



Computational Biology Lecture #7: Probabilistic Analysis

Bud Mishra
Professor of Computer Science, Mathematics, & Cell Biology
Oct 31 2005

10/30/2005

© Bud Mishra, 2005

L7-1



Bioinformatics Databases of Interest

10/30/2005

© Bud Mishra, 2005

L7-2



Bioinformatics DataSources

- ◇ Database interfaces
 - Genbank/EMBL/DDBJ, Medline, SwissProt, PDB, ...
- ◇ Sequence alignment
 - BLAST, FASTA
- ◇ Multiple sequence alignment
 - Clustal, MultAlin, DiAlign
- ◇ Gene finding
 - Genscan, GenomeScan, GeneMark, GRAIL
- ◇ Protein Domain analysis and identification
 - pfam, BLOCKS, ProDom,
- ◇ Pattern Identification/
- ◇ Characterization
 - Gibbs Sampler, AlignACE, MEME
- ◇ Protein Folding prediction
 - PredictProtein, SwissModeler

10/30/2005

© Bud Mishra, 2005

L7-3



Five Important Websites

- ◇ NCBI (The National Center for Biotechnology Information;
<http://www.ncbi.nlm.nih.gov/>)
- ◇ EBI (The European Bioinformatics Institute)
 - <http://www.ebi.ac.uk/>
- ◇ The Canadian Bioinformatics Resource
 - <http://www.cbr.nrc.ca/>
- ◇ SwissProt/ExPASy (Swiss Bioinformatics Resource)
 - <http://expasy.cbr.nrc.ca/sprot/>
- ◇ PDB (The Protein Databank)
 - <http://www.rcsb.org/PDB/>

10/30/2005

© Bud Mishra, 2005

L7-4



NCBI (<http://www.ncbi.nlm.nih.gov/>)

- ◇ Entrez interface to databases
 - Medline/OMIM
 - Genbank/Genpept/Structures
- ◇ BLAST server(s)
 - Five-plus flavors of blast
- ◇ Draft Human Genome
- ◇ Much, much more...

10/30/2005

© Bud Mishra, 2005

L7-5



EBI (<http://www.ebi.ac.uk/>)

- ◇ SRS database interface
 - EMBL, SwissProt, and many more
- ◇ Many server-based tools
 - ClustalW, DALI, ...

10/30/2005

© Bud Mishra, 2005

L7-6



SwissProt

(<http://expasy.cbr.nrc.ca/sprot/>)

- ◇ Curation...
 - Error rate in the information is greatly reduced in comparison to most other databases.
- ◇ Extensive cross-linking to other data sources
- ◇ SwissProt is the 'gold-standard' by which other databases can be measured, and is the best place to start if you have a specific protein to investigate

10/30/2005

© Bud Mishra, 2005

L7-7



A few more resources

- ◇ Human Genome Working Draft
<http://genome.ucsc.edu/>
- ◇ TIGR (The Institute for Genomics Research)
<http://www.tigr.org/>
- ◇ Celera
<http://www.celera.com/>
- ◇ (Model) Organism specific information:
 - Yeast: <http://genome-www.stanford.edu/Saccharomyces/>
 - Arabidopsis: <http://www.tair.org/>
 - Mouse: <http://www.jax.org/>
 - Fruitfly: <http://www.fruitfly.org/>
 - Nematode: <http://www.wormbase.org/>
- ◇ Nucleic Acids Research Database Issue
<http://nar.oupjournals.org/>

10/30/2005

© Bud Mishra, 2005

L7-8



Example 1:

- ◇ Searching a new genome for a specific protein
- ◇ Specific problem:
 - We want to find the closest match in *C. elegans* of *D. melanogaster* protein NTF1, a transcription factor
- ◇ First- understanding the different forms of blast

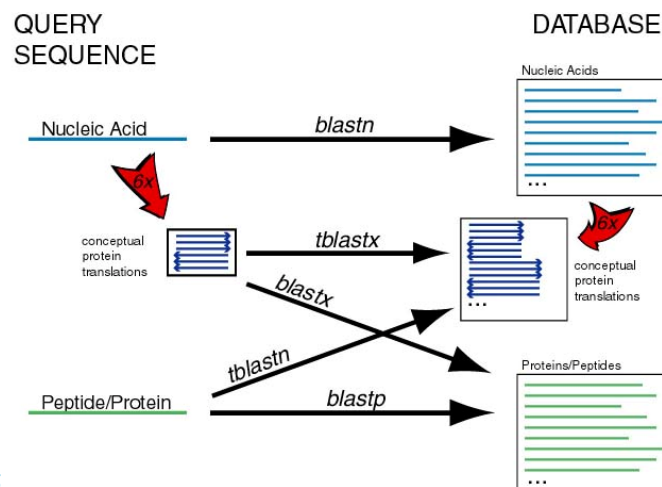
10/30/2005

© Bud Mishra, 2005

L7-9



The different versions of BLAST



10/30/2005

© Bud Mishra, 2005



Some possible methods

- ◇ If the domain is a known domain:
- ◇ SwissProt
 - text search capabilities
 - good annotation of known domains
 - crosslinks to other databases (domains)
- ◇ Databases of known domains:
 - BLOCKS (<http://blocks.fhcrc.org/>)
 - Pfam (<http://pfam.wustl.edu/>)
 - Others (ProDom, ProSite, DOMO,...)

10/30/2005

© Bud Mishra, 2005

L7-11



Nature of conservation in a domain

- ◇ For new domains, multiple alignment is your best option
 - Global: clustalw
 - Local: DiAlign
 - Hidden Markov Model: HMMER
- ◇ For known domains, this work has largely been done for you
 - BLOCKS
 - Pfam

10/30/2005

© Bud Mishra, 2005

L7-12



Protein Tools

◇ Search/Analysis tools

- Pfam
- BLOCKS
- PredictProtein

(<http://cubic.bioc.columbia.edu/predictprotein/predictprotein.html>)

10/30/2005

© Bud Mishra, 2005

L7-13



Different representations of conserved domains

◇ BLOCKS

- Gapless regions
- Often multiple blocks for one domain

◇ PFAM

- Statistical model, based on HMM
- Since gaps are allowed, most domains have only one pfam model

10/30/2005

© Bud Mishra, 2005

L7-14



Bayesian Probabilities

10/30/2005

© Bud Mishra, 2005

L7-15



Probabilities Overview

◇ **Ensemble:**

- 'X' is a random variable x with a set possible outcomes $A_x = \{a_1, a_2, \dots, a_i, \dots, a_l\}$, having probabilities $\{p_1, p_2, \dots, p_i, \dots, p_l\}$. $p_i \geq 0$ and $\sum_x p_x = 1$.

◇ **Joint Ensemble**

- 'XY' is an ensemble with ordered outcomes x and y .
- $x \in A_x = \{a_1, a_2, \dots, a_i, \dots, a_l\}$, and
- $y \in A_y = \{b_1, b_2, \dots, b_j, \dots, b_k\}$.

10/30/2005

© Bud Mishra, 2005

L7-16



Marginal & Conditional Probabilities

◇ **Product Rule:**

- $P(x,y | H) = P(x | y, H) P(y | H)$

◇ **Sum Rule:**

- $P(x | H) = \sum_y P(x,y | H) = \sum_y P(x | y, H) P(y | H)$

◇ **Bayes' Rule:**

- $P(y|x, H) = P(x | y, H) P(y | H) / P(x | H)$

- $P(y | x, H)$

- $= P(x | y, H) P(y | H) / \sum_{y'} P(x | y', H) P(y' | H)$

10/30/2005

© Bud Mishra, 2005

L7-17



Bayesian Interpretation

◇ **Probability $P(e)$**

- \mapsto our uncertainty about whether e is true or false in the real world

- (given whatever information we have available)

◇ **"Degree of Belief"**

◇ **More rigorously, we should write**

- conditional probability $P(e | L) \mapsto$ represents degree of belief, where L is the background information on which our belief is based

10/30/2005

© Bud Mishra, 2005

L7-18



Probability as a Dynamic Entity

- ◇ "degree of belief"
 - Update the "degree of belief" as more data arrives:
- ◇ Bayes Theorem: $P(e | \mathbf{D}) = P(\mathbf{D} | e) P(e) / P(\mathbf{D})$
- ◇ Posterior is proportional to the prior.

10/30/2005

© Bud Mishra, 2005

L7-19



Probability as a Dynamic Entity

- ◇ Bayes Theorem: $P(e | \mathbf{D}) = P(\mathbf{D} | e) P(e) / P(\mathbf{D})$
- ◇ *Prior Probability:*
 - $P(e)$ is your belief in the event e before you see any data at all
- ◇ *Posterior:*
 - $P(e | \mathbf{D})$ is the updated posterior belief in e given the observed data.
- ◇ *Likelihood:*
 - $P(\mathbf{D} | e) \mapsto$ probability of the data under the assumption e .

10/30/2005

© Bud Mishra, 2005

L7-20



Dynamics

- $P(e | D_1, D_2) = P(D_2 | e, D_1) P(e | D_1) / P(D_2 | D_1)$
- ◇ *Important Observation:*
 - The effects of prior diminish as the number of data points increases.
- ◇ *The Law of Large Number:*
 - With large number of data points, Bayesian and frequentist viewpoints become indistinguishable.

10/30/2005

© Bud Mishra, 2005

L7-21



Parameter Estimation

- ◇ *Functional form for a model M*
 - Depends on parameters Θ
 - Best estimation for Θ ?
- ◇ *Typically our parameters Θ are a set of real-valued numbers*
 - Both prior $P(\Theta)$ and the posterior $P(\Theta | D)$ are defining probability density functions

10/30/2005

© Bud Mishra, 2005

L7-22



Maximum A Posteriori (MAP)

- ◊ Find the set of parameters Θ
 - maximizing the posterior $P(\Theta | \mathbf{D})$ or minimizing a score $-\log P(\Theta | \mathbf{D})$
 - $E'(\Theta) = -\log P(\Theta | \mathbf{D})$
 $= -\log P(\mathbf{D} | \Theta) - \log P(\Theta) + \log P(\mathbf{D})$
 - Same as minimizing $E(\Theta) = -\log P(\mathbf{D} | \Theta) - \log P(\Theta)$
 - If the prior $P(\Theta)$ is uniform over the entire parameter space (uninformative):
Minimize $E_L(\Theta) = -\log P(\mathbf{D} | \Theta)$
 - *Maximum likelihood solution*

10/30/2005

© Bud Mishra, 2005

L7-23



Information Theory

10/30/2005

© Bud Mishra, 2005

L7-24



Entropy

- ◇ **X = r.v.; Entropy of X**
 - $H(X) = \sum_x P(x) \log(1/P(x)) = E_x [-\log P(x)]$
- ◇ **Entropy measures the information content or "uncertainty" of x**
 - $0 \leq H(X) \leq \log(|X|)$.
 - $H(X) = 0$, if $\exists x, P(x) = 1$; It's minimal if the probability is concentrated at one value (no uncertainty)
 - $H(X) = \log(|X|)$, if $\forall x, P(x) = 1/|X|$; It's maximal if the probability is distributed uniformly (complete uncertainty)

10/30/2005

© Bud Mishra, 2005

L7-25



Joint Entropy

- ◇ **Joint entropy of X, Y:**
 - $H(X, Y) = \sum_{x,y \in A_x \times A_y} P(x, y) \log(1/P(x, y))$
 - Entropy is additive for independent r.v.'s.
 - $H(X, Y) = H(X) + H(Y)$ iff $P(x, y) = P(x) P(y)$.
- ◇ **Conditional Entropy of X given Y:**
 - $H(X|Y) = \sum_{x \in A_x} P(x|y) \log(1/P(x|y))$
 - $H(X|Y) = E_y H(X|y)$
 $= \sum_y P(y) \sum_x P(x|y) \log(1/P(x|y))$
 $= \sum_{x,y} P(x, y) \log(1/P(x|y))$

10/30/2005

© Bud Mishra, 2005

L7-26



Chain Rule

- ◇ Chain Rule for Entropy
 - $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- ◇ Mutual Information
 - ◇ It measures the average reduction in uncertainty about x that results from learning y or vice versa.
 - $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
 $= H(X) + H(Y) - H(X,Y)$
 - $I(X;Y) = \sum_{x,y} P(x,y) \log [P(x,y)/P(x)P(y)]$
 - ◇ Properties:
 - $I(X;Y) = I(Y;X); I(X;Y) \geq 0$

10/30/2005

© Bud Mishra, 2005

L7-27



Distance

- ◇ Distance between two r.v.'s:
 - $D(X, Y) = H(X,Y) - I(X;Y)$
 $= 2 H(X,Y) - H(X) - H(Y)$
 - $D(X,Y) \geq 0$.
- ◇ Idempotent:
 - $D(X,X) = 0$.
- ◇ Symmetry:
 - $D(X,Y) = D(Y,X)$
- ◇ Triangle Inequality:
 - $D(X,Z) \leq D(X,Y) + D(Y, Z)$

10/30/2005

© Bud Mishra, 2005

L7-28



Data Processing Inequality

- ◇ **Markov Chain**
 - $X \rightarrow Y \rightarrow Z$
 - $P(z | x, y) = P(z | y)$
 - OR $P(x, y, z) = P(x, y) P(z | x, y) = P(x) P(x | y) P(z | y)$
- ◇ **Then $I(X; Y) \geq I(X; Z)$**
- ◇ **Corollary:**
 - $I(X; Y) \geq I(X; g(Y))$

10/30/2005

© Bud Mishra, 2005

L7-29



KL Distance

- ◇ **Kullback-Leibler (KL) Distance (Relative Entropy):**
 - Given two probability distributions $p(x)$ and $q(x)$ [defined over the same $x \in \mathcal{A}_x$]
 - $D_{KL}(p||q) = \mathbf{E}_x \log p(x)/q(x)$
 $= \sum_x p(x) \log p(x)/q(x)$
- ◇ **Properties:**
 - Gibb's Inequality: $D_{KL}(p||q) \geq 0$
 - $D_{KL}(p||q) \neq D_{KL}(q||p)$

10/30/2005

© Bud Mishra, 2005

L7-30



What Good is Information Theory

- ◇ **Family of genes:**
 - Can genes be grouped to explain how they work together? Information Compression.
- ◇ **Relation between groups of genes and their effect on the traits:**
 - How does a group of genes code the information about a complex trait?
 - Which gene affects the trait more directly than another gene?

10/30/2005

© Bud Mishra, 2005

L7-31



Rate Distortion Theories Kolmogorov-Shannon Theorem

(Also called "information bottleneck.")

10/30/2005

© Bud Mishra, 2005

L7-32



Strangeness of RDT

- ◇ **An intriguing aspect of Rate Distortion Theory:**
 - Joint descriptions are more efficient than individual descriptions.
 - This is true even for independent random variables.
 - It is simpler to describe an elephant and a penguin with one description than to describe each alone.

10/30/2005

© Bud Mishra, 2005

L7-33



Rate Distortion Theorem

- ◇ **Due to Kolmogorov & Shannon:**
 - X = Dictionary (think of all the genes)
 - \mathbf{X} = Codebook
(think of families of coregulated genes)
 - Rate = $I(X; \mathbf{X}) = H(X) - H(X, \mathbf{X})$
 $= \int \int p(x, x') \log [p(x, x') / p(x) p(x')] dx dx'$
 - Distortion = $\mathbf{h} \delta(X, \mathbf{X})$
 $= \int \int p(x, x') \delta(x, x') dx dx'$

10/30/2005

© Bud Mishra, 2005

L7-34



Succinct Theory

- We want highest rate (maximum compression) with least amount of distortion:
- ◇ Optimization Problem:
 - Min $I(X; \mathbf{X})$
 - Subject to $\mathbf{h} \delta(X, \mathbf{X}) \mathbf{i} \cdot D$
- ◇ Lagrangian of a Constrained Optimization Problem

$$F[p(x | x'), \beta] = I(X; \mathbf{X}) + \beta \mathbf{h} \delta(X, \mathbf{X}) \mathbf{i}$$

10/30/2005

© Bud Mishra, 2005

L7-35



Solution to Lagrangian

- ◇ The variational problem is solved at:

$$p(x | x') = [1/Z(x, \beta)] p(x') \text{Exp}[-\beta \delta(x, x')]$$
 - ◇ In other words:

$$p(x, x')/p(x) p(x') / \text{Exp}[-\beta \delta(x, x')]$$
 - ◇ Thus,

$$I(X, \mathbf{X}) = \mathbf{s}_{x, x} p(x, x') [-\beta \delta(x, x')] dx dx' = -\beta \mathbf{h} \delta(X, \mathbf{X}) \mathbf{i}$$
- $$F[p(x | x'), \beta] = I(X; \mathbf{X}) + \beta \mathbf{h} \delta(X, \mathbf{X}) \mathbf{i} = 0$$

10/30/2005

© Bud Mishra, 2005

L7-36



Blahut-Arimoto Algorithm

Fixed point:

- $p(x') = \sum_x p(x, x') = \sum_x p(x) p(x' | x)$
- $p(x' | x) = p(x') \text{Exp}[-\beta \delta(x, x')] / Z(x, \beta)$
- $w(x, x') = p(x, x') / p(x) p(x') = \text{Exp}[-\beta \delta(x, x')] / Z(x, \beta)$

Computation:

- Start with some K randomly chosen code words $\mapsto X$;
 $\sum_{x' \in X} p_0(x') = 1/K$
- $p_{t+1}(x' | x) = p_t(x') \text{Exp}[-\beta \delta(x, x')] / Z_t(x, \beta)$
- Choose new code words: $\sum_x x p_{t+1}(x' | x)$
- Thus, $p_{t+1}(x') = \sum_x p(x) p_{t+1}(x' | x)$

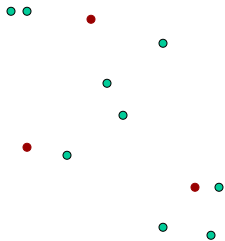
10/30/2005

© Bud Mishra, 2005

L7-37



Clustering



- ◇ Set of green points on the plane... g_1, g_2, \dots, g_9 . We wish to encode them succinctly with three brown points... b_1, b_2, b_3
- ◇ Choose three brown points at random.
- ◇ For each brown point, compute $p(b_i | g_j)$ depending on the current distance between b_i & g_j

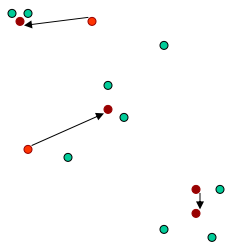
10/30/2005

© Bud Mishra, 2005

L7-38



Clustering



- Recompute the new positions of b_1, b_2, b_3 :
- ◇ **New $(b_i) = \sum g_j p(\text{Old}(b_i) | g_j)$**
 - Weighted Centroids of green points g_j 's
- ◇ **REPEAT**
- ◇ **UNTIL the brown points do not move:**
- ◇ **Soft_k_means Clustering**

10/30/2005

© Bud Mishra, 2005

L7-39



A "Harder" Version

- ◇ Choose K "centroid" positions
 - b_1, b_2, \dots, b_K
- ◇ Using b_i 's partition the green points g_1, g_2, \dots, g_n into K classes:
 - $G_i = \{ g_j \text{ closer to } b_i \text{ than any other } b_j \}$
- ◇ Update K centroid positions:
 - $\text{New}(b_i) = \text{Centroid of } G_i$

10/30/2005

© Bud Mishra, 2005

L7-40



Information Bottleneck

- ◇ Markov Chain: $\mathbf{X} \rightarrow \mathbf{X}' \rightarrow \mathbf{Y}$
- ◇ Measure Distortion by KL (Kullback-Leibler) distance:
 - $D_{KL}(p(y|x) \parallel p(y|x')) = \delta_y(x, x')$
- ◇ Minimize rate without "much" KL-distortion:
- ◇ Optimization Problem
 - Min $I(\mathbf{X}; \mathbf{X}')$
 - Subject to $\mathbf{h} \delta_y(x, x') \mathbf{i} \cdot D$
- ◇ Lagrangian
 - $F(p(x|x') \parallel p(y|x')) = I(\mathbf{X}; \mathbf{X}') + \beta \mathbf{h} D_{KL}(p(y|x) \parallel p(y|x')) \mathbf{i}$

10/30/2005

© Bud Mishra, 2005

L7-41



Fixed Point Solution

- ◇ $p(x, x') / p(x)p(x') = \text{Exp}[-\beta \delta_y(x, x')] / Z(x, \beta)$
- ◇ $\delta_y(x, x') = D_{KL}(p(y|x) \parallel p(y|x'))$
 - $p(x'|x) = p(x') \text{Exp}[-\beta \delta_y(x, x')] / Z(x, \beta)$
 - $p(x') = \sum_x p(x'|x)p(x)$
 - $p(x|x') = p(x'|x)p(x) / p(x')$
 - $p(y|x') = \sum_{y,x} p(y|x, x') p(x|x')$
 $= \sum_{y,x} p(y|x) p(x|x')$
 - $\delta_y(x, x') = D_{KL}(p(y|x) \parallel p(y|x'))$

10/30/2005

© Bud Mishra, 2005

L7-42



Blahut-Arimoto Algorithm

- $p_{t+1}(x' | x) = p_t(x') \text{Exp}[-\beta \delta_y(x, x')] / Z(x, \beta)$
- $p_{t+1}(x') = \sum_x p_{t+1}(x' | x) p(x)$
- $p_{t+1}(x | x') = p_{t+1}(x' | x) p(x) / p_{t+1}(x')$
- $p_{t+1}(y | x') = \sum_{y,x} p(y | x) p_{t+1}(x | x')$
- $\delta_{y,t+1}(x, x') = D_{KL}(p(y|x) \parallel p_{t+1}(y|x'))$

10/30/2005

© Bud Mishra, 2005

L7-43



How Can this Help Us:

- ◇ Think of the Markov Chain: $X \rightarrow X \rightarrow Y$ as
- ◇ GeneFamilies ! GeneExpressions ! Pathophysiology
- In other words, we wish to cluster the genes so that they explain various aspects of the pathophysiology...
- You may take other metadata into account in this picture...
- **HOMEWORK:** Try to make these ideas less abstract!!!
- Translate the algorithm directly to our "CFS problem."

10/30/2005

© Bud Mishra, 2005

L7-44



GRAPHICAL MODELS

10/30/2005

© Bud Mishra, 2005

L7-45

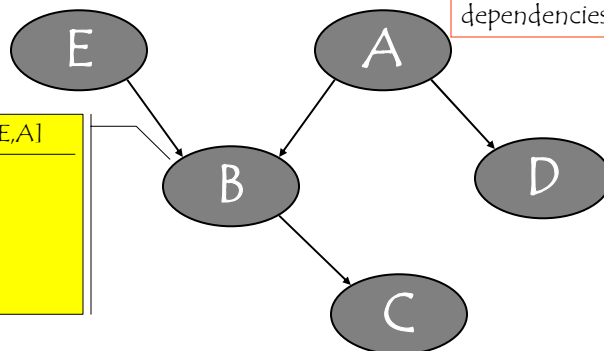


Bayesian Network: Example

Nodes represent gene activities

Edges represent dependencies

| E | A | $\Pr[B E,A]$ | $\Pr[C B,E,A]$ |
|---|---|--------------|----------------|
| 0 | 0 | 0.3 | 0.7 |
| 0 | 1 | 0.4 | 0.6 |
| 1 | 0 | 0.7 | 0.3 |
| 1 | 1 | 0.1 | 0.9 |



10/30/2005

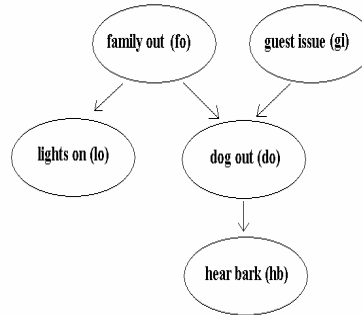
© Bud Mishra, 2005

L7-46



Bayesian networks (BN) in brief

- ◇ Graphs in which nodes represent random variables
- ◇ (Lack of) Arcs represent conditional independence assumptions
- ◇ Present & absent arcs provide compact representation of joint probability distributions
- ◇ BNs have complicated notion of independence, which takes into account the directionality of the arcs



10/30/2005

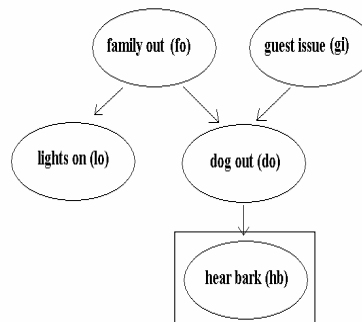
© Bud Mishra, 2005

L7-47



Bayesian network example

- ◇ $P(\text{hear your dog bark as you get home}) = P(\text{hb}) = ?$



10/30/2005

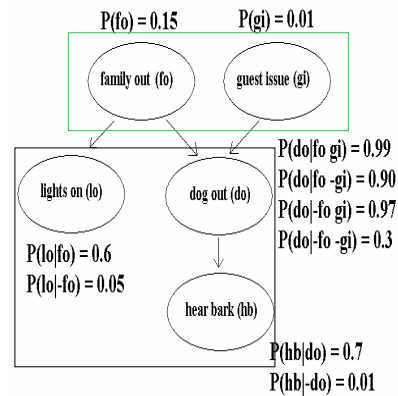
© Bud Mishra, 2005

L7-48



Belief Propagation

- Need prior P for root nodes and conditional Ps, that consider all possible values of parent nodes, for nonroot nodes



10/30/2005

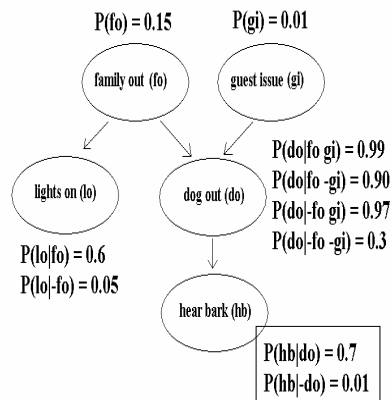
© Bud Mishra, 2005

L7-49



Major benefit of BN

- We can know $P(hb)$ based only on the conditional probabilities of hb and its parent node.
- We don't need to know/include all the ancestor probabilities between hb and the root nodes.



10/30/2005

© Bud Mishra, 2005

L7-50



Applications to Diverse Problems

PANG et al.: COMPUTERIZED TONGUE DIAGNOSIS BASED ON BAYESIAN NETWORKS

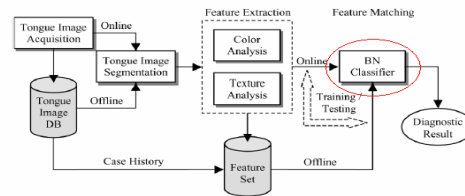


Fig. 1. The outline of the computerized tongue diagnosis system.

"Computerized tongue diagnosis based on Bayesian networks": devising expert system for Chinese medical method (supplementary reference 3)



Fig. 2. Four tongue image samples of patients suffering associated infections (top-left), liver cancer (top-right), appendicitis (bottom-left), and pneumonia (bottom-right).

10/30/2005

© Bud Mishra, 2005



Bayesian Networks

- ◇ Bayesian Network Model M consists of a set of random variables:
 - X_1, X_2, \dots, X_n
- ◇ and an underlying directed acyclic graph (DAG)
 - $G = (V, E)$
- ◇ such that each random variable is uniquely associated with a vertex of DAG

10/30/2005

© Bud Mishra, 2005

L7-52



Parameters

- ◇ The parameters Θ of the model are the numbers that specify the local conditional probability distributions
 - $P(X_i | X_{\text{pa}(i)})$, $1 \cdot 1 \cdot n$
 - where $X_{\text{pa}(i)}$ denotes the parent of node i in the graph
- ◇ Global probability distribution must equal the local conditional probability distributions:
 - $P(X_1, \dots, X_n) = \prod_i P(X_i | X_{\text{pa}(i)})$.
- ◇ Learning Bayesian network
 - Belief Propagation:
 - In general, NP-complete.

10/30/2005

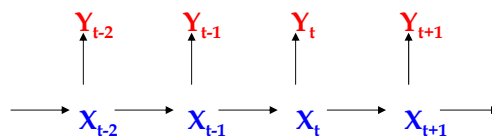
© Bud Mishra, 2005

L7-53



Markov Model

- ◇ Bayesian network structure for both
 - Hidden Markov Model
 - Kalman Filter 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



10/30/2005

© Bud Mishra, 2005

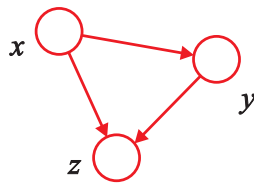
L7-54



Decomposition

- ◇ Consider an arbitrary joint distribution

- $p(x,y,z)$



- ◇ By successive application of the product rule

- $P(x,y,z) = p(x) p(y,z | x)$
 $= p(x) p(y|x) p(z | x, y)$

10/30/2005

© Bud Mishra, 2005

L7-55

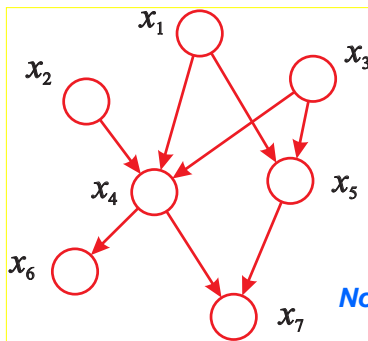


Directed Acyclic Graphs

- ◇ Joint distribution

- $P(x_1, \dots, x_D) = \prod_{i=1}^D p(x_i | pa_i)$
 where pa_i denotes the parents of i .

- ◇ $p(x_1, \dots, x_7) = p(x_1) p(x_2) p(x_3)$
 $p(x_4 | x_1, x_2, x_3) p(x_5 | x_1, x_3)$
 $p(x_6 | x_4) p(x_7 | x_4, x_5)$



No directed cycles

10/30/2005

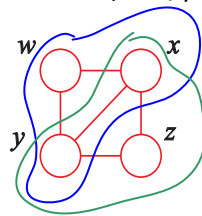
© Bud Mishra, 2005

L7-56



Undirected Graphs

- Provided $p(x) > 0$ then joint distribution is product of non-negative functions over the cliques of the graph
- $P(x) = (1/Z) \prod_C \psi_C(x_C)$
- Where $\psi_C(x_C)$ are the clique potentials, and Z is a normalization constant



$$p(w,x,y,z) = (1/Z) \psi_A(w,x,y) \psi_B(x,y,z)$$

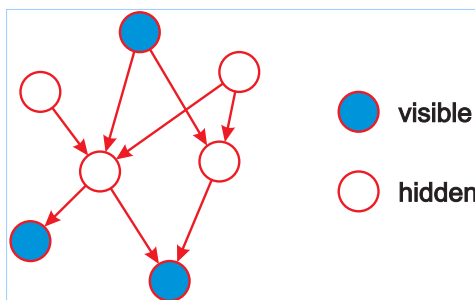
10/30/2005

© Bud Mishra, 2005

L7-57



Conditioning on Evidence



- Variables may be hidden (latent) or visible (observed)
- Latent variables may have a specific interpretation, or may be introduced to permit a richer class of distribution
- Recall HMM

10/30/2005

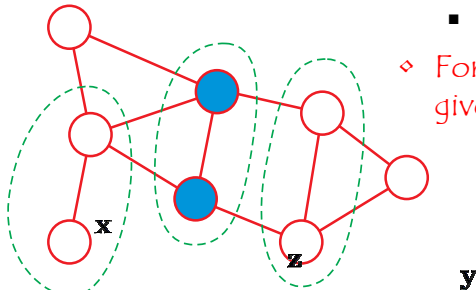
© Bud Mishra, 2005

L7-58



Conditional Independences

- ◇ x independent of y given z if, for all values of z ,
 - $(x \perp y \mid z) \iff p(x, y \mid z) = p(x \mid z) p(y \mid z)$
- ◇ For undirected graphs this is given by graph separation!



10/30/2005

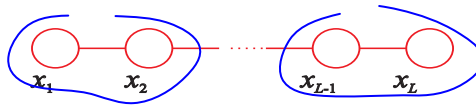
© Bud Mishra, 2005

L7-59



Message Passing

- ◇ Example



- ◇ Find marginal for a particular node

$$p(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_L} p(x_1, \dots, x_L)$$

- for M -state nodes, cost is $O(M^L)$
 - ❖ exponential in length of chain
 - ❖ but, we can exploit the graphical structure (conditional independences)

10/30/2005

© Bud Mishra, 2005

L7-60



Message Passing

- Joint distribution

$$p(x_1, \dots, x_L) = (1/Z) \psi(x_1, x_2) \dots \psi(x_{L-1}, x_L)$$

- Exchange sums and products

$$p(x_i) = (1/Z) \dots \sum_{x_2} \psi(x_2, x_3) [\sum_{x_1} \psi(x_1, x_2)] \dots \sum_{x_{L-1}} \psi(x_{L-2}, x_{L-1}) [\sum_{x_L} \psi(x_{L-1}, x_L)]$$

$m_\alpha(x_i)$ (above the first bracket) and $m_\beta(x_i)$ (below the second bracket)

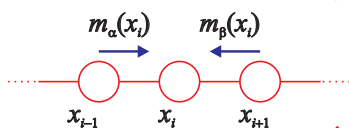
10/30/2005

© Bud Mishra, 2005

L7-61



Message Passing



- Express as product of messages

$$p(x_i) = (1/Z) m_\alpha(x_i) m_\beta(x_i)$$

- Recursive evaluation of messages

$$m_\alpha(x_i) = \sum_{x_{i-1}} \psi(x_{i-1}, x_i) m_\alpha(x_{i-1})$$

$$m_\beta(x_i) = \sum_{x_{i+1}} \psi(x_{i+1}, x_i) m_\beta(x_{i+1})$$

- Find **Z** by normalizing $p(x_i)$

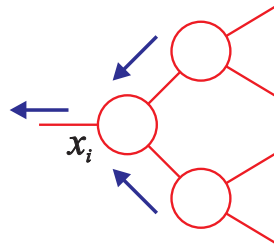
10/30/2005

© Bud Mishra, 2005

L7-62



Belief Propagation



- ◇ Extension to general tree-structured graphs
- ◇ At each node:
 - form product of *incoming* messages and local evidence
 - marginalize to give *outgoing* message
 - one message in each direction across every link
- ◇ Fails if there are loops

10/30/2005

© Bud Mishra, 2005

L7-63



Junction Tree Algorithm

- ◇ An efficient exact algorithm for a general graph
 - applies to both directed and undirected graphs
 - compile original graph into a tree of cliques
 - then perform message passing on this tree
- ◇ Problem:
 - cost is exponential in size of largest clique
 - many vision models have intractably large cliques

10/30/2005

© Bud Mishra, 2005

L7-64



Junction Tree

- ◇ **Marry parents:**
 - Add undirected edges to all co-parents which are not currently joined
- ◇ **Moralize**
 - Drop all directions in the graph. a moral graph
- ◇ **Triangulate the Moral Graph**
 - Add additional links so that there is no cycle of length 4 or more
- ◇ **Identify and Join Cliques to form the Junction Tree**
- ◇ **Perform Message passing on the Junction Tree**

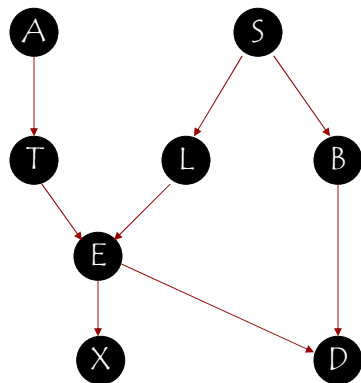
10/30/2005

© Bud Mishra, 2005

L7-65



Example



- **D**yspnoea (shortness of breath) may be due to **T**uberculosis, **L**ung cancer or **B**ronchitis, or none of them or more than one of them. A recent visit to **A**sia increases the chance of **T**uberculosis, while **S**moking is known to be a risk factor for both **L**ung cancer and **B**ronchitis. The results of a single **X**-ray do not discriminate between **L**ung cancer and **T**uberculosis, as neither does the presence or absence of **D**yspnoea.

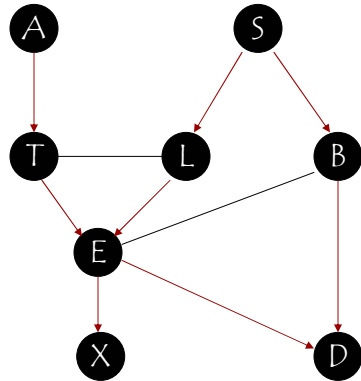
10/30/2005

© Bud Mishra, 2005

L7-66



Example



◇ Marry parents

- Connect T & L
- Connect E & B

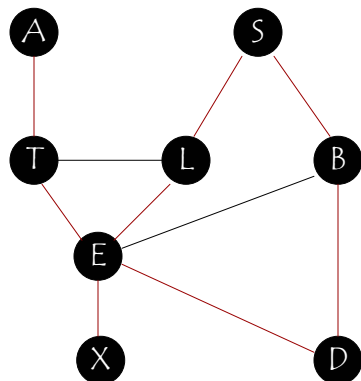
10/30/2005

© Bud Mishra, 2005

L7-67



Example



◇ Moralize

◇ Next

- ◇ Triangulate the Moral Graph
- ◇ Find Cliques
- ◇ Form the Junction Tree.

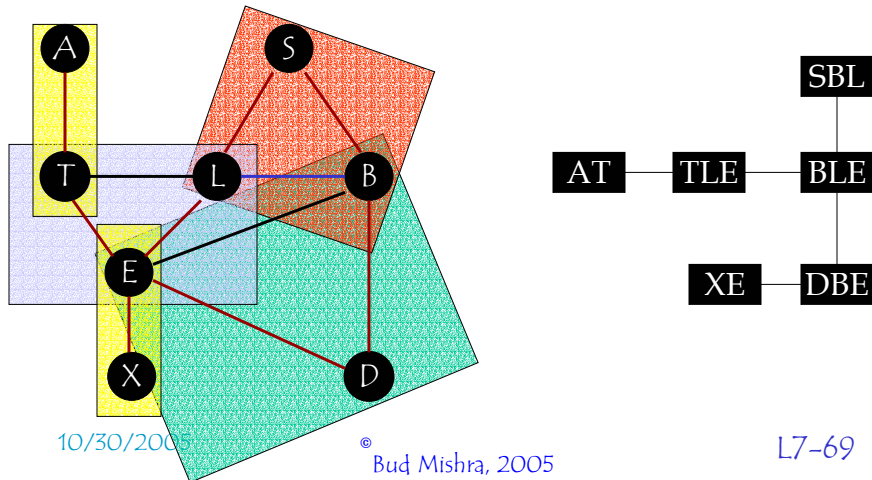
10/30/2005

© Bud Mishra, 2005

L7-68



Example



Loopy Belief Propagation

- ◇ Apply belief propagation directly to general graph
 - need to keep iterating
 - might not converge
- ◇ State-of-the-art performance in error-correcting codes

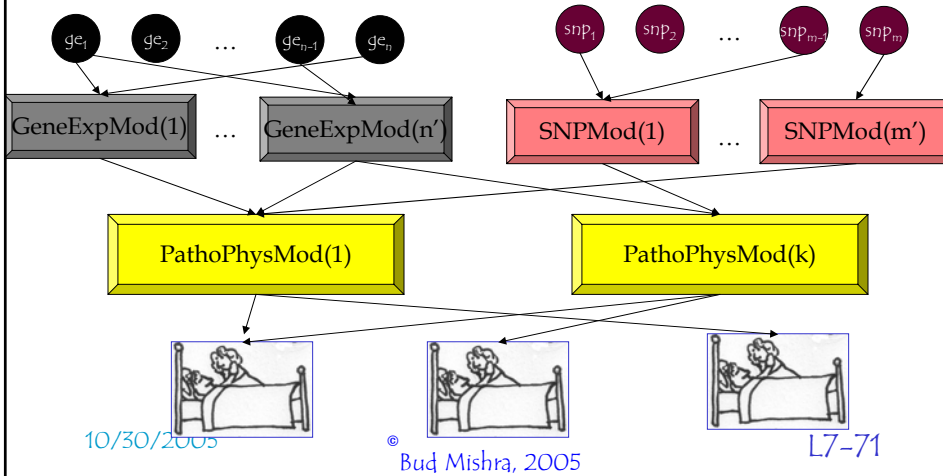
10/30/2005

© Bud Mishra, 2005

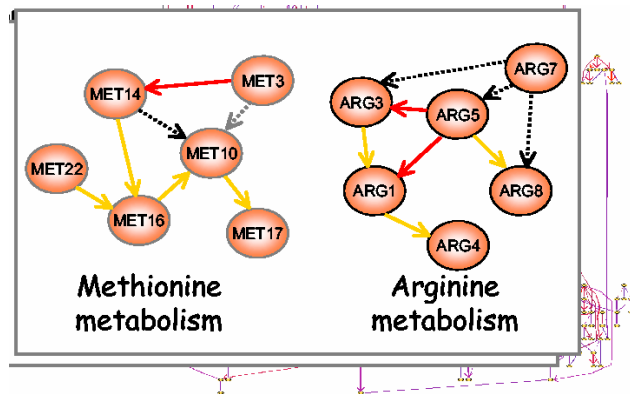
L7-70



Graphical Model for CFS



Gene Interaction Maps



10/30/2005

© Bud Mishra, 2005

[Pe'er et al, ISMB 2001]

L7-72



Graphical Models for Biology

- Rich modeling language for biological systems

- Based on probabilistic graphical models

- Classes of objects:

- Genes, experiments, tissues, patients

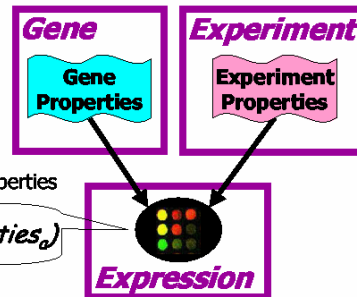
- Properties

- Observed: gene sequence, experiment conditions

- Hidden: gene function

- Interactions

- Expression level is function of gene and experiment properties



$$P(\text{Level}_{g,a} \mid \text{Properties}_g, \text{Properties}_a)$$

Segal *et al.* (ISMB 2001)

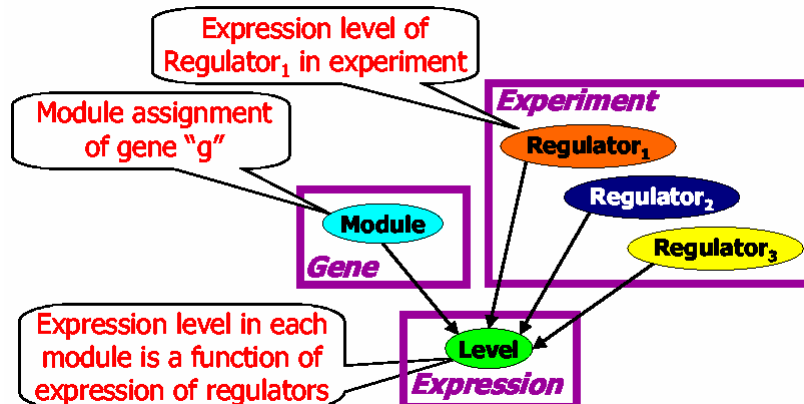
10/30/2005

© Bud Mishra, 2005

L7-73



Gene Regulation Model



10/30/2005

© Bud Mishra, 2005

L7-74



REDESCRIPTION

10/30/2005

© Bud Mishra, 2005

L7-75



What is redescription?

- ◇ **Shift of vocabulary**
 - from one language (descriptor family) to another to describe the same entity
 - Descriptor is any meaningful way of defining a subset within a universal set of entities
 - Set theoretic operations used on basic descriptors to define derived descriptors
- ◇ **Evaluated on the basis of Jaccard's coefficient**
 - $(A, B) = (A \cap B) / (A \cup B)$

10/30/2005

© Bud Mishra, 2005

L7-76



Examples of redescription

◇ Universal set: Countries of the world

- Countries with > 200 Nobel prize winners {USA}
 - , Countries with > 150 billionaires {USA}
- (Jaccard's = 1.0)

Universal set: Words in English language

- Words with 6 letters **AND NOT** Words with vowels {Rhythm, Syzygy}
 - , Words with 6 letters **AND** Words with 3 y's {Syzygy}
- (Jaccard's = 0.5)

10/30/2005

© Bud Mishra, 2005

L7-77



Why Redescribe?

◇ Advantages

- Allows feature construction
- Can handle any kind of data in terms of descriptors – no data specific mining required
- Can find commonalities and differences between various descriptors/descriptor families at the same time
- Can look for stories using a series of inexact redescriptions

10/30/2005

© Bud Mishra, 2005

L7-78



CARTwheels algorithm for redescription

$$\begin{aligned}
 X_1 &= \{ \quad \quad \quad o_2, o_3 \quad \quad \quad \} & Y_1 &= \{ o_1, o_2, \quad \quad \quad \} \\
 X_2 &= \{ \quad \quad \quad o_3, o_4 \quad \quad \quad \} & Y_2 &= \{ \quad \quad \quad o_2, o_3, o_4 \quad \quad \quad \} \\
 X_3 &= \{ \quad \quad \quad o_2, \quad \quad \quad o_4 \quad \quad \quad \} & Y_3 &= \{ \quad \quad \quad o_3, \quad \quad \quad o_5 \quad \quad \quad \} \\
 X_4 &= \{ o_1, \quad \quad \quad \quad \quad \quad o_5 \quad \quad \quad \} & Y_4 &= \{ o_1, o_2, \quad \quad \quad o_5 \quad \quad \quad \}
 \end{aligned}$$

Figure 1: Example data for illustrating operation of CARTwheels algorithm.

| object | Y ₁ | Y ₂ | Y ₃ | Y ₄ | class |
|----------------|----------------|----------------|----------------|----------------|----------------|
| o ₁ | √ | × | × | √ | X ₄ |
| o ₂ | √ | √ | × | √ | X ₁ |
| o ₃ | × | √ | √ | × | X ₁ |
| o ₄ | × | √ | × | × | X ₂ |
| o ₅ | × | × | √ | √ | X ₄ |

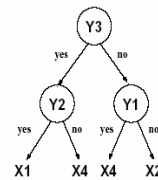


Figure 2: (left) Dataset to initialize CARTwheels algorithm. (right) induced classification tree.

10/30/2005

© Bud Mishra, 2005

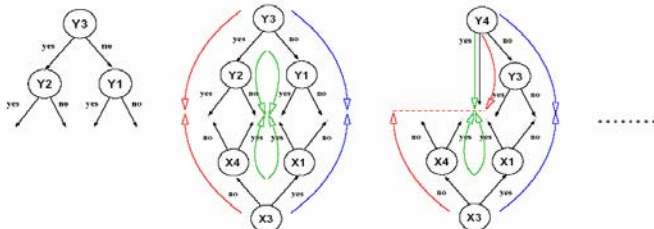
L7-79



CARTwheels algorithm for redescription (contd.)

| obj. | X ₁ | X ₂ | X ₃ | X ₄ | class | obj. | Y ₁ | Y ₂ | Y ₃ | Y ₄ | class |
|----------------|----------------|----------------|----------------|----------------|---|----------------|----------------|----------------|----------------|----------------|---|
| o ₁ | × | × | × | √ | (Y ₃ - Y ₂) ∪ (Y ₁ - Y ₃) | o ₁ | √ | × | × | √ | (X ₃ ∩ X ₁) ∪ (X ₄ - X ₃) |
| o ₂ | √ | × | √ | × | (Y ₃ - Y ₂) ∪ (Y ₁ - Y ₃) | o ₂ | √ | √ | × | √ | (X ₃ ∩ X ₁) ∪ (X ₄ - X ₃) |
| o ₃ | √ | √ | × | × | Y ₃ ∩ Y ₂ | o ₃ | × | √ | √ | × | (O - X ₃ - X ₄) |
| o ₄ | × | √ | √ | × | O - Y ₃ - Y ₁ | o ₄ | × | √ | × | × | (X ₃ - X ₁) |
| o ₅ | × | × | × | √ | (Y ₃ - Y ₂) ∪ (Y ₁ - Y ₃) | o ₅ | × | × | √ | √ | (X ₃ ∩ X ₁) ∪ (X ₄ - X ₃) |

Figure 3: (left) Dataset for second iteration of CARTwheels algorithm. Notice that class labels are now set-theoretic expressions involving Y_i's. (right) Dataset for third iteration of CARTwheels algorithm.



10/30/2005

© Bud Mishra, 2005

L7-80



Implementation details – descriptors used

- ◇ **Experimental (microarray) data**
 - for yeast from Gasch et al. Descriptors constructed of the form μ, \cdot
 - 9 different stress used from Gasch et al. data
 - GO category assignments for genes (biological process, cellular component, molecular function)

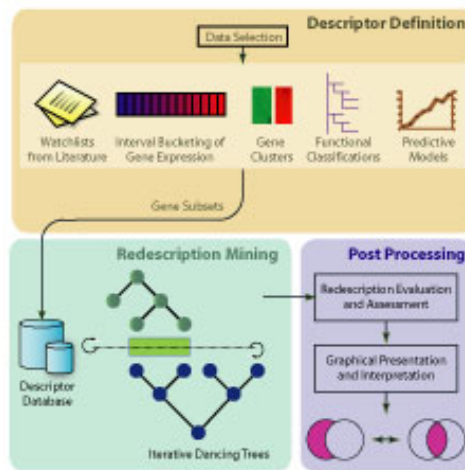
10/30/2005

© Bud Mishra, 2005

L7-81



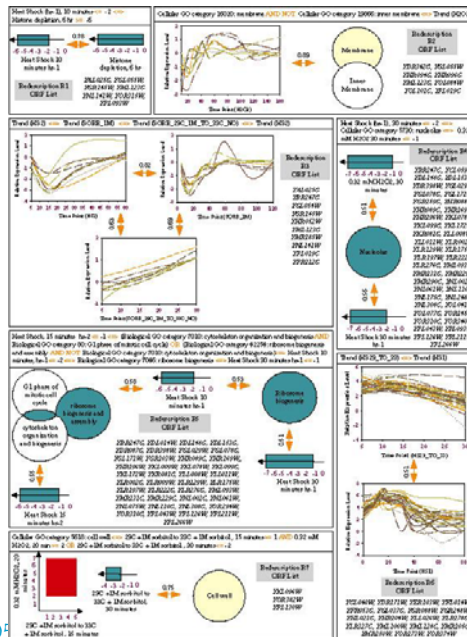
Design of System



10/30/2005

© Bud Mishra, 2005

L7-82



10/30/2005

Bud Mishra, 2005

L7-83



To be continued...

...

10/30/2005

© Bud Mishra, 2005

L7-84