Computational problems in functional genomics

David K. Gifford
(Adapted from a lecture by Tommi S. Jaakkola)
MIT LCS and AI Lab
gifford@mit.edu

Computational functional genomics

- What does computational functional genomics wish to achieve?
  1. Prediction
     - e.g., tumor identification, pathogens, etc.
  2. Modeling
     - e.g., simulation, model induction and verification
  3. Understanding
     - e.g., organizing/functional principles

Functional genomics: data analysis

- We can organize the (preprocessed, normalized) experimental data into a matrix

	Population 1	Population 2	...
Gene 1	181	1	137
Gene 2	499	229	218
Gene 3	167	147	120
...	296	110	380

- We are now ready for data analysis...

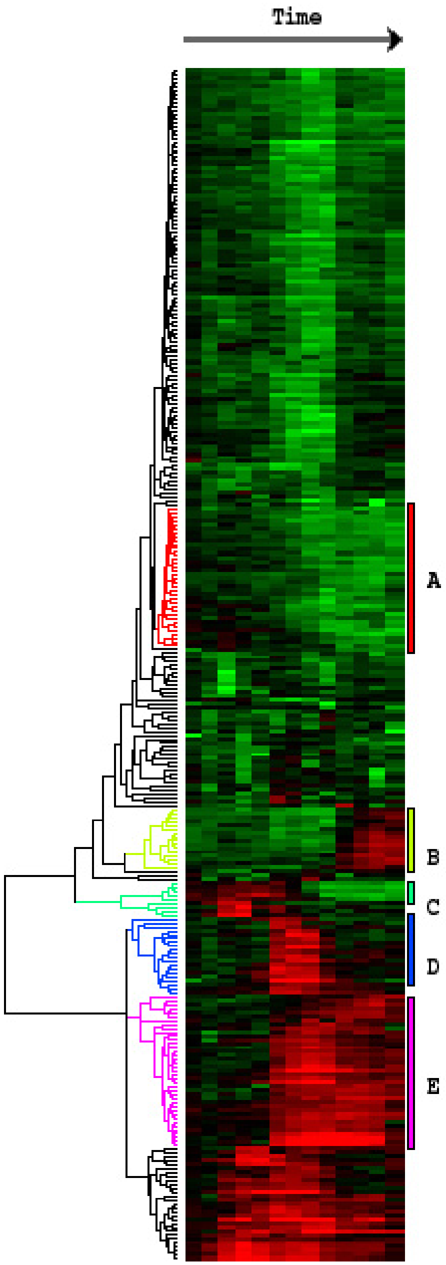
Functional genomics: computational problems

Types of computational problems we are interested in solving in functional genomics:

1. Clustering
   – identifying functionally related genes (via co-expression)
   – identifying functional subclasses of samples
2. Classification
   – classification of tissue samples, diagnosis of diseases
   – classification of functional classes of genes
3. Feature selection
   - identification of relevant genes
4. Time series analysis
   – pathogen infection time course analysis
5. Induction and verification of regulatory network models
6. Combining multiple sources of information
   – supplementing expression analysis with DNA sequence analysis

Clustering cont'd

- Hierarchical agglomerative clustering: sequentially merge the pair of "closest" points/clusters
- There are many types of clustering algorithms but the main issue is to decide what the similarity measure is between any two gene profiles or experiments
- Why is there only a single gene per cluster?
- How "significant" are the clusters?
- How do we know that the resulting clusters are real?



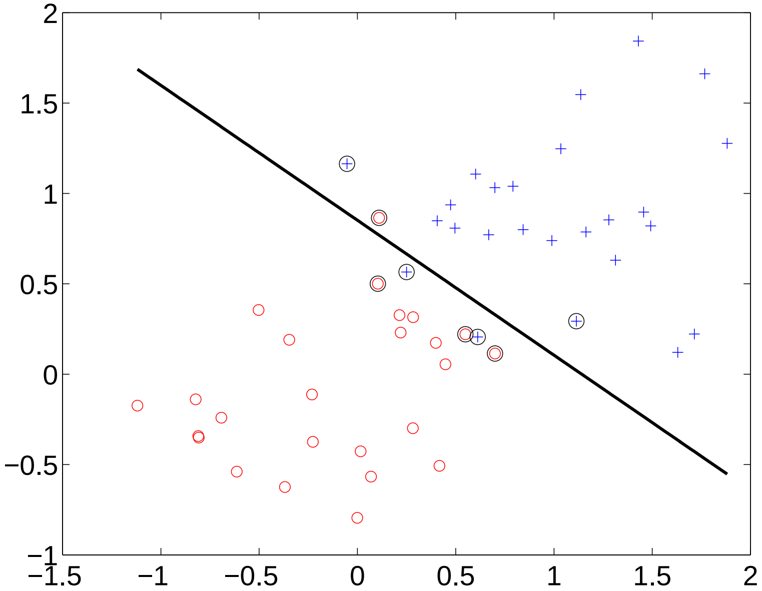
Classification

- For clinical purposes, we would like to differentiate between tumor and normal cells or between tumor cells of different type

- Assuming we have available a set of tissue samples with known class labels, the problem is to estimate a classifier based on such a training set to be able to correctly classify unseen tissue samples

Classification cont'd

- We estimate a "decision" boundary that separates normal cells from tumor cells based on the overall expression pattern across the genes



- Some computational issues:
  1. What type of decision boundaries should we use?
  2. How do we find the decision boundary?
  3. How much confidence do we have in the resulting classification decisions?

Gene identification (feature selection)

- Not only do we want to make accurate predictions but also identify the set of genes whose differential expression in the tumor/normal cells underlies the class distinction
- This problem is known as “feature selection” or “subset selection”

  A simple approach would select genes that are highly correlated with the class label

	tumor	normal	...
Gene 1	181	1	137
Gene 2	499	229	218
Gene 3	167	147	120
...	296	110	380

- Gene indentification also carries a computational benefit: reducing the dimensionality of the problem leads to more accurate classification decisions

Time series analysis

- Many interesting biological processes involve time
  - yeast cell cycle
  - pathogen infection

  etc.
- We need a computational approach to
  1. cluster
  2. classify
  3. characterize

  time course profiles

  Note: the fact that we KNOW the data comes from a time series permits us to make stronger assumptions

Pathogen infection

- Differential time course response of cells to pathogen infection (HIV, ebola, TB, ...)

  The data "cube" now involves three relevant directions of variation: pathogen type, gene id, measurement time

- Computational questions:
  1. disentangling pathogen specific/generic responses
  2. pathogen identification based on (time course) observations
  3. modeling cell response dynamics (latency etc.)
  4. cell donor dependent/independent response
- These are not new computational problems...

Modeling/uncovering gene regulation

- Ultimately we wish to understand the regulatory network underlying the behavior of the cell
- Genes are regulated in a combinatorial fashion; the effect of transcriptional activators can be context sensitive

  For example, transcription initiation relies selectively on the components of the RNA polymerase holoenzyme

Modeling/uncovering gene regulation

- We need a computational language for specifying and verifying (incomplete) models of gene regulation

- Differential equation models
  - analogous to chemical reaction equations
- Stochastic circuit models
  - simulation approach
- Statistical graph models
  - robustness, verification, abstractions, ...

Combining multiple sources of information

- Expression data alone is not sufficient
  - literature, sequence, proteomics, location analysis
- Combining multiple sources of information yields complementary constraints
  - e.g., gene identification with protein homology assessments
  - e.g., we may combine expression data with constraints from conserved promoter regions to generate more reliable regulatory network models

Summary

- Computational approach relies heavily on the problem formulation and the assumptions that we can make
- The areas of application of computational methods in functional genomics are practically limitless
- The purpose of this course is to furnish you with some basic computational tools as well as the ability to use them in a biological context