Bioinformatics: from sequence to structure Module 2

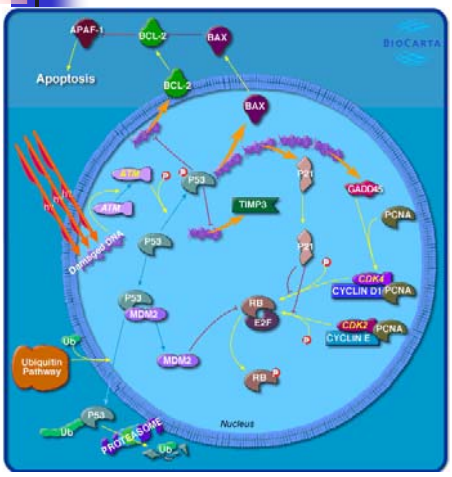
Statistical machine learning

Devika Subramanian
Comp 470

Module design inspiration

- "….Deciphering how a mere 10^7 nucleotides result in a yeast cell, let alone how 3×10^9 nucleotides result in a human – cannot begin until the genes have been annotated. This step includes figuring out the proteins these genes encode and what they do for a living. But understanding how all of these proteins collaborate to carry out cellular processes is the real enterprise at hand."
  - -- ----- Stanley Fields (Science:Feb 16 2001: 1221-1224)

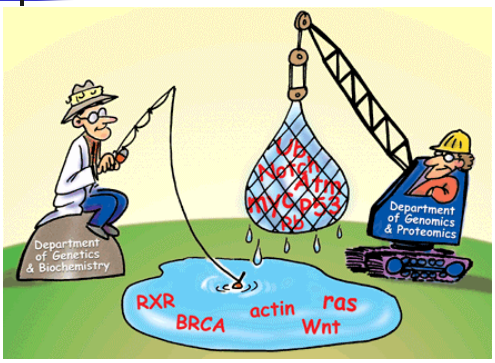
Signaling & metabolic networks



- Consist of interacting proteins, genes and, small molecules.
- Underlie the major functions of living cells.

The quest:
The wiring diagrams of life, particularly how they are altered in diseased cells.

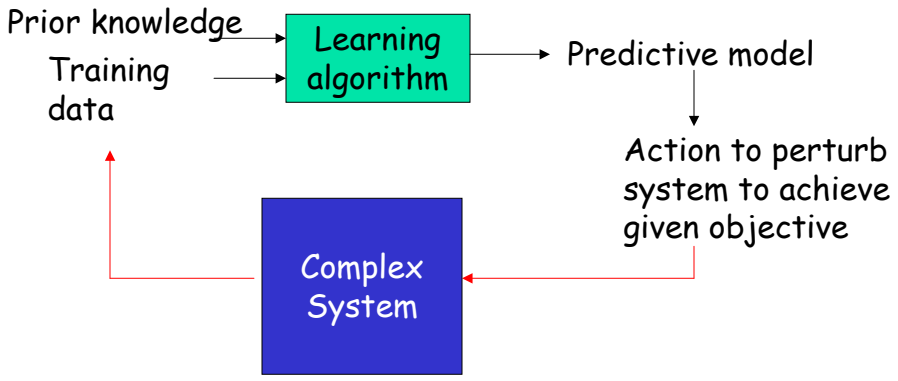
Building models from data



- 3 billion base pairs in human genome.
- 1.5 million known proteins.
- 10^6 to 10^9 (projected) protein-protein interactions.

High throughput assays: mRNA expression levels of 15,000 genes in 1 shot, flow cytometry, MALDI-TOF proteomics assays, allow us access to cellular processes

What is machine learning?



How to use data to build models useful for "fixing" the system.

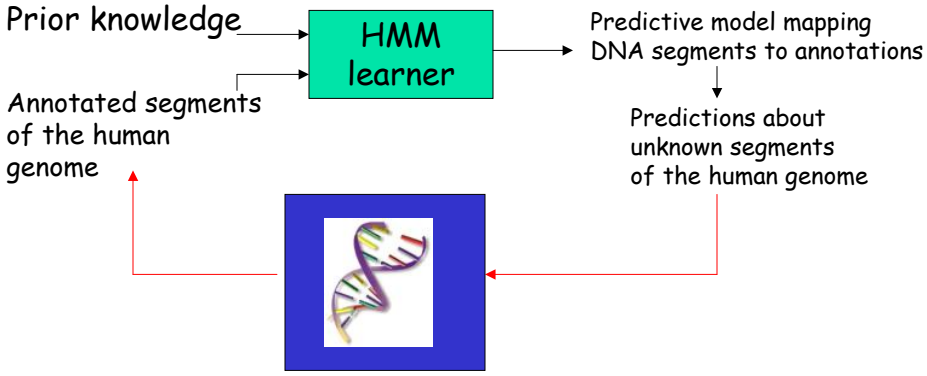
Fundamental questions in machine learning

- What aspects of the system to observe? (Feature selection)
- What class of models to build from observed data and prior knowledge? (Model selection)
- How to evaluate efficacy of the learned model? (Model validation)

Three illustrative problems

- Given a DNA sequence, find and annotate genes in it.
- Given gene expression data, determine biologically significant genes that are differentially expressed.
- Given flow cytometry data, learn signaling networks in normal and diseased cells.

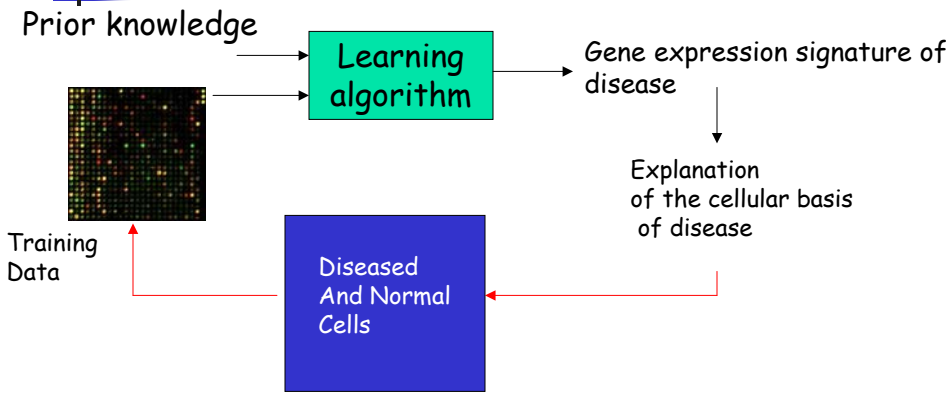
Computational Genefinding



Three illustrative problems

- Given a DNA sequence, find and annotate genes in it.
- Given gene expression data, determine biologically significant genes that are differentially expressed.
- Given flow cytometry data, learn signaling networks in normal and diseased cells.

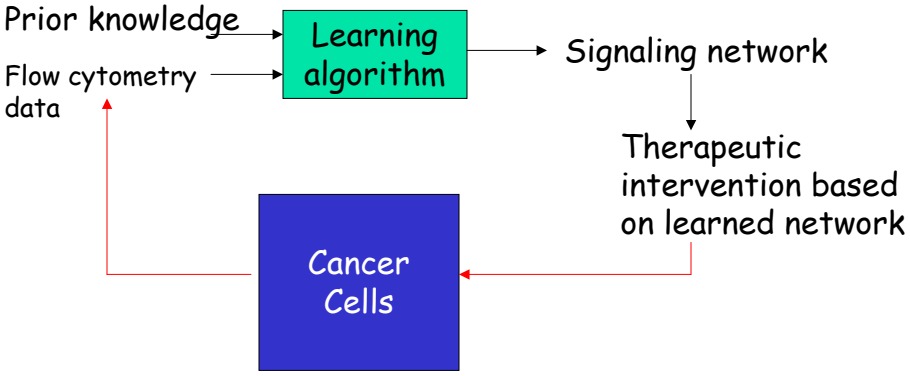
Molecular fingerprinting of disease



Three illustrative problems

- Given a DNA sequence, find and annotate genes in it.
- Given gene expression data, determine biologically significant genes that are differentially expressed.
- Given flow cytometry data, learn signaling networks in normal and diseased cells.

Learning cell signaling networks from data



Three statistical learning algorithms

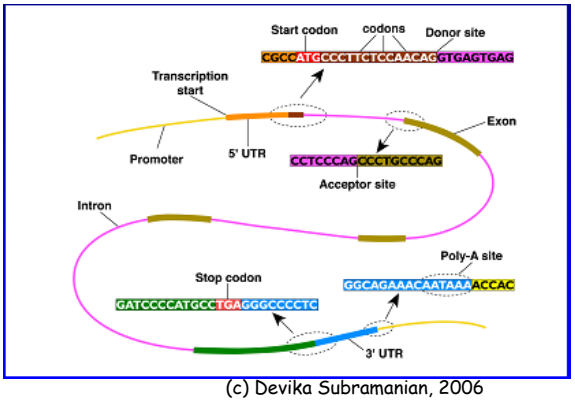
- Hidden Markov Models and variants (Conditional Random Fields).
- Naïve Bayes classifiers and support vector machines.
- Bayesian network learning: parameter and structure learning.

Module objectives

- Learn to model heterogeneous biological data and choose appropriate statistical machine learning algorithms.
- Understand the basics of supervised and sequential machine learning algorithms with particular focus on hidden markov models, naïve Bayes classifiers, kernel-based methods and Bayesian networks.
- Apply these techniques in the context of real data (human chromosome 22, prostate cancer gene expression data, flow cytometry data from T-cell signaling).

Computational gene finding

- Gene finding in eukaryotic DNA

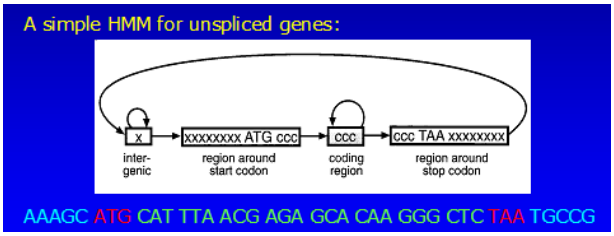


Mathematical model

- Hidden Markov models
  - Structure of HMMs
  - Viterbi algorithm for annotation
  - Baum-Welch (EM) algorithm for learning models
  - Extensions: pair HMMs

Ab initio methods

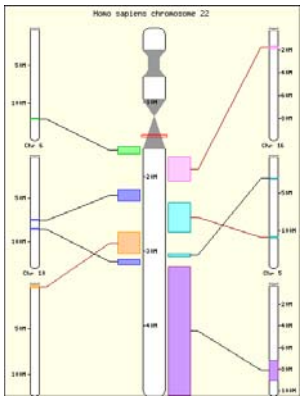
- Genscan (Burge et. al., JMB 1997)



- Intrinsic limits on performance of *ab initio* methods; evaluation studies (Rogic, Gen. Res. 2001)

Comparative methods

- SLAM (Pachter et. al. Gen. Res. 2003); simultaneous gene prediction and sequence alignment of two syntenic genomic regions.
- Paired HMMs

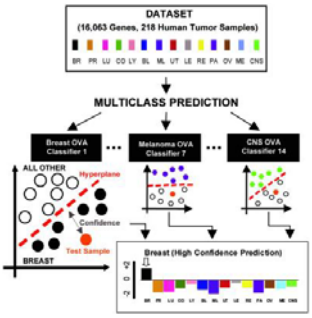


Exercise

- CpG island detection on human chromosome 22 using learned HMMs.
- Analyze similarities and differences in prediction between Viterbi and posterior decoding.

Molecular fingerprinting of cancers

- Work of Golub et. al (Science 1999 (AML/ALL), Bioinformatics 2001, Nature 2003), Lee & Lee (Bioinformatics 2003)



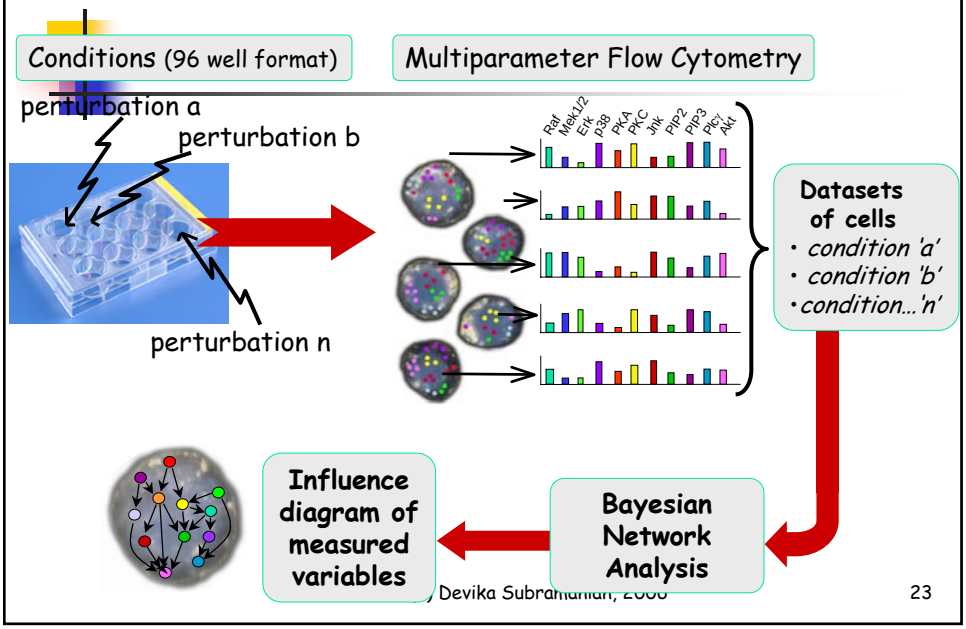
Mathematical model

- Naïve Bayes classifiers
  - Ensemble methods: boosting and bagging
- Support vector machines (SVM)
  - Maximum margin separating hyperplane
  - Linear SVMs and soft margin hyperplanes
  - Non-linear SVMs and the kernel trick

Exercise

- From Singh prostate cancer data, determine which genes are differentially expressed using Naïve Bayes and SVM classifiers.
- Experiment with various feature selection techniques, compare predictions against the latest theories of compromised cellular processes in prostate cancer (Science 2004).

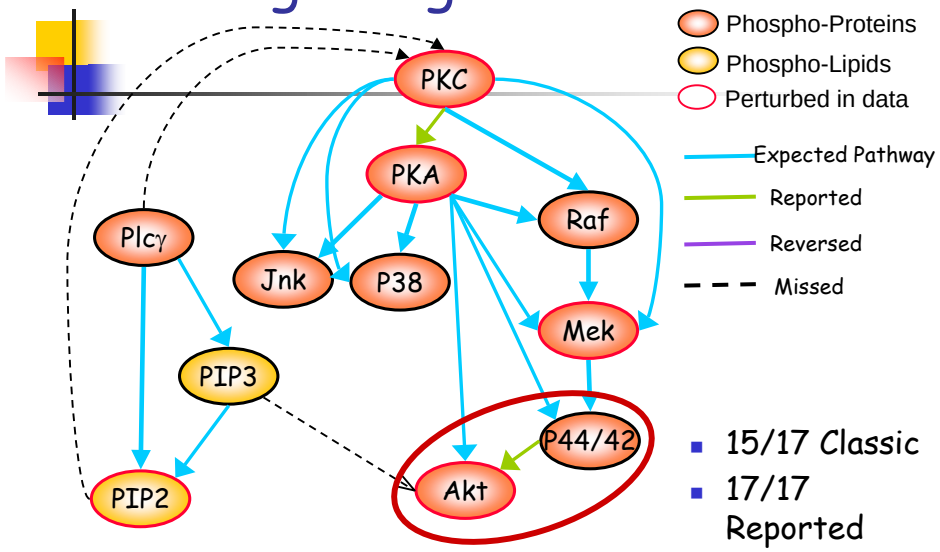
Learning networks from data



Mathematical model

- Probabilistic models : bayesian network representations of signaling networks.
- The sparse candidate algorithm for learning Bayesian networks from high-throughput data .

T-cell signaling network



- 15/17 Classic
- 17/17 Reported
- 3 Missed

Science 2005, Sachs et. al

Summary

- How to use the underlying biology to constrain model selection and feature selection.
- How to choose and adapt machine learning algorithms for biological problems.
- How to design learning protocols to deal with incomplete, noisy data.
- How to interpret the results of machine learning algorithms.