Data Mining Research Directions Raised by Biological Data

David Page
University of Wisconsin
Biostatistics & Medical Informatics,
Computer Science Departments

Outline

- Overview of High-Throughput Biological Data Types
  - Motivate by drug design process
  - Examples of data mining tasks for each
- Ten Observations About Bio Data
- Focus on Three Research Directions

Outline

- Overview of High-Throughput Biological Data Types
  - Motivate by drug design process
  - Examples of data mining tasks for each
- Ten Observations About Bio Data
- Focus on Three Research Directions

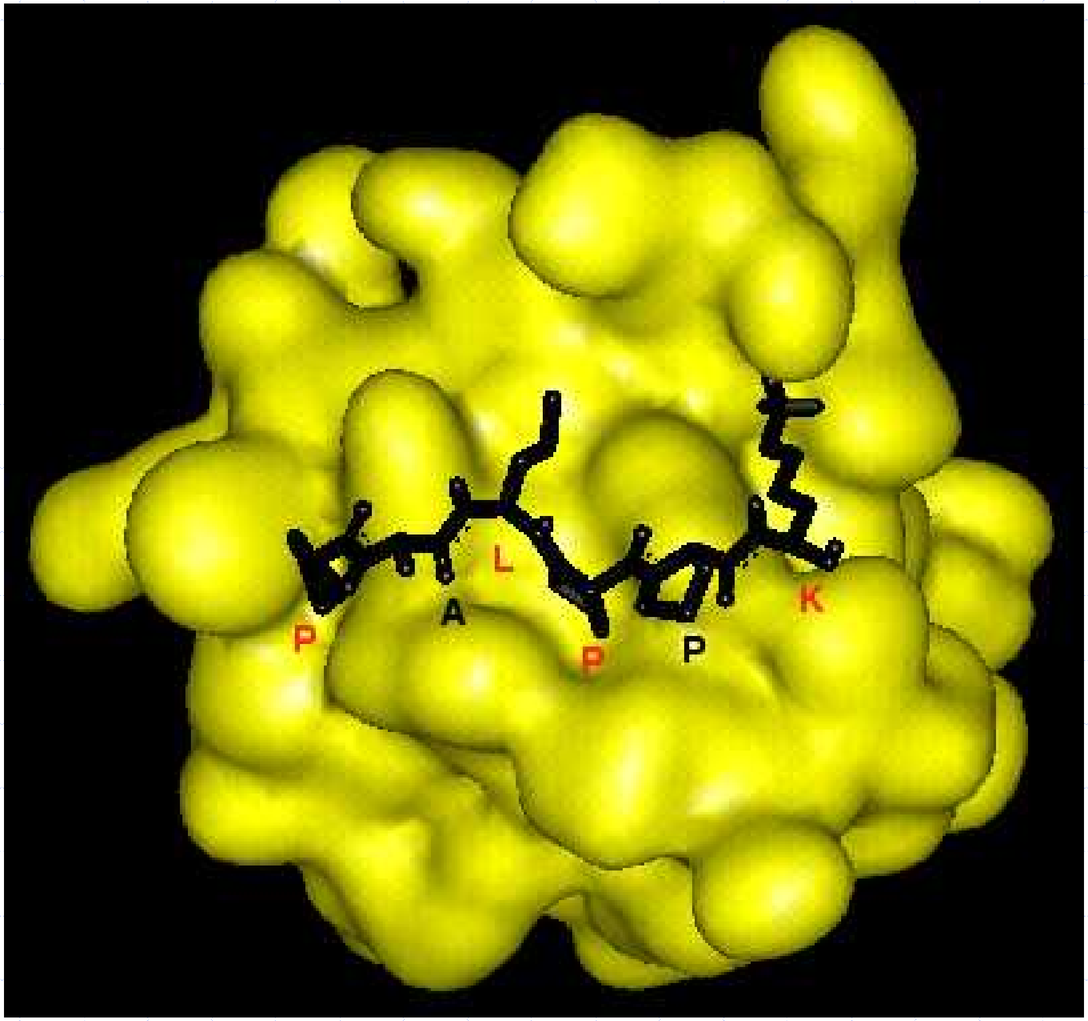
Outline

- Overview of High-Throughput Biological Data Types
  - Motivate by drug design process
  - Examples of data mining tasks for each
- Ten Observations About Bio Data
- Focus on Three Research Directions

Drugs Typically Are...

- Small organic molecules that...
- Modulate disease by binding to some target protein...
- At a location that alters the protein's behavior (e.g., antagonist or agonist).
- Target protein might be human (e.g., ACE for blood pressure) or belong to invading organism (e.g., surface protein of a bacterium).

Example of Binding (thanks Brian Kay)



So To Design a Drug:

Identify Target Protein

Knowledge of proteome/genome
Relevant biochemical pathways

↓

Determine Target Site Structure

Crystallography, NMR
Difficult if Membrane-Bound

↓

Synthesize a Molecule that Will Bind

Imperfect modeling of structure
Structures may change at binding
And even then…

Molecule Binds Target But May:

- Bind too tightly or not tightly enough.
- Be toxic.
- Have other effects (side-effects) in the body.
- Break down as soon as it gets into the body, or may not leave the body soon enough.
- It may not get to where it should in the body (e.g., crossing blood-brain barrier).
- Not diffuse from gut to bloodstream.

And Every Body is Different:

- Even if a molecule works in the test tube and works in animal studies, it may not work in people (will fail in clinical trials).
- A molecule may work for some people but not others.
- A molecule may cause harmful side-effects in some people but not others.

Places to Use Data Mining

- Finding target proteins
  - Signaling pathways
  - Regulatory pathways
  - Metabolic pathways
- Inferring target site structure
- Predicting who will respond positively: pharmacogenetics, pharmacogenomics

High-Throughput Biological Data

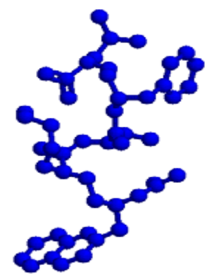
- Robotic high-throughput screening of molecules for bio activities
- Gene Chips (Microarrays)
- Single-nucleotide polymorphisms (SNPs)
- Proteomics
  - Detecting proteins in sample
  - Protein-protein interactions
- Metabolomics (metabonomics), lipomics

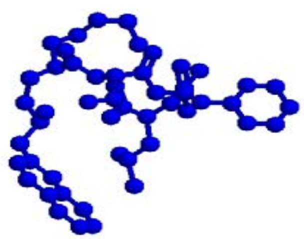
Low-Throughput Biological Data (High-Throughput Future?)

- Sequencing
  - Amount of data may seem high-throughput, but...
  - only haploid sequence, no one knows his/her sequence currently (SNPs are surrogate)
- Protein complexes, post-translational modifications
- Protein structures
  - X-ray crystallography
  - NMR

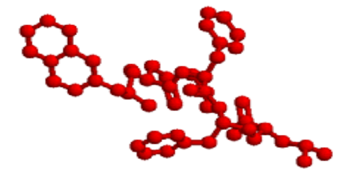
High-Throughput Screen Results

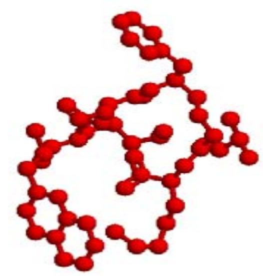
Active





Inactive





Data Mining Task

- Predict active vs. inactive from structure
- Why need data mining?
  - 100,000 molecules instead of 4.
  - Each molecule can take multiple “stable” shapes (low-energy conformers) by rotating single bonds... only one might permit it to bind to target protein.
  - For each molecule, only a few atoms are responsible for activity (don’t know which).

Need Target Proteins, so Need:

- More complete knowledge of biological pathways for signaling, regulation, metabolism, etc.
- Which proteins change with disease (more or less of the protein, change in what it does).

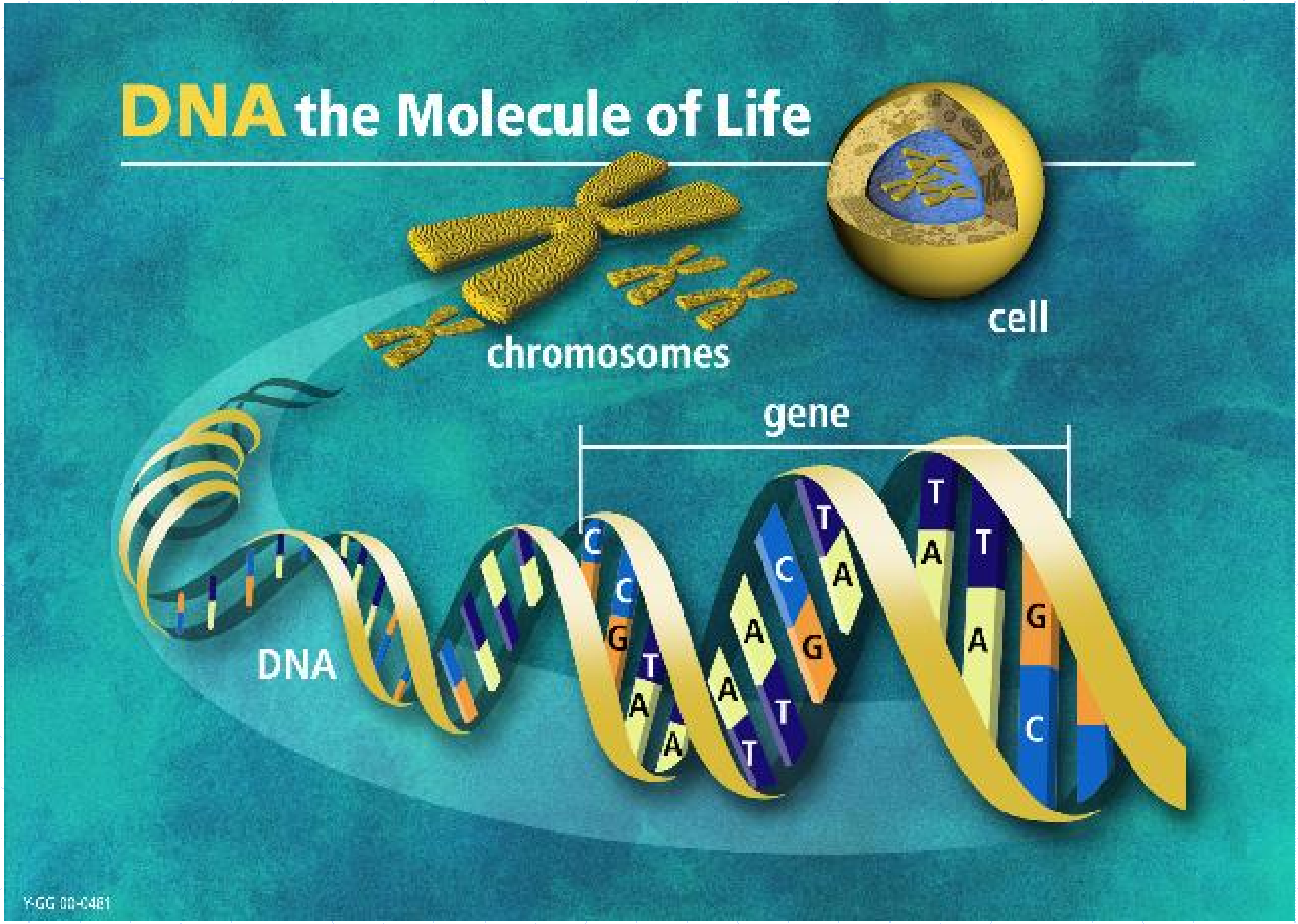
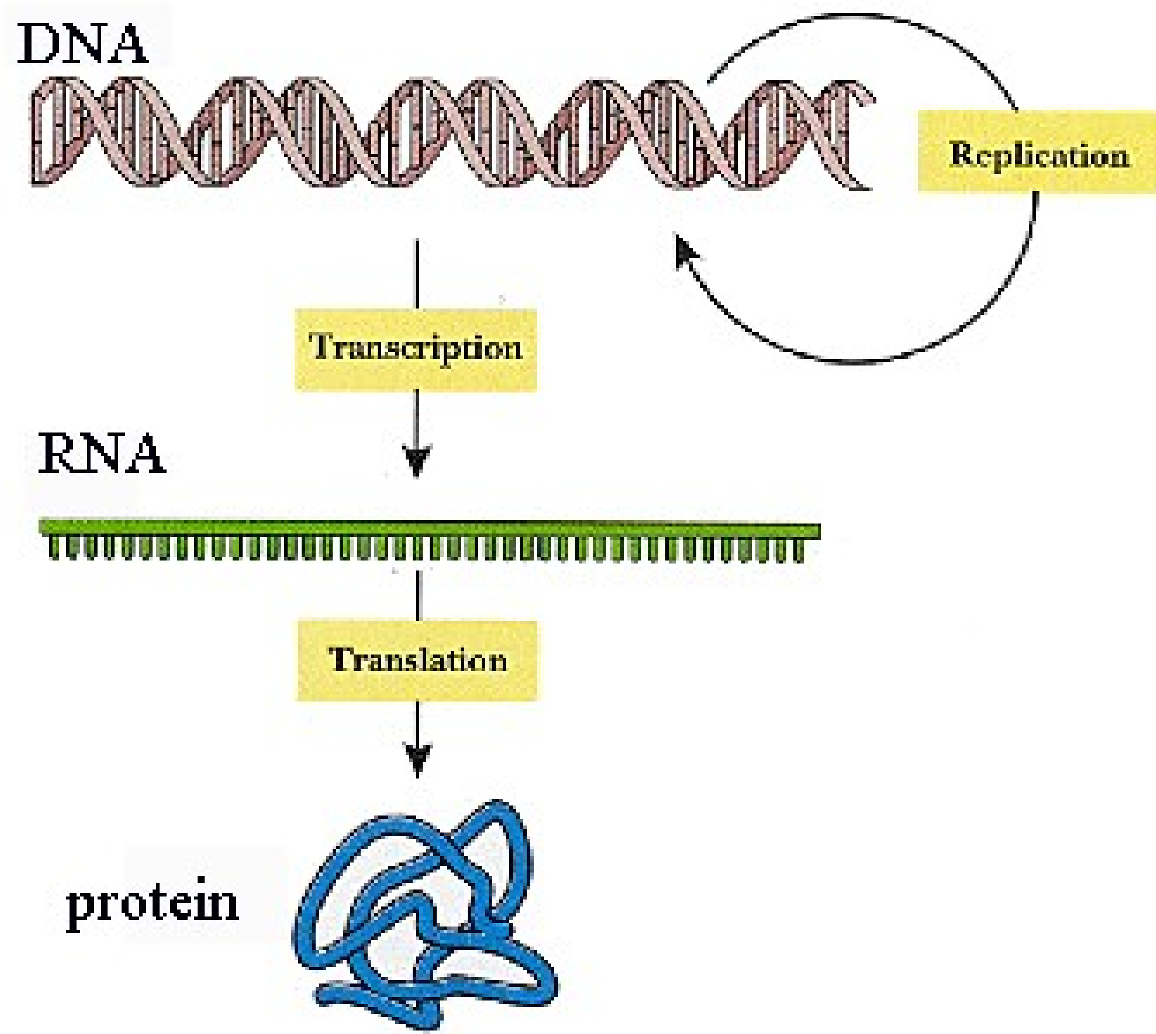


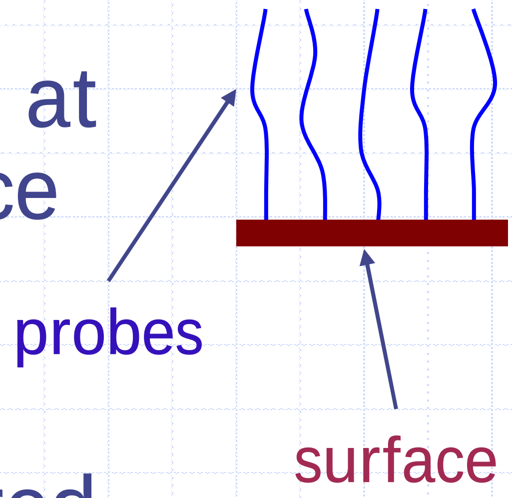
image from the DOE Human Genome Program
http://www.ornl.gov/hgmis

The “Central Dogma” of Mol Bio



Microarrays ("Gene Chips")

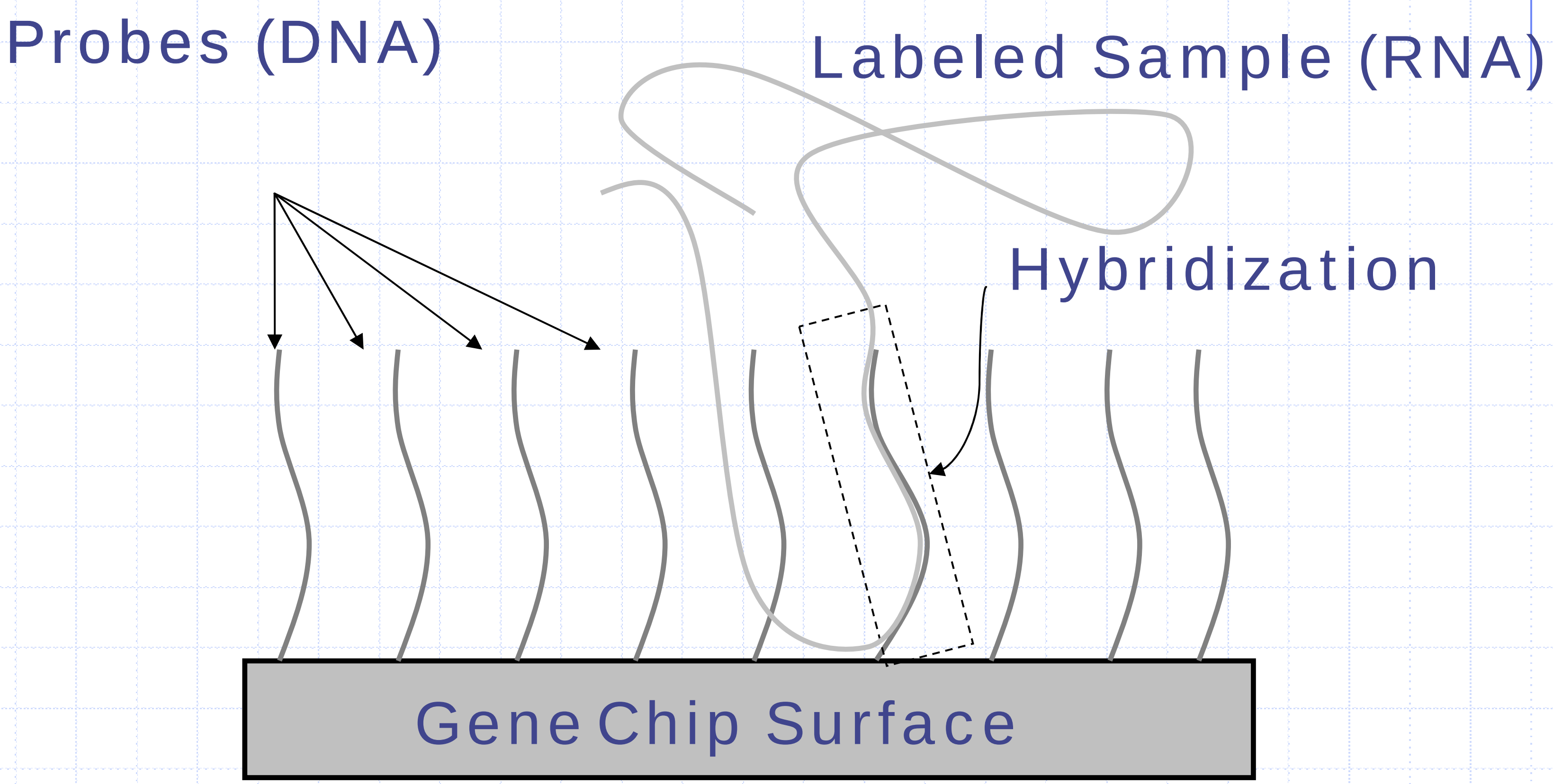
- Specific probes synthesized at known spot on chip's surface



- Probes complementary to RNA of genes to be measured

- Typical gene (1kb+) MUCH longer than typical probe (24 bases)

How Microarrays Work



Example of Microarray Data

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

View 1: Data Points are Genes

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

View 2: Data Points are Samples

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

Supervision: Add Classes

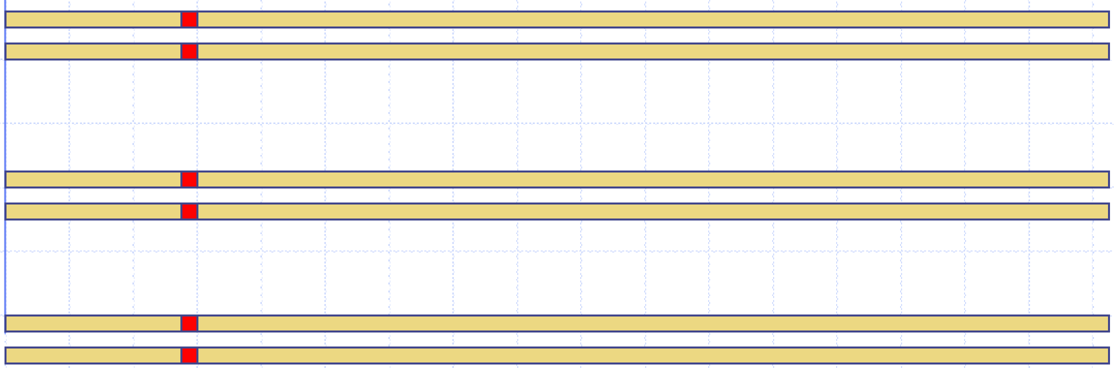
Person ↓ Gene →	A28202_ac		AB00014_at		AB00015_at		. . .	CLASS
Person 1	P	1142.0	A	321.0	P	2567.2	. . .	myeloma
Person 2	A	-586.3	P	586.1	P	759.0	. . .	normal
Person 3	A	105.2	A	559.3	P	3210.7	. . .	myeloma
Person 4	P	-42.8	P	692.1	P	812.0	. . .	normal
. . .		. .		. . . .		.		
. . .		. .		. . . .		.		
. . .		. .		. . . .		.		

Some Problems

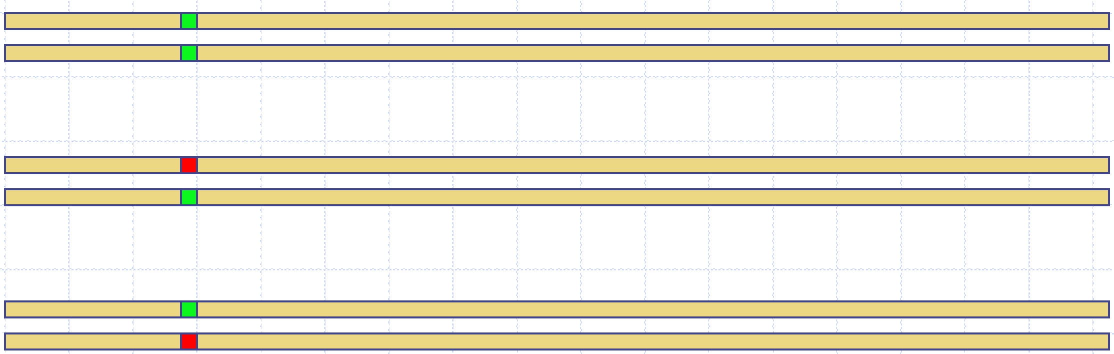
- For insight into diseases, say cancer: many changes. Can get nearly 100% accuracy but little idea of the key changes and little knowledge of succeptibility.
- For regulatory networks: much missing information... proteins, complexes, post-translational modifications... hard to get insight into causality.

One day we all will know our sequences (if we wish)...

Succeptible to Disease D or Responds to Treatment T



Not Succeptible or Not Responding



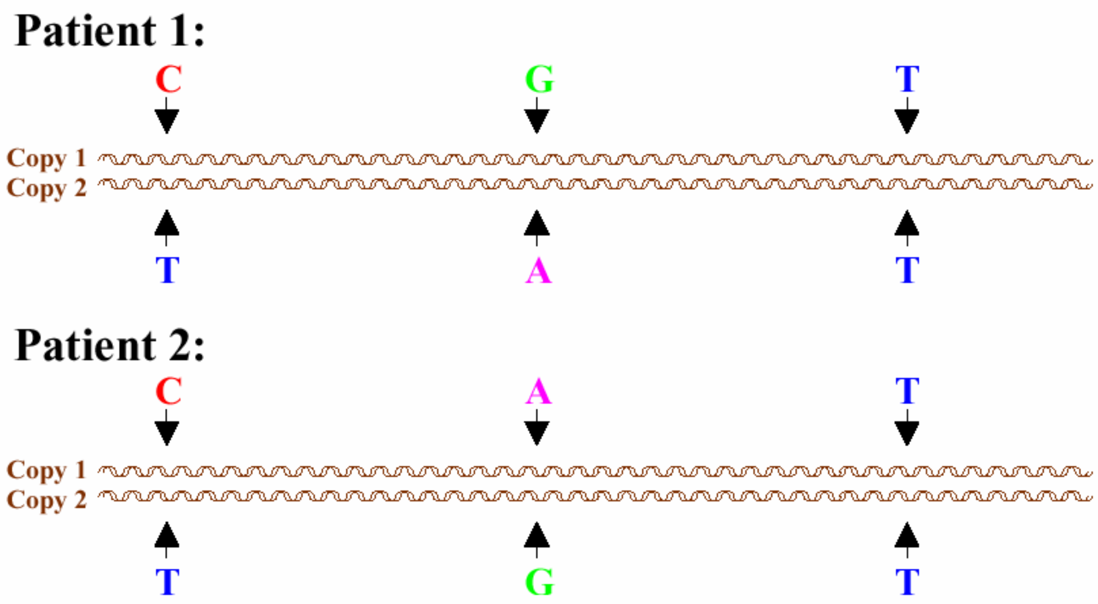
Single-Nucleotide Polymorphisms

- 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

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 Task from SNP Data

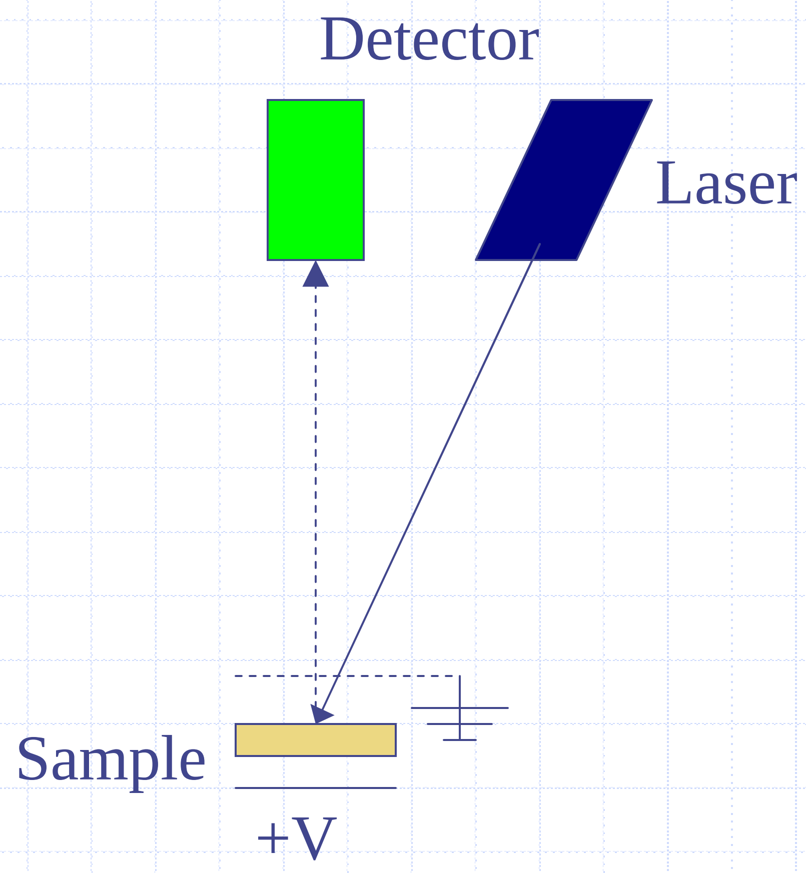
- Distinguish disease from normal by SNP pattern.
- Probably cannot do this near 100% accuracy, because SNP pattern does not change with disease or with time.
- But if can do this significantly better than chance, it suggests a genetic predisposition to the disease.

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

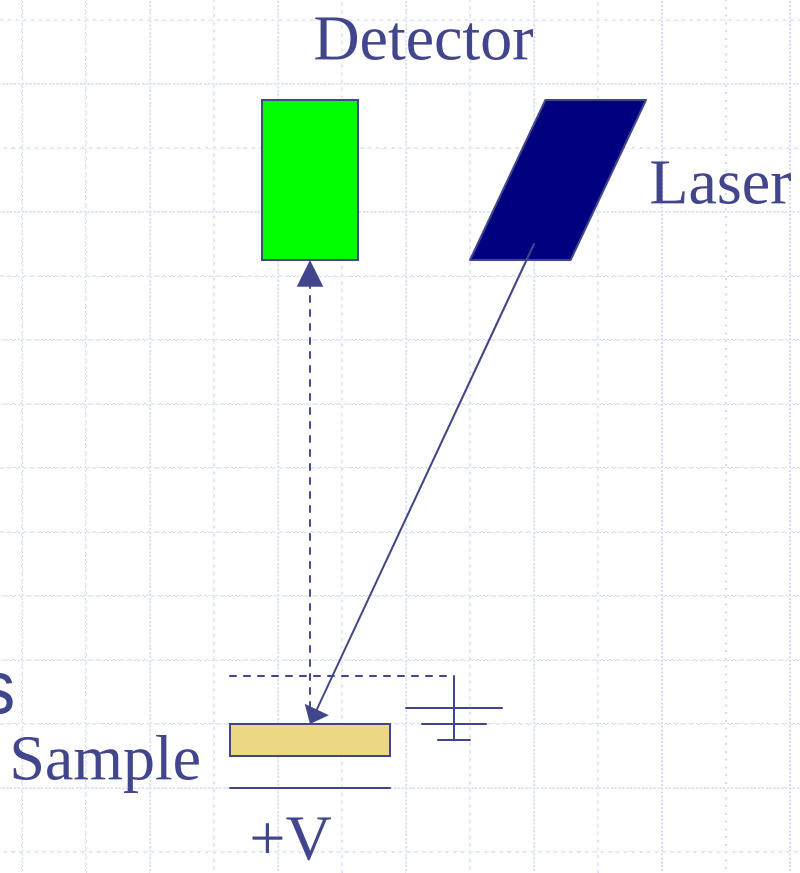
Time-of-Flight (TOF) Mass Spectrometry

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

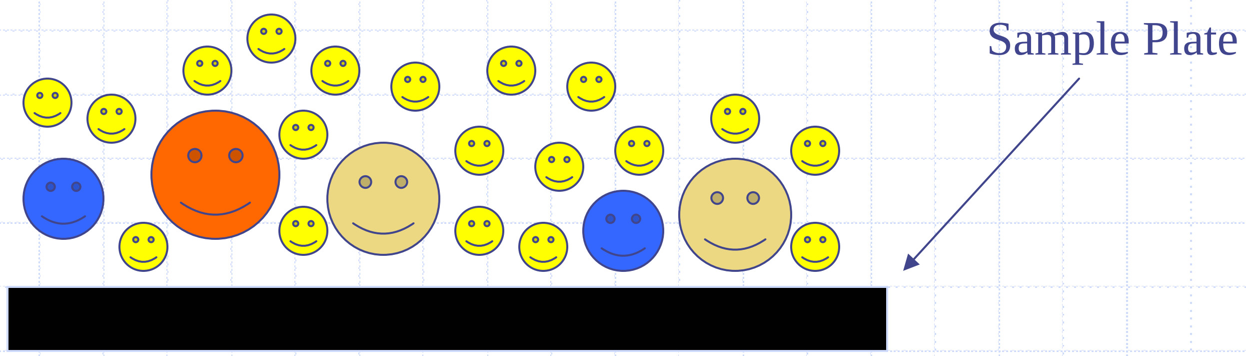


Time-of-Flight (TOF) Mass Spectrometry 2

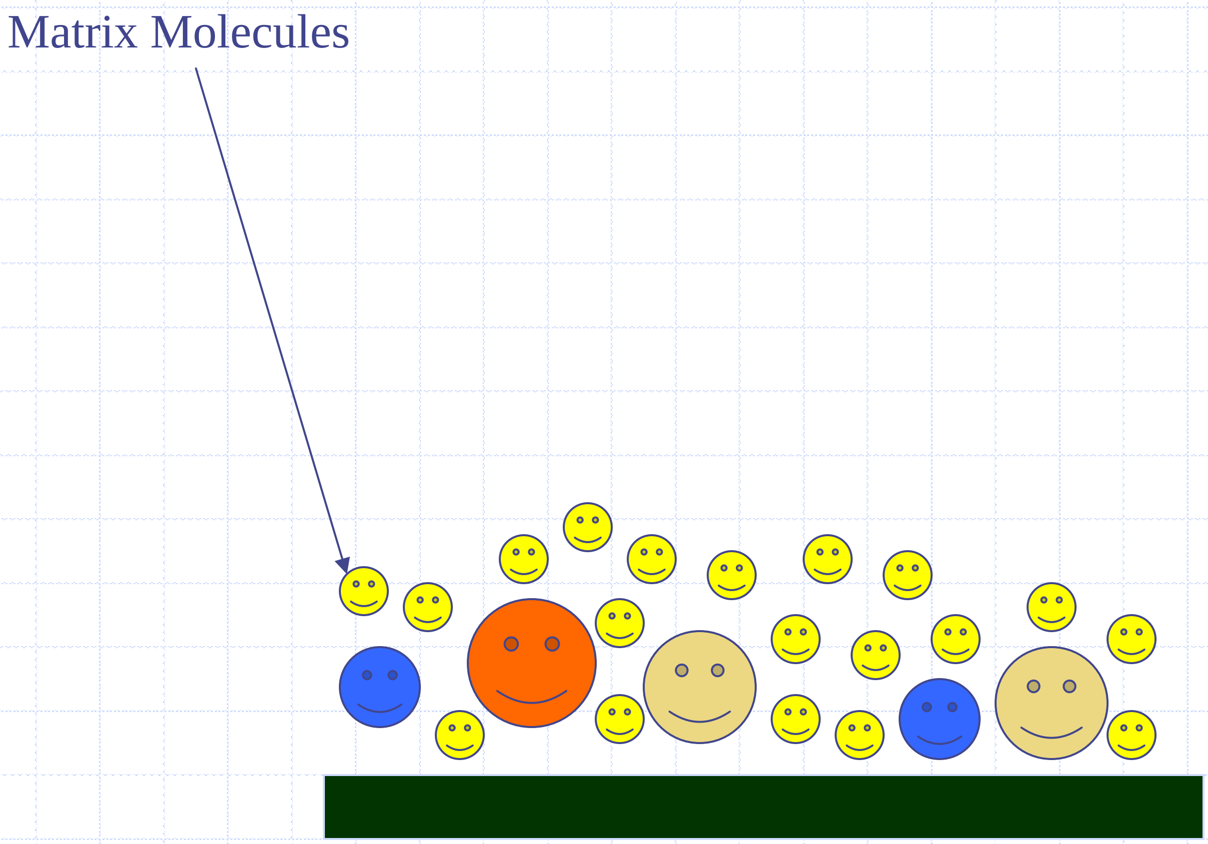
- *Matrix-Assisted Laser Desorption-Ionization* (MALDI)
- Crystalloid structures made using proton-rich matrix molecule
- Hitting crystalloid with laser causes molecules to ionize and "fly" towards detector



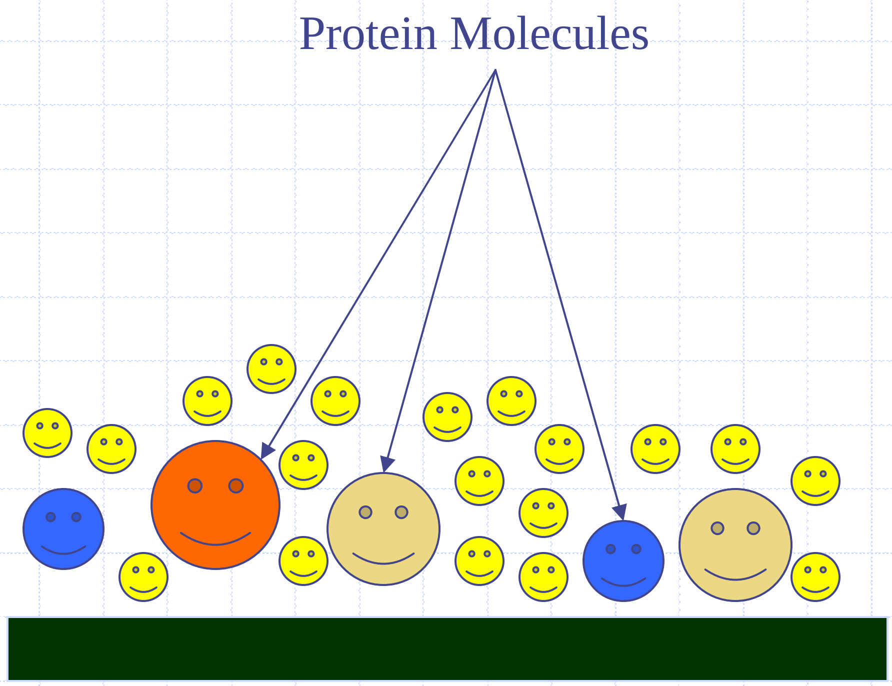
Time-of-Flight Demonstration 0 (thanks Sean McIlwain)



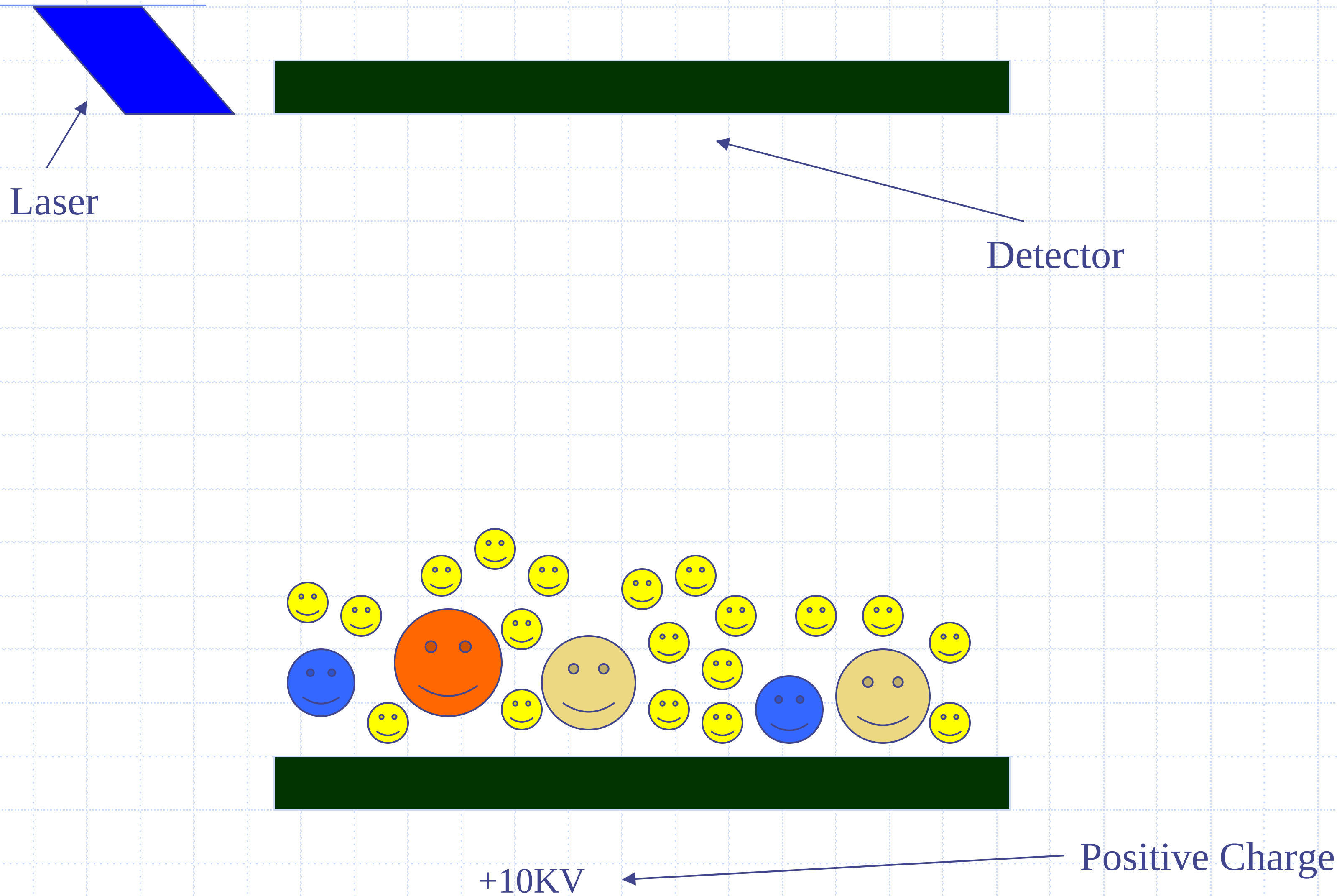
Time-of-Flight Demonstration 1



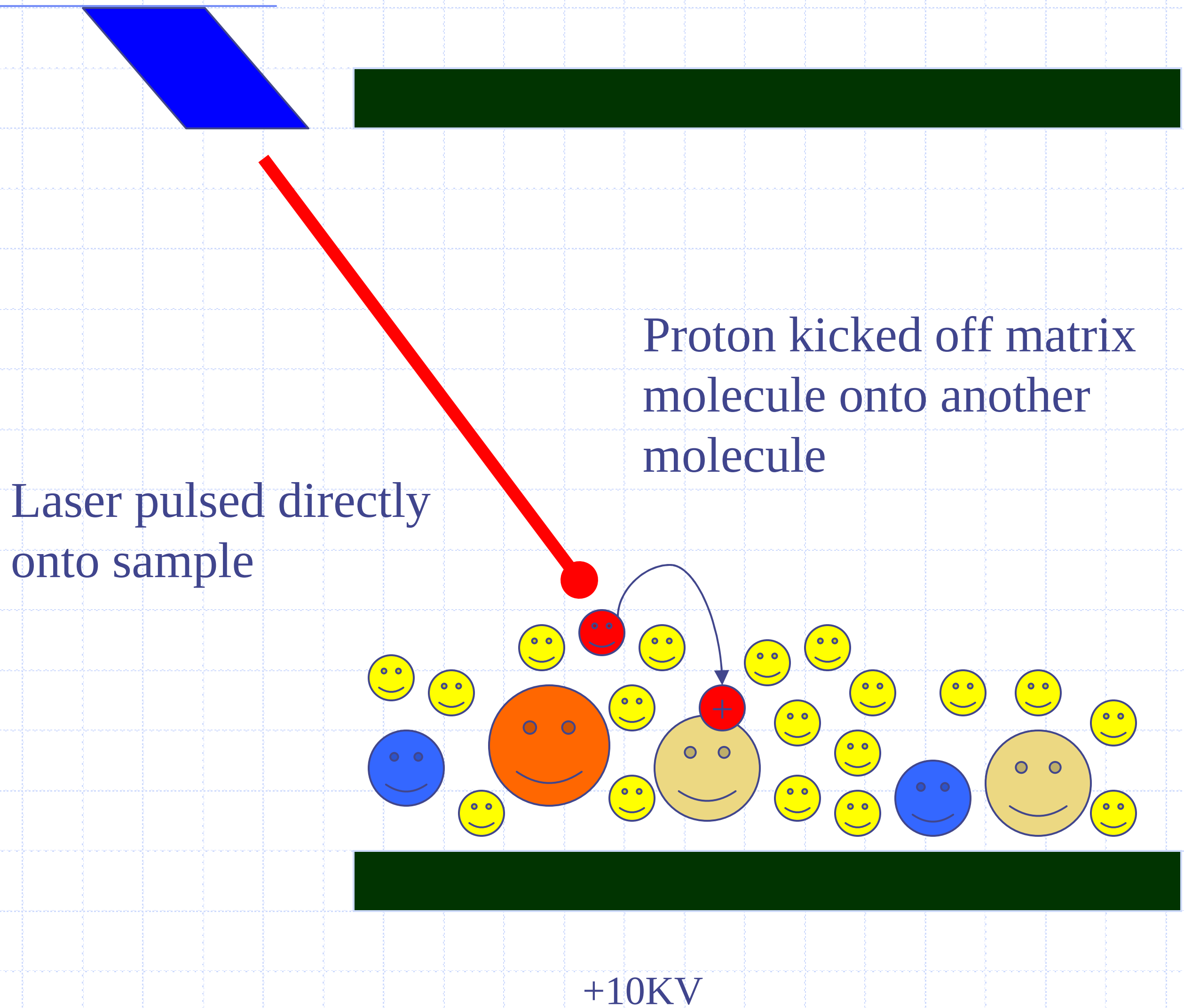
Time-of-Flight Demonstration 2



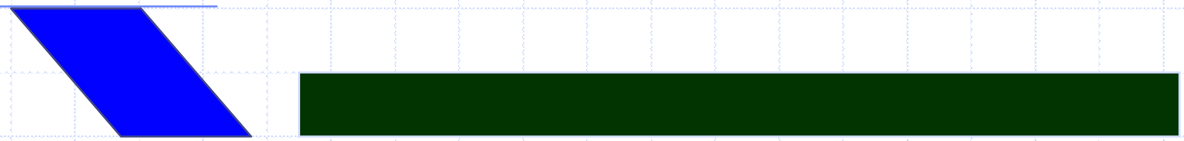
Time-of-Flight Demonstration 3



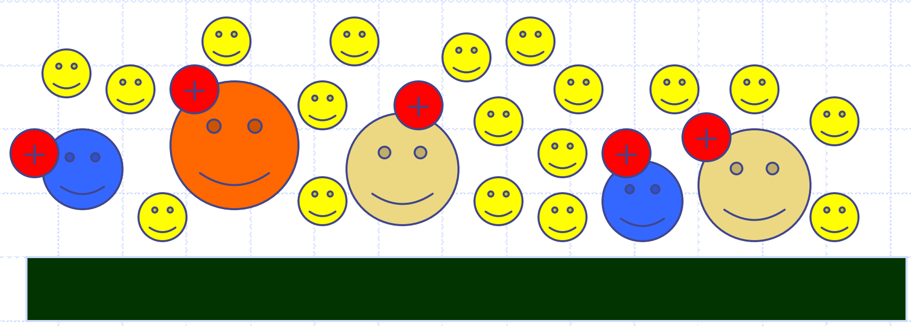
Time-of-Flight Demonstration 4



Time-of-Flight Demonstration 5

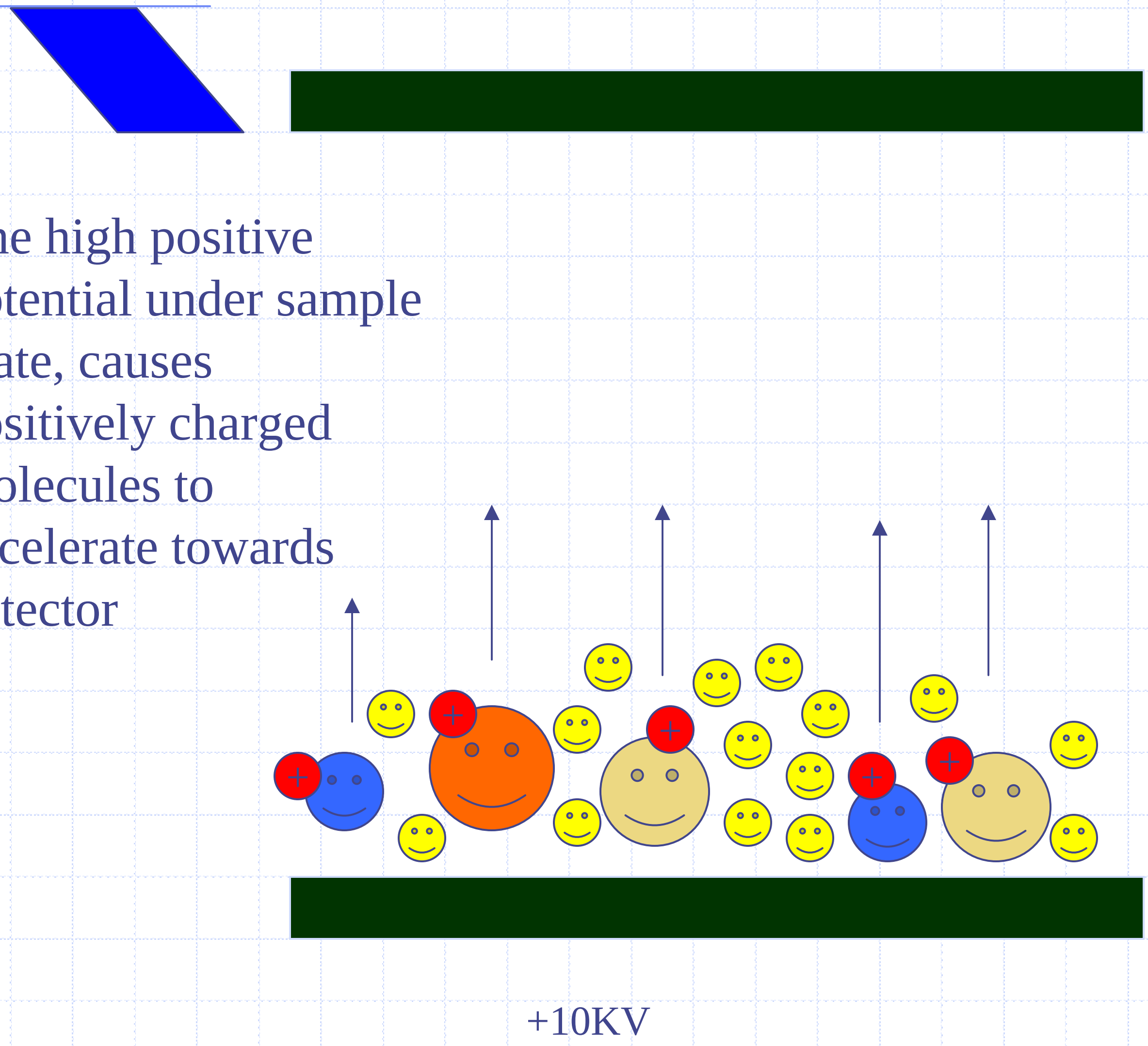


Lots of protons kicked off matrix ions, giving rise to more positively charged molecules

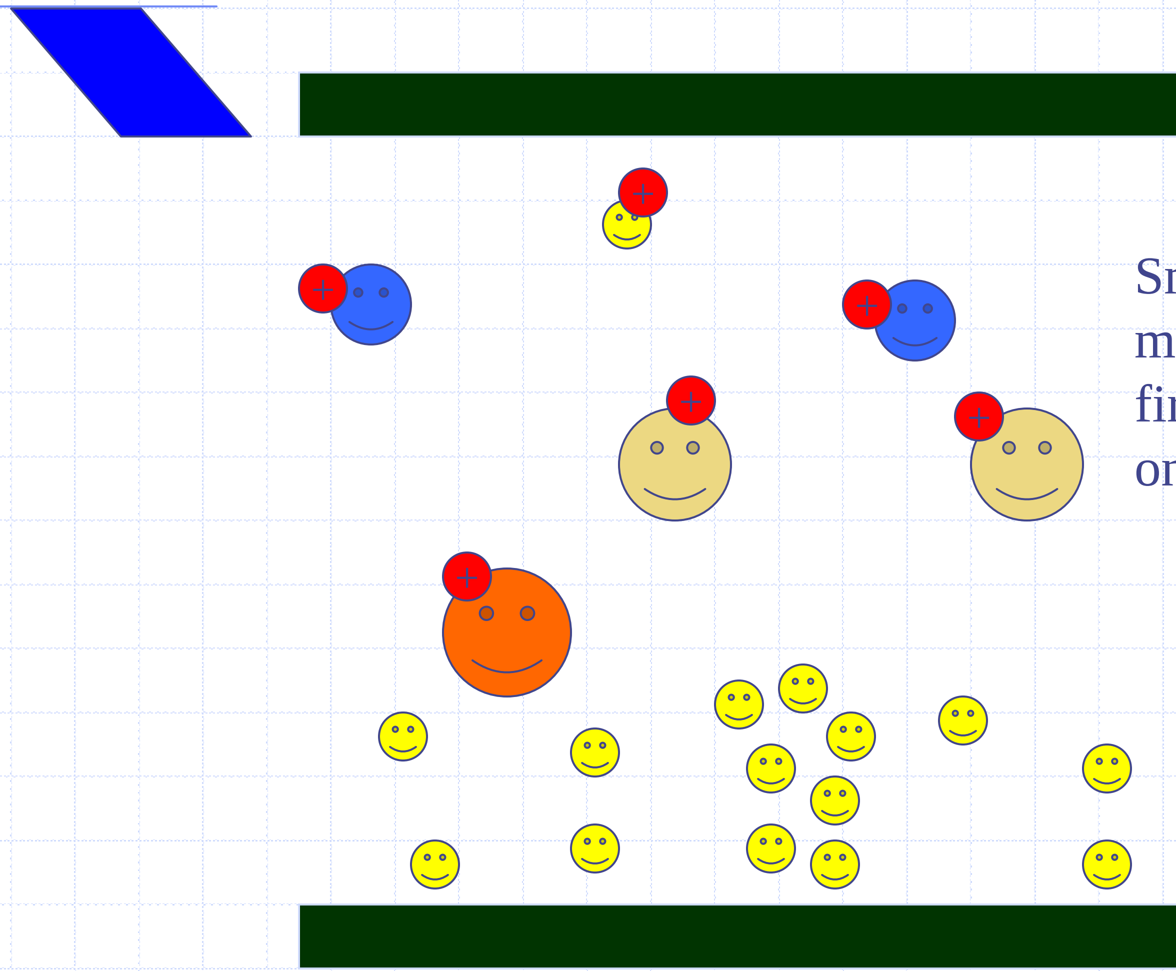


Time-of-Flight Demonstration 6

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



Time-of-Flight Demonstration 7

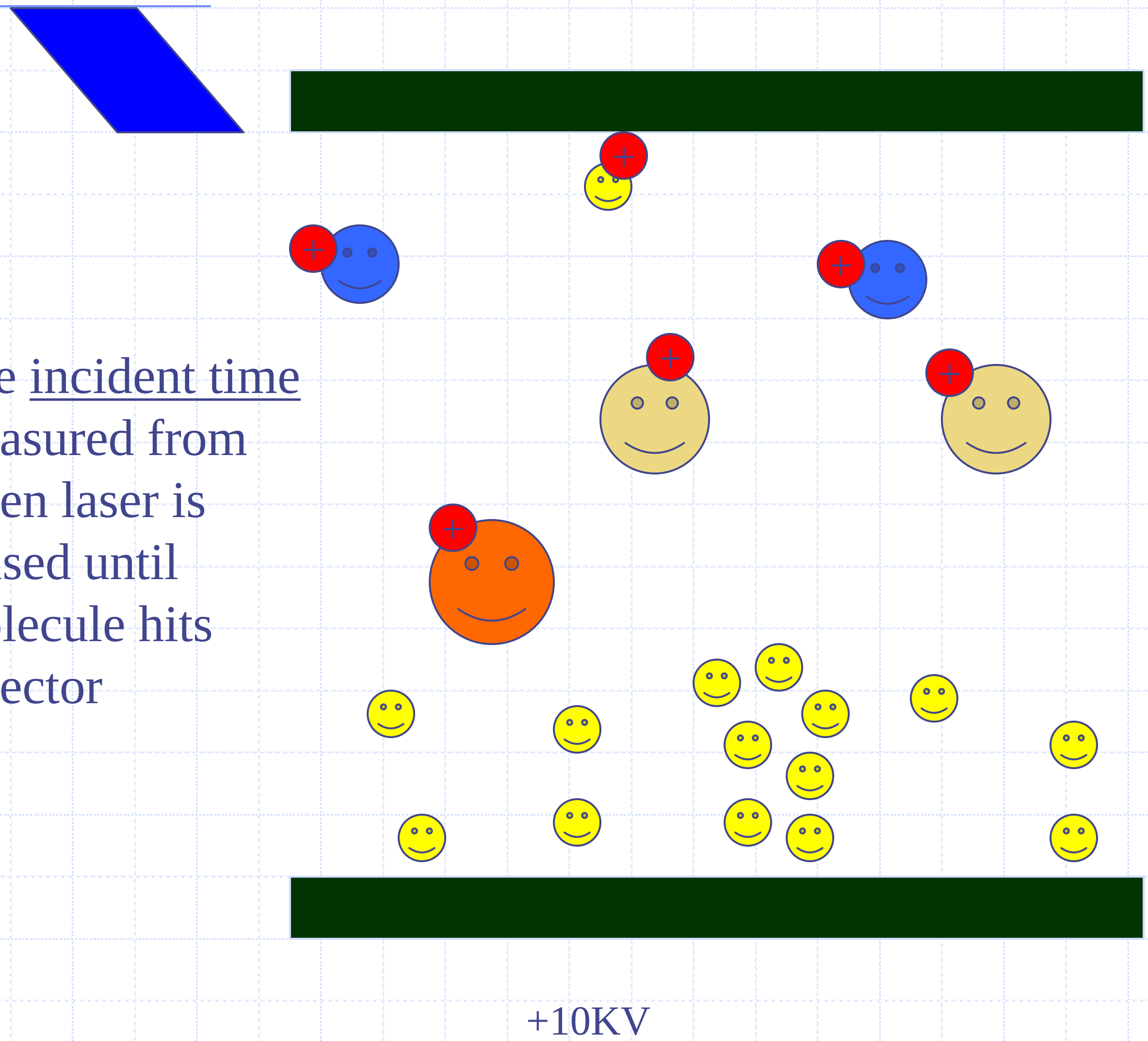


Smaller mass molecules hit detector first, while heavier ones detected later

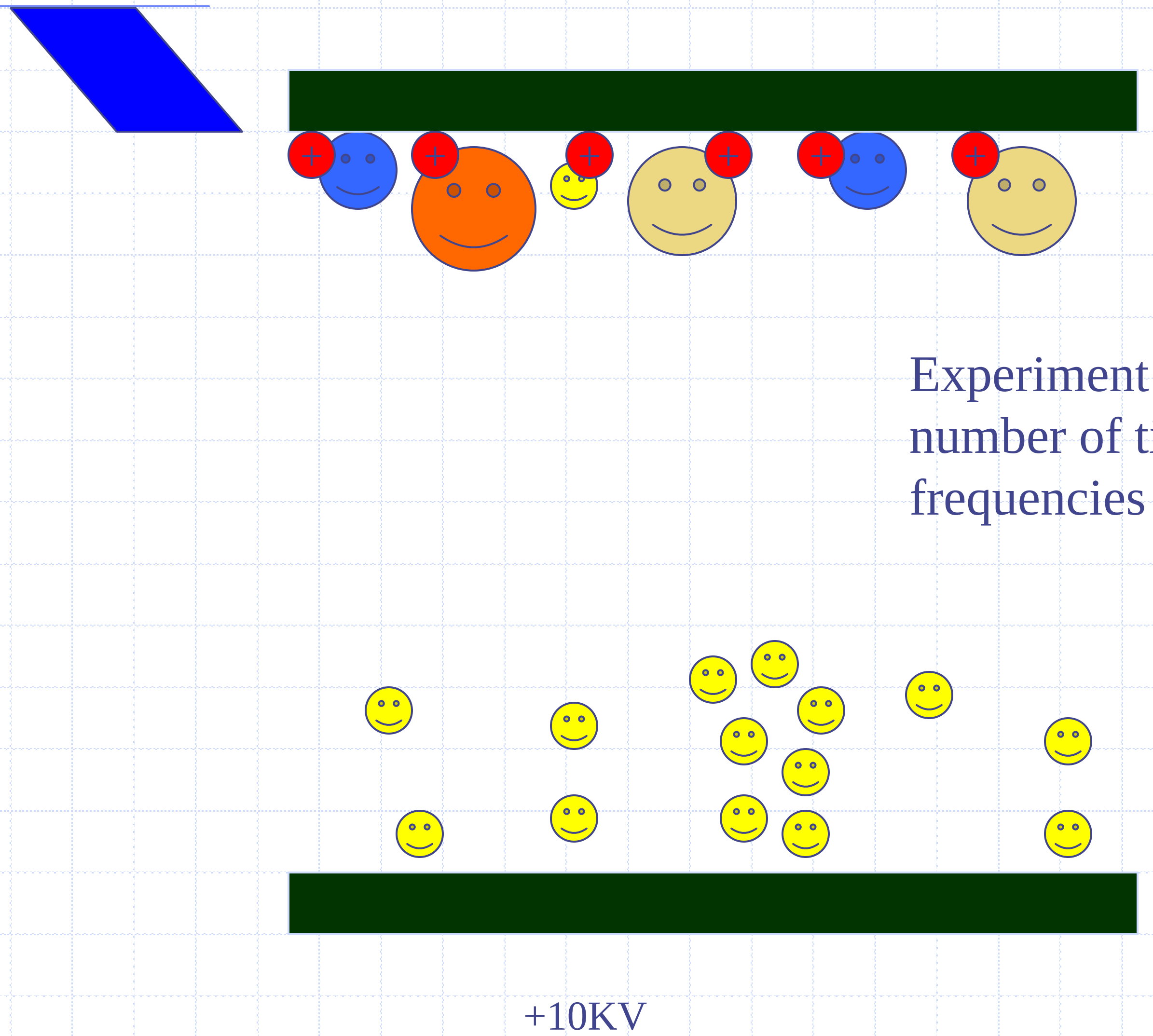
+10Kv

Time-of-Flight Demonstration 8

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

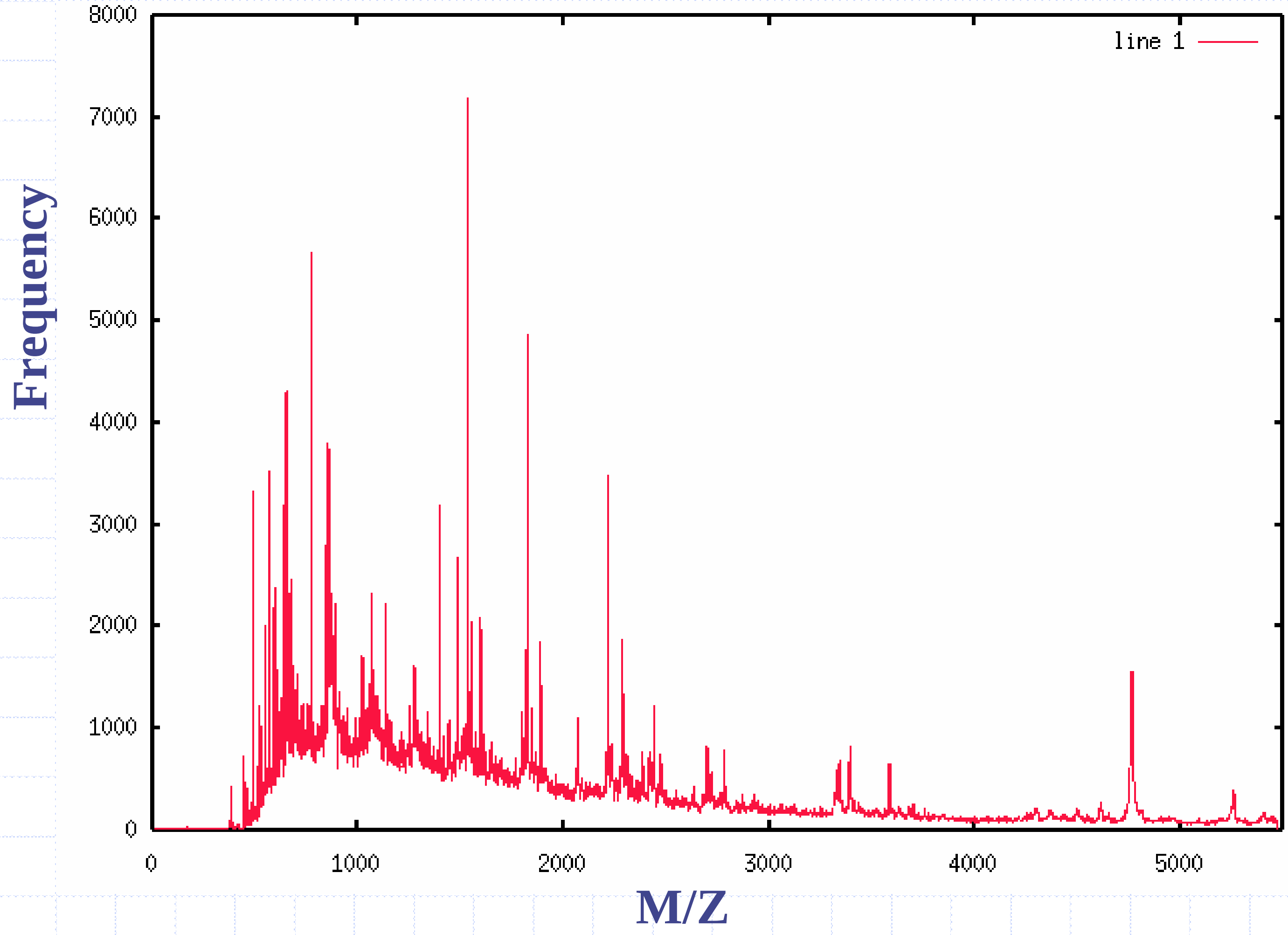


Time-of-Flight Demonstration 9



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

Example Spectrum



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

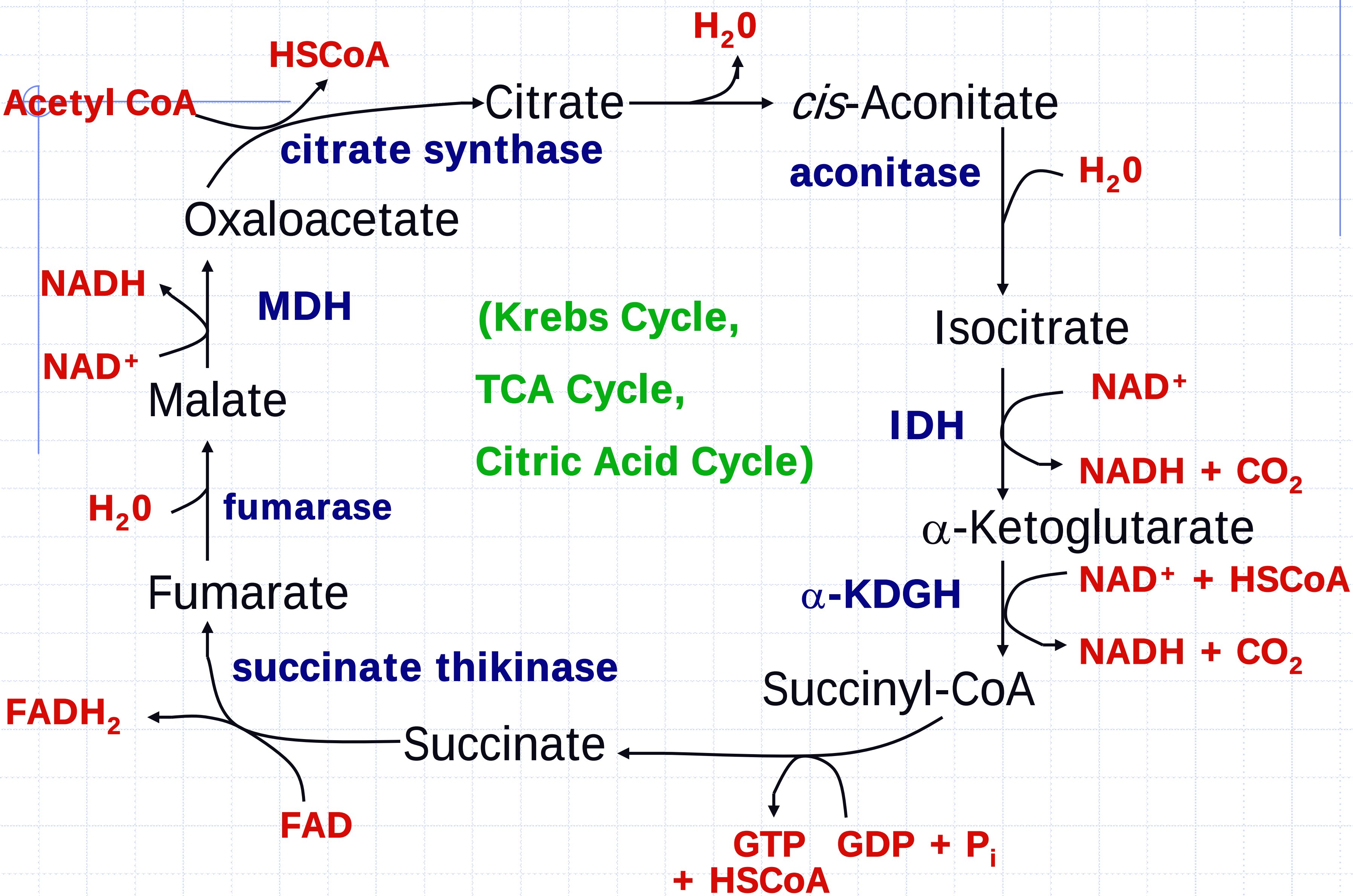
Tasks from Mass Spec Data

- Determine which proteins are present in a sample.
- Distinguish samples based on spectra, e.g. distinguish disease from normal based on spectrum.

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)

Metabolic Pathway Example



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

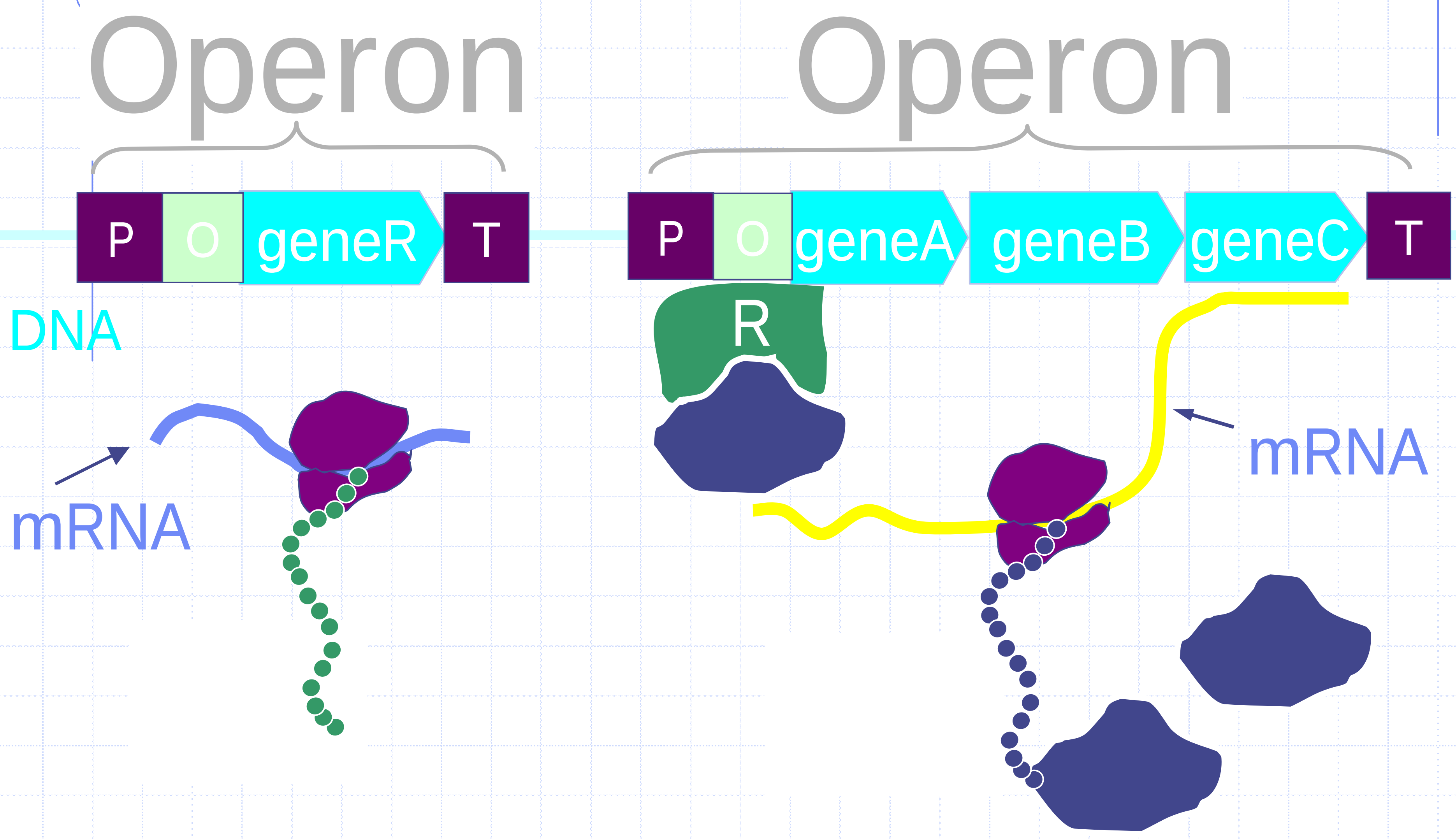
Auxotrophic Growth Experiments

- Which knock-outs (organisms with a gene removed or incapacitated) will grow on which media?
- Example: yeast with one gene knocked out, growing on media with some nutrients.
- Can be carried out robotically.
- Example: King et al., Nature 2004.

Low-Throughput Biological Data (High-Throughput Future?)

- Sequencing
  - Amount of data may seem high-throughput, but...
  - only haploid sequence, no one knows his/her sequence currently (SNPs are surrogate)
- Protein complexes, post-translational modifications
- Protein structures
  - X-ray crystallography
  - NMR

In E. coli (thanks Irene Ong)



Outline

- Overview of High-Throughput Biological Data Types
  - Motivate by drug design process
  - Examples of data mining tasks for each
- Ten Observations About Bio Data
- Focus on Three Research Directions

Observation 1: Noise

- Much work in reducing noise in microarray data (mostly by statisticians)
- Noise even worse in mass spec data
- Noise issues in every data type discussed

Obs 2: Missing Data or Info

- Missing data common in SNP data sets
- For inducing regulatory models from microarrays, much of the work is carried out by (modified) proteins such as transcription factors (TFs)… levels of TFs don’t change must, modifications not measured.

Obs 3: Exceptions Rule

- Biology is full of exceptions to (almost) every general statement.
- Therefore often need probabilities in the models we build.

Obs 4: Wide, not Deep

- Each of the high-throughput data types typically yields thousands to millions of features.
- All but molecule screening typically are run on at most a few hundred samples.
- And molecule screening typically yields less than a hundred positive examples (active molecules).

Obs 5: Comprehensibility is Key

- Structure-Activity models for molecules are useless unless they give chemists insight into what to make next.
- New biological pathway models need to be tested, published, and used to gain clues to potential target proteins.
- Physicians and patients will want to know what, in a SNP pattern, indicates the patient's succeptibility to a disease.

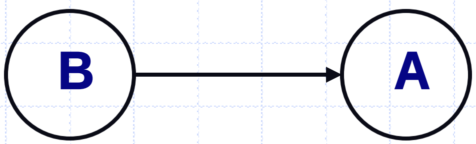
Obs 6: Time Often Important

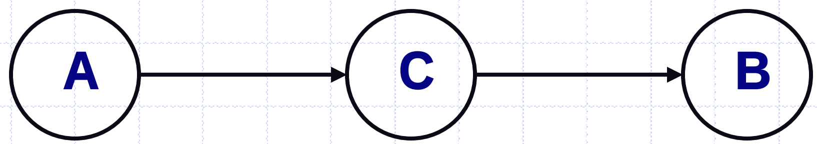
- Would like to see how the set of proteins present change over time after a change in condition.
- Time-series microarray data can give more insight into causality for network modeling.
  - How do we model time?
  - How rapidly should we sample?

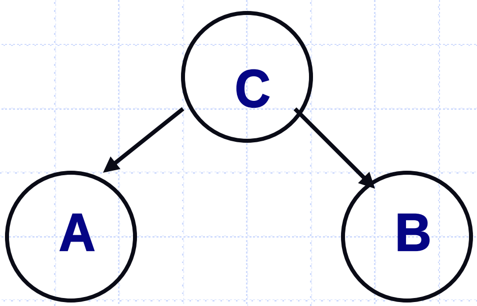
Without Time (or other help), Hard to Get Causality

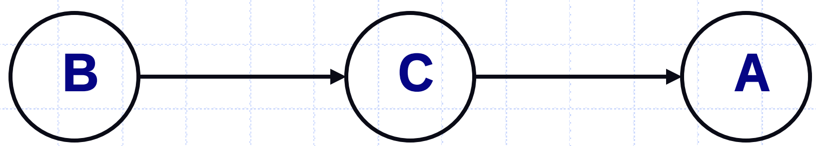
A is a good predictor of B. But is A <u>regulating</u> B??

Ground truth might be:









Or a more complicated variant

Obs 7: Opportunities to Perturb and Observe

- Can subject a cell to various conditions.
- For some organisms (e.g., yeast), can knock-out a gene. May become easier with RNAi.
- Harder to do multi-gene knock-outs.
- Should open the door to more applications of active learning.

Obs 8: Background Knowledge

- Partial models often are available.
- Because number of data points often is limited, try revising an existing model from data rather than constructing it from scratch.

Obs 9: Diverse Data Types Relevant to any Given System

- To study a biological pathway, it would be ideal to have data on:
  - Expression of related genes
  - Protein levels, protein-protein interactions, post-translational modifications
  - Structures of proteins and the small molecules that also interact with them
  - Levels of metabolites, etc.
- Systems Biology becoming prominent

Obs 10: Models and Data Points often are “Multi-Relational”

- A pathway consist of proteins and other biomolecules and the interactions among them.
- A protein-protein interaction consists of several pairs of atoms, one from each protein, that interact in one of several ways (charge, hydrophobicity, steric).
- A molecule consists of atoms and relations among them (bonds, distances).
- Such relationships are most easily represented in a database with multiple relational tables. The same is true of diverse data types related to a single system. If collapsing to a single table, other issues arise.

Outline

- Overview of High-Throughput Biological Data Types
  - Motivate by drug design process
  - Examples of data mining tasks for each
- Ten Observations About Bio Data
- Focus on Three Research Directions

1. Fully-Automated Discovery

- Update/revise an existing theory based on results of experiments.
- Active learning: propose experiments.
- Given automated (robotic) high-throughput data collection techniques, human may not be needed in the process at all.

Robot Scientist

- R.D. King, K.E. Whelan, F.M. Jones, P.K.G. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247-252, 2004.

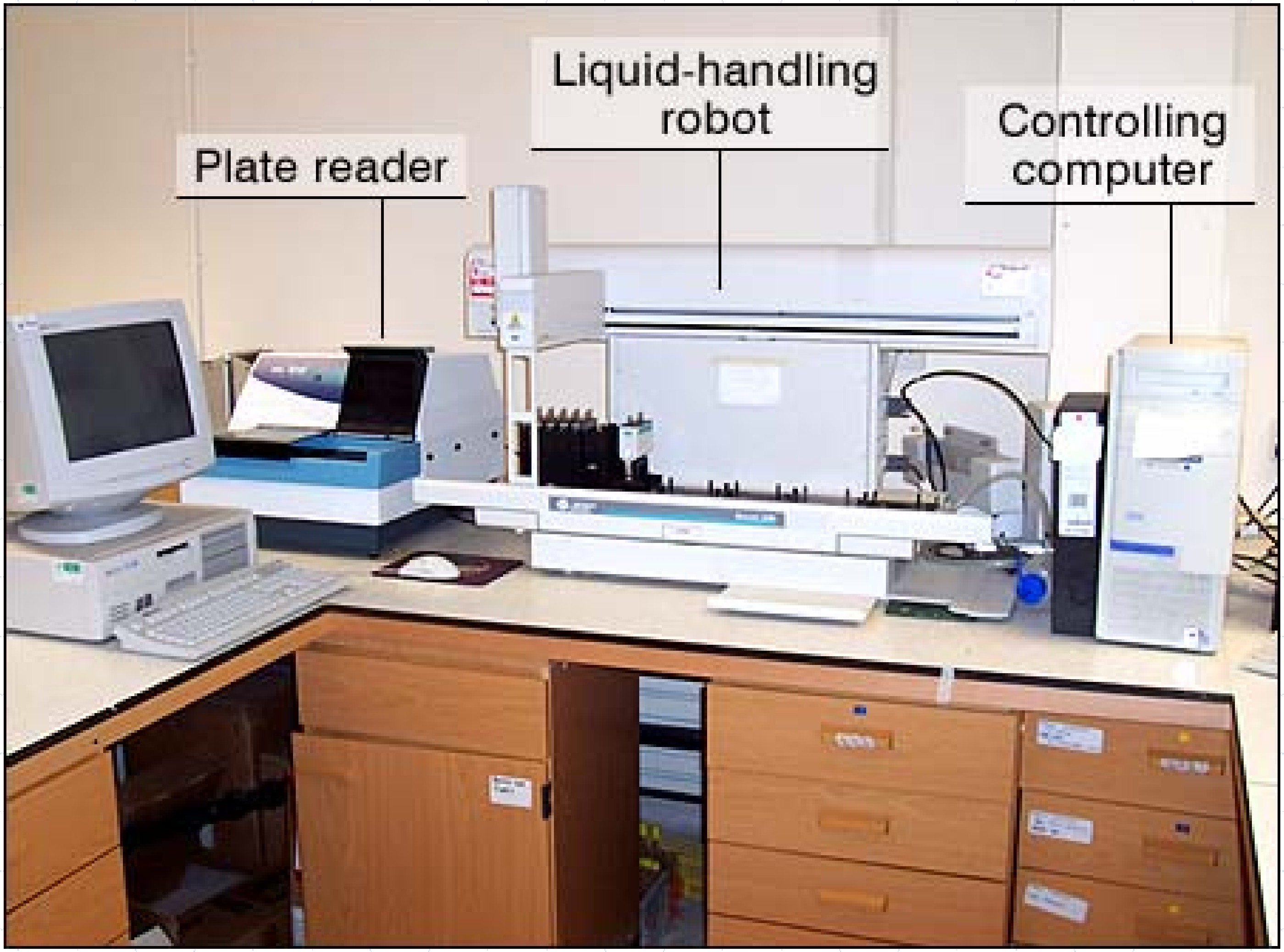
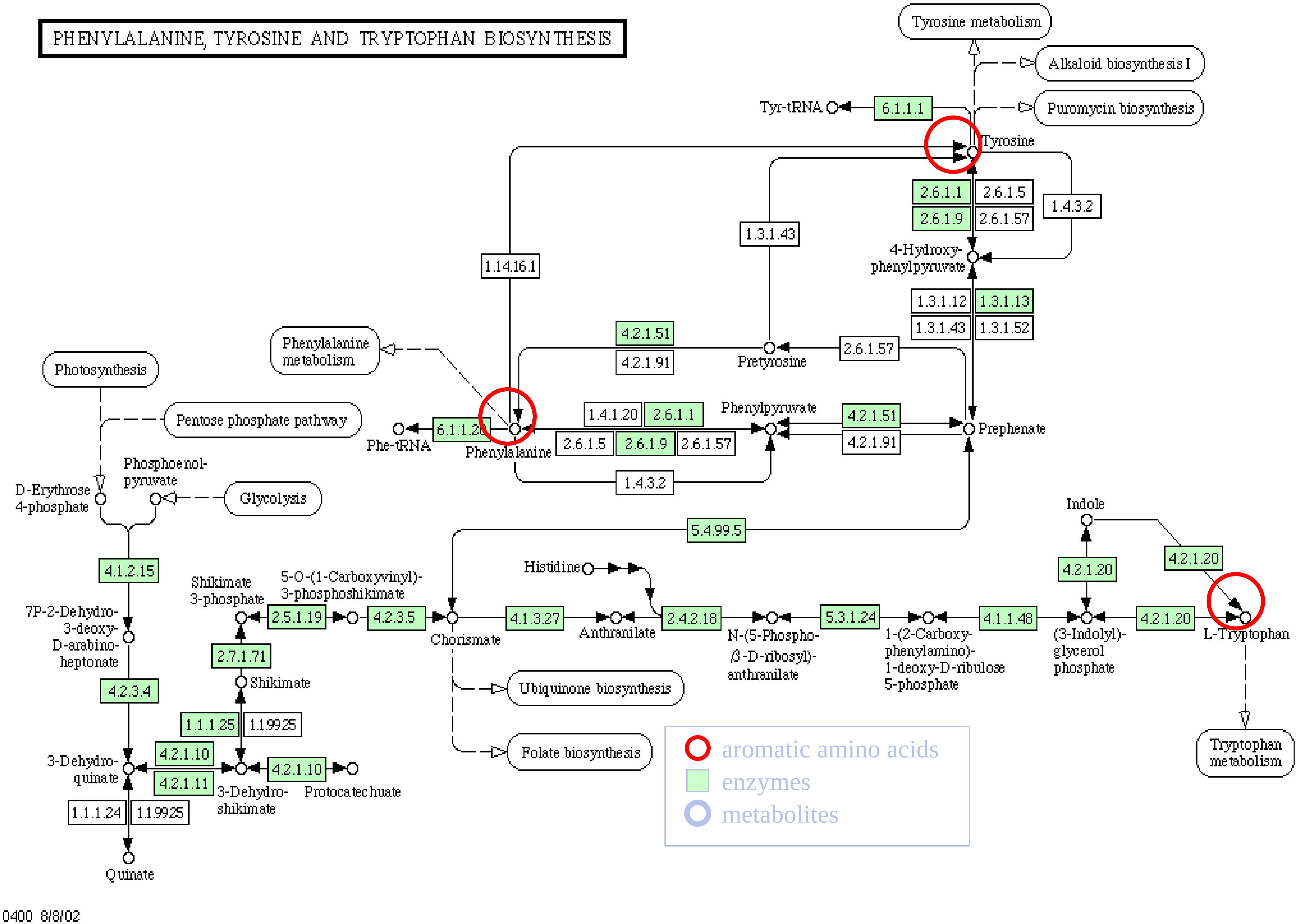


Plate reader
Liquid-handling robot
Controlling computer

Their Experiment

- Began with a partial model of amino acid biosynthesis in yeast.
- Learning algorithm:
  - Proposed experiments to test the model
  - Modified the model based on the experiments
  - When several alternative modifications competed, proposed experiments to distinguish between them.


PHENYLALANINE, TYROSINE AND TRYPTOPHAN BIOSYNTHESIS
Tyrosine metabolism
Alkaloid biosynthesis I
Puromycin biosynthesis
Tyr-tRNA
6.1.1.1
Tyrosine
2.6.1.1
2.6.1.5
2.6.1.9
2.6.1.57
1.4.3.2
1.3.1.43
1.14.16.1
4-Hydroxy-phenylpyruvate
1.3.1.12
1.3.1.13
1.3.1.43
1.3.1.52
Phenylalanine metabolism
4.2.1.51
4.2.1.91
Pretyrosine
2.6.1.57
Photosynthesis
Pentose phosphate pathway
1.4.1.20
2.6.1.1
Phenylpyruvate
4.2.1.51
4.2.1.91
Prephenate
Phe-tRNA
6.1.1.20
Phenylalanine
2.6.1.5
2.6.1.9
2.6.1.57
1.4.3.2
Phosphoenol-pyruvate
D-Erythrose 4-phosphate
Glycolysis
5.4.99.5
Indole
4.1.2.15
Histidine
4.2.1.20
4.2.1.20
Shikimate 3-phosphate
5-O-(1-Carboxyvinyl)-3-phosphoshikimate
7P-2-Dehydro-3-deoxy-D-arabino-heptonate
2.5.1.19
4.2.3.5
Chorismate
4.1.3.27
Anthranilate
2.4.2.18
N-(5-Phospho-β-D-ribosyl)-anthranilate
5.3.1.24
1-(2-Carboxy-phenylamino)-1-deoxy-D-ribulose 5-phosphate
4.1.1.48
(3-Indolyl)-glycerol phosphate
4.2.1.20
L-Tryptophan
2.7.1.71
Shikimate
4.2.3.4
Ubiquinone biosynthesis
1.1.1.25
1.1.99.25
4.2.1.10
4.2.1.10
3-Dehydro-quinate
4.2.1.11
3-Dehydro-shikimate
Protocatechuate
Folate biosynthesis
aromatic amino acids
enzymes
metabolites
Tryptophan metabolism
1.1.1.24
1.1.99.25
Quinate
0400 8/8/02

Much Room to Carry Further

- Auxotrophic growth experiments are relatively simple. More complex experiments?
- This was rediscovery experiment. Can we discover something new?
- Advances in active learning and theory revision can have major impact here.

2. Multi-Relational Data Mining and Statistical Relational Learning

- MRDM: Mine databases with multiple relational tables.
- SRL: Combine multi-relational and probabilistic approaches. E.g., PRMs (Koller, Pfeffer, Friedman, Getoor, etc.) overlay Bayes nets on relational databases, BLPs (Kersting and De Raedt) use logic programs to build Bayes nets dynamically. Many others... See www.biostat.wisc.edu/~page/838.html

Example: Molecular Database

bioactivity

mol	ACE	SSRI
m1	0	0
m2	1	0
m3	0	1
⋮	⋮	⋮

atoms

mol	name	type	X	Y	Z
m1	a1	C	2.1	-1.3	3.4
m1	a2	O	3.0	-1.0	3.0
⋮	⋮	⋮	⋮	⋮	⋮

bonds

mol	atom1	atom2	type
m1	a1	a2	2
m1	a1	a3	1
⋮	⋮	⋮	⋮

Why Can't We Just Merge Into a Single Table? Show me how...

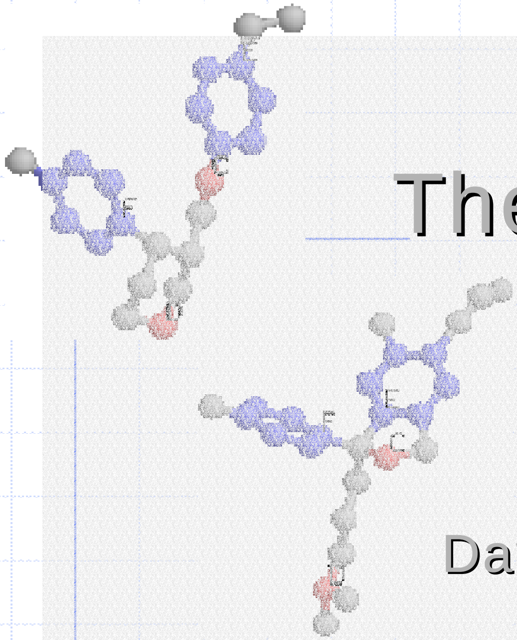
- Joins: will get many more rows (examples) for molecules with many atoms and bonds... altered distribution... and broken examples.
- Features for each possible atom and pair of atoms (for bonds). Poor feature matches... how do we know which atoms to align with which?

Alternative for Getting a Single File: Construct New Features

- Propositionalization: much work, started with ILP system LINUS (Dzeroski & Lavrac, 1991). Automated creation of new features.
- Pharmaceutical companies: features for shapes, combinations of atoms, e.g., carbon double-bonded to oxygen and single bonded to two other carbons. Then learn trees, etc.
- For Molecules: millions of features and still details are lost.

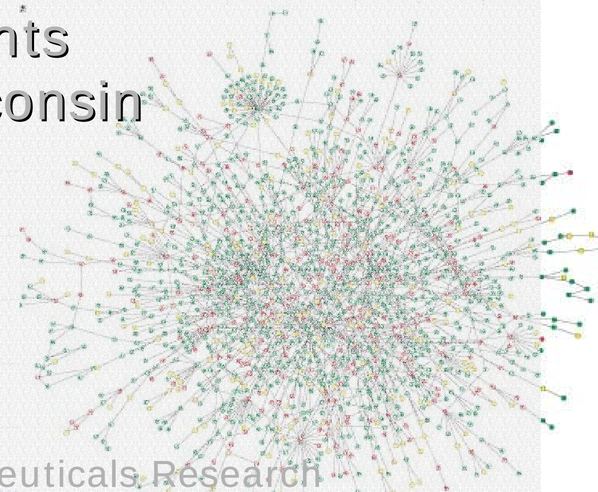
Many other issues...

- Data may not be i.i.d.
  - Not a problem with molecules... each structure is independent of the others.
  - What about predicting protein function from data about protein-protein interactions, etc.? Examples interact with one another.
- Difficult to construct test/train splits.
- This issue came up with predicting protein function from a protein interaction network and other data in KDD Cup 2001. It also arises with many other relational databases.

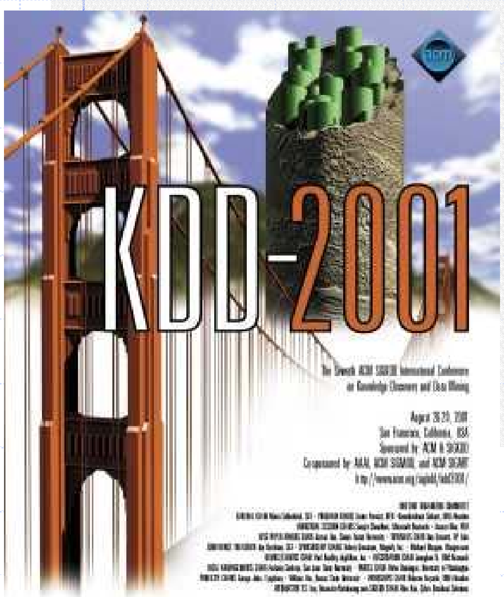


KDD-2001 Cup
The Genomics Challenge

Christos Hatzis, Silico Insights
David Page, University of Wisconsin
Co-chairs



August 26, 2001



Special thanks: DuPont Pharmaceuticals Research Laboratories for providing data set 1, Chris Kostas from Silico Insights for cleaning and organizing data sets 2 and 3

http://www.cs.wisc.edu/~dpage/kddcup2001/

SIGKDD
Knowledge Discovery and Data Mining

3. Complex Interactions of Features

- Many of our data mining algorithms rely on relevant features having some value by themselves.
  - Greedy tree learners (CART, C5.0, etc.)
  - Candidate Elimination for Bayes nets.
  - Many feature selection methods.
- Even when we can get a single table, biology often presents us with trick problems.

Example: Genetics

Female	*Sxl* gene active	Survival
0	0	0
0	1	1
1	0	1
1	1	0

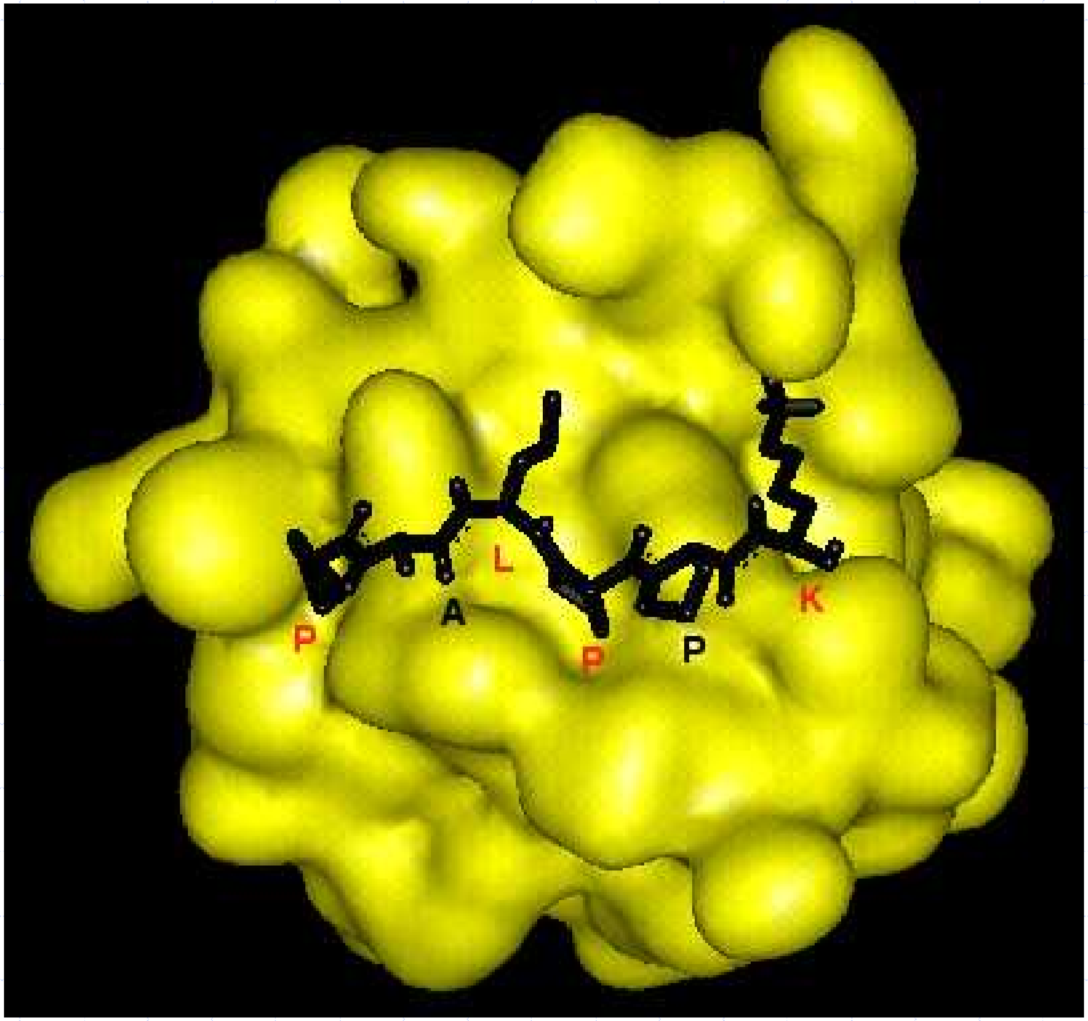
Drosophila survival based on gender and *Sxl* gene activity

Example: Binding

Prot1 Charge	Prot2 Charge	Binds
-	-	0
-	+	1
+	-	1
+	+	0

Binding based on complementary charges of nearby atoms

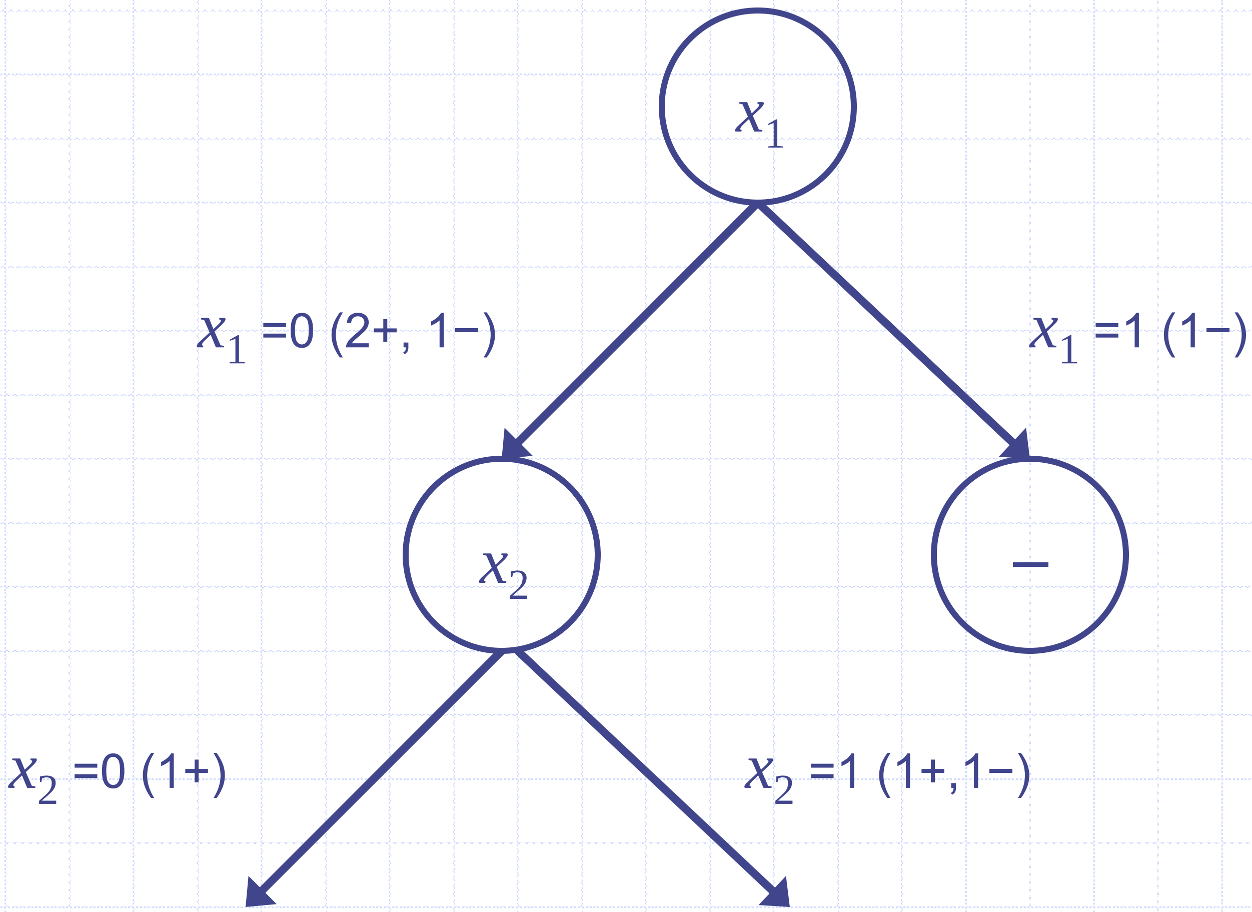
K is positive, contact is negative



Tree Learner (TDIDT) Example

x_1	x_2	x_3	Value
0	0	0	+
1	0	1	−
0	1	0	−
0	1	1	+

TDIDT Example



Learning Hard Functions

- Standard method of learning hard functions with TDIDT: depth-*k* lookahead
  - $O(mn^{2^k-1})$ for *m* examples in *n* variables

- Can we devise a technique that allows TDIDT algorithms to *efficiently* learn hard functions?

Skewing (IJCAI'03, ICML'04) Joint work with Soumya Ray

Hard functions aren't – if the data distribution is significantly different from uniform

Example

- Uniform distribution can be sampled by setting each variable (feature) independently of all others, with probability 0.5 of being set to 1.
- Consider same distribution, but with each variable having probability 0.75 of being set to 1.

Example

x_1	x_2	$x_3 \ldots x_{100}$	f
0	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	0
0	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	1
1	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	1
1	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	0

$$GINI(f) = 0.25$$

$$GINI(f; x_i = 0) = 0.25$$

$$GINI(f; x_i = 1) = 0.25$$

$$\Downarrow$$

$$GAIN(x_i) = 0$$

Example

x_1	x_2	x_3 … x_{100}	f	Wt
0	0	0…0000000 0…0000001 0…0000010 … 1…1111111	0	$\frac{1}{16}$
0	1	0…0000000 0…0000001 0…0000010 … 1…1111111	1	$\frac{3}{16}$
1	0	0…0000000 0…0000001 0…0000010 … 1…1111111	1	$\frac{3}{16}$
1	1	0…0000000 0…0000001 0…0000010 … 1…1111111	0	$\frac{9}{16}$

$$GINI(f) = \frac{60}{256}$$

$$GINI(f; x_1 = 0) = \frac{48}{256}$$

$$GINI(f; x_1 = 1) = \frac{48}{256}$$

$$\Downarrow$$

$$GAIN(x_1) = \frac{(60-48)}{256} = \frac{12}{256}$$

$$GINI(f; x_4 = 0) = \frac{60}{256}$$

$$GINI(f; x_4 = 1) = \frac{60}{256}$$

$$\Downarrow$$

$$GAIN(x_4) = 0$$

More Detailed Example

x_1	x_2	$x_3 \ldots x_{100}$	f	Wt
0	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	0	$\frac{1}{16}$
0	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	1	$\frac{3}{16}$
1	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	1	$\frac{3}{16}$
1	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	0	$\frac{9}{16}$

$$GINI(f) = \frac{6}{16}\frac{10}{16} = \frac{60}{256}$$

More Detailed Example

x_1	x_2	$x_3 \ldots x_{100}$	f	Wt
0	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	0	$\frac{1}{16}$
0	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	1	$\frac{3}{16}$
1	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	1	$\frac{3}{16}$
1	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	0	$\frac{9}{16}$

$$GINI\,(f\,;x_1=0)=\frac{1}{4}\frac{3}{4}=\frac{48}{256}$$

$$GINI\,(f\,;x_1=1)=\frac{1}{4}\frac{3}{4}=\frac{48}{256}$$

More Detailed Example

x_4	x_1	$x_2 x_3 x_5 \ldots x_{10}$	f	Wt
0	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	.25:0 .75:1	$\frac{1}{16}$
0	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	.75:0 .25:1	$\frac{3}{16}$
1	0	0...0000000 0...0000001 0...0000010 ... 1...1111111	.25:0 .75:1	$\frac{3}{16}$
1	1	0...0000000 0...0000001 0...0000010 ... 1...1111111	.75:0 .25:1	$\frac{9}{16}$

$$GINI\,(f\,;x_4=0)=\left[\frac{1}{4}\frac{1}{4}+\frac{3}{4}\frac{3}{4}\right]\left[\frac{1}{4}\frac{3}{4}+\frac{3}{4}\frac{1}{4}\right]=\frac{10}{16}\frac{6}{16}=\frac{60}{256}$$

$$GINI\,(f\,;x_4=1)=\frac{60}{256}$$

Key Idea

- Given
  - a *large enough sample* and
  - a second distribution *sufficiently different from the first*,

we can learn functions that are hard for TDIDT algorithms under the original distribution.

Issues to Address

- How can we get a “sufficiently different” distribution?
  - Our approach: “skew” the given sample by choosing “favored settings” for the variables

- Not-large-enough sample effects?
  - Our approach: Average “goodness” of any variable over multiple skews

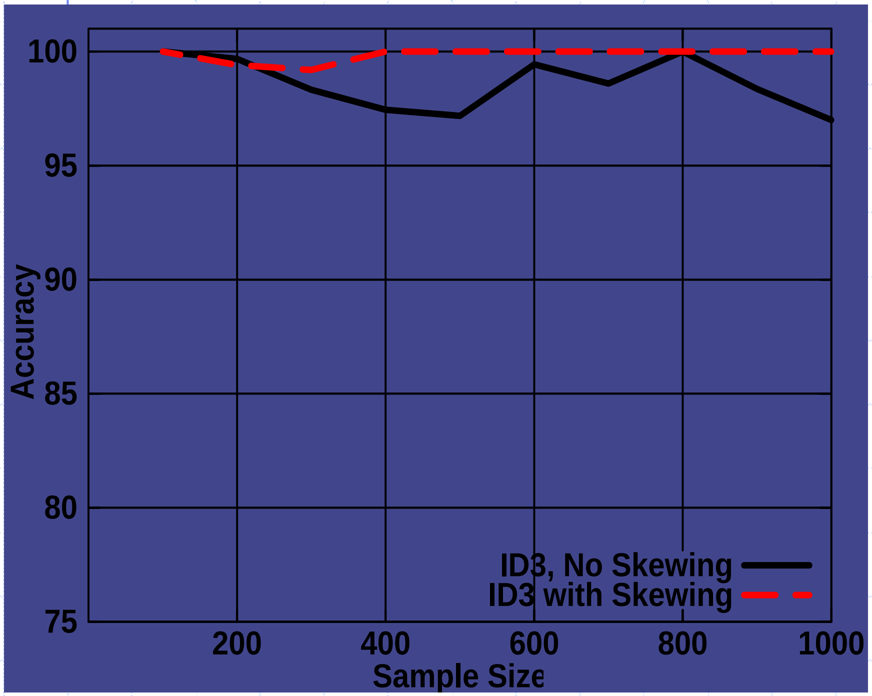
Skewing Algorithm

- For *T* trials do
  - Choose a favored setting for each variable
  - Reweight the sample
  - Calculate entropy of each variable split under this weighting
  - For each variable that has sufficient gain, increment a counter
- Split on the variable with the highest count

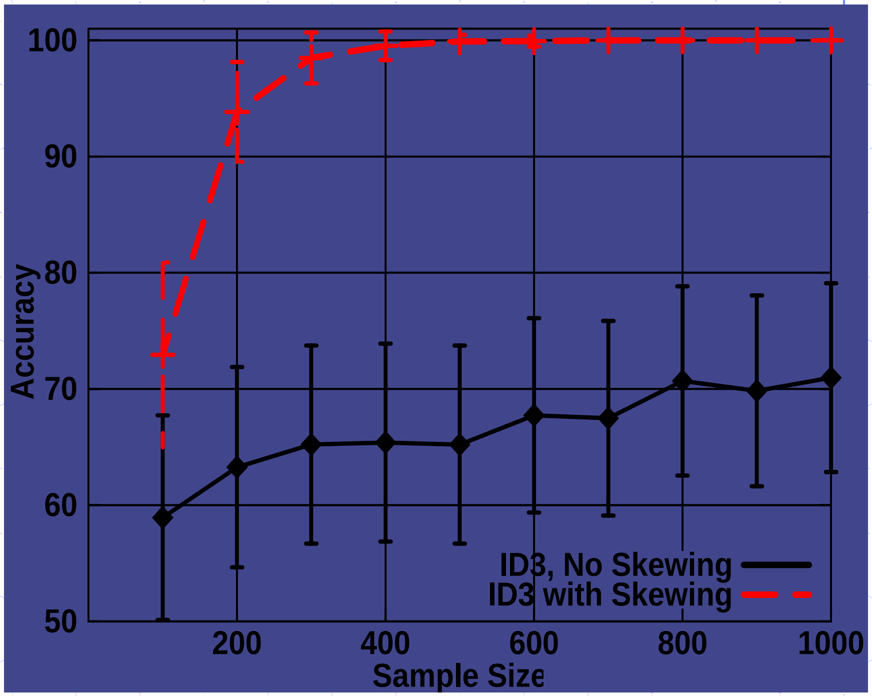
Experiments

- Synthetic data: random and random hard Boolean functions, random uniform data.
- Binding data for studying protein-protein interactions. Involving a family of proteins for which enough is known to allow us to describe them by feature vectors.

Results (3-variable Boolean functions)

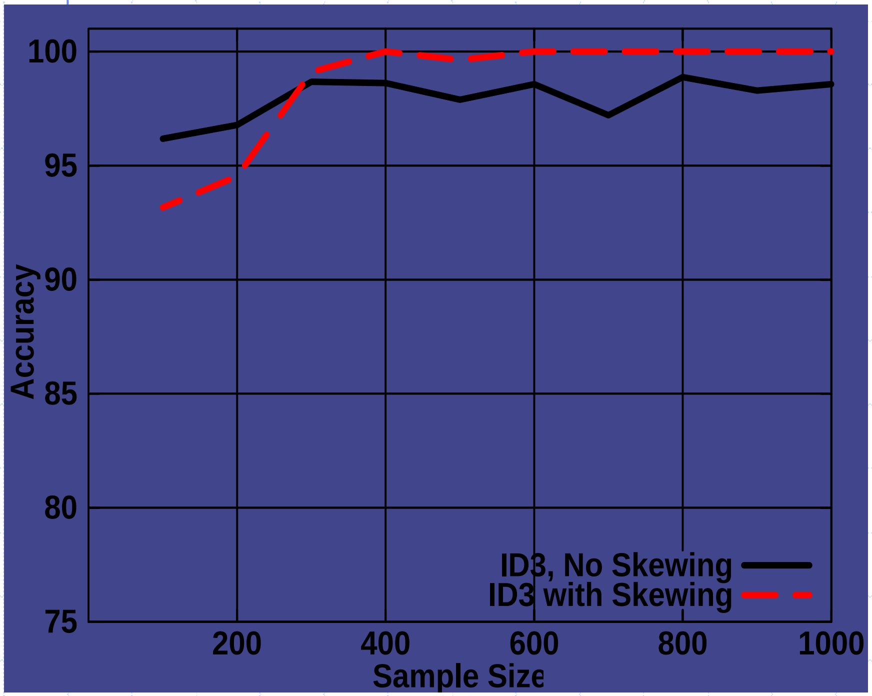


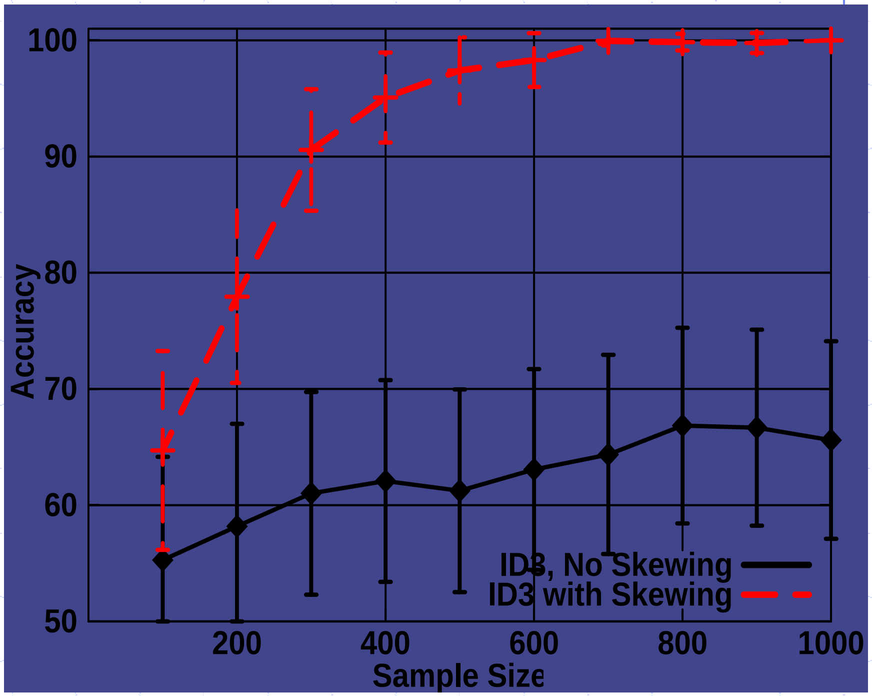
Random functions



Hard functions

Results (4-variable Boolean functions)

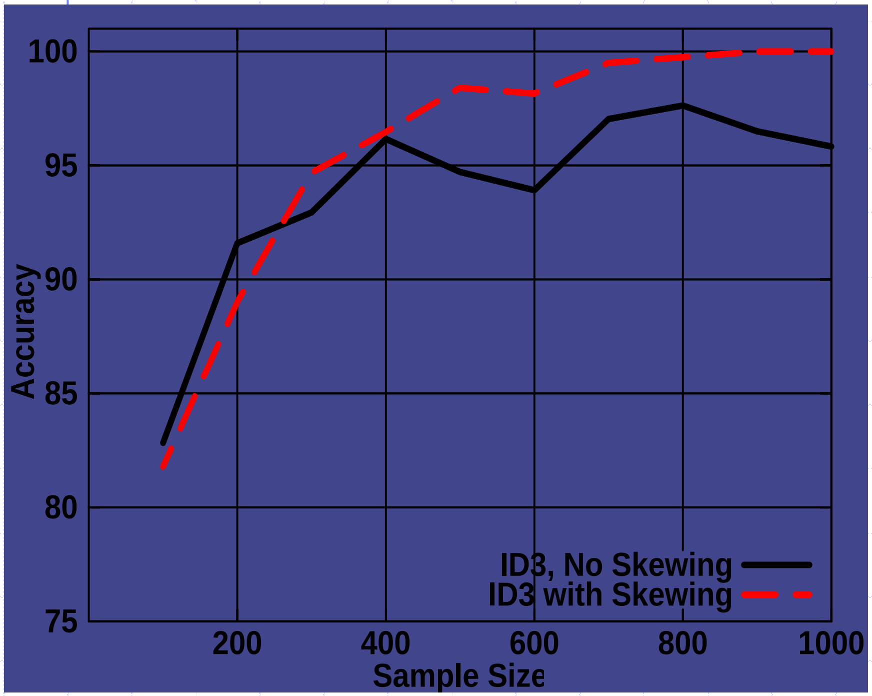




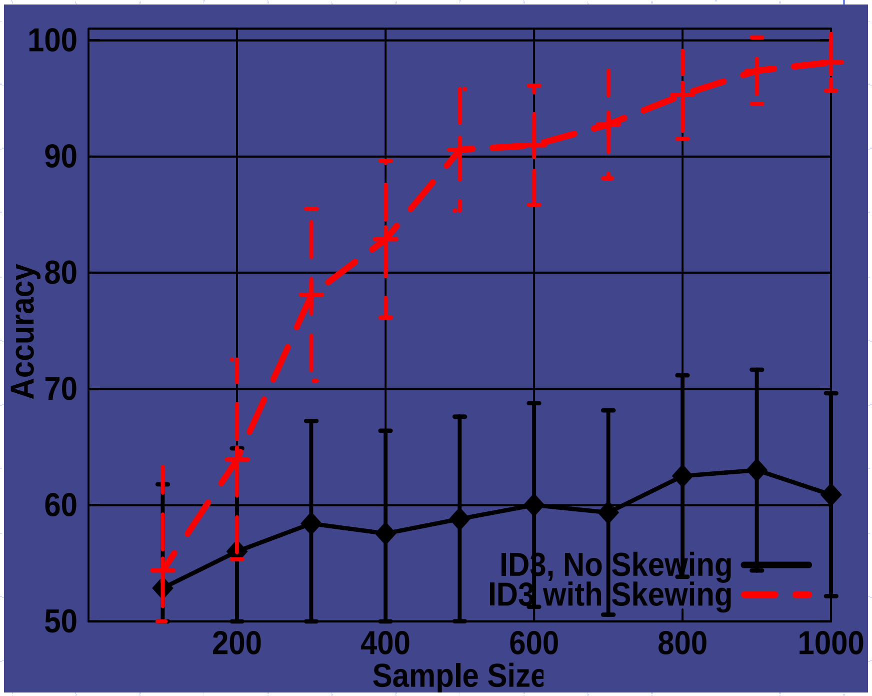
Random functions

Hard functions

Results (5-variable Boolean functions)

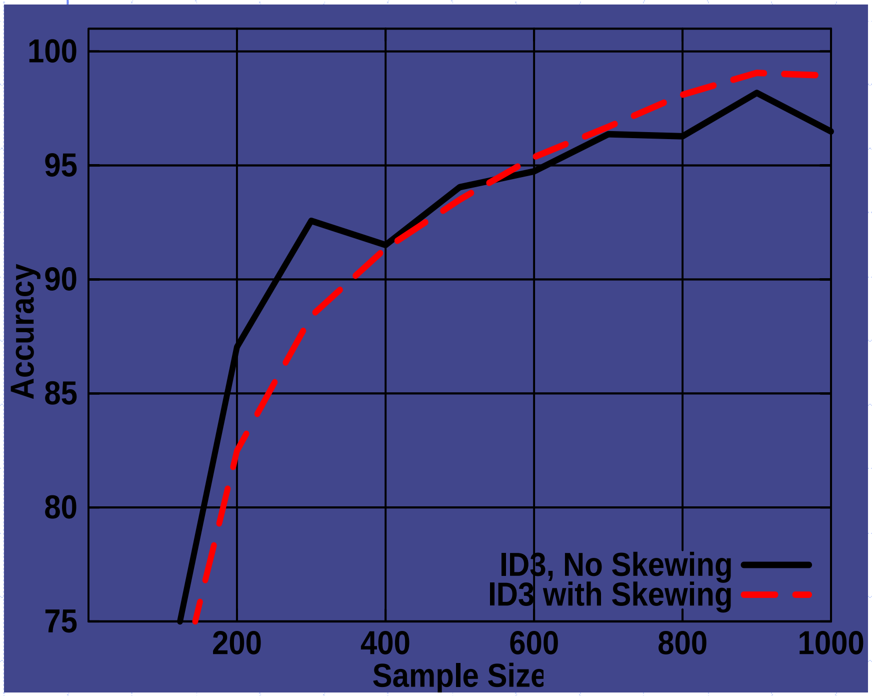


Random functions

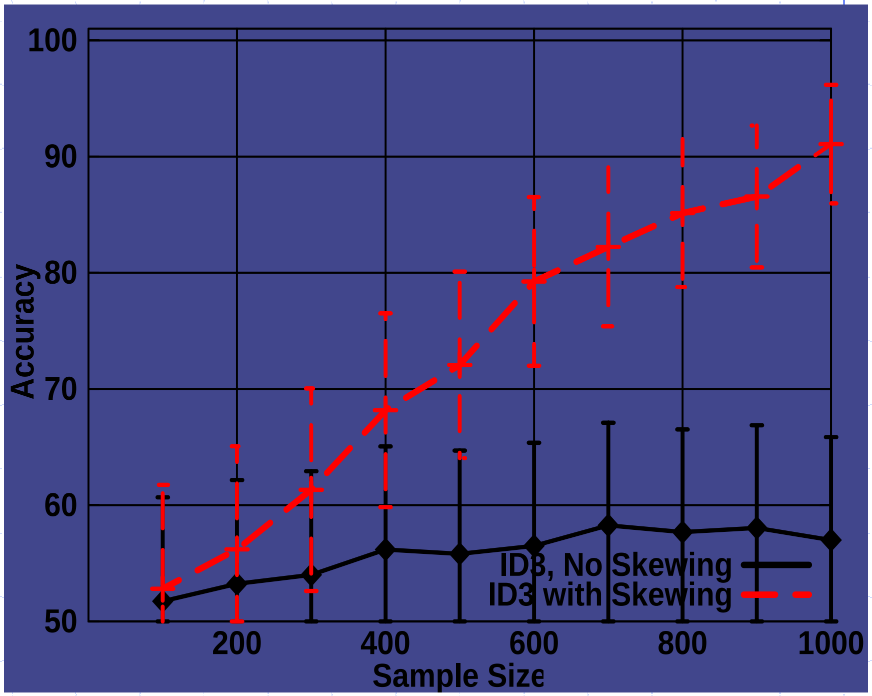


Hard functions

Results (6-variable Boolean functions)



Random functions



Hard functions

On Protein-Protein Interactions

- Skewing yields significantly more accurate predictions than ordinary tree learners (by accuracy and by weighted accuracy).
- Very hard data set... skewing is the only one significantly better than chance.

Conclusion

- Biological Data will continue to grow in importance
- Raises many interesting research issues for data mining
- Great application area even if you’re not all that interested in biology

More...

- ICML tutorial, www.cs.wisc.edu/~dpage
- *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

- Jude Shavlik
- Mark Craven
- Soumya Ray
- Sean McIlwain
- Michael Molla
- Michael Waddell
- Irene Ong
- Brian Kay