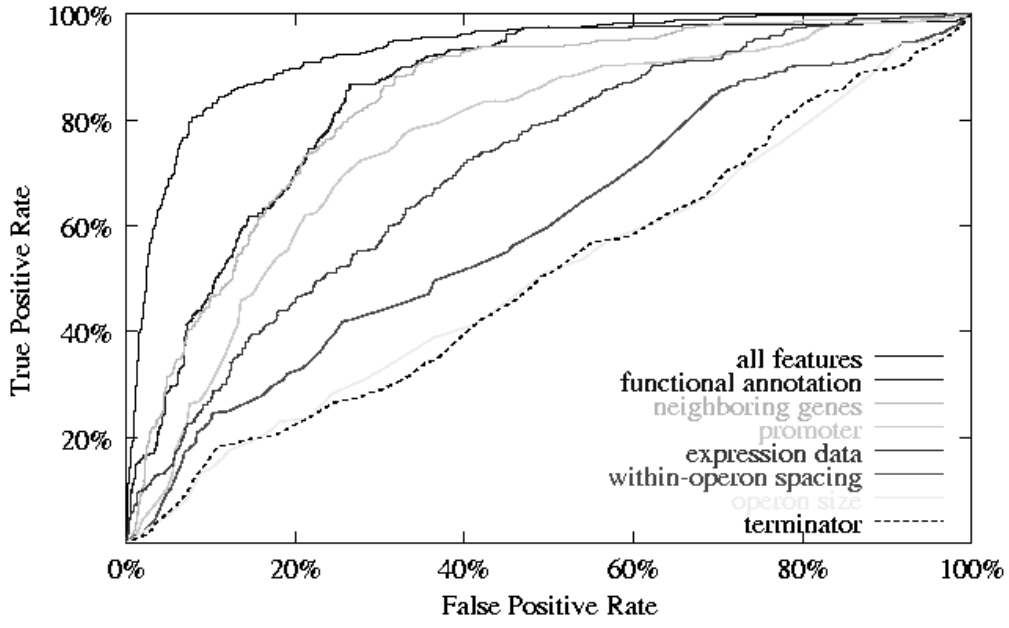
  Do: predict all operons in *E. coli*

- Additional Sources of Information
  - gene-expression data
  - functional annotation

Comparing Naive Bayes and Decision Trees (C5.0)



Using **Only** Individual Features



Outline

- Molecular Biology and Microtechnology
- Machine Learning Applications
  - Technological: Designing Microarrays
  - Medical: Predicting Disease (Diagnosis, Prognosis, & Treatment)
  - Biological: Constructing Pathway Models
- Looking Ahead: Related Technologies

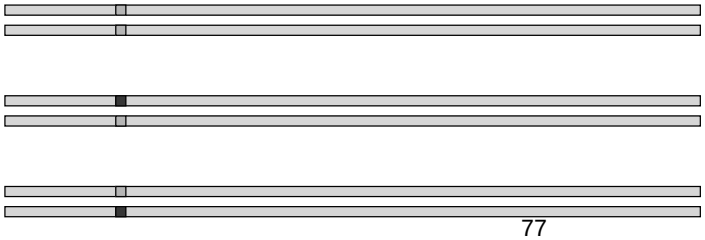
Single-Nucleotide Polymorphisms

- SNPs: Individual positions in DNA where variation is common
- Now 1.8 million known SNPs in humans
- Easier/faster/cheaper to measure SNPs than to completely sequence everyone
- Motivation ...

If We Sequenced Everyone...

Succeptible to Disease D or Responds to Treatment T

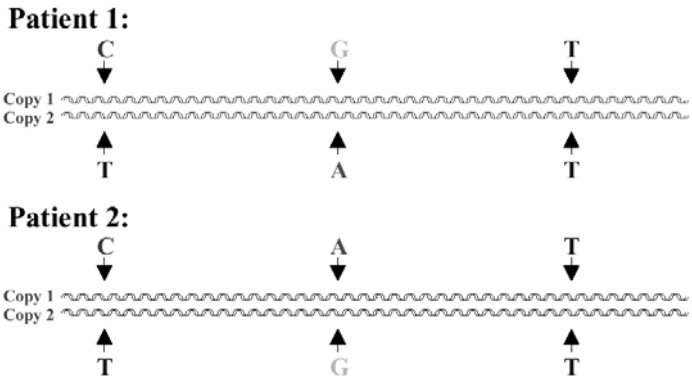
Not Succeptible or Not Responding



Example of SNP Data

Person SNP►	1	2	3	. . .	CLASS
Person 1	C T	A G	T T	. . .	old
Person 2	C C	A G	C T	. . .	young
Person 3	T T	A A	C C	. . .	old
Person 4	C T	G G	T T	. . .	young
. . .	. .	. . .	.		
. . .	. .	. . .	.		
. . .	. .	. . .	.		

Phasing (Haplotyping)



	SNP 1		SNP 2		SNP 3		...	class
Patient 1	C	T	A	G	T	T	...	Diseased
Patient 2	C	T	A	G	T	T	...	Healthy
...	...	...	...	...	...	...	...	...

Advantages of SNP Data

- Person's SNP pattern does not change with time or disease, so it can give more insight into susceptibility
- Easier to collect samples (can simply use blood rather than affected tissue)

Challenges of SNP Data

- Unphased
  Algorithms exist for phasing (haplotyping), but they make errors and typically need related individuals, dense coverage
- Missing values are more common than in microarray data
- More expensive than microarray data if we want similar level of completeness

Example

- Multiple Myeloma, 3000 SNPs,
  Young (susceptible) vs. Old (less susceptible)
- SVM^{light} with feature selection
  (repeated on every fold of cross-validation)
- Result significantly better than chance

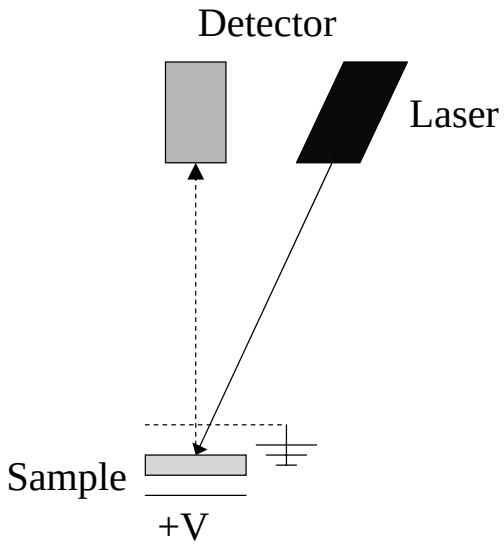
		Old	Young
Actual	Old	31	9
	Young	14	26

Proteomics

- Microarrays are useful primarily because mRNA concentrations serve as surrogate for protein concentrations
- Like to measure protein concentrations directly, but at present cannot do so in same high-throughput manner
- Proteins do not have obvious direct complements
- Could build molecules that bind, but binding greatly affected by protein structure

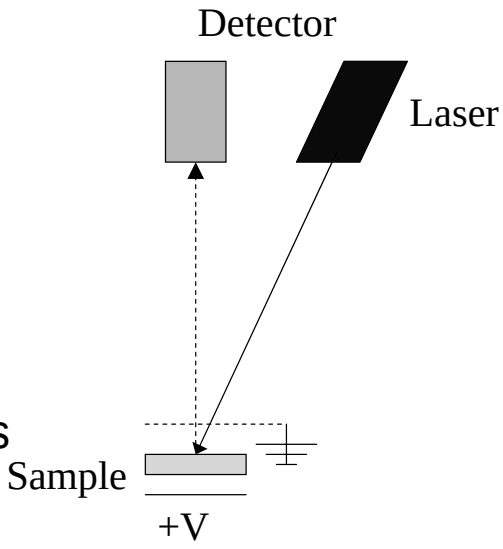
Time-of-Flight (TOF) Mass Spectrometry

- Measures the time for an ionized particle, starting from the sample plate, to hit the detector

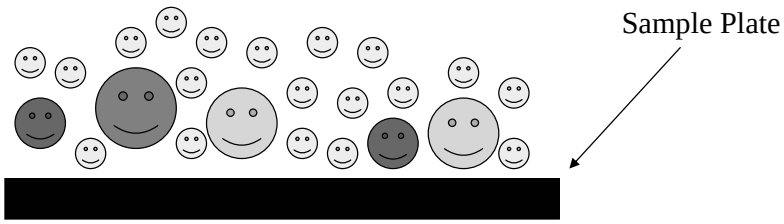


Time-of-Flight (TOF) Mass Spectrometry 2

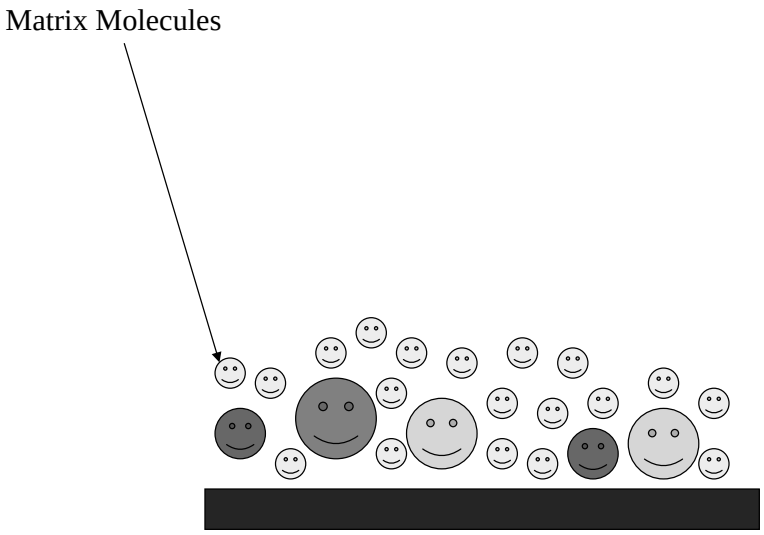
- *Matrix-Assisted Laser Desorption-Ionization* (MALDI)
- Crystalloid structures made using proton-rich matrix molecule
- Hitting crystalloid with laser causes molecules to ionize and “fly” towards detector



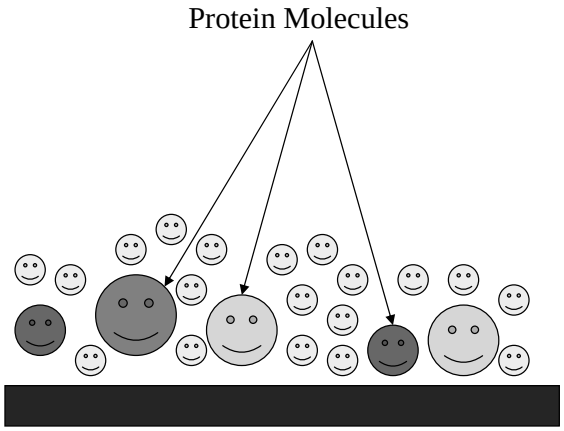
Time-of-Flight Demonstration 0



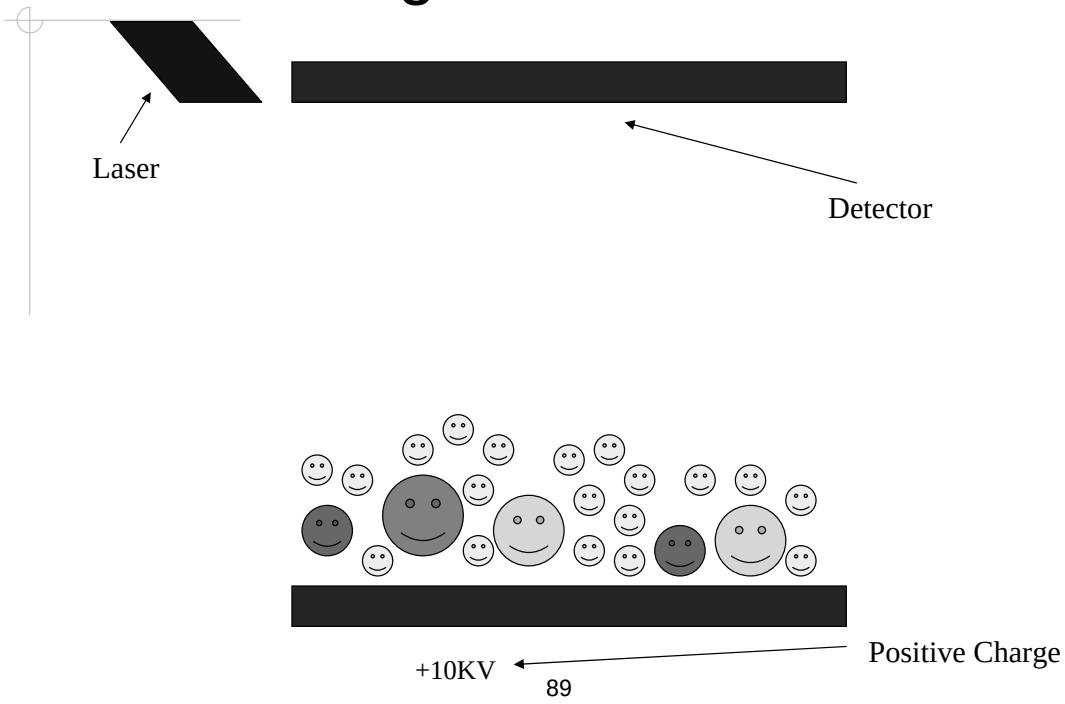
Time-of-Flight Demonstration 1



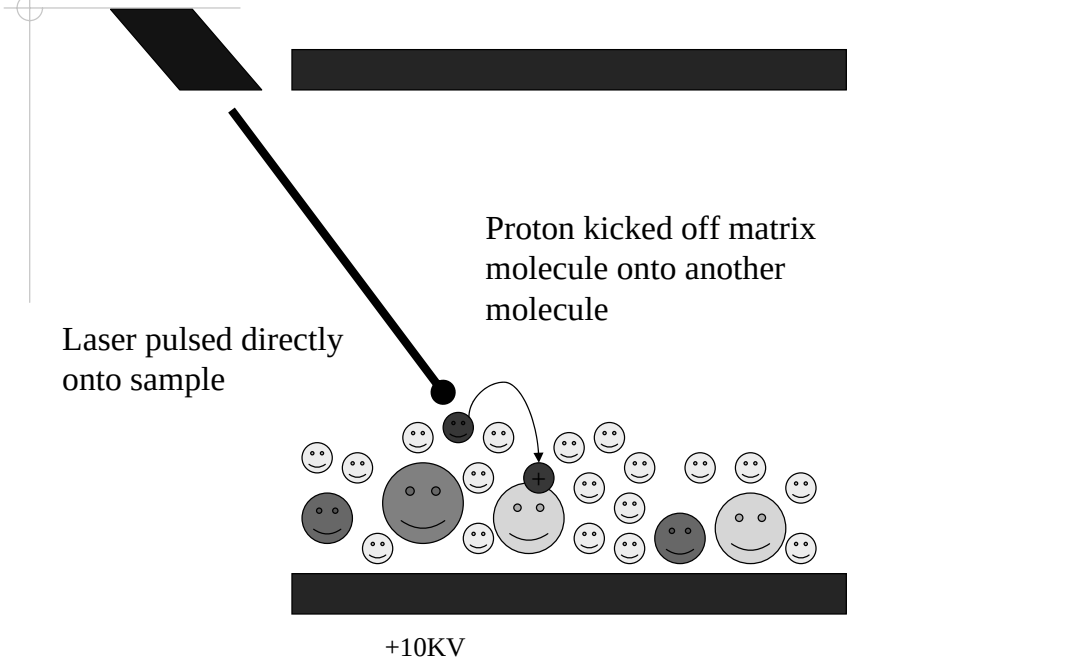
Time-of-Flight Demonstration 2



Time-of-Flight Demonstration 3

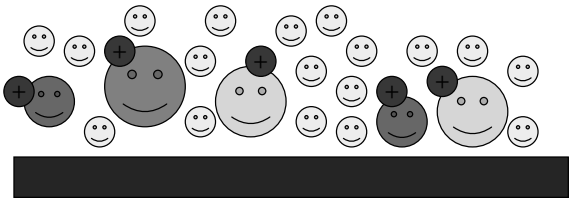


Time-of-Flight Demonstration 4



Time-of-Flight Demonstration 5

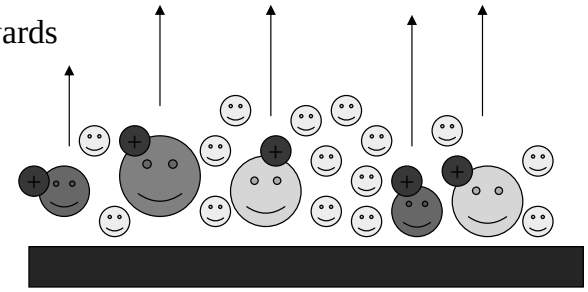
Lots of protons kicked off matrix ions, giving rise to more positively charged molecules



+10KV

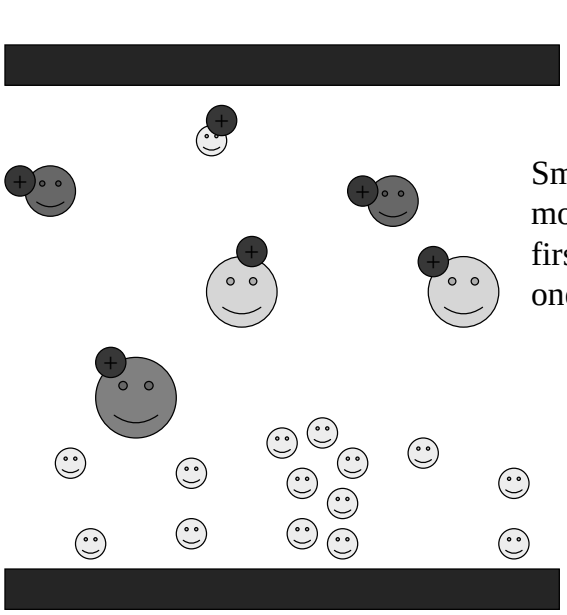
Time-of-Flight Demonstration 6

The high positive potential under sample plate, causes positively charged molecules to accelerate towards detector



+10KV

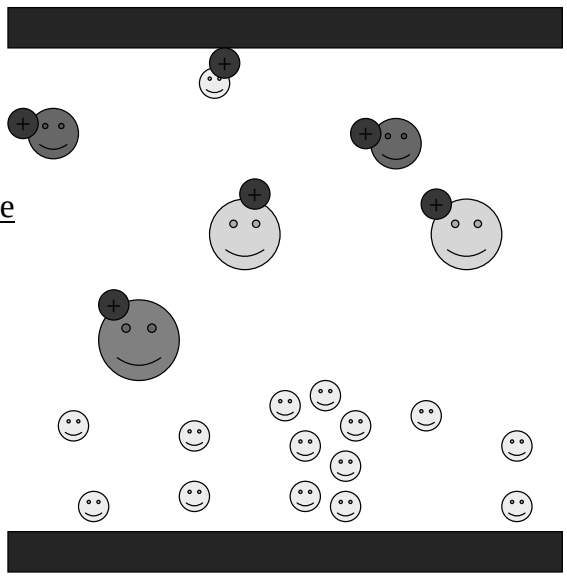
Time-of-Flight Demonstration 7



Smaller mass molecules hit detector first, while heavier ones detected later

+10Kv

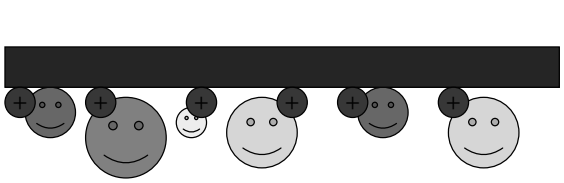
Time-of-Flight Demonstration 8



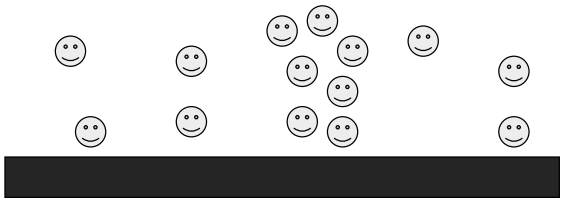
The <u>incident time</u> measured from when laser is pulsed until molecule hits detector

+10KV

Time-of-Flight Demonstration 9

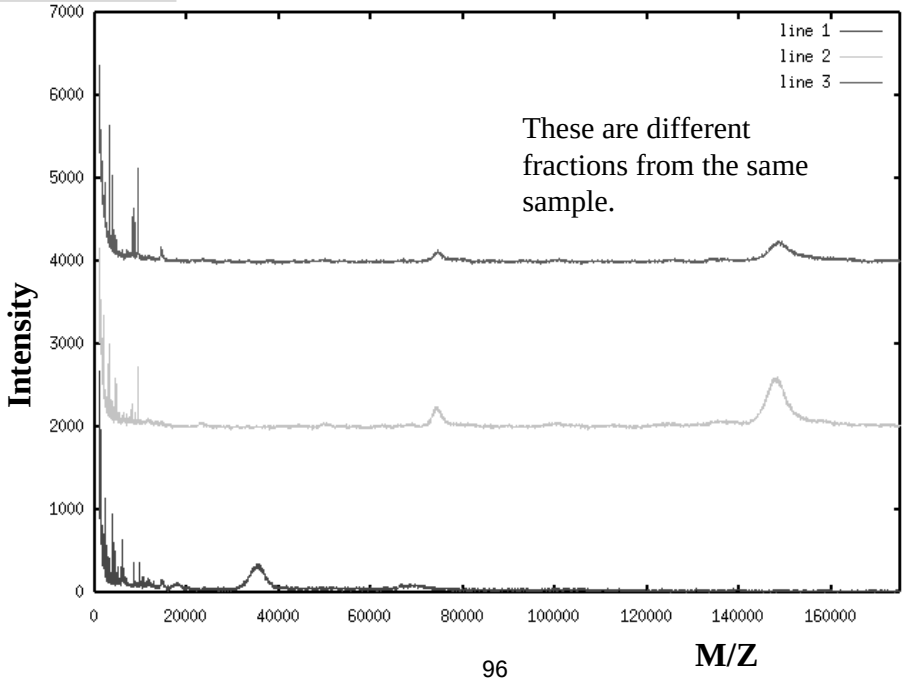


Experiment repeated a number of times, counting frequencies of “flight-times”

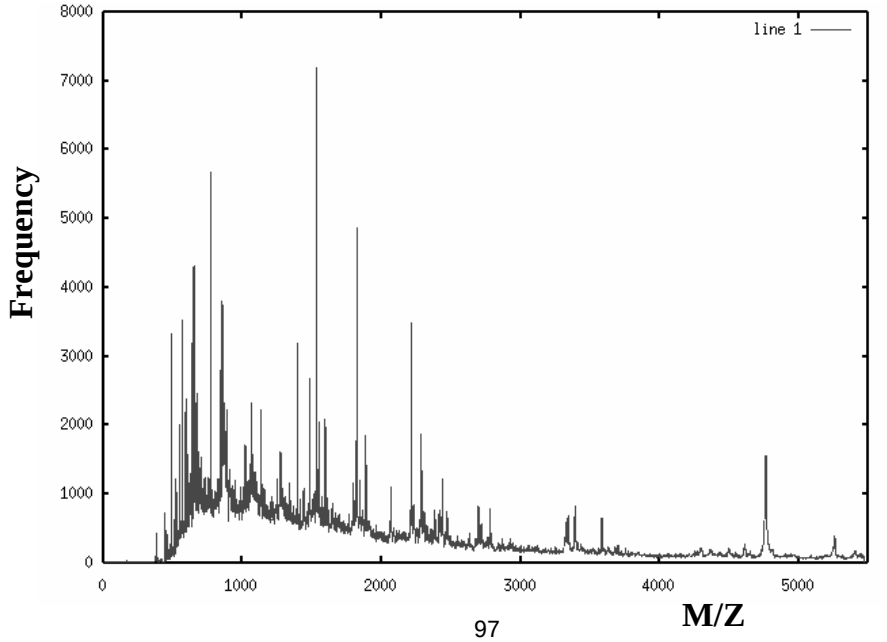


+10KV

Example Spectra from Duke



Trypsin-Treated Spectra



Challenges of Proteomics Data

- Noise
  - M/Z values may not align exactly across spectra (resolution ~0.1%)
  - Intensities not calibrated across spectra
- Must identify proteins from “signatures” ... best results if proteins broken down
- Cannot get all proteins... typically several hundred

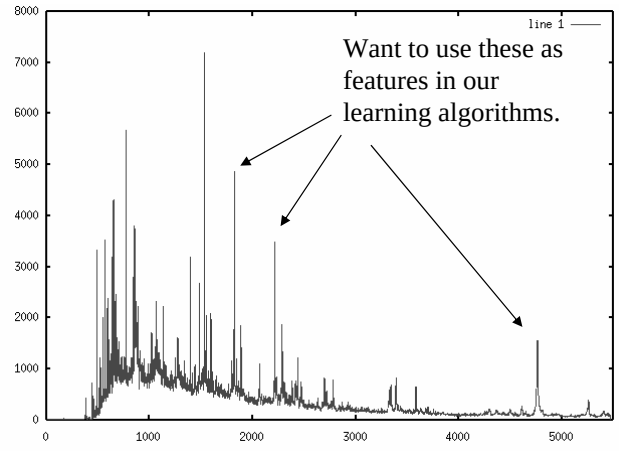
Peak Picking

- Want to pick peaks that are statistically significant from the noise signal

• Fortunately, data from Duke had peaks picked from spectra already

• Page Group working on a peak-picking algorithm

• Want sensitivity to peaks, while filtering out peaks tdue to noise



Metabolomics

- Measures concentration of each low-molecular weight molecule in sample
- These typically are "*metabolites*," or small molecules produced or consumed by reactions in biochemical pathways
- These reactions typically catalyzed by proteins (specifically, enzymes)

Lipomics

- Analogous to metabolomics, but measuring concentrations of lipids rather than metabolites
- Potentially help induce biochemical pathway information or to help disease diagnosis or treatment choice

Final Wrapup

- Molecular biology collecting lots and lots of data in post-genome era
- Opportunity to “connect” molecular-level information to diseases and treatment
- Need analysis tools to interpret
- Machine learning opportunities abound
- Hopefully this tutorial provided solid start toward applying ML to biological data

Some Additional Readings

- **Molla, Waddell, Page & Shavlik,** Using Machine Learning to Design and Interpret Gene-Expression Microarrays (to appear in the *AI Magazine* special issue on Bioinformatics)
- Special issue of *Machine Learning* journal (Volume 52:1/2, 2003) on Machine Learning in the Genomics Era

Thanks To

- Mark Craven
- Michael Molla
- Michael Waddell
- Sean McIlwain
- Irene Ong
- Roland Green
- John Tobler

Some Useful Datasets	Brief Description
www.ebi.ac.uk/arrayexpress/	EBI microarray data repository
www.ncbi.nlm.nih.gov/geo/	NCBI microarray data repository
genome-www5.stanford.edu/MicroArray/SMD/	Stanford microarray database
rana.lbl.gov/EisenData.htm	Eisen-lab's yeast data, (Spellman *et al.* 1998)
www.genome.wisc.edu/functional/microarray.htm	University of Wisconsin *E. coli* Genome Project
llmpp.nih.gov/lymphoma/data.shtml	Diffuse large B-cell lymphoma (Alizadeh *et al.* 2000)
llmpp.nih.gov/DLBCL/	Molecular profiling (Rosenwald *et al.* 2002)
www.rii.com/publications/2002/vantveer.htm	Breast cancer prognosis (Van't Veer *et al.* 2002)
www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi	MIT Whitehead Center for Genome Research, including data in Golub *et al.* (1999)
lambertlab.uams.edu/publicdata.htm	Lambert Laboratory data for multiple myeloma
www.cs.wisc.edu/~dpage/kddcup2001/	KDD Cup 2001 data; Task 2 includes correlations in genes' expression levels
clinicalproteomics.steem.com/	Proteomics data (mass spectrometry of proteins)
snp.cshl.org/	Single nucleotide polymorphism (SNP) data

Bibliography from *AI Mag* Article

- Alizadeh, A.; Eisen, M.; Davis, R.; Ma, C.; Lossos, I. , Rosenwald, A.; Boldrick, J.; Hajeer, S.;Tran, T.; Yu, X.; Powell, J.; Yang, L.; Marti, G.; Moore, T.; Hudson, J. Jr; Lu, L.; Lewis, D.; Tibshirani, R.; Sherlock, G; Chan, W.; Greiner, T.; Weisenburger, D.; Armitage, J.; Warnke, R.; Levy, R.; Wyndham Wilson, W.; Grever, M.; Byrd, J.; Botstein, D.; Brown, P.; and Staudt, L. 2000. Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling. *Nature* 403:503-511.
- Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000. *Nucleic Acids Research* 28:45-48.
- Breslauer, K.; Frank, R.; Blocker, H.; and Marky, L. 1986. Predicting DNA Duplex Stability from the Base Sequence. *Proceedings of the National Academy of Science USA* 83:3746-3750.
- Brown, M.; Grundy, W.; Lin, D.; Cristianini, N.; Sugnet, C.; Furey, T.; Ares M. Jr.; and Haussler, D. 2000. Knowledge-based Analysis of Microarray Gene Expression Data by using Support Vector Machines. *Proceedings of the National Academy of Science USA* 97(1):262-267.
- Cheng, J.; Hatzis, C.; Hayashi, H.; Krogel, M.; Morishita, S.; Page, D. and Sese, J. 2002. Report on KDD Cup 2001. *SIGKDD Explorations* 3(2):47-64.
- Craven, M.; Page, D.; Shavlik, J.; Bockhorst J.; and Glasner J. 2000. Using Multiple Levels of Learning and Diverse Evidence Sources to Uncover Coordinately Controlled Genes. *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, Palo Alto, CA.
- Davidson, E.; Rast, J.; Oliveri, P.; Ransik, A.; Calestani, C.; Yuh, C.; Amore, G.; Minokawa, T.; Hynman, V.,; Arenas-Mena, C.; Otim, O.; Brown, C.; Livi, C.; Lee, P.; Revilla, R.; Alistair R.; Pan Z.; Schilstra M.; Clarke, P.; Arnone, M.; Rowen, L.; Cameron, R.; McClay, D.; Hood, L. and Bolouri, H. 2002. A Genomic Regulatory Network for Development. *Science* 295:1669-1678.
- Eisen M.; Spellman P.; Brown P.; and Botstein D. 1998. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proceedings of the National Academy of Science USA* 95:14863-14868.
- Friedman, N. and Halpern J. 1999. Modeling Beliefs in Dynamic Systems. Part II: Revision and Update. *Journal of AI Research* 10:117-167.
- Golub T.; Slonim D.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.; Coller, H.; Loh, M.; Downing, J.; Caligiuri, M.; Bloomfield, C; and Lander, E. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286:531-537.
- Hanisch, D.; Zien, A.; Zimmer, R.; and Lengauer, T. 2002. Co-Clustering of Biological Networks and Gene Expression Data. *Bioinformatics* 18:S145-S1554.
- Hood, L. and Galas, D. 2003. The Digital Code of DNA. *Nature* 421:444-448.
- Hunter, L. 2003. An Introduction to Molecular Biology for Computer Scientists. *AI Magazine,* this issue.
- Khodursky, A.; Peter, B.; Cozzarelli, N.; Botstein, D.; Brown, P. and Yanofsky, C. 2000. DNA Microarray Analysis of Gene Expression in Response to Physiological and Genetic Changes that Affect Tryptophan in *Escheria Coli. Proceedings of the National Academy of Science USA* 97:12170-12175.
- Lazarou, J.; Pomeranz, B. and Corey, P. 1998. Incidence of Adverse Drug Reactions in Hospitalized Patients. *Journal of the American Medical Association* 279(15):1200-1205.

More Bibliography

- Li, C. and Wong, W. 2001. Model-based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. *Proceedings of the National Academy of Science USA* 98(1):31-36.
- Mancinelli, L.; Cronin, M. and Sadee W. 2000. Pharmacogenomics: The Promise of Personalized Medicine. *AAPS PharmSci* 2(1): article 4.
- Molla, M; Andrae, P; Glasner, J; Blattner, F. and Shavlik, J. 2002. Interpreting Microarray Expression Data Using Text Annotating the Genes. *Information Sciences* 146:75-88.
- Mitchell, T. 1997. *Machine Learning*. McGraw-Hill, Boston, MA.
- Oliver, S.; Winson, M.; Kell, D. and Baganz, F. 1998. Systematic Functional Analysis of the Yeast Genome. *Trends in Biotechnology* 16(9):373-378.
- Ong, I.; Glassner, J. and Page, D. 2002. Modelling Regulatory Pathways in *E.coli* from Time Series Expression Profiles. *Bioinformatics* 18:241S-248S.
- Newton, M.; Kendziorski C.; Richmond, C.; Blattner, F. and Tsui, K. 2001. On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. *Journal of Computational Biology* 8:37-52.
- Nuwaysir, E. F.;Huang, W.; Albert, T.; Singh, J.; Nuwaysir, K.; Pitas, A.; Richmond, T.; Gorski, T.; Berg, J.; Ballin, J.; McCormick, M.; Norton, J.; Pollock, T.; Sumwalt, T.; Butcher, L.; Porter, D.; Molla, M.; Hall, C.; Blattner, F.; Sussman, M.; Wallace, R.; Cerrina, F. and Green, R. 2002. Gene Expression Analysis Using Oligonucleotide Arrays Produced by Maskless Lithography. *Genome Research* 12(11):1749-1755.
- Pe'er, D.; Regev, A.; Elidan, G. and Friedman, N. 2001. Inferring Subnetworks from Perturbed Expression Profiles. *Bioinformatics* 17:S215-S224
- Rosenwald, A.; Wright, G.; Chan, W.; Connors, J.; Campo, E.; Fisher, R.; Gascoyne, R.; Muller-Hermelink, H.; Smeland, E. and Staudt, L. 2002. The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine* 346(25):1937-1947.
- Segal, E.; Taskar, B.; Gasch, A.; Friedman, N. and Koller, D. 2001. Rich Probabilistic Models for Gene Expression. *Bioinformatics* 1(1):1-10.
- Shrager, J.; Langley, P.; and Pohorille, A. 2002. Guiding Revision of Regulatory Models with Expression Data. *Proceedings of the Pacific Symposium on Biocomputing,* 486-497, World Scientific, Lihue, Hawaii.
- Spellman, P.; Sherlock, G.; Zhang, M.; Iyer, V.; Anders, K.; Eisen, M.; Brown, P.; Botstein, D. and Futcher. B. 1998. Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* 9:3273-3297.
- Thomas, R.; Rank, D.; Penn, S.; Zastrow, G.; Hayes, K.; Pande, K.; Glover, E.; Silander, T.; Craven, M.; Reddy, J.; Jovanovich, S. and Bradfield, C. 2001. Identification of Toxicologically Predictive Gene Sets using cDNA Microarrays. *Molecular Pharmacology* 60:1189-1194.
- Tobler J.; Molla M.; Nuwaysir, E.; Green R. and Shavlik J. 2002. Evaluating Machine Learning Approaches for Aiding Probe Selection for Gene-Expression Arrays. *Bioinformatics*, 18:S164-S171.
- Van 't Veer, L.; Dai, H.; van de Vijver, M.; He, Y.; Hart, A.; Mao, M.; Peterse, H.; van der Kooy, K.; Marton, M.; Witteveen, A.; Schreiber, G.; Kerkhoven, R.; Roberts, C.; Linsley, P.; Bernards, R. and Friend, S. 2002. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* 415:530-536.

Additional Citations

- Akutsu, T.; Kuhara, S.; Maruyama, O. and Miyano, S. 1998. Identification of Gene Regulatory Networks by Strategic Gene Disruptions and Gene Overexpressions. ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 695-702
- Chrisman, L.; Langley, P.; Bay, S. and Pohorille, A. 2003. Incorporating Biological Knowledge into Evaluation of Causal Regulatory Hypotheses. *Pacific Symposium on Biocomputing*, pp. 128-139.
- Ideker, T.; Thorsson, V. and Karp, R. 2000. Discovery of Regulatory Interactions Through Perturbation: Inference and Experimental Design. *Pacific Symposium on Biocomputing*, pp. 302-313.
- Segal, E.; Taskar, B.; Gasch, A.; Friedman, N. and Koller, D. 2002. Rich Probabilistic Models for Gene Expression. *Proc. Ninth International Conference on Intelligent Systems for Molecular Biology (ISMB), Bioinformatics*, 17 (Suppl 1), pp. 243--252.
- Shamir, R. and Sharan, R. 2000. CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis. *Currents in Computational Molecular Biology*, pages 6--7, S. Miyano, R. Shamir and T. Takagi (editors) Universal Academy Press, 2000). Proc. ISMB '00, pp., 307--316, AAAI Press, Menlo Park, CA.
- Tanay, A. and Shamir, R. 2001. Computational Expansion of Genetic Networks. *Proc. Ninth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 270-278.