

Bioinformatics: from sequence to structure Module 2

Statistical machine learning

Devika Subramanian
Comp 470

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, SELDI-TOF proteomics assays, allow us access to cellular processes
(c) Devika Subramanian, 2007

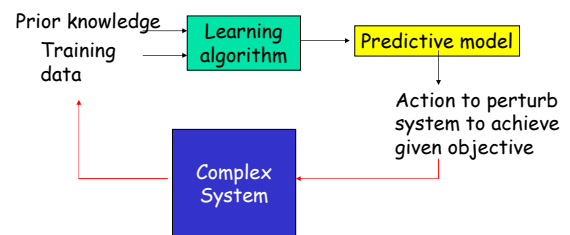
Module design inspiration

- "....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)

(c) Devika Subramanian, 2007

2

What is machine learning?



Observe complex system and build models to predict its response.

(c) Devika Subramanian, 2007

5

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.

(c) Devika Subramanian, 2007

3

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)

(c) Devika Subramanian, 2007

6

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.

(c) Devika Subramanian, 2007 7

Molecular fingerprinting of disease

(c) Devika Subramanian, 2007 10

Computational Genefinding

(c) Devika Subramanian, 2007 8

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.

(c) Devika Subramanian, 2007 11

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.

(c) Devika Subramanian, 2007 9

Learning cell signaling networks from data

(c) Devika Subramanian, 2007 12

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.

(c) Devika Subramanian, 2007

13

Mathematical model

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

(c) Devika Subramanian, 2007

16

Module objectives

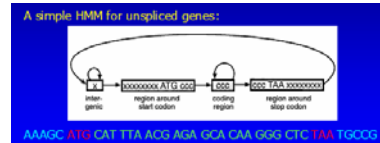
- Learn to model heterogeneous biological data and choose appropriate statistical machine learning algorithms.
- Understand the basics of supervised and sequential machine learning algorithms with particular focus on hidden Markov models, naïve Bayes classifiers, kernel-based methods and Bayesian networks.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

(c) Devika Subramanian, 2007

14

Ab initio methods

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



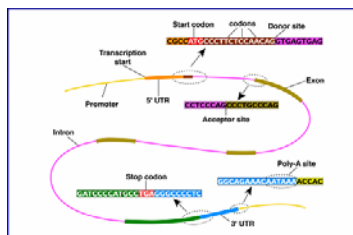
- Intrinsic limits on performance of *ab initio* methods; evaluation studies (Rogic, Gen. Res. 2001)

(c) Devika Subramanian, 2007

17

Computational gene finding

- Gene finding in eukaryotic DNA

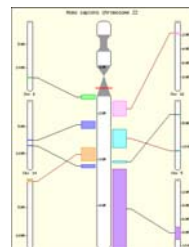


(c) Devika Subramanian, 2007

15

Comparative methods

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



(c) Devika Subramanian, 2007

18

Exercise

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

(c) Devika Subramanian, 2007

19

Exercise

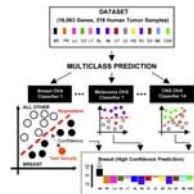
- From Singh prostate cancer data, determine which genes are differentially expressed using Naïve Bayes and SVM classifiers.
- Experiment with various feature selection techniques, compare predictions against the latest theories of compromised cellular processes in prostate cancer (Science 2004).

(c) Devika Subramanian, 2007

22

Molecular fingerprinting of cancers

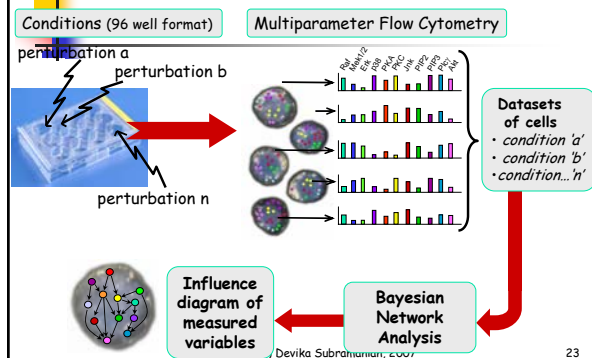
- Work of Golub et. al (Science 1999 (AML/ALL), Bioinformatics 2001, Nature 2003), Lee & Lee (Bioinformatics 2003)



(c) Devika Subramanian, 2007

20

Learning networks from data



Devika Subramanian, 2007

23

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

(c) Devika Subramanian, 2007

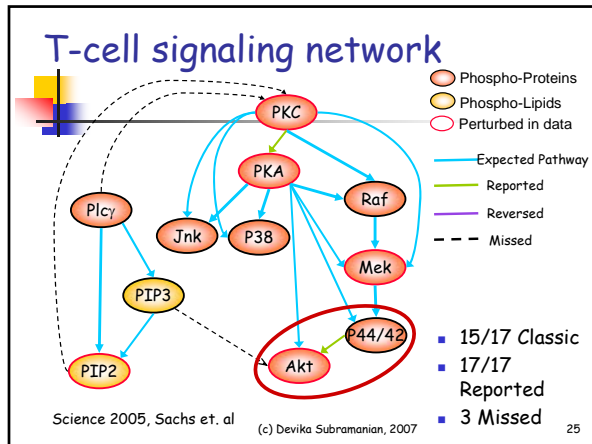
21

Mathematical model

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

(c) Devika Subramanian, 2007

24



- ### 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.
- (c) Devika Subramanian, 2007 26