Computational Biology
Lecture #9: Transcriptomics

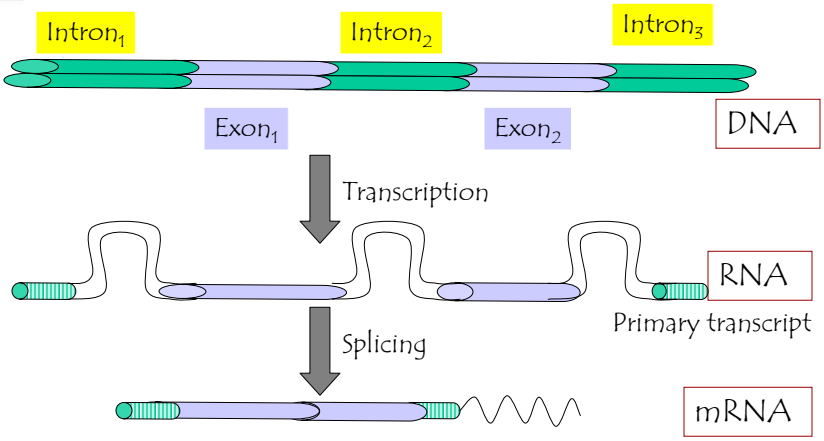
Bud Mishra
Professor of Computer Science, Mathematics, & Cell Biology
Nov 14 2005

Interrupted Genes

Interrupted Genes:

- An open reading frame (containing a gene) consists of
  - INTRONS: Intervening sequences a Noncoding regions
  - EXONS: Protein coding regions
- Introns are abundant in eukaryotes and certain animal viruses.

Interrupted Genes:



Some Genes…

Gene Product	Organism	Exon Length	#Introns	Intron Length
Adenoshine deaminase	Human	1500	11	30,000
Apolipoprotein B	Human	14,000	28	29,000
Erythropoietin	Human	582	4	1562
Thyroglobulin	Human	8500	= 40	100,000
α-interferon	Human	600	0	0
Fibroin	Silk Worm	18,000	1	970
Phaseolin	French Bean	1263	5	515

Organization of Genetic Information

- Bacterial Genome:
  - Genes are closely spaced along the DNA.
  - The sequences of genes may overlap.
  - Related genes (encoding enzymes whose functions are part of the same pathway or whose activities are related) are linked as a single transcription unit.

Organization of Genetic Information

- Eukaryotic Genome:
  - Genes are separated by long stretches of noncoding DNA sequences.
  - Multiple genes in a single transcription unit is extremely rare.
  - Multiple chromosomes ↦ Linear
  - Chloroplasts and mitochondria ↦ Circular
  - Genes appearing on the same chromosome are syntenic.

Location of Some Genes on Human Chromosome.

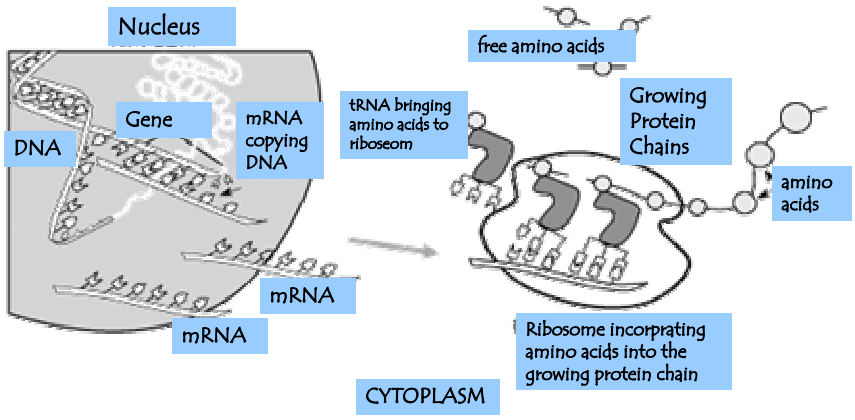
Genes	chromosomes
α-globin cluster	16
β-globin cluster	11
Immunoglobulin	
κ (light chain)	2
λ (light chain)	22
Heavy Chain	14
Pseudogenes	9,32,15,18
Growth Hormone gene cluster	17
Thymidine kinase	17

Genes	chromosomes
Insulin	11
Galactokinase	11
Viral oncogene homologues	
C-sis	22
C-mos	8
C-Ha-Ras-1	11
C-myb	6
Interferons	
α & β cluster	9
γ	12

Eukaryotic Genome

- Multiple copies of the same gene
  - Solve "supply problem"
  - There are several hundred ribosomal RNA genes I mammals
- Pseudogenes
  - Nonfunctional copies of genes…(Deletions or alterations in the DNA sequence)
  - Number of pseudo genes for a particular gene varies greatly…Different from one organism to another.

Gene Expression



Transcription

- A gene consists of a coding region and a regulatory region.
  - The coding region is the part that is transcribed into an mRNA and translated into a finished protein.
  - The regulatory region is the part of the DNA that contributes to the control of the gene.
- The regulatory region contains
  - Binding sites for transcription factors (TF), which act by binding to the DNA (directly or with other transcription factors in a small complex);

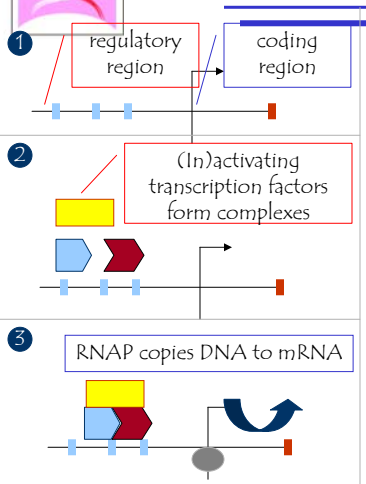
Regulatory Regions

- In prokaryotes,
  - The regulatory region is short (10-100 bases) and contains binding sites for small number of TFs.
- In eukaryotes
  - The regulatory regions can be very long (up to 10,000 or 100,000 bases), and contains binding sites for multiple TFs.
  - TFs may act positively or negatively.
  - Another input mechanism is phosphorylation or dephosphorylation of a bound TF by other proteins.

Terminology

- TFs are sometimes called *trans-regulatory elements*, and DNA sites where TFs bind are called cis-regulatory elements.
  - The collection of cis-regulatory elements upstream of the coding region is called the **promoter**.

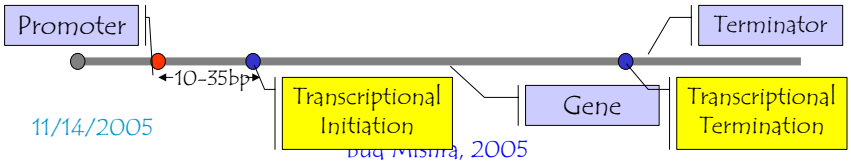
Transcription Initiation



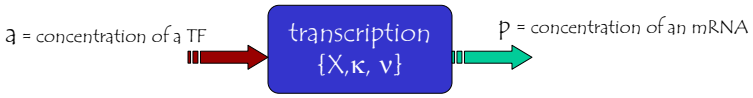
- Typically, TFs do not bind singly, but in complexes:
- Once bound to the DNA, TF complex allows RNA polymerase (RNAP) to bind to the DNA upstream of the coding region.
- RNAP forms a transcriptional complex that separates the two strands of DNA, thus forming an open complex, then moves along one strand of the DNA, step by step and transcribes the coding region into mRNA.

Regulation of Gene Expression

- Motifs (short DNA sequences) that regulate transcription:
  - Promoter & Terminator
  - The rate of transcription varies according to experimental conditions
- Motifs that modulate transcription
  - Repressor, Activator, Antiterminator



Model of transcription



- ν = Cooperativity coefficient
- κ = Concentration of a at which transcription of m is "half-maximally" activated.

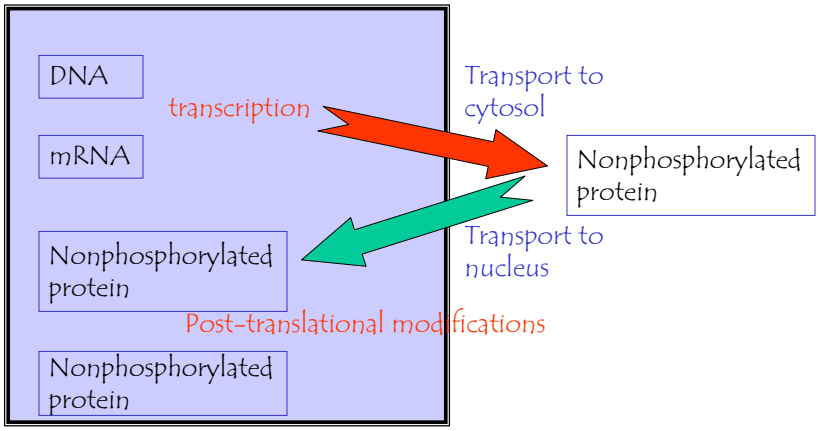
- $dp/dt = \Phi(a, \kappa, \nu) = V a^{\nu}/[\kappa^{\nu} + a^{\nu}]$
- A graph of function Φ = Sigmoid Function
- If $\nu = 1$ then, the transcription activation function resembles the classical Michaelis-Menten!

Regulatory Networks

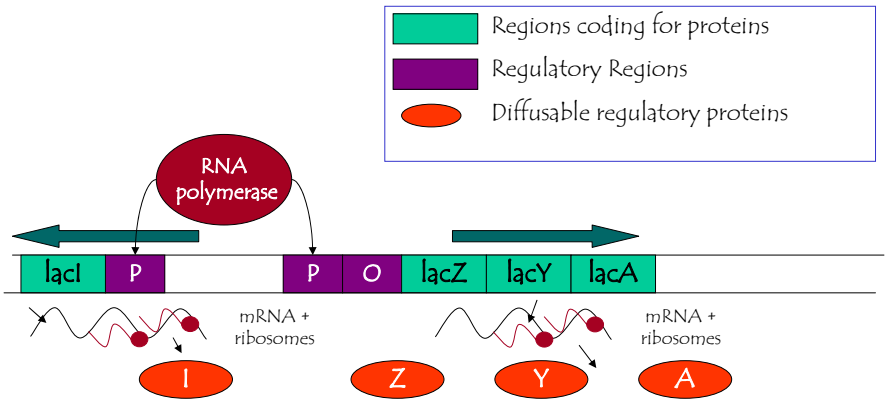
Regulatory Networks

- Variations among Cells:
  - All cells in an organism have the same genomic data, but the proteins synthesized in each vary according to cell type, time and environmental factors
  - There are network of interactions among various biochemical entities in a cell (DNA RNA, protein, small molecules)

Gene Regulation



Transcriptional Regulation: Example: The lac Operon



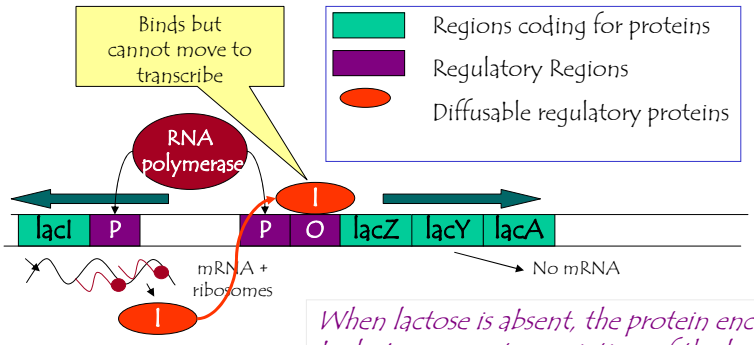
The lac Operon

- Regulates utilization of lactose by the bacterium E. coli.
  - Lactose is not generally available to *E. coli* as a food substrate, so the bacterium does not usually synthesize the enzymes necessary for its metabolic use.
- There is an operon, called the lac operaon, normally turned off, that codes for three enzymes:
  - b-galactoside permease, b-galactosidase and b-thiogalactoside acetyl transferase.

Activation of the lac operon

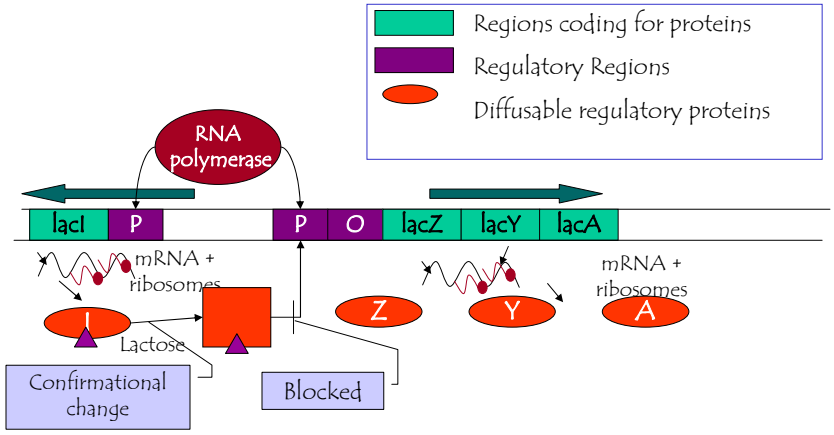
- An autoctalyitic reaction..
  - If the bacterium is exposed to lactose, these enzymes work together to (1) transport lactose into the cell and (2) isomerizes lactose into allolactose (an allosteric isomer of lactose.
    - The allolactose binds with a repressor molecule to keep it from repressing the production of mRNA.
    - Production of allolactose turns on the production of mRNA, which then leads to production of more enzyme, enabling production of more lactose to allolactose…

Transcriptional Regulation: Example: The lac Operon



When lactose is absent, the protein encoded by lacI represses transcription of the lac operon

Transcriptional Regulation: Example: The lac Operon



Mathematical Model

$$G + mP \underset{k_{-1}}{\overset{k1}{\rightleftharpoons}} X$$

- Production of enzyme is turned on by m molecules of the product allolactose P…
  - G=Inactive state of the gene
  - X=Active state of the gene
  - In a large population of genes, the percentage of active genes is given by the chemical equilibrium:

$$p = [P]^m/(k_{eq}^m + [P]^m)$$

Production of mRNA

- The differential equation governing the (average) production of mRNA

$$dM/dt = M_0 + k_1 [P]^m/(k_{eq}^m + [P]^m) - k_2 M,$$

- where M is the concentration of mRNA that codes for the enzyme.
  - Production of the enzymes (responsible for transforming into allolactose substrate):

$$dE_1/dt = c_1 M - d_1 E_1;$$
$$dE_2/dt = c_2 M - d_2 E_2.$$

Lactose states

- S_0 = Concentration of the lactose, exterior to the cell.
- S = Concentration of the lactose interior to the cell.

◇ [P] = Concentration of allolactose.

$$dS_0/dt = -\sigma_0 E_1 S_0/(k_0 + S_0)$$

$$dS/dt = \sigma_0 E_1 S_0/(k_0 + S_0) - \sigma_1 E_2 S/(k_s + S)$$

$$d[P]/dt = \sigma_1 E_2 S/(k_s + S) - \sigma_2 E_2 [P]/(k_p + [P])$$

Simplification

◇ Assume:

- mRNA is in quasi-steady state:

$$M = (k_1/k_2)[P]^m/(k_{eq}^m+[P]^m) + M_0/k_2;$$

- $d_1 = d_2$. Degradation is slow compared to cell growth. Also, $E_1 = E_2$.

$$dE_1/dt = c_1M_0/k_2 + (c_1k_1/k_2)[P]^m/(k_{eq}^m+[P]^m) - d_1 E_1;$$

- No delay in conversion from lactose to allolactose:

$$d[P]/dt = \sigma_0 E_1 S_0/(k_0+S_0) - \sigma_2 E_1 [P]/(k_p + [P]).$$

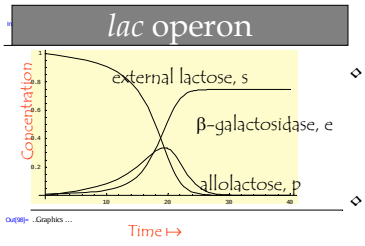
Dimensionless Form

- Dimensionless variables:
  - $S_0 = k_0\, s$, $[P] = k_p\, p$, $E_1 = e_0\, e$, and $t = t_0\, \tau$…

$$de/d\tau = m_0 + p^m/(\kappa^m + p^m) - \varepsilon\, e,$$
$$dp/d\tau = \mu\, e[s/(s+1) - \lambda\, p/(p+1)],$$
$$ds/d\tau = -e\, s/(s+1),$$

  - where $e_0^2 = c_1 k_0 k_1/(\sigma_0 k_2)$, $t_0 = k+O/(e_{0\sigma 0})$, $\lambda = \sigma_2/\sigma_0$, $\mu = k_0/k_p$, $\kappa = k/k_p$, $m_0 = M_0/k_1$, and $\varepsilon = t_0\, d_1$…

The *lac* operon

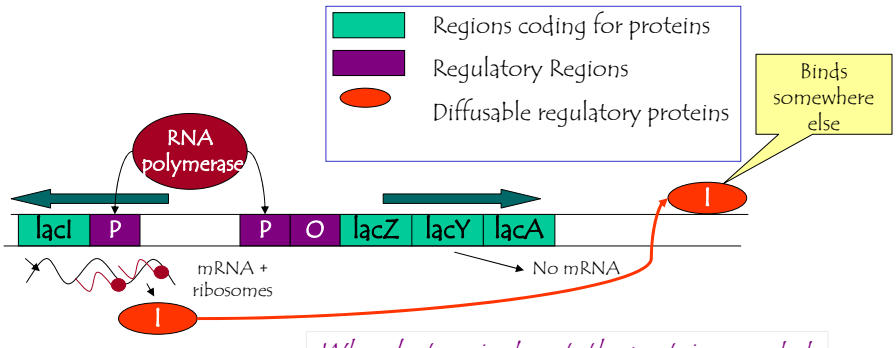


- If the amount of lactose is too small, then the lactose is gradually depleted, although there is no increase in enzyme concentration.
- However, if the lactose dose is sufficiently large, then there is an autocatalytic response, as the lac operon is turned on and enzyme is produced.
- The production of enzyme shuts down when the lactose stimuls is consumed, and the enzyme concentration gradually declines…

Example of Competition

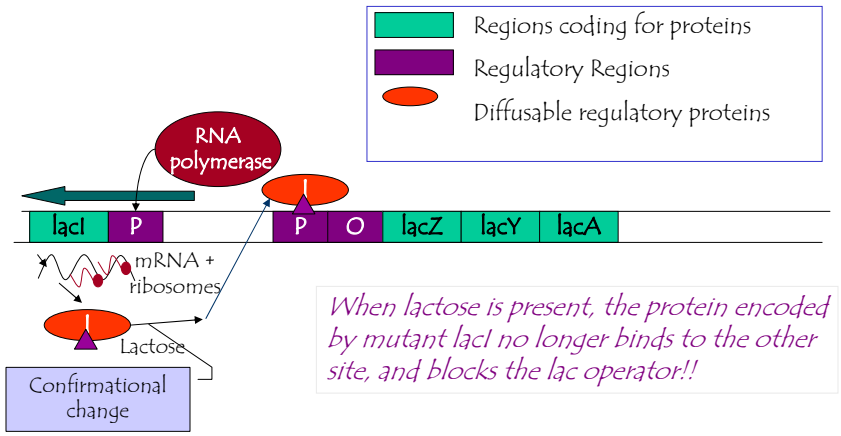
- The mutant Lac repressor X186:
  - This mutant represses transcription of the *lac* genes in the presence of lactose…
  - The mutant binds DNA so tightly that, in the absence of inducer (allolactose), it is sequestered on non-operator DNA sites.
  - The inducer weakens the binding of the mutant repressor; thus, allowing it bind to the *lac* operon.

Lac repressor X 186



When lactose is absent, the protein encoded by mutant lacI binds to some other site.

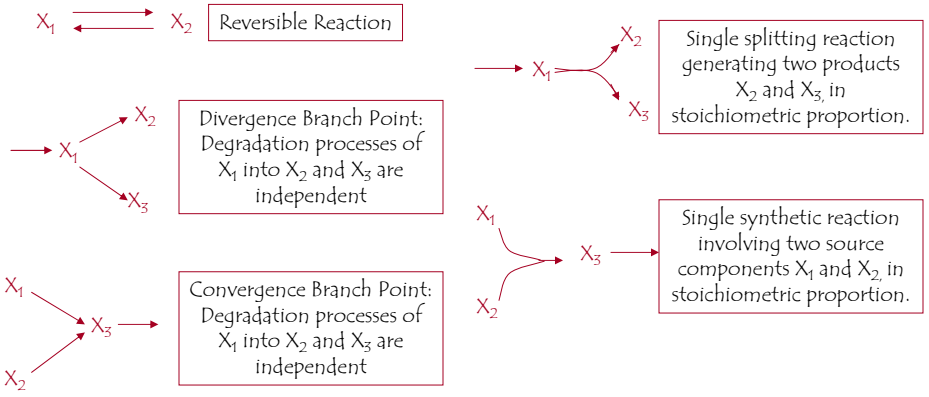
Lac repressor X 186



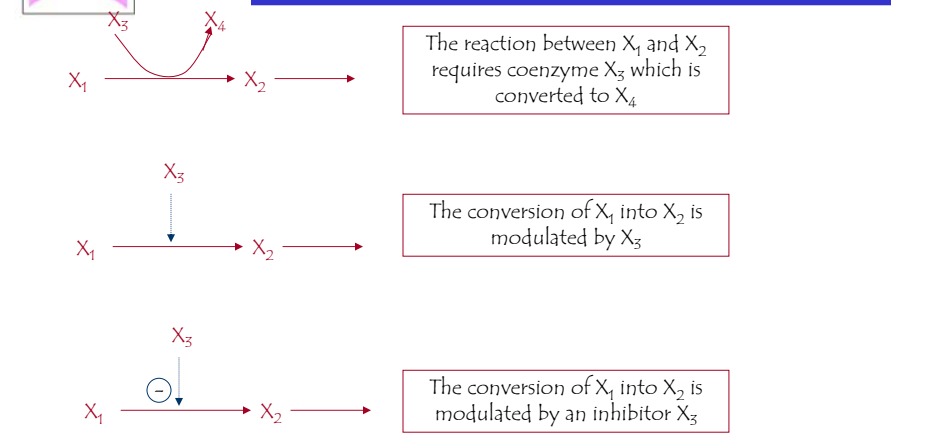
When lactose is present, the protein encoded by mutant lacI no longer binds to the other site, and blocks the lac operator!!

S-Systems

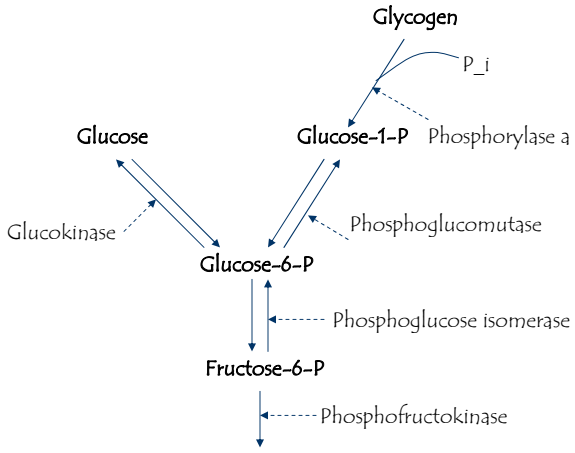
Graphical Representation



Graphical Representation



Glycolysis



S-Systems

- Dependent Variables: $X_i(t)$, i=1,…,n, 0 **5** t.
- System is described in terms of the temporal changes in dependent variables:
  - E.g., Instantaneous product formation in response to changes in the exogenous substrate, inhibitor or enzyme concentration…
  - Kinetic Laws: Relate a reaction rate to concentrations.
  - Reaction Rate = Instantaneous temporal rate of change in concentration of substrate or product.
- Is this information sufficient to deduce the dynamics of a biochemical system? Yes.

Systems of Differential Equations

- dX_i/dt = (instantaneous) rate of change in X_i at time t = Function of substrate concentrations, enzymes, factors and products:

$dX_i/dt = f(S_1, S_2, \ldots, E_1, E_2, \ldots, F_1, F_2, \ldots, P_1, P_2, \ldots)$

- E.g. Michaelis-Menten for substrate S & product P:

1. $dS/dt = - V_{max}\, S/(K_M + S)$
2. $dP/dt = V_{max}\, S/(K_M + S)$

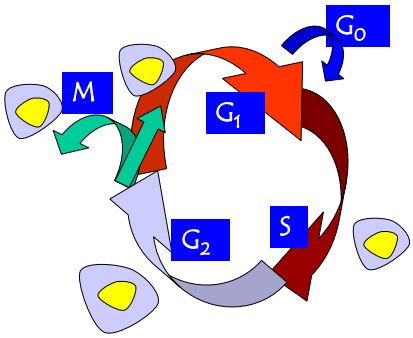
Cell Informatics

The dynamics of cell:

- The cell cycle) the set of events that occur within a cell between its birth by mitosis and its division into daughter cells again by mitosis
  - interphase period when DNA is synthesized and
  - mitotic phase
    - The cell division by mitosis (into 2 daughter cells) and meiosis (into 4 gametes from germ-line cells);

The Cell Cycle:

- In growing cells, the four phases proceed successively, taking from 10-20 hrs.
- Interphase: comprises the G_1, S, and G_2 phases. DNA is synthesized in S and other cellular macromolecules are synthesized throughout interphase, roughly doubling cell's mass.
- During G_2 the cell is prepared for mitotic (M) phase when the genetic material is evenly proportioned and the cell divides.
- Nondividing cells exit the normal cycle, entering the quiesecent G_0 state.



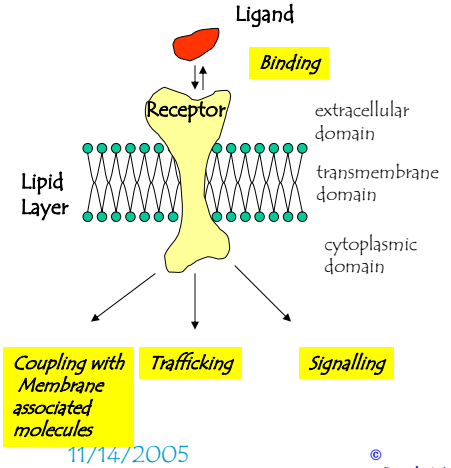
Differentiation

- Cellular dynamics controls how a cell changes (or differentiates) to carry out a specialized functions
  - Structural or morphological changes (muscles, neural, skin..)
  - Immune systems: Many cell types come together in organized tissues designed to let the body distinguish self from non-self.

Suicide

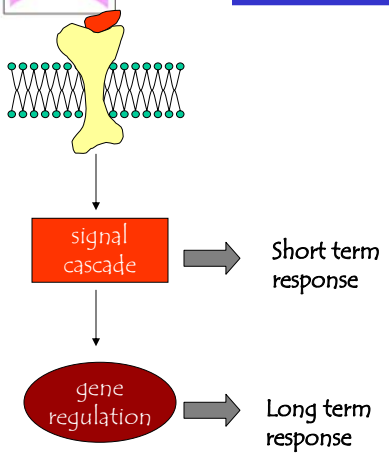
- Programmed Cell Death/Apoptosis:
  - Condensation of the nucleus.
  - Fragmentation of the DNA.
  - Morphological changes followed by consumption by macrophages.

Cell Talk



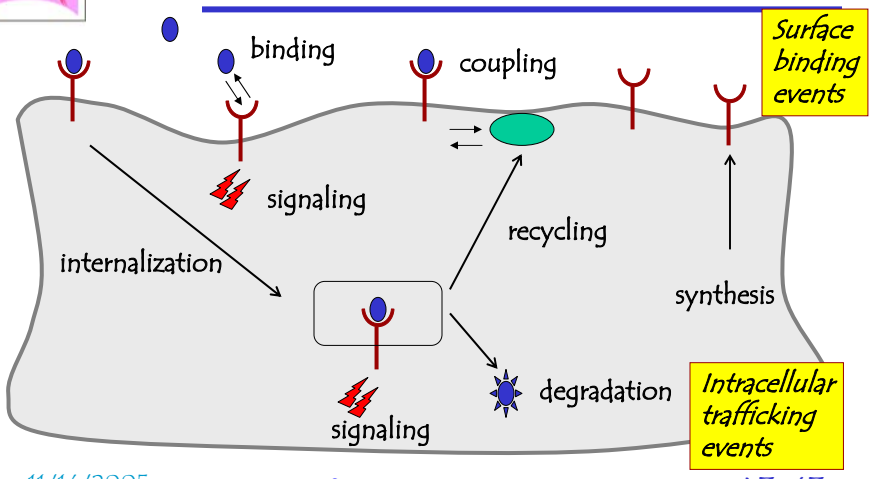
- Cell Surface Receptors
  - Extracellular domain for binding ligands (e.g., growth factors, adhesion molecules, etc.)
  - Transmembrane domain
  - Intracellular cytoplasmic domain
- Receptor driven cellular behavior are extremely important
  - E.g., Growth, Secretion, Contraction, Motility and Adhesion

Receptors and Gene Regulation



- Ligands bind to receptors at the cell surface.
- Bound receptors activate various intracellular enzymes and initiate entire cascades of intracellular reactions
  - Some of these regions trigger short term (of the order of milliseconds to minutes) responses.
  - Some eventually trigger long-term responses..e.g., requiring protein synthesis and additional molecular interactions

A Complex Picture



A Complex Picture

- Trafficking
  - Receptor population undergoes many complex events of coupling with other cell surface molecules
  - Internalization (RME: receptor-mediated endocytosis)
  - Recycling
  - Degradation
  - Synthesis

Gene Expression Data

- Microarrays enable one
  - To simultaneously measure the activity of up to 30,000 (» 10^4—10^5) genes.
  - In particular, the amount of mRNA for each gene in a given sample (or a pair of samples) can be measured.

Gene Expression Data

- Microarrays provide a tool for answering a wide variety of questions:
  - In which cells is each gene active?
  - Under what environmental conditions is each gene active?
  - How does the activity level of a gene change under different conditions?
    - Stage of a cell cycle? Environmental conditions? Diseases?

Gene Expression Data

- Functional genomics with microarrays:
  - What genes can be inferred to be regulated together?
  - What happens to the expression level of every gene when a (candidate) gene is mutated?
  - What can be inferred about the regulatory structure?

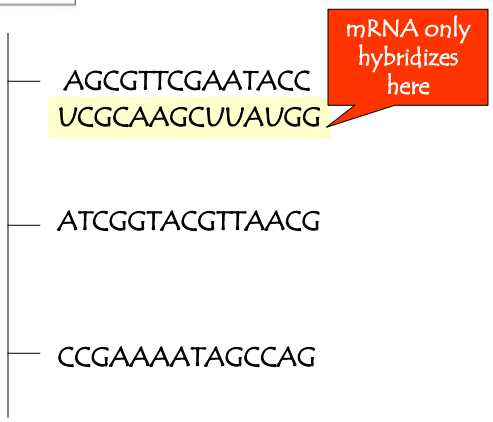
The Computational Tasks

- Clustering Genes:
  - Genes co-regulated together
- Classifying Genes:
  - Functional class a particular gene fall into?
- Classifying Gene Expressions:
  - Disease classification from the set of all mRNA expressed in a cell
- Inferring Regulatory Networks:
  - The "circuitry" of the cell

Microarrays

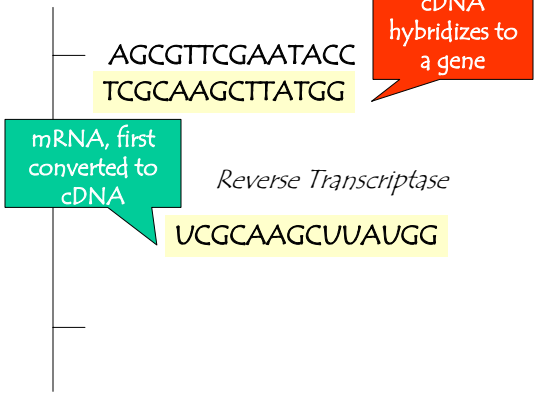
- Two general types currently popular…
  - Spotted Arrays (Pat Brown, Stanford)
  - Oligonucleotide Arrays (Affymetrix)
  - Other variations (Agilent, Incyte, NGS, …)
- The key idea is to query a genome for a particular pattern by complementary hybridization.

Complementary Hybridization



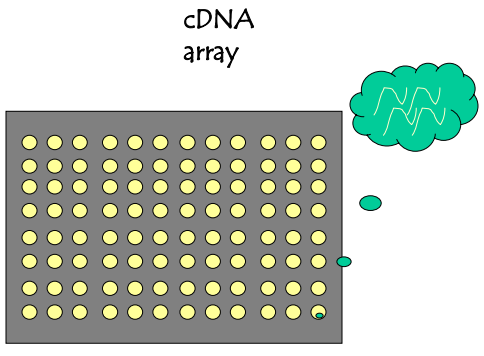
- Due to Watson-Crick base pairing, an mRNA molecule will hybridize to a omplementary DNA molecule.

Complementary Hybridization



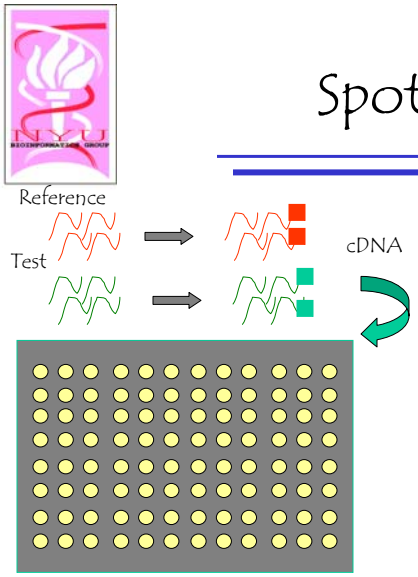
- Practical implementation:
  - Put the actual gene sequence on array
  - Convert mRNA to cDNA using *reverse transcriptase*
  - Hybridize cDNA to the sequence on the array

Spotted Array



- Robots array microscopic sized spots of DNA on glass slides
  - Each spot is DNA analog (cDNA) of one of the mRNA's we wish to measure…

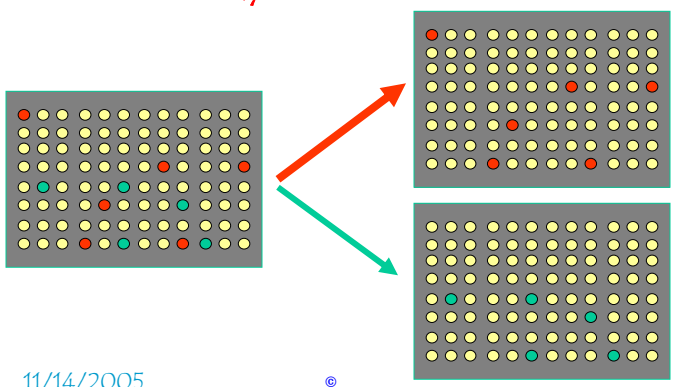
Spotted Arrays



- Two samples (reference and test) of mRNA are converted to cDNA, labeled with fluorochrome dyes and allowed to hybridize to the array.

Spotted Arrays

- Lasers applied to the arrays yield an emission for each fluorescent dye.



Oligonucleotide Arrays

- "Gene Chips"
  - Instead of putting entire genes on array, put sets of DNA 25-mers (synthesized oligonucleotides)
  - Produced using a photolithography process similar to the ones used to create semiconductor chips
  - mRNA samples are processed separately instead of in pairs (of reference/control and test)

Oligonucleotide Arrays

- Given a gene to be queried/measured, select a large number (» 20) 25-mers for that gene.
- Selection criteria
  - Specificity
  - Hybridization properties
  - Ease of manufacturing

Oligonucleotide Arrays

- Each of these probes is put on the chip
  - Additionally a slight variant (that differs only at the 13th base) of each oligo is put next to it.
  - This helps factor out false hybridization (pm[perfect match] vs. mm[mismatch])
- The measurement for a gene is derived from these 40 separate measurements.

Genome-wide Cluster Analysis

- Put all genes (» 6200) of *S. cerevisae* (yeast) on a single microarray
- Measure experiment across *m* independent experiments
- Group together genes that have similar expression profiles.
  - Eisen et al. PNAS 1998

Genome-wide Cluster Analysis

- Each measurement G_i represents

  $\log (red_i/green_i)$

  Where red is the test expression level and green is the reference expression level for gene G in the i^{th} experiment.
- The expression profile of a gene is the vector of measurements across all experiments:

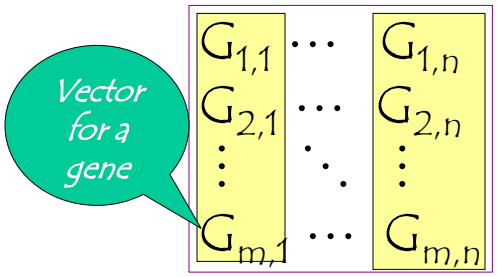
  $\mathbf{h}\ G_1, \ldots, G_m\ \mathbf{i}.$

The Data

- 79 measurements for each of 2467 genes
- Data collected at various times during
  - Diauxic shift (shutting down genes for metabolizing sugar, activating genes for metabolizing ethanol)
  - Mitotic cell division cycle
  - Sporulation
  - Temperature shock
  - Reducing Shock

The Data

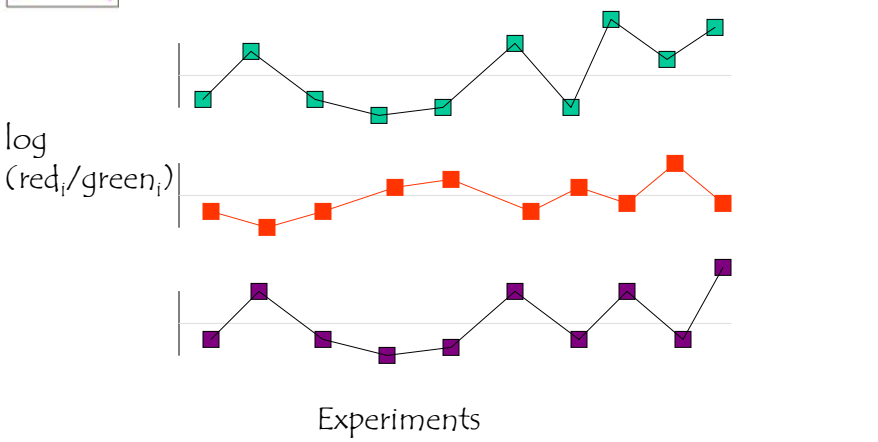
- *n* genes measured in *m* experiments:



The Task

- Given
  - Expression profiles for a set of genes.
- Compute
  - An organization of genes into clusters such that genes within a cluster have similar profiles.

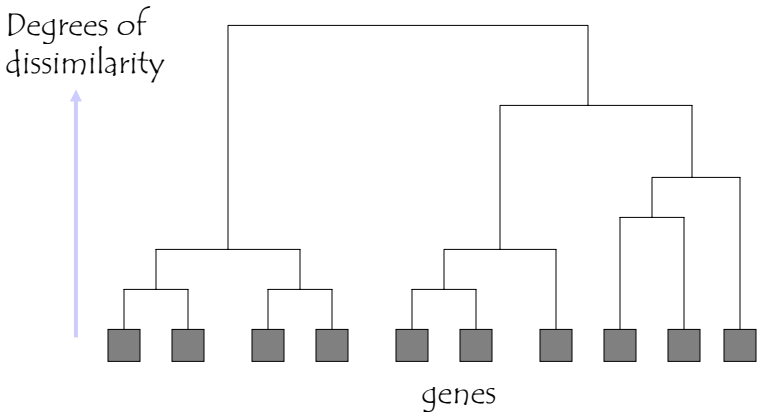
The Task



Approaches

- Eisen et al.: Hierarchical clustering.
- Other clustering methods have been applied to this gene expression data:
  - EM with Gaussian Clusters [Mjolsness et al. '99]
  - Self Organizing Maps [Tamayo et al. '99]
  - Graph Theory Algorithms [Ben-Dor & Yakhini '98, Hartuv et al. '99]

Hierarchical Clustering



Hierarchical Clustering

- *P* = set of genes
- While more than one subtree in *P*
  - Pick the most similar pair *i*, *j* in *P*
  - Define a new subtree *k* joining *i* and *j*
  - Remove *i* and *j* from *P* and insert *k*

Gene Similarity Metric

- Similarity between two genes: X and Y

$S(X, Y) =$

$$(1/N) \sum_{i=1}^{N} (X_i - X_{offset}/\Phi_X) (Y_i - Y_{offset}/\Phi_Y)$$

where

$$\Phi_G = [\sum_{i=1}^{N} (G_i - G_{offset})^2/N]^{1/2}$$

Gene Similarity Metric

- Since there is an assumed reference state (the gene's expression level did not change), G_{offset} is set to 0 for all genes

- $S(X, Y)$

$= (1/N) \sum_{i=1}^{N} [X_i/ \{\sum_{i=1}^{N} X_i^2/N\}^{1/2}] [Y_i/ \{\sum_{i=1}^{N} Y_i^2/N\}^{1/2}]$

$= (1/N) \{\sum_{i=1}^{N} X_i Y_i\}/\{SD(X) SD(Y)\}$

$= (1/N) \{X ¢ Y\}/\{SD(X) SD(Y)\}$

Results

- Redundant representations of genes cluster together.
  - But individual genes can be distinguished from related genes by subtle differences in expression.

Results

- Genes of similar function cluster together.
  - E.g., 126 genes were found strongly down-regulated in response to stress.
    - ❖112 of these genes encode ribosomal and other proteins related to translation
    - ❖The result agrees with previously known result that yeast responds to favorable growth conditions by increasing the production of ribosomes.

Molecular Classification of Cancer

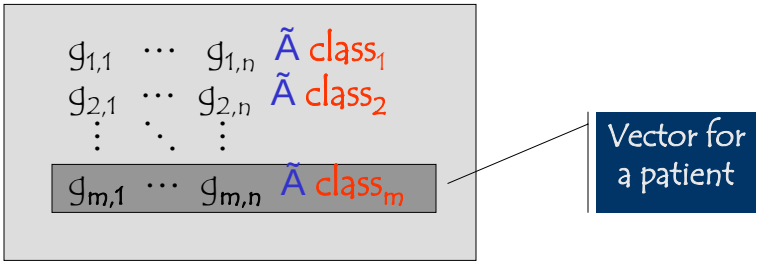
- Measure activity of 6817 genes in 38 leukemia patients
- Two tasks:
  - ❖Class Prediction
  - ❖Class Discovery
  - Golub et al., Science '99.
  - Slonim et al. '99.

Cancer Class Prediction

- Learning Task
  - Given: Expression profiles of leukemia patients
  - Compute: A model distinguishing disease classes (e.g., AML vs. ALL patients) from expression data.
- Classification Task
  - Given: Expression profile of a new patient + A learned model (e.g., one computed in a learning task)
  - Determine: The disease class of the patient (e.g., whether the patient has AML or ALL)

Cancer Class Prediction

- *n* genes measured in *m* patients

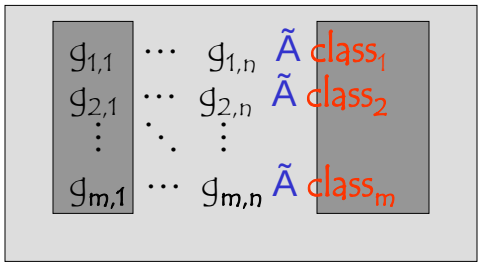


Cancer Class Prediction Approach

- Rank genes by their correlation with class variable (AML/ALL)
- Select subset of "informative" genes
- Have these genes do a weighted vote to classify a previously unclassified patient.

Ranking Genes

- Rank genes by how predictive they are (individually) of the class…



Ranking Genes

- Split the expression values for a given gene g into two pools – one for each class (AML vs. ALL)
- Determine their mean μ and standard deviation σ of each pool

- Rank genes by

$$P(g, class) = (\mu_{ALL} - \mu_{AML})/(\sigma_{ALL} + \sigma_{AML})$$

Selecting Genes

- Select the kALL top ranked genes (highly expressed in ALL) and the kAML bottom ranked genes (highly expressed in AML)

$$P(g, class) = (mALL - mAML)/(sALL + sAML)$$

Weighted Voting

- Given a new patient to classify,
  - Each of the selected genes casts a weighted vote for only one class.
  - The class that gets the most vote is the prediction.

Weighted Voting

- Suppose that x is the expression level measured for gene g in the patient

$$V = P(g,class) \pounds (x - [\mu_{ALL} + \mu_{AML}]/2)$$

Weight for gene g

Distance from the measurement to the class boundary

Prediction Strength

- Can assess the "strength" of a prediction as follows:

$$PS = (V_{winner} - V_{loser})/(V_{winner} + V_{loser})$$

where V_{winner} is the summed vote from the winning class, and V_{loser} is the summed vote for the losing class

Prediction Strength

- When classifying new cases, the algorithm ignores those cases where the strength of the prediction is below a threshold…
- Prediction =
  - [ALL, if $V_{ALL} > V_{AML}$ Æ $PS > \theta$
  - [AML, if $V_{AML} > V_{ALL}$ Æ $PS > \theta$
  - [No-call, otherwise.

Experiments

- Cross validation with the original set of patients
  - For i = 1 to 38
    - ❖Hold the i^{th} gene aside
    - ❖Use the other 37 genes to determine weights
    - ❖With this set of weights, make prediction on the i^{th} gene
- Testing with another set of 34 patients…

Results

- Cross-validation experiments
  - All trials that used at least 3 genes had 0 prediction error, with 1—4 no-calls.
- Using the 50 gene model on a test set of 34 additional patients
  - 29 correct predictions
  - 5 no-calls.

Comments on Molecular Classification of Cancer

- Gene expression profiling appears to be a promising tool for molecular medicine
  - Screening
  - Diagnosis
  - Prognosis
  - Highly targeted genome-based therapy

Cancer Class Discovery

- Given
  - Expression profiles of leukemia patients
- Do
  - Cluster the profiles, leading to discovery of the subclasses of leukemia represented by the set of patients

Cancer Class Discovery Experiment

- Cluster the expression profiles of 38 patients in the training set
  - Using self-organizing maps with a predefined number of clusters (say, k)
- Run with k = 2
  - Cluster 1 contained 1 AML, 24 ALL
  - Cluster 2 contained 10 AML, 3 ALL

Cancer Class Discovery Experiment

- Run with k = 4
  - Cluster 1 contained mostly AML
  - Cluster 2 contained mostly T-cell ALL
  - Cluster 3 contained mostly B-cell ALL
  - Cluster 4 contained mostly B-cell ALL
- It is unlikely that the clustering algorithm was able to discover the distinction between T-cell and B-cell ALL cases

Comments on Cancer Class Discovery

- Potential to discover unknown but clinically significant classes
  - One may still be able to take advantage of class labels to guide subclass discovery
  - Room for novel statistical algorithms…

To be continued…

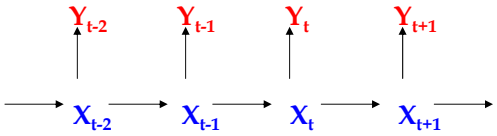
. . .

Hidden Markov Models

HMM

Hidden Markov Model

- Bayesian network structure for both
  - Hidden Markov Model
- Important Independence Assumptions:
  - Current state X_t depends only on the past state X_{t-1}
  - Current output Y_t only depends on the state X_t



Hidden Markov Models (HMM)

- Defined by an alphabet S,
  - A set of (hidden) states Q,
  - A matrix of state transition probabilities A,
  - and a matrix of emission probabilities E.

States

- S = An alphabet of symbols
- Q = A set of states that emit symbols from the alphabet S
- A = (akl) = |Q| £ |Q| matrix of state transition probabilities
- E = (eK(B)) = |Q| £ |S| matrix of emission probabilities

A Path in the HMM

- $\pi = \pi_1 \pi_2 \cdots \pi_n$
  = a sequence of states 2 Q* in the hidden Markov model M
  - x 2 Σ^* = sequence generated by the path π, determined by the model M
  - $P(x| \pi) = P(\pi_1)[\prod_{i=1}^{n} P(x_i | \pi_i) P(\pi_i | \pi_{i+1})]$

A Path in the HMM

- $P(x|\pi) = [\prod_{i=1}^{n} P(x_i \mid \pi_i)\, P(\pi_i \mid \pi_{i+1})]\, P(\pi_1)$
- $P(x_i \mid \pi_i) = e_{\pi i}(x_i)$
- $P(\pi_i \mid \pi_{i+1}) = a_{\pi i, \pi i+1}$
- π_0 = Initial state "begin"
- π_{n+1} = Final state "end"

◇ $P(x|\pi)$

$= a_{\pi 0, \pi 1}\, e_{\pi 1}(x_1)\, a_{\pi 1, \pi 2}\, e_{\pi 2}(x_2) \cdots e_{\pi n}(x_n)\, a_{\pi n, \pi n+1}$

$= a_{\pi 0, \pi 1} \prod_{i=1}^{n} e_{\pi i}(x_i)\, a_{\pi i, \pi i+1}$

Decoding Problem

◇ For a given sequence x, and a given path π,
- The model (Markovian) defines the probability
- $P(x \mid \pi)$
- The dealer knows π and x
- The player knows x but not π

"The path of x is hidden."

◇ Decoding Problem: Find an optimal path π^* for x such that $P(x|\pi)$ is maximized.

$$\pi^* = \arg\max_{\pi} P(x|\pi)$$

Dynamic Programming Approach

- Principle of Optimality:
- Optimal path for the (i+1)-prefix of x

$$x_1 \cdots x_{i+1}$$

  - uses a path for an i-prefix of x that is optimal among the paths ending in an (unknown) state $\pi_i = k \in Q$

Dynamic Programming Approach

- $s_k(i) =$
  - The probability of the most probable path for the i-prefix ending in state k.

$$\forall k \in Q \; \forall 1 \le i \le n$$
$$s_l(i+1) = e_l(x_{i+1}) \cdot \max_{k \in Q} [s_k(i) \cdot a_{kl}]$$

Dynamic Programming

- i=0

$$s_{begin}(0) = 1,\ s_k(0) = 0,\ \forall_{k \neq begin}$$

- $0 < i \leq n$

$$s_l(i+1) = e_l(x_{i+1}) \cdot \max_{k \in Q} [\, s_k(i) \cdot a_{kl} \,]$$

- $i = n+1$

$$P(x \mid \pi^*) = \max_{k \in Q} s_k(n)\, a_{k,\,end}$$

Viterbi Algorithm

- Dynamic Programming
  - with log-score function

$$S_l(i) = \log s_l(i)$$

  - Space complexity = $O(n\,|Q|)$
  - Time complexity = $O(n\,|Q|)$
  - $S_l(i+1) = \log e_l(x_{i+1}) + \max_{k \in Q} [\, S_k(i) + \log a_{kl} \,]$

Estimating the i^{th} State

- $P(\pi_i = k \mid x) =$
  - Given a sequence $x \in \Sigma^*$, the probability that the HMM was in state k at instant i.

Forward Estimate:

- $f_k(i) = P(x_1 \cdots x_i, \pi_i = k) =$
  - Probability of emitting the prefix $x_1 \cdots x_i$ and reaching the state $\pi_i = k$
- $f_k(i) = e_k(x_i) \cdot \sum_{l \in Q} f_l(i-1)\, a_{lk}$

Backward Estimate:

- $b_k(i) = P(x_{i+1} \cdots x_n, \pi_i = k) =$
  - Probability of being at the state $\pi_i = k$ and emitting the suffix $x_{i+1} \cdots x_n$.
- $b_k(i) = \sum_{l \in Q} e_k(x_{i+1}) \cdot b_l(i+1) a_{kl}$

Applying Bayes' Rule

- $P(\pi_i = k \mid x) = (1/P(x))$
  $\cdot P(x_1 \ldots x_i, \pi_i = k)$
  $\cdot P(x_{i+1} \ldots x_n, \pi_i = k)$
  $= [f_k(i)\, b_k(i)] / P(x)$
- $P(x) = \sum_{\pi} P(x \mid \pi)$

Sequence Motifs

- A Sequence of patterns of biological significance.
- Examples:
  - DNA: Protein binding sites
    - (e.g. promoters, regulatory sequences)
  - Protein: sequences corresponding to conserved pieces of structure
    - (e.g. Local features, At various scales: blocks, domains & families)

MEME Algorithm

- Uses EM (Expectation Minimization) algorithm to find multiple motifs in a set of sequences.
- Description of a motif:
  - W = (Fixed) width of a motif
  - $P = (p_{lc})_{l \in \Sigma, c \in 1..W}$
    = Matrix of probabilities that letter l occurs at position c
    = $|\Sigma| \times W$ matrix

Example

- DNA motif of width
  - W = 3,
  - Σ = { A, T, C, G}
- ρ = motif) W_ρ = 3,
- P_ρ = 4 £ 3 stochastic matrix

$$P_\rho = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{Bmatrix} 0.1 & 0.5 & 0.2 \\ 0.2 & 0.2 & 0.1 \\ 0.4 & 0.2 & 0.1 \\ 0.3 & 0.1 & 0.6 \end{Bmatrix} \end{array}$$

Computational Problem

- Given:
  - A set of sequences, G
  - A width parameter W
- Find:
  - Motifs of width W common to sequences G and present their probabilistic descriptions.
  - Note that motif start sites in each sequence are unknown (hidden).

Basic EM Approach

$$Z = \begin{array}{c} \\ seq1 \\ seq2 \\ seq3 \\ seq4 \end{array} \begin{array}{c} [1] \quad [2] \quad [3] \quad [4] \\ \begin{pmatrix} 0.1 & 0.1 & 0.2 & 0.6 \\ 0.4 & 0.2 & 0.1 & 0.3 \\ 0.3 & 0.1 & 0.5 & 0.1 \\ 0.1 & 0.5 & 0.1 & 0.3 \end{pmatrix} \end{array}$$

- Γ =
  - Training sequences.
- $|\Gamma|$ =
  - Total number of sequences = m;
  - Minimum length of a sequence in Γ = l.
- **Z = m £ l matrix of probabilities**
  - z_{ij} = Probability that the motif starts in position j in sequence i.

EM Algorithm

- Set initial values for P
- do
  - Re-estimate Z from P
  - Re-estimate P from Z
- until change in P < ε
- return P

EM Algorithm

- Maximize the likelihood of a motif in the training sequence:
  - $S_i \in \Gamma$ i^{th} sequence
  - I_{ij} = {1, if motif starts at posn. j in seq. i
    {0, otherwise.
  - l_k = the char. at posn. j+k-1 in seq. S_i
- $Pr(S_i \mid I_{ij} = 1, \rho) = \prod_{k=1}^{W} \rho_{l\{k,k\}}$.

Example

S_i = AGGCTGTAGACAC

$P_\rho =$

	1	2	3
A	0.1	0.5	0.2
T	0.2	0.2	0.1
C	0.4	0.2	0.1
G	0.3	0.1	0.6

- $Pr(S_i = TGT \mid I_{i5} = 1, \rho)$
  $= \rho_{T,1} \times \rho_{G,2} \times \rho_{T,3}$
  $= 0.2 \times 0.1 \times 0.1$
  $= 2 \times 10^{-3}$

Estimating Z

- $z_{ij} = Pr(I_{ij} = 1 \mid \rho, S_i)$
  = Estimates the starting position in S_i **2** Γ.
- $z_{ij} = Pr(I_{ij} = 1 \mid \rho, S_i)$
  $= Pr(S_i, I_{ij} = 1 \mid \rho) / Pr(S_i \mid \rho)$
  $= Pr(S_i \mid I_{ij} = 1, \rho)\, Pr(I_{ij} = 1) / \sum_k Pr(S_i \mid I_{ik} = 1, \rho)\, Pr(I_{ik} = 1)$
  $= Pr(S_i \mid I_{ij} = 1, \rho) / \sum_k Pr(S_i \mid I_{ik} = 1, \rho)$
- Follows from an application of the Bayes' rule and the assumption that "it is equally likely that the motif will start in any position."

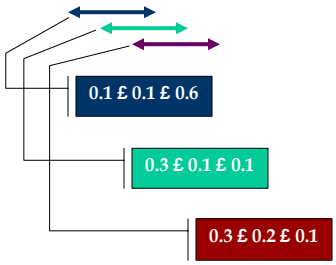
$$\forall_{jk}\ Pr(I_{ij} = 1) = Pr(I_{ik} = 1)$$

Example

S_i = AGGCTGTAGACAC

$P_\rho =$

	1	2	3
A	0.1	0.5	0.2
T	0.2	0.2	0.1
C	0.4	0.2	0.1
G	0.3	0.1	0.6



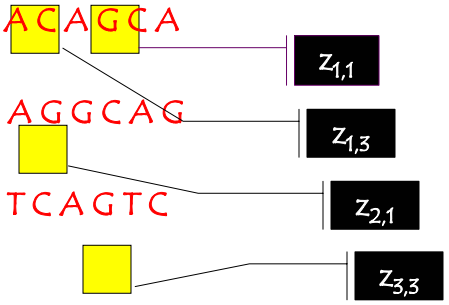
$z'_{i1} = 6\ £\ 10^{-3}$
$z'_{i2} = 3\ £\ 10^{-3}$
$z'_{i3} = 6\ £\ 10^{-3}$
...
NORMALIZE

Estimating P_ρ

- Given **Z**, estimate the probability that the character c occurs at the k^{th} position of a motif.
- $n_{ck} = \Sigma_{s(i) \in \Gamma; \{j \mid s(i, j+k-1) = c\}} z_{ij}$
- Expected number of occurrences of the character c at the k^{th} position of a motif ρ (assuming that the motif "start position" is known.)
- $p_{ck} = (n_{ck} + 1) / \Sigma_d (n_{dk} + 1)$

Example

- s_1 : A C A G C A
- s_2 : A G G C A G
- s_3 : T C A G T C



$$p_{A,1} = (z_{11} + z_{13} + z_{21} + z_{33} + 1) / (z_{11} + z_{12} + \cdots + z_{33} + z_{34} + 4)$$

Meme

- Uses the basic EM approach
  - Try many starting points.
  - Allow multiple occurrences of a motif per sequence
  - Allows multiple motifs to be learned simultaneously.

Meme

- Initial set of possible motifs:
  - Take every distinct subsequences of length W in the training set
  - Derive an initial matrix P
  - $p_{ck} = \{ \alpha$ if c occurs in position k in the subsequence

    $\{(1-\alpha)/(|\Sigma| - 1)$ Otherwise

Example

- W = 3,
  ρ = T A T,
  α = 0.5
  - Choose the motif model with the highest likelihood.
  - Run EM to convergence

$P_\rho =$

	1	2	3
A	0.17	0.50	0.17
T	0.17	0.17	0.17
C	0.17	0.17	0.17
G	0.50	0.17	0.50