Machine Learning for Bioinformatics

(COMP 766-001)

Prof. Ted Perkins

www.mcb.mcgill.ca/~perkins/COMP766001_Fall2006

TR 1:05pm-2:25pm

McTavish 3438, Room 4

Fall Session, 2006

What is machine learning?

(or data mining, pattern recognition, knowledge discovery, signal processing, system identification. . . ?)

From “Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations” by Ian H. Witten and Eibe Frank:

> *If* **data** *is characterized as recorded facts, then* **information** *is the set of patterns, or expectations that underlie the data. . . information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth.*

ML in a bioinformatics context. . .

- . . . is computer-aided *discovery science.*

Exploration	Visualization
Summarization	Generalization
Prediction	Estimation
Modeling	Hypothesis generation

- It's usually *not* about testing a specific hypothesis, as is most prototypical in statistics—though modern ML borrows heavily from statistics.

Our four main topics

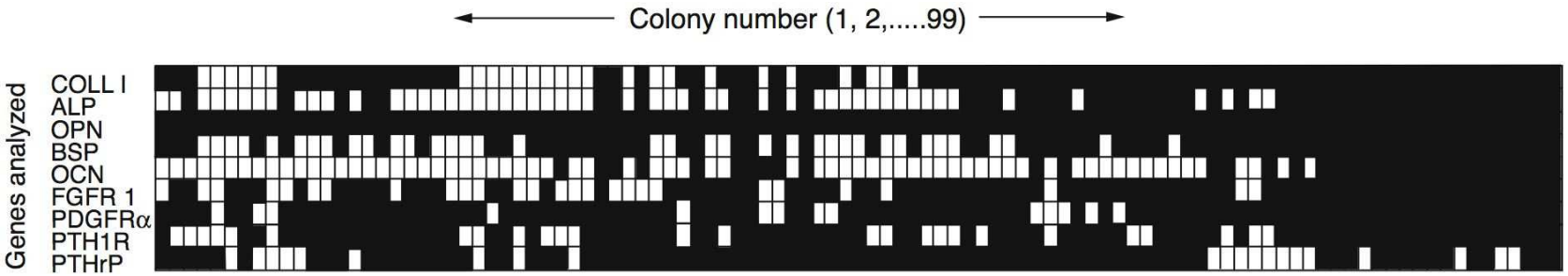
1. Probabilistic modeling
2. Unsupervised learning
3. Supservised learning
4. Modeling dynamical systems

Probabilistic modeling

- Maximum likelihood
- Bayes's rule: for inference and for model-fitting
- Density estimation
- Testing for associations between variables
- Bayesian networks

Probabilistic modeling examples

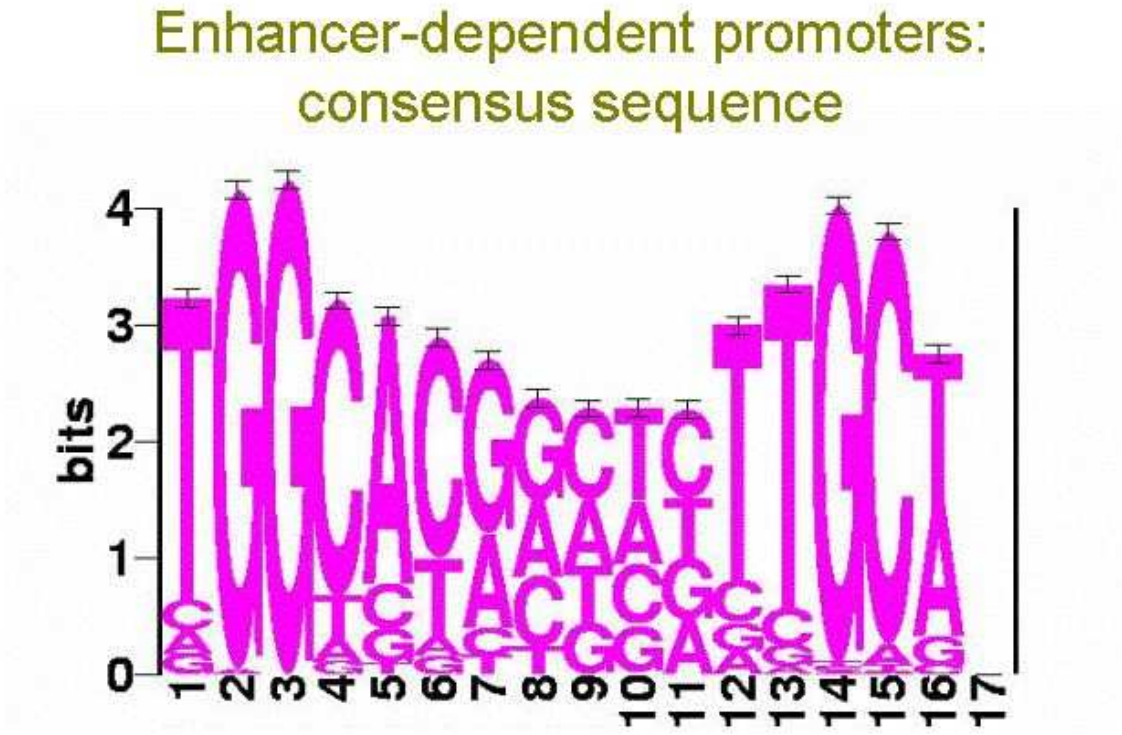
Genetic network inference:



From Madras et al. in *Stem Cells* (2002).

Probabilistic modeling examples

Where does a transcription factor bind to the DNA?



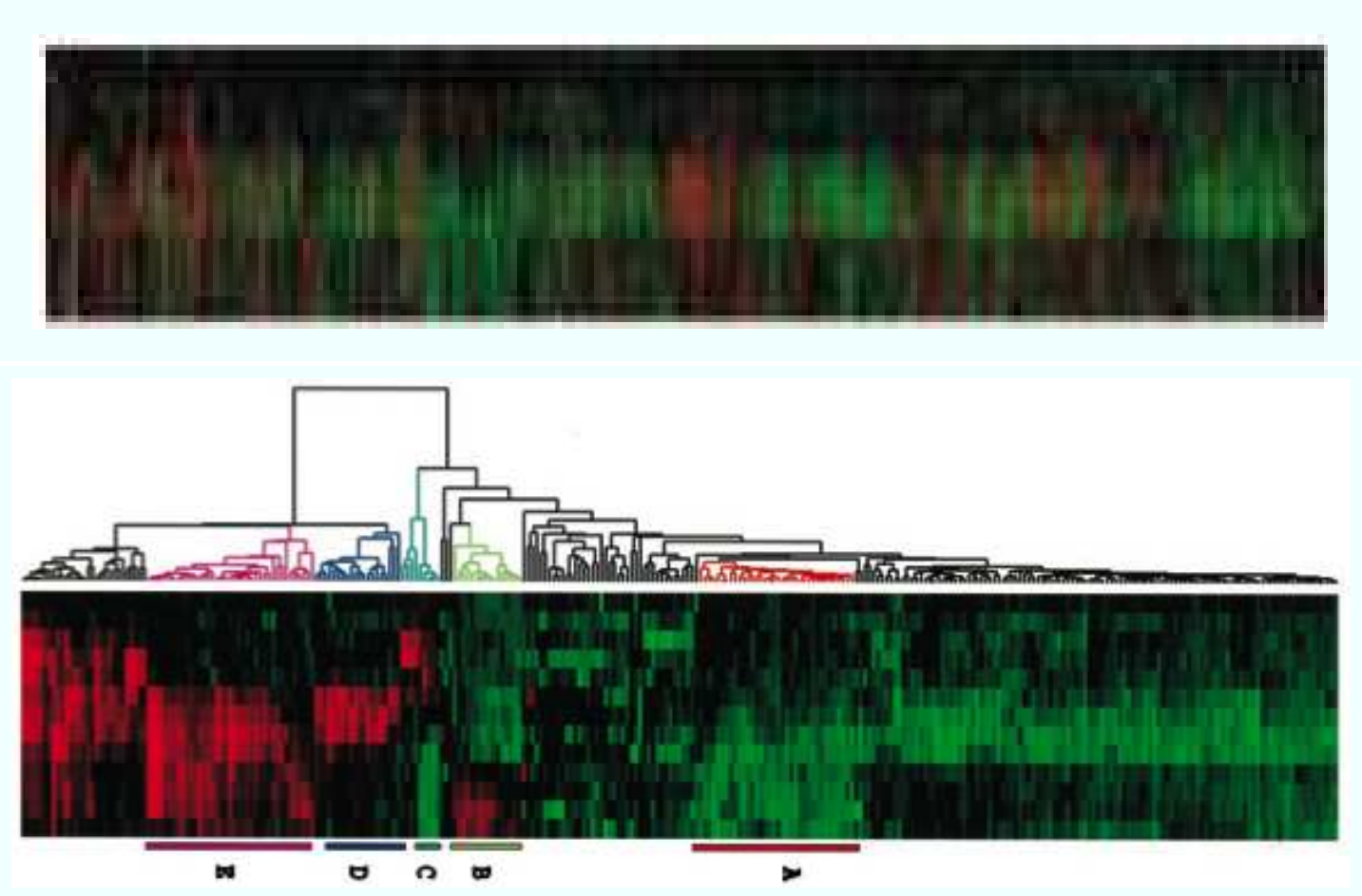
Adapted from **Barrios *et al.***, *Nucl. Ac. Res.* **27**:4305-4313.

Unsupervised learning

- Clustering - “flat” and hierarchical
- Semi-parametric density estimation?
- Dimensionality reduction
  - Principle components analysis
  - Possibly: multidimensional scaling
  - Possibly: self-organizing maps

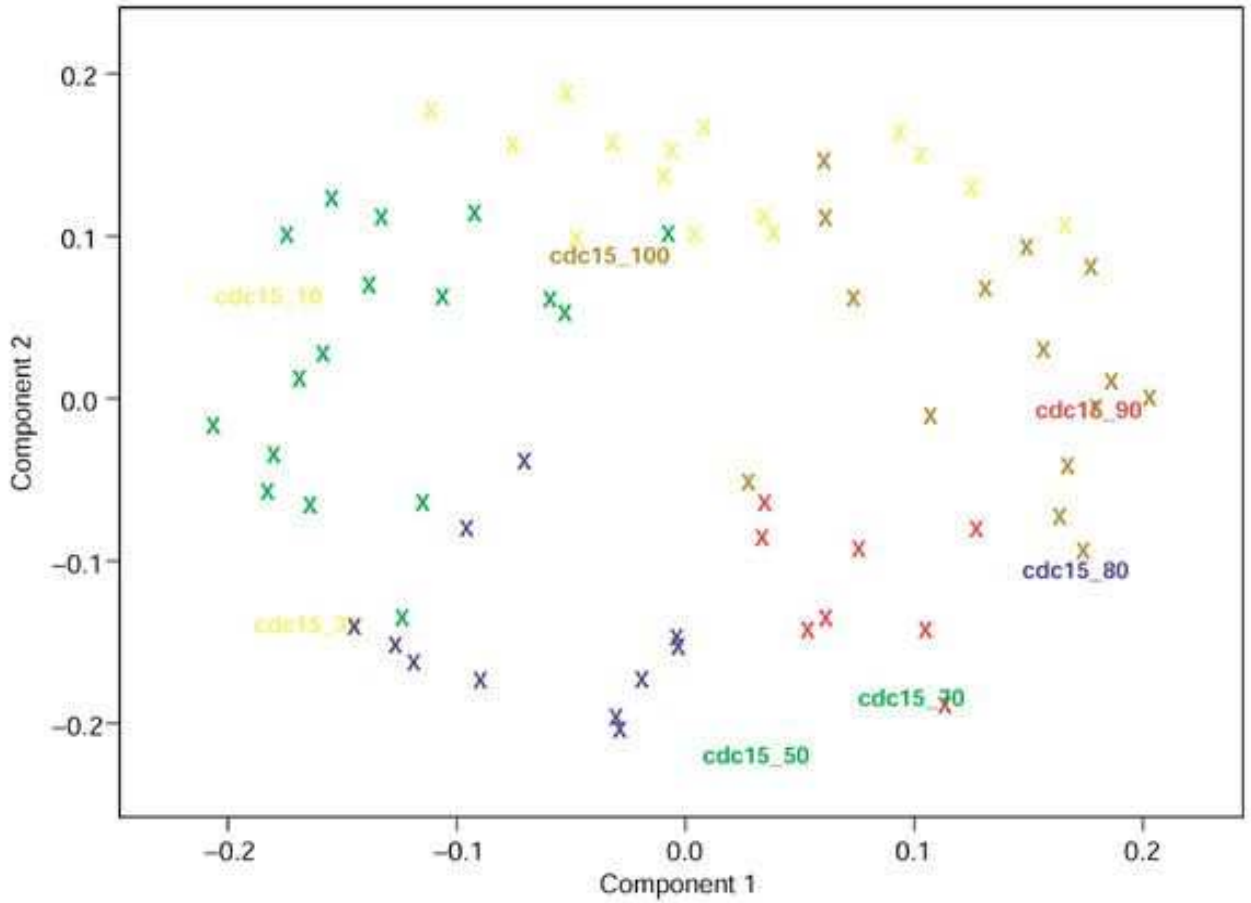
Unsupervised learning examples

Time-series of microarray data (from Eisen et al. *PNAS* (1998)):



Unsupervised learning examples

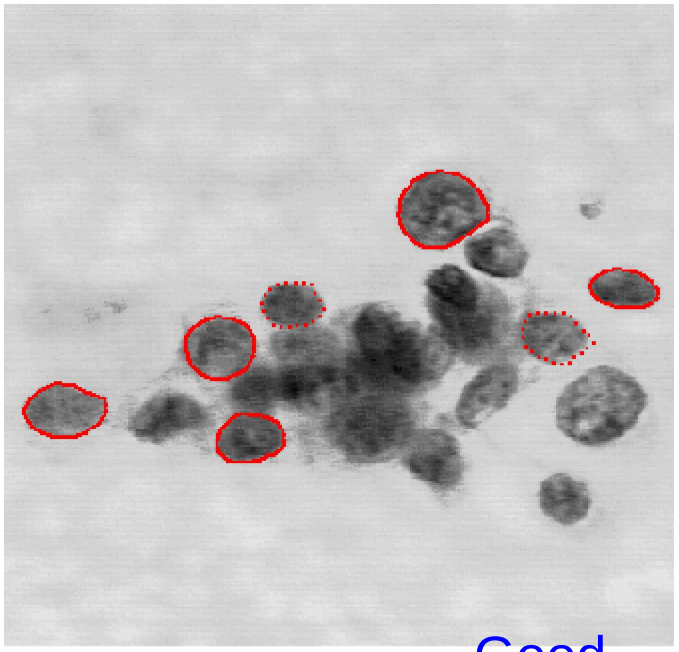
PCA of Spellman's cell-cycle data (from Landgrebe et al. *Genome Biology* (2002)):



Supervised learning

- Linear and logistic regression
- Nearest neighbor
- Tree-based techniques
- And others? Possibly: Artificial neural networks, support vector machines

From http://www.cs.wisc.edu/∼olvi



$\Rightarrow$ Features such as tumor size (from surgery), and cell area, perimeter, texture (from image).

	Good	no chemo recommended
$\Rightarrow$	Intermediate	chemo likely to prolong survival
	Poor	chemo may or may not enhance survival

More supervised learning examples

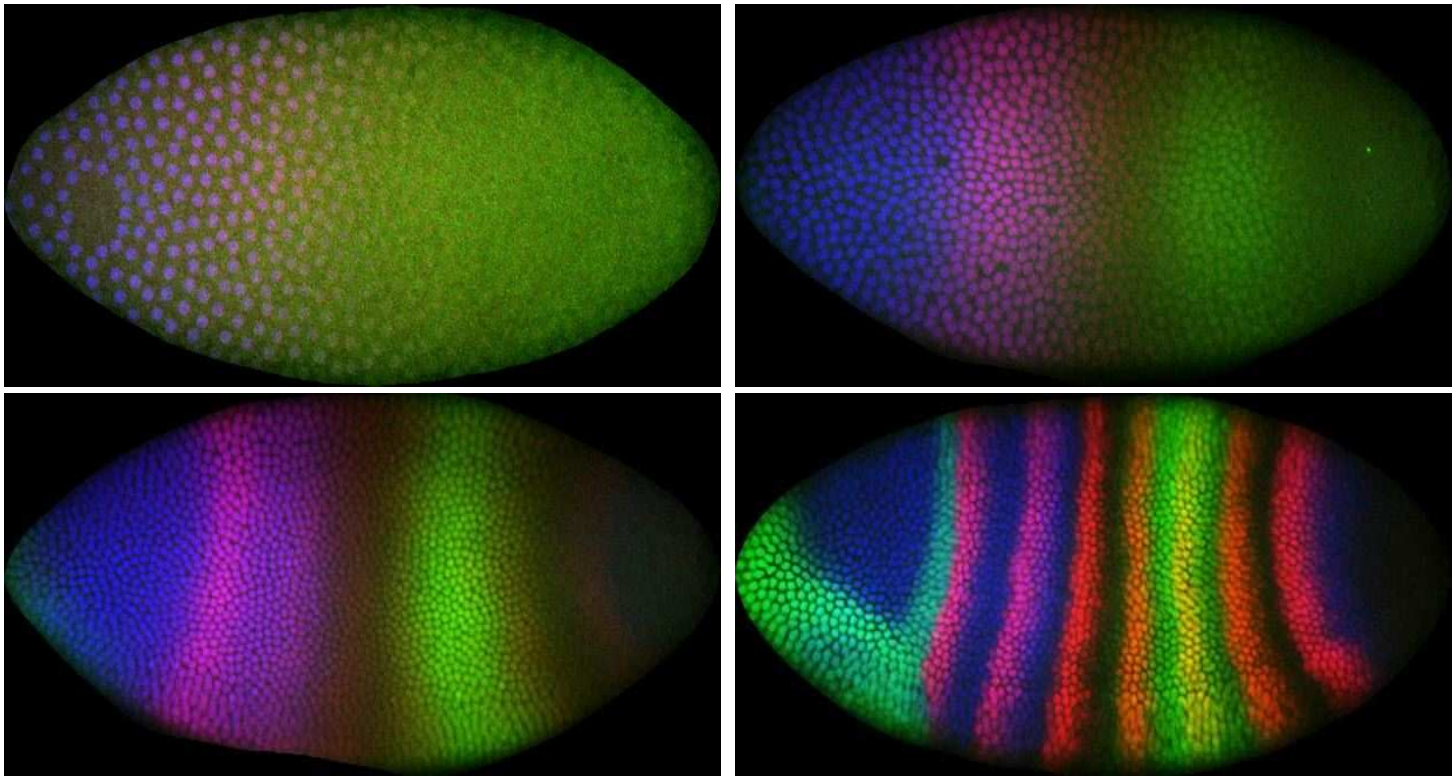
- Given medical test results X, how long does the patient live?
- Is DNA sequence X a transcription factor binding site?
- Does amino acid sequence X fold into α-helix, β-sheet, ...
- Do proteins X and Y interact?

Modeling dynamical systems

- Differential equation models
- Dynamic Bayesian networks

Dynamical modeling examples

Genetic network inference again:



From FlyEx on-line database (http://flyex.ams.sunysb.edu/flyex/).

Course philosophy and goals

Emphasis is on:

- Principles behind machine learning algorithms
- Practical techniques
- Correct methodology

“Learning outcomes”—You should be able to:

- Select appropriate machine learning techniques for data analysis problems you face, and apply them correctly
- Understand and critique the techniques and methodology used in research papers
- Delve deeper into ML, if simple approaches fail
- Derive new ML algorithms for specific problems