CS612 - Algorithms in Bioinformatics

Misc

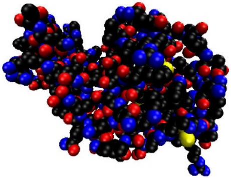
May 6, 2019

Overview of Systems Biology

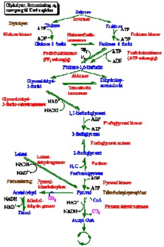
- Investigate Biology at a systems level
  - Consider the organism as a whole rather than looking at
  - proteins or cells at an individual level
- Airplane analogy:
  - Merely looking at a list of the parts for an airplane does not
  - tell you that it can fly

What is a Biological System?

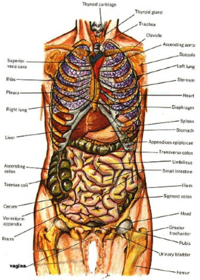
- Set of components that interact to form functional unit
- System can be defined at different hierarchical levels (enzyme, glycolysis, cellular, tissue, organ, whole organism, ecosystems)



www4.liber.se/kemionline/gymkeb/bilder/12_a.jpg



http://biologi.uio.no/plfys/haa/gif/form142.gif



http://www.acuhealthzone.com/images/

Why Study a Biological System?

- Genome-wide data sets allow us to find patterns and fill in missing information – Protein Networks
- Having a mathematical model to represent a biological system has many benefits:
- Experimental Validation – validate experimental results by comparing them to the model
- Experimental Planning – Help plan more effective experiments based on expected outcomes
- Remove parts of a model more easily than removing a part of a real biological system
- Test robustness of a system

Why Study a Biological System?

- Ok, I've invented a new drug...
- What dosage should I use?
- How long will the effects of drug last?
- Where is the best place to inject it?
- Etc...

Four Properties of Systems Biology

- **Systems Structures** – Networks of interactions and pathways
- **System Dynamics** – How the system behaves over time, Metabolic analysis
- **The Control Method** – Modulate mechanisms that control the state of a cell
- **The Design Method** – Modify or construct biological systems

Systems Biology = Components and Interactions

- Approach to biology where organisms and biological processes are analyzed and described in terms of their components and their interactions
- Description is through a framework of mathematical models
- Broader picture
    - Every molecule is just a component in a larger system
    - So, even a statement gene X produces protein Y is misleading, as so many other things have to happen for protein Y to be synthesized

Goals of Systems Biology

- Understand in detail to control
- How does a particular system behave?
- What kind of perturbations can it tolerate?
- What kind of perturbations can it not tolerate?
- Control
  - How can I control this system to make it do something interesting?
  - For instance, how can I make a bacterium produce propanol instead of ethanol?
  - How can I change the metabolic network of the bacterium?

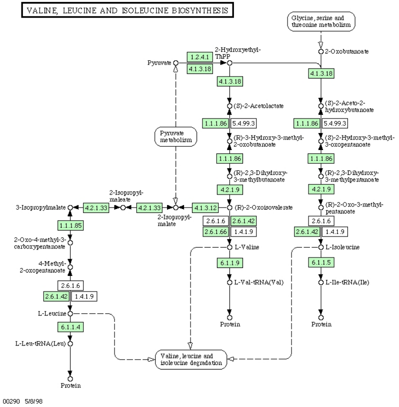
Control in Systems Biology for Health

- want to find best treatment for type II diabetes (obesity-related)
- What does the insulin regulatory system look like in detail?
- What is different in this system in patients with type II diabetes from normal individuals?
- What components of the system should I focus on?
- What are the best drug targets for these components?

Networks

- Systems Biology is the study of life processes
- Most fundamental life process = metabolism
- Ability to synthesize sugars, amino acids, lipids, and to create the energy required to synthesize these components
- System of connected chemical reactions = metabolic network

Metabolic Network – Example



- A portion of the metabolic network for the synthesis of amino acids valine, leucine, and isoleucine in E. Coli
- Produced from KEGG database
- Can click on components to navigate to connected metabolic subsystem
- Green boxes indicate enzymes that have been identified
- Four letter code is (enzyme classification) EC designation

Other Kinds of Networks

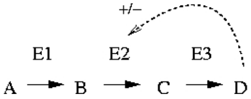
- **Signaling Networks** – Interactions among components propagate signal
- **Genetic (Regulatory) Networks** – Signaling networks where components are genes
- **Protein-protein interaction Networks** – Components are proteins

Signals and Feedback

- Living cell = signal processor
  - Cell receives signal from environment in form of specific molecules, such as hormones, transmitted substances, nutrients (glucose, lipids) or stress (osmotic stress, poison, heat, cold)
  - Cell produces signals internally, e.g. in its cycle, where events such as DNA damage or failures in chromosome duplication generate signals that stop cell division, start repair processes, or being cell death
- Signals and machinery that processes them form a regulatory network
  - Purpose of the network is to control life

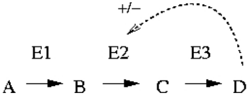
Signals and Feedback

- Key mechanism: Negative Feedback
  - The result of a process feeds back signal to control the process itself
  - The process is self-monitoring
  - There is positive and negative feedback



Important Problems in Regulatory Networks

- Reverse engineering
  - Given some microarray data that give you a picture of how much protein is around in a cell at a given time, can one figure out the underlying regulatory network?
  - Is it possible to identify building blocks of regulatory components that occur more often in biological systems?



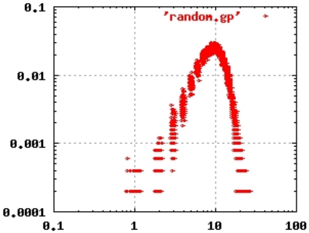
Some Well-Characterized Network Properties

- Overall structure of a network can be described by several different parameters, such as average number of connections per node, probability that a node has a given number of connections.
- Theoretical work has shown that different models for how a network has been created will give different values for these parameters.
  - Classical random network – Erdős and Renyi (1960)
  - Scale-free network – Reka, Jeong, Barabasi (2000)

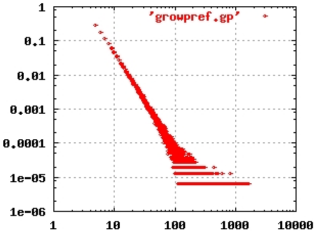
Some Well-Characterized Network Properties

- **Classical random networks** – Erdős and Renyi (1960)
- Given a set of nodes, connections are made randomly
- Gives same average number of connections per node
- Does not capture degree distribution observed in biological networks
- **Scale-free networks** – Reka, Jeong, Barabasi (2000)
- Most nodes have only a few connections, but a few nodes (called hubs) have a very large number of connections
- World-Wide Web, power grids, social networks, regulatory networks seem to be scale free
- Explains reason for robustness of biological networks
- Robust to random connection deletions

Some Well-Characterized Network Properties



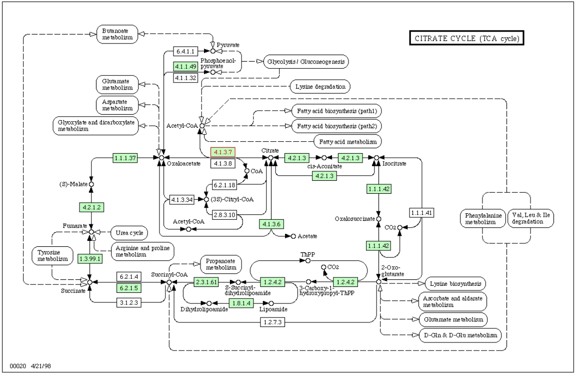
Log-plot of degree distribution in random Erdos-Renyi network



Log-plot of degree distribution in scale-free network

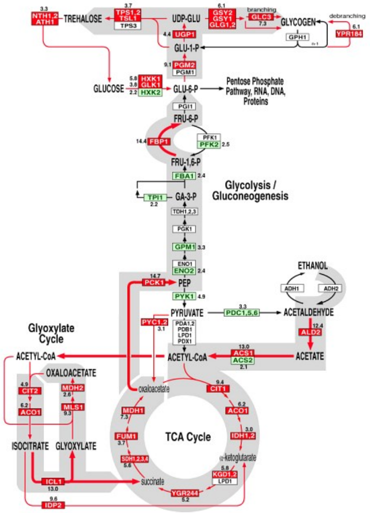
- An organism has a higher chance of creating new, useful functions by making copies of already existing systems and modifying these, rather than create new functions from scratch.
- It is possible that certain classes of genes are easier to cut-and-paste than others, and this may result in scale-free networks.

Example: Krebbs Cycle



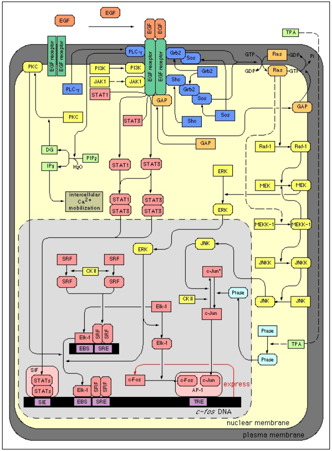
- Krebs cycle, or TCA (tricarboxylic acid) cycle, is a central part of the respiratory energy metabolism of many organisms.
- A number of enzymes act on substrates in a cycle, where an acetyl group from acetyl-coenzyme A is covalently added to oxaloacetate and then converted into two molecules of CO2 and reducing equivalents

Example: Krebbs Cycle



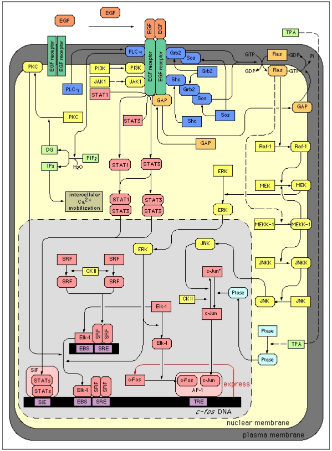
- In one of the first experiments using microarrays for monitoring gene expression changes (DeRisi, Iyer and Brown, Science (1997)), a yeast culture was allowed to grow in glucose-rich medium until glucose ran out.
- The metabolism then changes from anaerobic fermentation of glucose to ethanol, to aerobic glucose metabolism of ethanol. This is the so-called diauxic shift.
- The gene expression data shows a concerted change, where the TCS cycle genes are up-regulated as the glucose level goes down (red labels), at the same time as the glycolysis genes are down-regulated (green).

Example: RAS Signaling



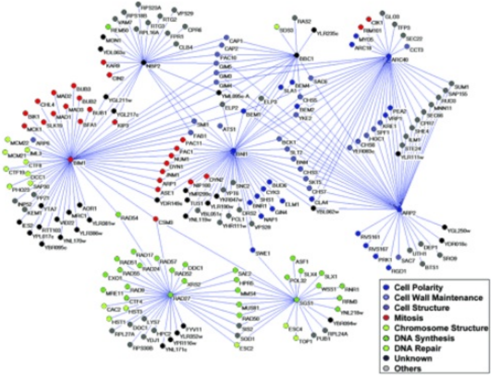
- Ras p21 (also called cH-ras p21) was discovered in the 80's as an oncogene – mutated gene in many tumors – presumably involved in making cells cancerous, or keeping them in a cancerous state.
- One large group of oncogens turned out to be genes that participate in the regulation of cell growth and division.
- Ras p21 is part of a signaling cascade that also contains proteins Shc, Grb2, SOS, GAP, Raf-1, MEK, ERK, and others.
- The purpose of the cascade is to transduce a signal from a growth-factor receptor in the cell membrane (there are several) to the transcriptional regulation of genes in the nucleus that control cellular growth and division.

Example: RAS Signaling



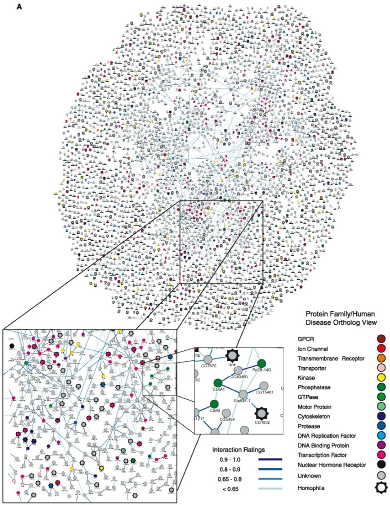
- This regulatory network can break down if a switch is stuck in 'on' position. This is what happens when certain mutations of Ras p21 gene occur. In a few Ras p21 mutants, Ras p21 is unable to hydrolyze GTP, stuck in active state.
- Ras p21 network is strongly conserved among species. The exact process that employs the pathway varies considerably. This is also the case within one organism.
- The Ras p21 pathway can be involved in different processes in different cell types, and may give different results depending on the nature of the original signal. This means that the Ras p21 pathway is to a certain extent non-specific: it usually implies cellular growth, but the exact nature of that cellular growth is dependent on other factors.

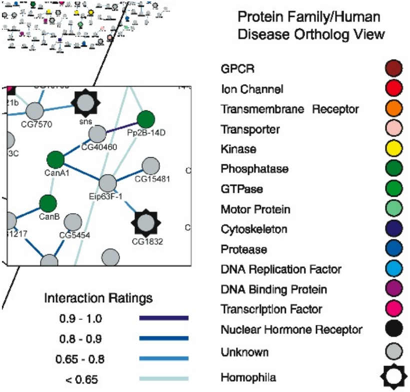
Protein Interaction Networks



Science, 14 December 2001: Vol. 294. no. 5550, pp. 2364 -2368

Protein Interaction Networks





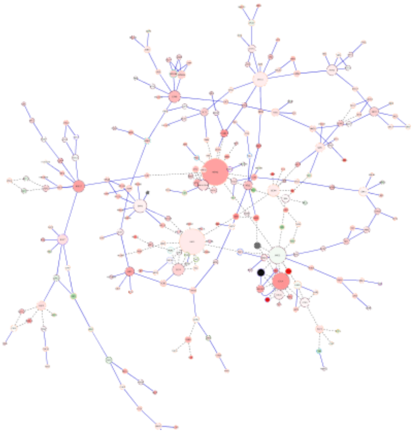
source: L. Giot, et al. 302 Science , 1727 (2003);

Network Alignment

- Overview of Network Alignment
- Definition
- Motivation
- Representation
- Applications:
  - Path-Network Alignment
  - Network-Network Alignment

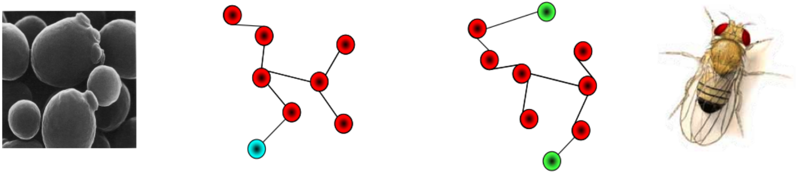
Network Alignment – Motivation

- Sequence alignment seeks to identify conserved regions of DNA or protein sequence
- Analogously, network alignment seeks to identify conserved subnetworks
- In both cases, the intuition is that conservation implies functionality



S. cerevisiae

Network Alignment – Motivation

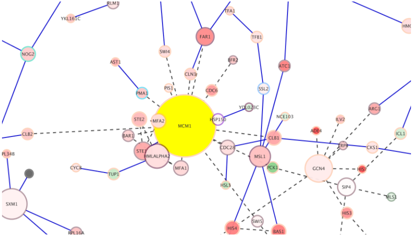


- Compare protein-protein interaction networks at the species level. Transfer annotation from known species to unknown species
- Identify functional orthologs
- Proteins performing the same function across species

Image from: http://www-math.mit.edu/womeninmath/slides/berger.pdf

Network Representation

- Undirected graph G = (V, E)
- $V = \{v_1, v_2, ..., v_N\}$ represent proteins
- $|V| = N$
- $E = e_{ij}$ represent interactions between v_i, v_j
- $|E| = M$

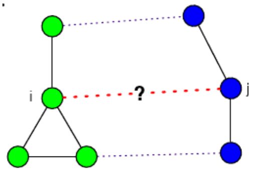


S. cerevisiae

IsoRank

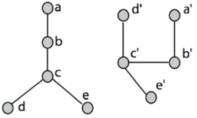
- Global alignment of multiple PPI networks.
- The guiding intuition – a protein in one PPI network is a good match for a protein in another network if their respective sequences and neighborhood topologies are a good match.
- Encoded as an eigenvalue problem in a manner analogous to Google's PageRank method.
- Maximum common subgraph – Given two graphs G_1 and G_2, map each node in G_1 to ≤ 1 node in G_2
- The mapping is good if the neighbors of i can be mapped to the neighbors of j

$$R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(v)||N(v)|} R_{uv}$$



Singh, Xu and Berger, RECOMB 2007

Example



R

	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

$R_{aa'} = \frac{1}{4} R_{bb'}$

$R_{bb'} = \frac{1}{3} R_{ac'} + \frac{1}{3} R_{a'c} + R_{aa'} + \frac{1}{9} R_{cc'}$

$R_{dd'} = \frac{1}{9} R_{cc'}$

$R_{cc'} = \frac{1}{4} R_{bb'} + \frac{1}{2} R_{be'} + \frac{1}{2} R_{bd'} + \frac{1}{2} R_{eb'} + \frac{1}{2} R_{db'} + R_{ee'} + R_{ed'} + R_{de'} + R_{dd'}$

(Figure from Singh, Xu, Berger, 2007)

$$R_{ij} := \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(u)||N(v)|} R_{uv}$$

Finding R Values

- Lots of R_{ij} values to find.
- Gather into a matrix of the form: $R = AR$ where

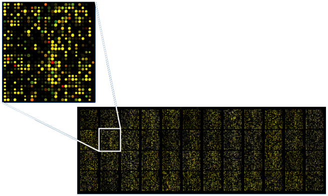
$$A[i,j][u,v] = \begin{cases} \frac{1}{|N(u)|N(v)|} & if (i,u) \in E_1, (j,v) \in E_2 \\ 0 & Otherwise \end{cases}$$

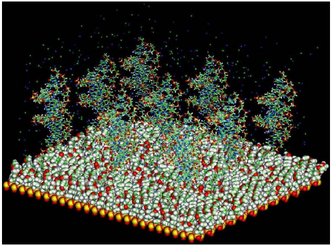
  And solve for R

- A is usually big but sparse.
- This is an eigenvalue problem.

Microarray Technology

- Microarrays allow "seeing" what genes are expressed in a patient
- Technology evolved from Southern Blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment
- A microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features
- Each feature contains picomoles of a specific DNA sequence





Applications of Microarray Technology

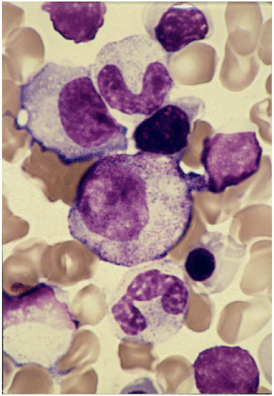
- Biological discovery
  - new and better molecular diagnostics
  - new molecular targets for therapy
  - finding and refining biological pathways
- Recent examples
  - molecular diagnosis of leukemia, breast cancer.
  - appropriate treatment for genetic signature
  - potential new drug targets

Classifying Gene Expressions (Microarray Data)

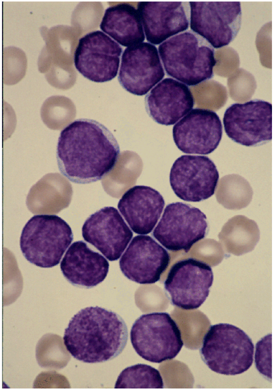
- Interesting questions:
  - What can be learned about a cell from the set of all mRNA expressed in a cell?
  - Classifying diseases: does a patient have benign prostate cancer or metastatic prostate cancer? ALL or AML?
- Problem: Classification of microarray data obtained from one patient into normal or disease categories

Classifying Gene Expressions (Microarray Data)

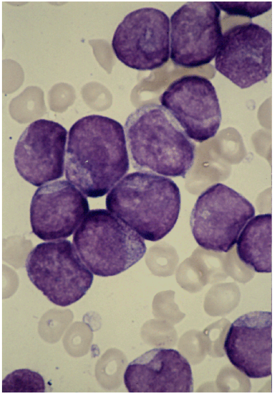
Can you tell the difference between ALL and AML?



Normal



Acute lymphoid leukemia (ALL)

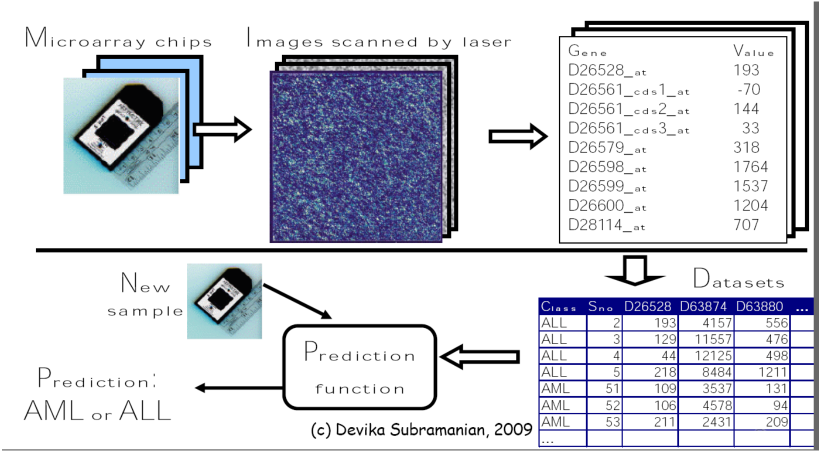


Acute myeloid leukemia (AML)

Classifying Gene Expressions (Microarray Data)

- Different leukemia classes have different genes expressed
- Is it possible to diagnose cancer classes using microarrays?
- Problem: microarrays contain a lot of data and noisy data
- Use knowledge of class values – e.g., myeloma vs. normal etc., to gain additional insight
- Find genes that are best predictors of a class
- Can provide useful tests, e.g. for choosing treatment
- If predictor is comprehensible, may provide novel insight – e.g., point to a new therapeutic target

Classifying Gene Expressions (Microarray Data)

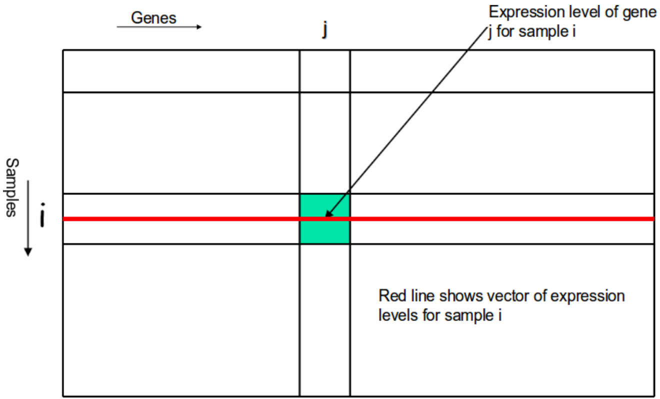


Gene	Value
D26528_at	193
D26561_cds1_at	-70
D26561_cds2_at	144
D26561_cds3_at	33
D26579_at	318
D26598_at	1764
D26599_at	1537
D26600_at	1204
D28114_at	707

Class	Sno	D26528	D63874	D63880	...
ALL	2	193	4157	556	
ALL	3	129	11557	476	
ALL	4	44	12125	498	
ALL	5	218	8484	1211	
AML	51	109	3537	131	
AML	52	106	4578	94	
AML	53	211	2431	209	
...					

(c) Devika Subramanian, 2009

What do the Data Look Like?

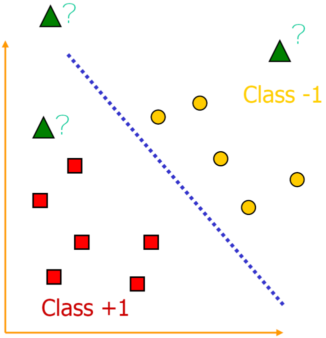

Genes
j
Expression level of gene j for sample i
Samples
i
Red line shows vector of expression levels for sample i

Challenges when Dealing with Microarray Data

- Microarray data inherit large experimental and biological variances
  - experimental bias + tissue heterogeneity
  - cross-hybridization
  - bad design: confounding effects
- Microarray data are sparse
  - high-dimensionality of genes
  - low number of samples/arrays
  - curse of dimensionality
- Microarray data are redundant
  - if are co-expressed, their expressions are strongly correlated.

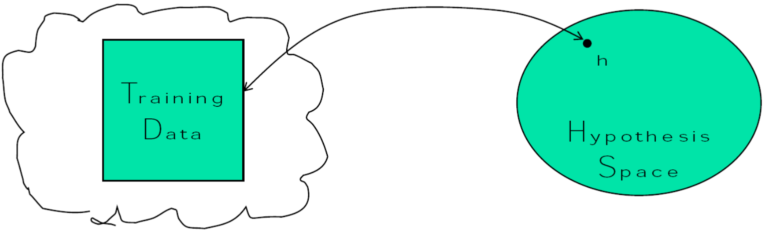
Back to Classification

- Given examples drawn from two classes, classify new examples into the correct class
- Each point represents a vector of gene expression levels
- In the Leukemia example, the two classes are AML and ALL
- Problem statement: Given training data $(x_1, y_1), \ldots, (x_m, y_m)$, $x_i \in \mathbb{R}^n$, $y_i \in \{+1, -1\}$: Estimate function $h : \mathbb{R}^n \to \{+1, -1\}$ such that h will correctly classify new unseen examples from the same underlying probability distribution as the training data



Classification as an Optimization Problem

- Set S of training data points
- Class H of hypotheses/models
- Optimization problem: Find hypothesis/model h in H that best fits all data

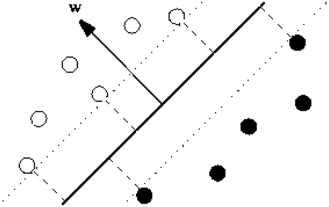


Statistical Machine Learning: Tradeoffs

- Provides tradeoff between complexity of model and amount of data needed to learn it
- With only a small amount of data, we can only discriminate between a small number of different hypotheses
- The more data, the more evidence; so one can consider more alternative hypotheses
- Complex hypotheses give better fit to the data

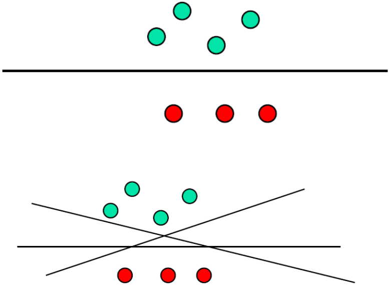
Linear Support Vector Machines

- We assume that the distribution of the two classes is such that they are linearly separable, i.e.
- one can find a linear function of the inputs x such that $f(x) < 0$ whenever the label $y = -1$ and $f(x) \geq 0$ otherwise
- This can be expressed as a hyperplane in the space x, since we are looking for a function f of the form $f(x) = w.x + b$



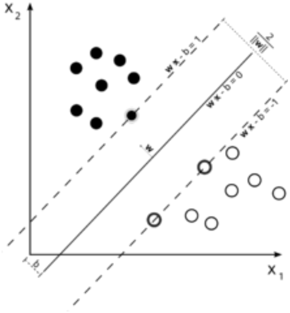
Linear Support Vector Machines

- Consider class of oriented hyperplanes in $\mathbb{R}^n$
- $h(x) = sign(w.x + b)$
- If data is linearly separable, then there is a function from this class that separates the +1 points from the -1 points
- BUT there is an infinite number of hyperplanes that can separate the data
- Among all separating hyperplanes, there is one with the maximum margin (minimizes the L2 norm)



What are Supporting Vectors?

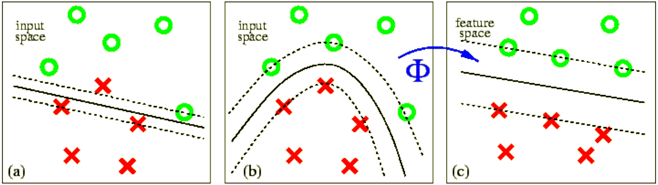
- If one works out the math, the vector w has an expansion in terms of a subset of the training data
- These data points are called support vectors. None of the other data points matter.
- The maximal margin hyperplane is completely determined by the support vectors



Non-linear Separators?

- What about the case when the decision function is known to be a non-linear function of the input x?
- Idea: feature spaces
- map x onto a higher dimensional feature space $\phi(x)$.
- Mapping transforms non-linear surface where data resides into linear surface on the higher dimensions
- Use linear support vector machines to obtain the optimal separating hyperplane in the high dimensional feature space

Non-linear Separators?

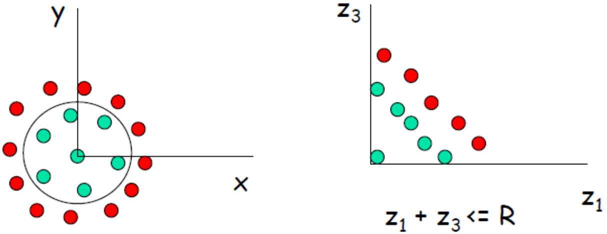


(a) Linear separation of the input points is not possible without errors. Even allowing misclassification of one data point results in a small margin.

(b) Better separation is provided by a non-linear surface in the input space.

(c) This non-linear surface corresponds to a linear surface in a feature space. Data points are mapped from input space to feature space by the function ϕ induced by the kernel function k

Non-linear Separators?



$$\varphi : \Re^2 \rightarrow \Re^3$$

$$\varphi((x, y)) = (x^2, \sqrt{2}xy, y^2)$$

More Reading

- William S Noble. What is a support vector machine? Nature Biotechnology, 24(16):1565-1567, 2006
- William S Noble. Support vector machine applications in computational biology. http://noble.gs.washington.edu/papers/noble_support.pdf
- S. Mukherjee, P. Tamayo D. Slonim, A. Verri, T. Golub, J.P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. MIT AI Lab, A.I. Memo No. 1677 - CBCL Paper N0.82 http://historical.ncstrl.org/tr/pdf/mitai/AIM-1677.pdf