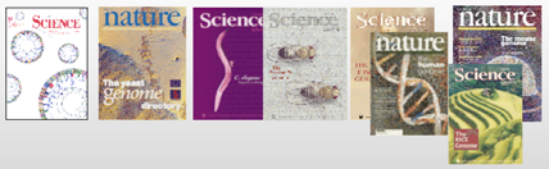
Graphical Models in Computational Molecular Biology

Nir Friedman
Hebrew University

Includes slides by:
Yoseph Barash, Nebojsa Jojic, Tommy Kaplan, Daphne Koller, Iftach Nachman, Dana Pe'er, Tal Pupko, Aviv Regev, Eran Segal

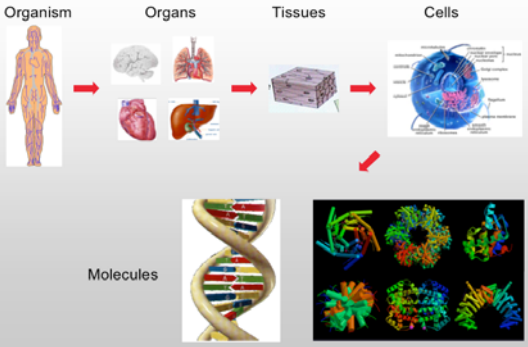
The Age of Genomes



Bacteria	Eukaryote	Animal	Human
1.6Mb	13Mb	100Mb	3Gb
1600 genes	~6000 genes	~20,000 genes	~30,000 genes?

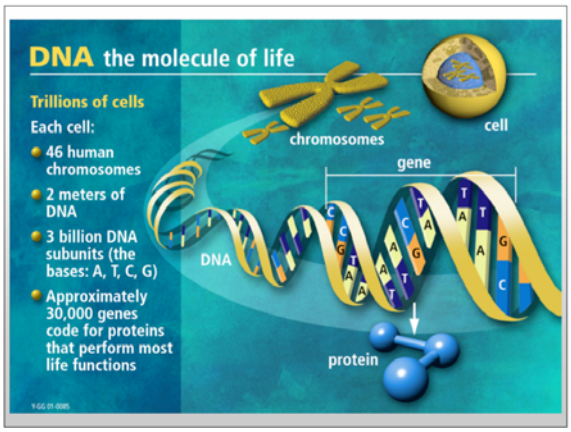
95 96 97 98 99 00 01 02 03

Biology → Molecular Biology

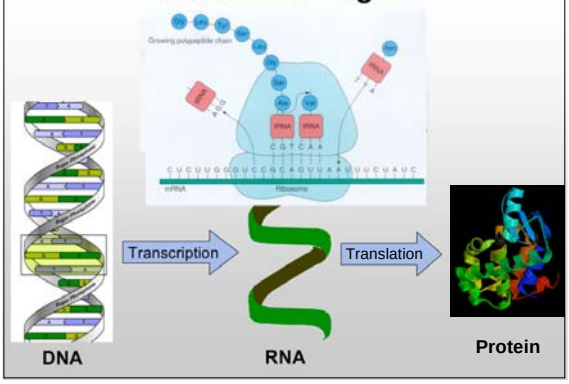


Macro Molecules as Sequences

Building blocks		Larger units
SUGARS	→	POLYSACCHARIDES
FATTY ACIDS	→	FATS, LIPIDS, MEMBRANES
AMINO ACIDS	→	PROTEINS
NUCLEOTIDES	→	NUCLEIC ACIDS (DNA, RNA)



The Central Dogma

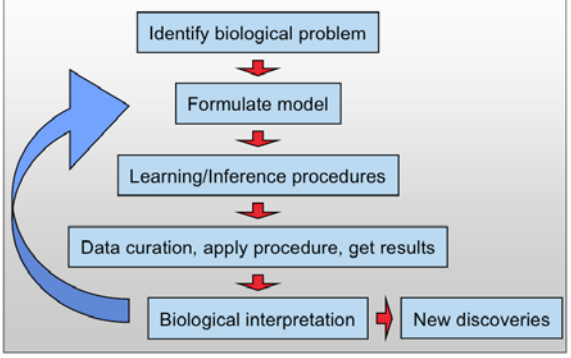


What Biologists Want?

- Identify Components
  - Find genes in DNA sequences
- Associate them with function
  - Understand what a gene does
- Interactions between components
  - Which genes work together?
- Dynamics of systems
  - When genes are activated, and by what?
- How did we get here
  - How did genes evolve, and why this way?

(And many other things)

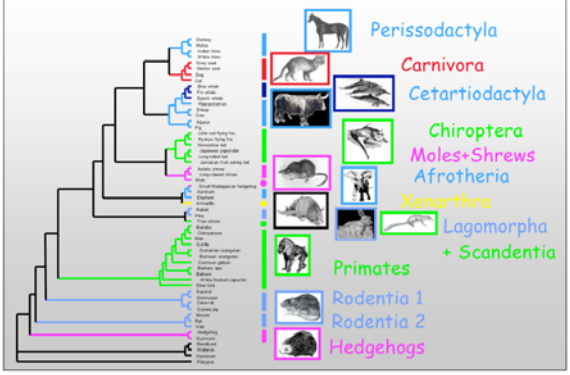
Philosophy



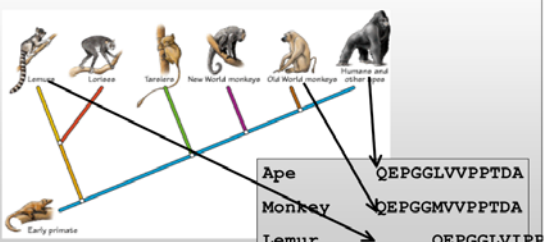
Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
- Gene expression
- Discussion

Evolutionary History

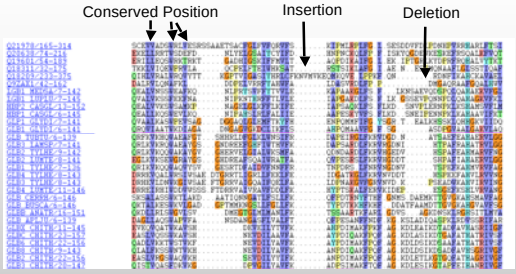


Phylogeny



Primate evolution

Related Proteins



Globins http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00042

Evolution

Evolution = Neutral Variation + Selection

Neutral variation
- Random changes
  Mutation, duplication, deletions, rearrangements, …

Selection
- Preference for variants with better fit

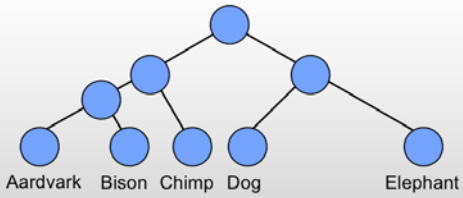
Challenges

- From sequences to phylogenies

Ape	QEPGGLVVPPTDA
Monkey	QEPGGMVVPPTDA
Lemur	QEPGGLVIPPTDA
Cat	REPQGLIVPPTEG

- Understand selection
  Selection ➔ Function & Structure
- Reconstruct history of evolutionary events

Probabilistic Model of Evolution



Random variables – sequence at current day taxa or at ancestors

Potentials/Conditional distribution – represent the probability of evolutionary changes along each branch

Parameterization of Phylogenies

Assumptions:
- Positions (columns) are independent of each other
- Each branch is a reversible continuous time discrete state Markov process

$$P(a \to c \mid t + t') = \sum_b P(a \to b \mid t) P(b \to c \mid t')$$

$$P(a) P(a \to b \mid t) = P(b) P(b \to a \mid t)$$

governed by a **rate matrix** Q

Parameterization:
- Rate matrix + tree topology + branch lengths

Computational Tasks

Likelihood computation, inference of ancestral states
- Inference (dynamic programming, belief propagation)

Branch length estimation:
- Parameter estimation (EM)

Reconstruction:
- Structure learning

Felsenstien, *JME* 1981

Maximum Likelihood Reconstruction

Observed data: (D)
- N sequences of length M
- Each position: an independent sample from the marginal distribution over N current day taxa

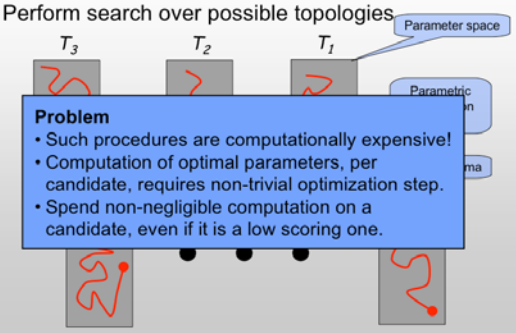
Likelihood:
- Given a tree (T,t):

$$l(T,t:D) = \log P(D \mid T,t)$$
$$= \sum_m \log P(x^m_{[1,\ldots,N]} \mid T,t)$$

Goal:
- Find a tree (T,t) that maximizes $l(T,t:D)$.

Current Approaches

- Perform search over possible topologies



Problem
- Such procedures are computationally expensive!
- Computation of optimal parameters, per candidate, requires non-trivial optimization step.
- Spend non-negligible computation on a candidate, even if it is a low scoring one.

SEMPHY: Structural EM Phylogenetic Reconstructoin

Borrow the idea of Structural EM:
Use parameters found for current topology to help evaluate new topologies.

Outline:

- Perform search in (T, t) space, using EM-like iterations:
  - **E-step**: use current solution to compute expected sufficient statistics for **all topologies**
  - **M-step**: select new topology based on these expected sufficient statistics

Friedman *et al*, *JCB* 2002

Algorithm Outline

- Compute: $E[S_{(i,j)}(a,b) \mid D, T^0, t^0]$
- Weights: $w_{i,j} = \max_t F(t, E[S_{i,j}])$

Original Tree (T^0, t^0)

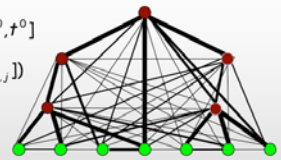
Unlike standard EM for trees, we compute all possible pairwise statistics

Time: $O(N^2M)$

Friedman *et al*, *JCB* 2002

Algorithm Outline

- Compute: $E[S_{(i,j)}(a,b) \mid D, T^0, t^0]$
- Weights: $w_{i,j} = \max_t F(t, E[S_{i,j}])$
- Find: $T' = \arg\max_T \sum_{(i,j) \in T} w_{i,j}$



Pairwise weights

This stage also computes the branch length for each pair (i,j)

Friedman *et al*, *JCB* 2002

Algorithm Outline

- Compute: $E[S_{(i,j)}(a,b) \mid D, T^0, t^0]$
- Weights: $w_{i,j} = \max_t F(t, E[S_{i,j}])$
- Find: $T' = \arg\max_T \sum_{(i,j) \in T} w_{i,j}$
- Construct bifurcation T_1

Max. Spanning Tree

Fast greedy procedure to find tree
By construction:

$$Q(T',t') \geq Q(T_0,t_0)$$

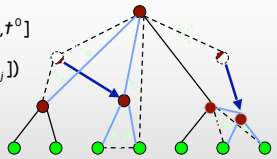
Thus,

$$l(T',t') \geq l(T_0,t_0)$$

Friedman *et al*, *JCB* 2002

Algorithm Outline

- Compute: $E[S_{(i,j)}(a,b) \mid D, T^0, t^0]$
- Weights: $w_{i,j} = \max_t F(t, E[S_{i,j}])$
- Find: $T' = \arg\max_T \sum_{(i,j) \in T} w_{i,j}$
- Construct bifurcation T_1



Fix Tree

Remove redundant nodes
Add nodes to break large degree

This operation preserves likelihood

$$l(T_1,t') = l(T',t') \geq l(T_0,t_0)$$

Friedman *et al*, *JCB* 2002

Algorithm Outline

- Compute: $E[S_{(i,j)}(a,b) \mid D, T^0, t^0]$
- Weights: $w_{i,j} = \max_t F(t, E[S_{i,j}])$
- Find: $T' = \arg\max_T \sum_{(i,j)\in T} w_{i,j}$
- Construct bifurcation T_1
- Thm: $l(T_1,t_1) \geq l(T_0,t_0)$

New Tree

These steps are then repeated until convergence

Friedman *et al*, *JCB* 2002

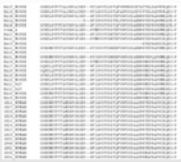
Beyond Vanilla Flavor

Rate variation:

- Different positions evolve at different rate

Yang, *JME* 1994

Application: Map Conservation



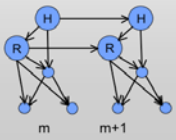
Calculate rates

http://consurf.tau.ac.il/

Beyond Vanilla Flavor

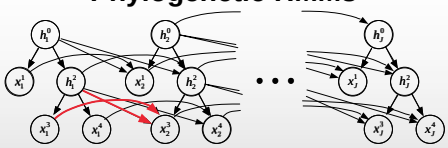
Dependencies between Positions:

- Dependent rate
- Dependent state



Felsenstein and Churchill, *MBE* 1996

Phylogenetic HMMs



- Each substitution depends on the substitution at the previous position
- This structure captures **context specific** effects during evolution

Problem:

- Resulting model is intractable

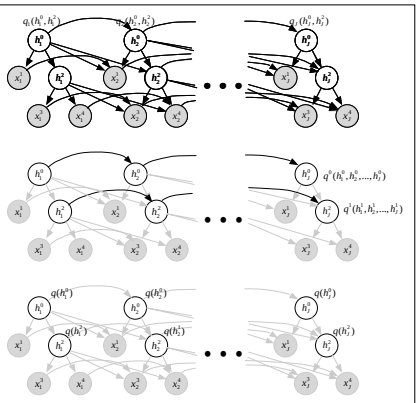
Siepel & Haussler, *MBE* 2004

Variational approximations

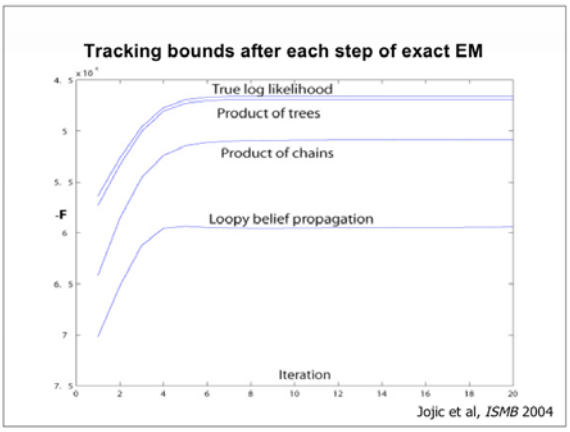
Product of trees

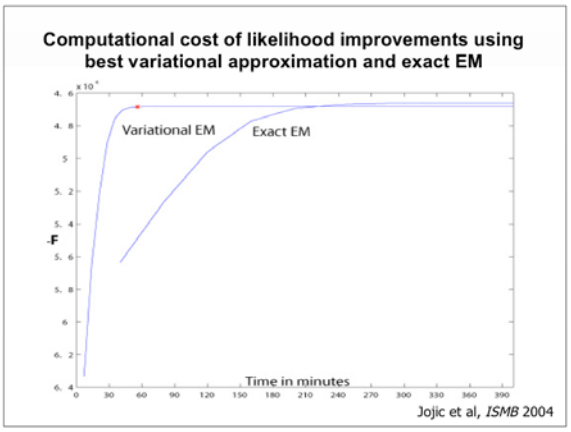
Product of chains

Mean field (product of "nodes")



Jojic et al, *ISMB* 2004





Remaining Challenges

- To learn phylogenies, need to align sequences
- For good alignment, need to know which positions are conserved

Can we do both at the same time?

Requires
- Dealing with insertions + deletions
- Different ways of shifting each sequence

➔Hard problem

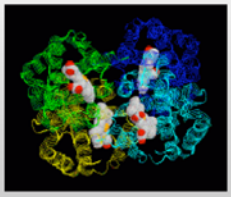
Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
- Gene expression
- Discussion

Proteins

Sequence
➔ Structure
➔ Function

Major question:
- How to annotate structure and function of new protein sequences?



Hemoglobin β

Protein Families

Idea:
- Use knowledge about function of known proteins
- Sequence conservation
  ➔ similar structure & function

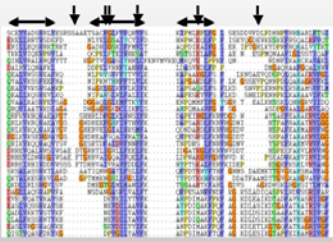
Protein Family:
A collection of proteins sequences that have a common
- Evolutionary ancestor
- Structure
- Function

Protein Family

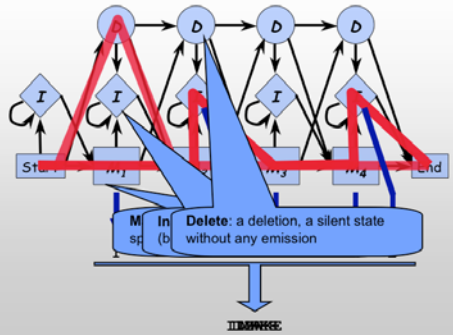
Multiple Sequence alignment ➔ detect new proteins

Need to represent
- Commonalities
- Variations

- Conserved stretches
- Potential insertions
- Potential deletions
- Preferred letters at each position



Profile HMMs



Using Profile HMMs

Training set of curated proteins (domains) from the same family

Manual curation: 20-200 seq

↓

Profile HMM

↓

Annotation of new protein sequences + alignment with original sequences

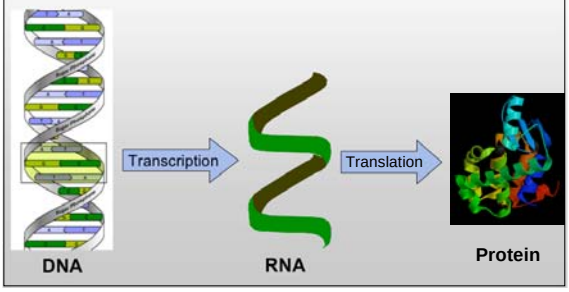
Search throughout all known protein sequences: 10^6 -- 10^8

PFAM Database

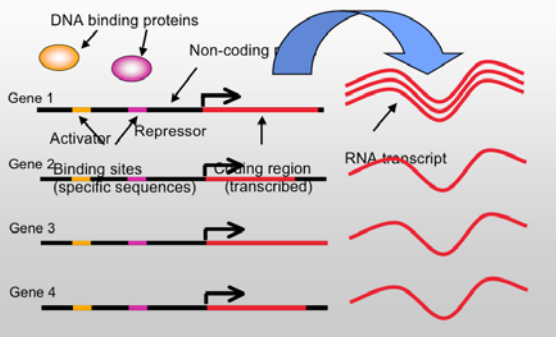
Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
- Discussion

The Central Dogma

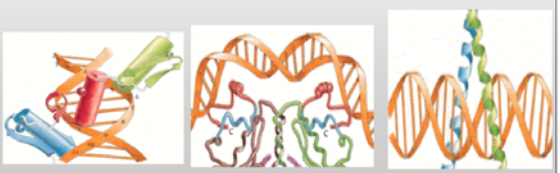


Transcriptional Regulation



Transcription Factor Binding Sites

- Gene regulatory proteins contain structural elements that can "read" DNA sequence "motifs"
- The amino acid – DNA recognition is not straightforward
- Experiments can pinpoint binding sites on DNA



Zinc finger | Helix-Turn-Helix | Leucine zipper

Modeling Binding Sites

Given a set of (aligned) binding sites …

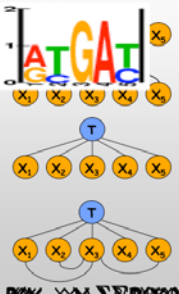
- Consensus sequence
- Probabilistic model (profile of a binding site)

A	4	3	1	1	0	0	1	0	0	1	0	1
C	1	3	0	0	0	0	13	6	0	0	1	9
G	5	5	13	13	14	14	0	8	14	12	13	1
T	4	3	0	0	0	0	0	0	0	1	0	3

Is this sufficient?

How to model binding sites ?

$P(X_1\ X_2\ X_3\ X_4\ X_5) = ?$ **represents a distribution of binding sites**



Profile: Independency model

Tree: Direct dependencies

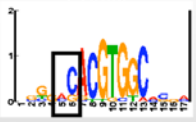
Mixture of Profiles: Global dependencies

Mixture of Trees: Both types of dependencies

Barash et al, *RECOMB* 2002

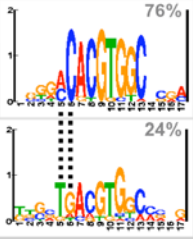
Arabidopsis ABA binding factor 1

Profile



Test LL per instance **-19.93**

Mixture of Profiles



Tree

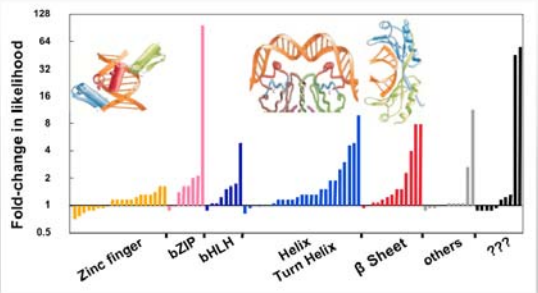
Test LL per instance **-18.47** (+1.46)
(improvement in likelihood > 2.5-fold)

Test LL per instance **-18.70** (+1.23)
(improvement in likelihood > 2-fold)

Barash et al, *RECOMB* 2002

Likelihood improvement over profiles

TRANSFAC: 95 aligned data sets



Barash et al, *RECOMB* 2002

Motif finding problem

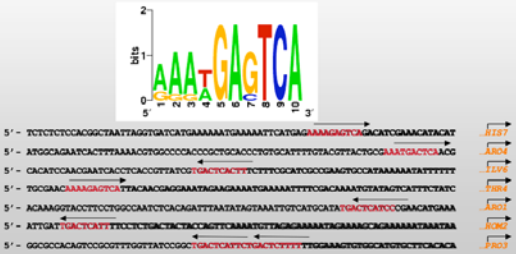
Input: A set of potentially co-regulated genes
Output: A common motif in their promoters

Sources of data:

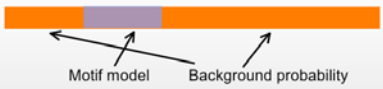
- Gene annotation (e.g. **Hughes et al**, 2000)
- Gene expression (e.g. **Spellman et al**, 1998; **Tavazoie et al**, 2000)
- ChIP (e.g. **Simon et al**, 2001; **Lee et al**, 2002)

Example

- Upstream regions from yeast *Sacharomyces cerevisiae* genes (300-600bp)



Probabilistic Model



- Background probability: given
- Motif model – parameters being learned

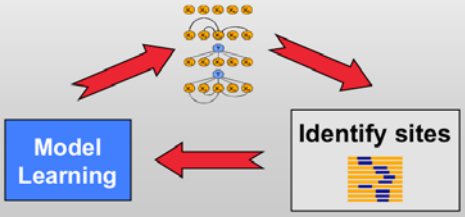
Hidden variable:

- Location of motif within each sequence

Learning models: unaligned data

EM (MEME)

- Identify binding site positions
- Learn a dependency model



Learning models: unaligned data

EM (MEME)

Gibbs Sampling (AlignACE)

Discriminative Sampling

- Find motif that best separates positive examples from rest of promoter sequences

Challenges

Small training sets

- 10-500 sequences (out of 1000s genes)

Short motifs within long sequences

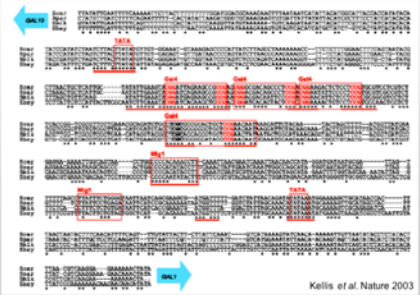
- Motifs are 6-20bp, promoters are 500-5000bps

Motifs are not perfect words

- Mismatches are allowed

Comparative Genomics

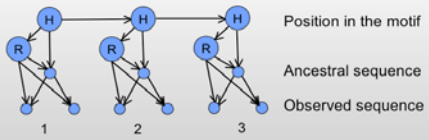
Functional areas should be conserved



Kellis *et al.* Nature 2003

Modeling Conserved Motifs

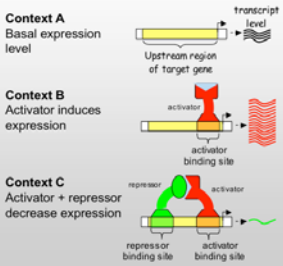
- Parallel search in orthologous promoters
- Model evolution at each position



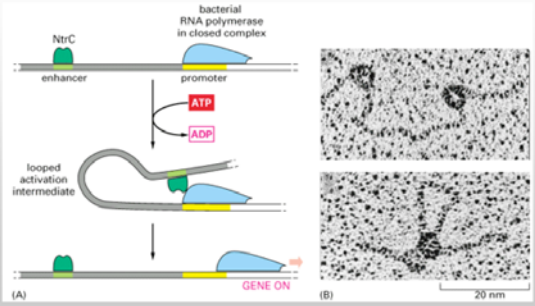
Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
- Discussion

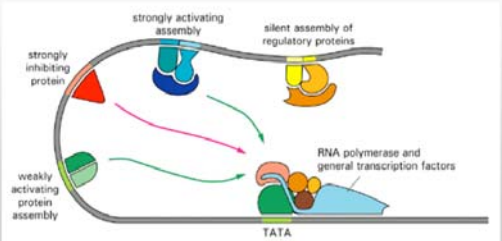
Combinatorial Regulation



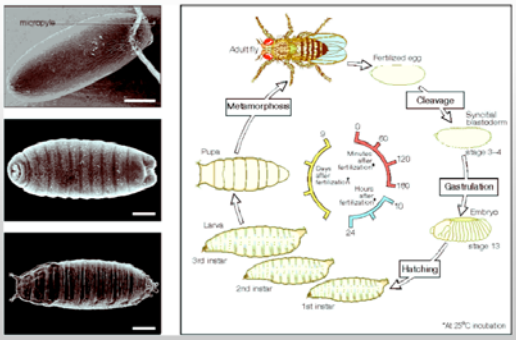
Effect at a Distance



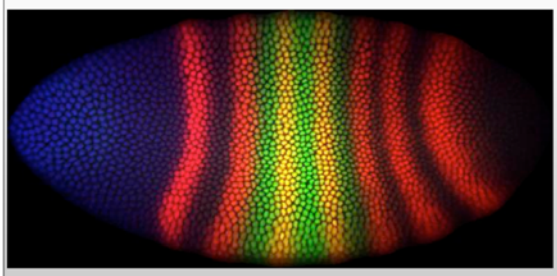
Eukaryotic Regulation



The Life Cycle of a Drosophila Fly



Segmentation Achieved by Spatial patterning of genes

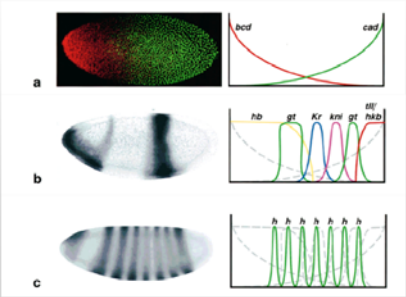


Bicoid –Blue Eve Stripped-Red Kruppel-Yellow

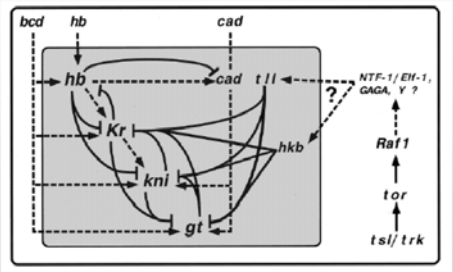
Gradual Refinement of Expression Domains

Coarse

Fine

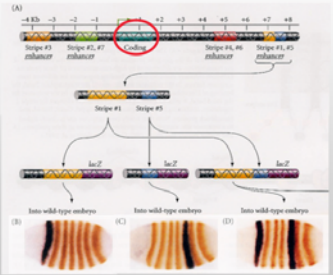


Partial gene network of the segmentation

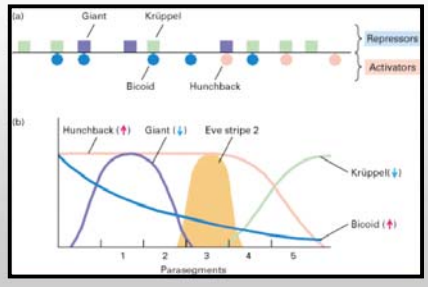


Each Stripe has its own "Enhancer"

Control of even-skipped expression

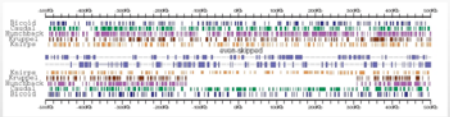


Example: Eve Stripe 2

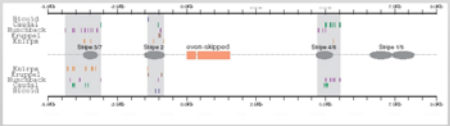


Eve Gene

- Search all sites 1MB around eve gene.



- Searching for "dense" windows



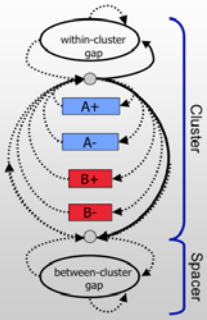
Berman et al, *PNAS* 2002

Probabilistic Model

Build (generalized HMM)

States:

- Gap (background)
  - Long gaps (between clusters)
  - Short gap (in cluster)
- Motifs
  - Emit motif sequence
  - Either positive strand or negative one



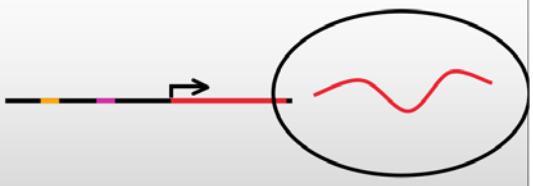
Parse long genomic sequences by Viterbi's algorithm

Frith *et al*, *NAR*, 2002; Rajewsky *et al*, *BMC Bioinf*, 2002; Baily & Noble, *Bioinf*, 2003

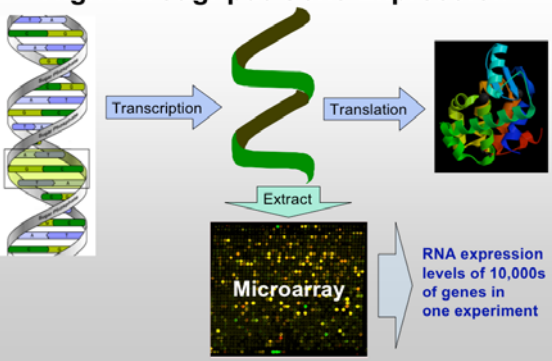
Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
  - General methods
  - Unified models of transcriptional regulation
  - Regulatory networks
- Discussion

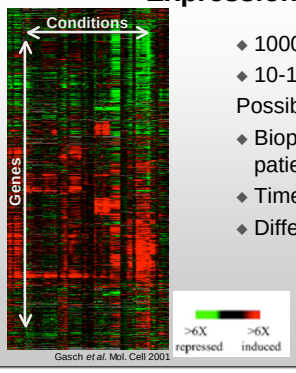
Transcriptional Regulation



High Throughput Gene Expression



Expression Data



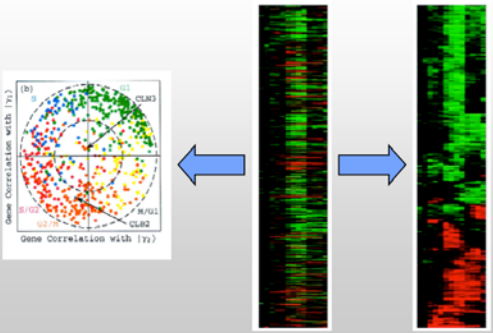
Gasch *et al.* Mol. Cell 2001

- 1000s of genes
- 10-100s of arrays

Possible designs

- Biopsies from different patient populations
- Time course
- Different perturbations

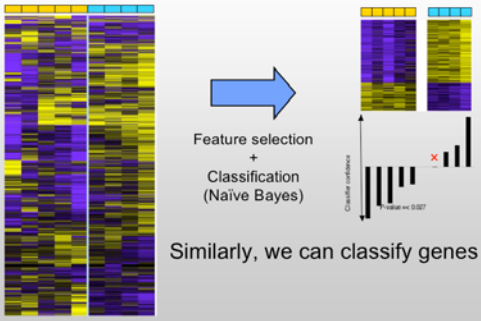
Expression: Unsupervised



Eisen *et al.* PNAS 1998; Alter *et al*, PNAS 2000

Expression: Supervised

Labeled samples



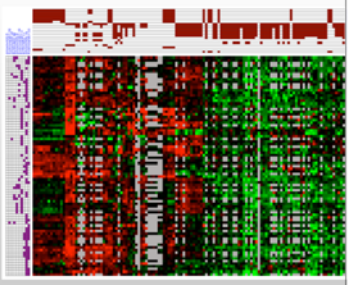
Feature selection
+
Classification
(Naïve Bayes)

Similarly, we can classify genes

Expression +Annotations

Array annotations:
Tissue type, Clinical conditions, Treatments, Prognosis

Gene annotations:
Function, Process, Regulatory regions Cellular location, protein family

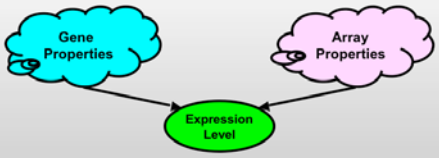


Relational models!

Relational Approach

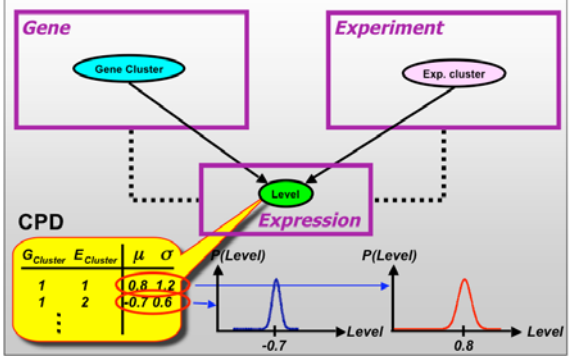
Key Idea:

Expression level is "explained" by properties of gene and properties of experiment

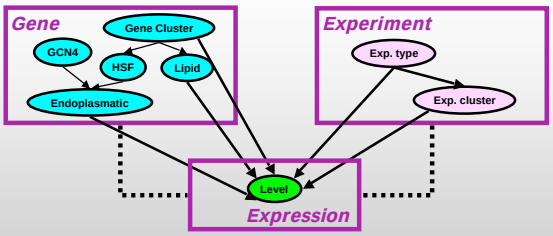


Segal *et al*, *ISMB* 2001

Probabilistic Relational Models



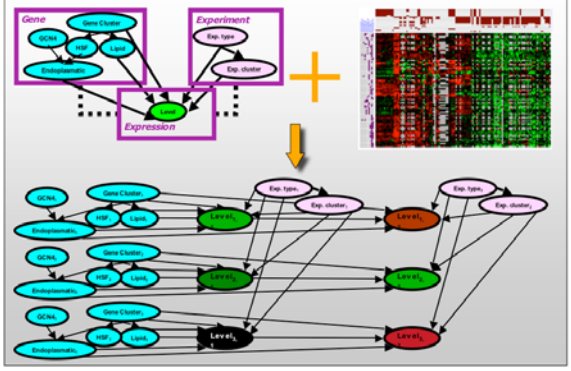
Adding Heterogeneous Data



- Annotations
- Binding sites
- Experimental details

Segal *et al*, *ISMB* 2001

Semantics



Outline

- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
  - General methods
  - Unified models of transcriptional regulation
  - Regulatory networks
- Discussion

TF to Expression

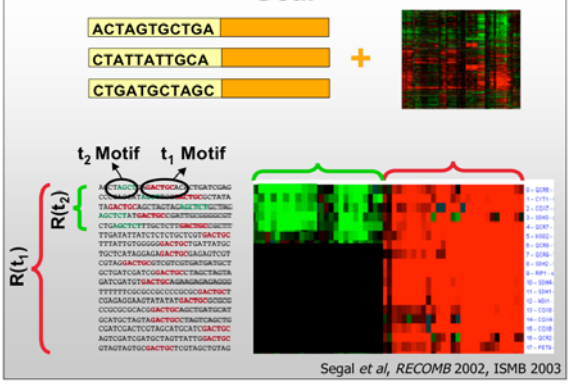
Key Question:

- Can we explain changes in expression?

General model:

- Transcription factor binding sites in promoter region should "explain" changes in transcription

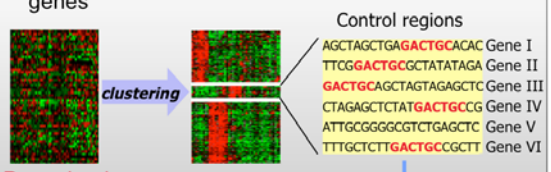
Goal



Segal *et al*, *RECOMB* 2002, ISMB 2003

"Classical" Approach

- Cluster gene expression profiles
- Search for motifs in control regions of clustered genes

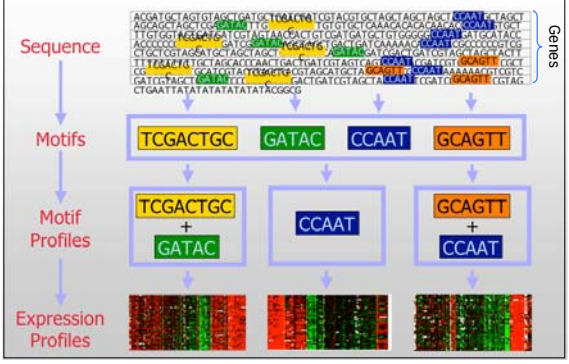


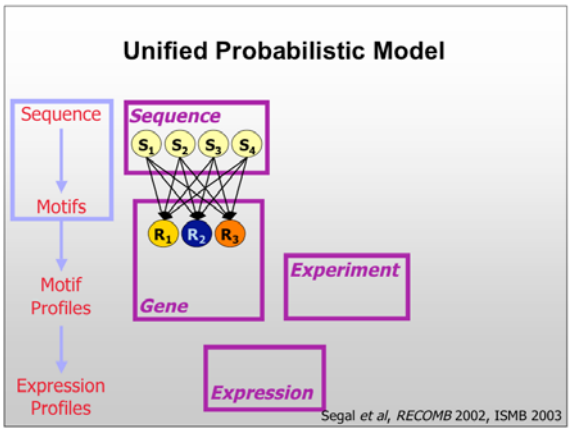
Procedural

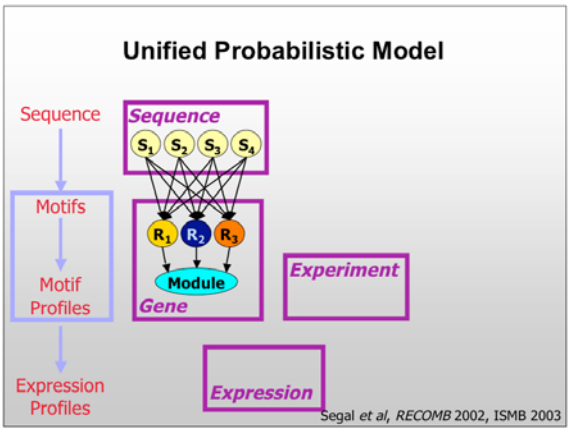
- Apply separate method to each type of data
- Use output of one method as input to the next
- Unidirectional information flow

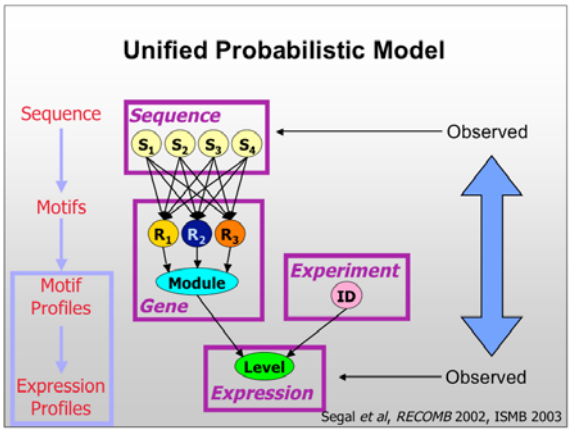
Motif → GACTGC

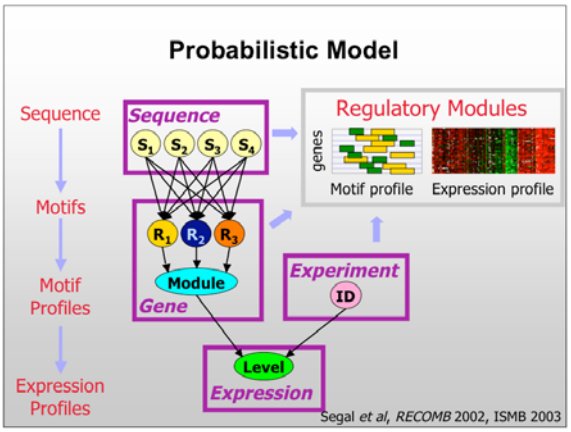
Flow of Information





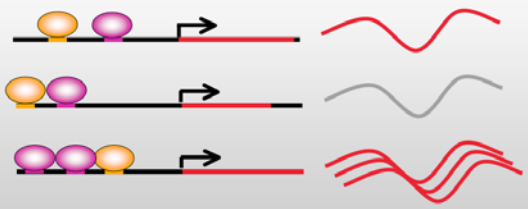






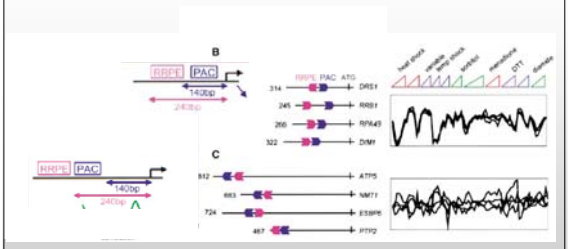
Promoters Revisited

- Promoter organization matters



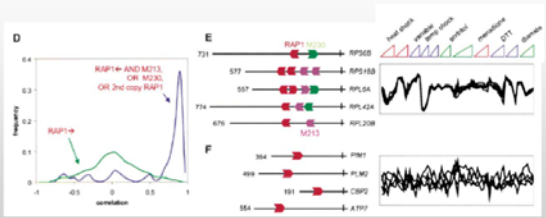
- Different logic for promoters
  - AND, OR, NAND, XOR, ….

More Refined Models



Beer & Tavazoie, *Cell, 2004*

More Refined Models

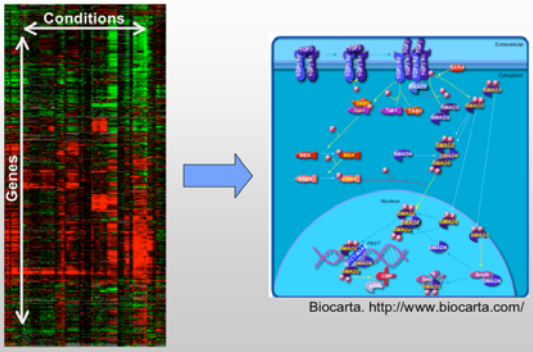


Beer & Tavazoie, *Cell, 2004*

Outline

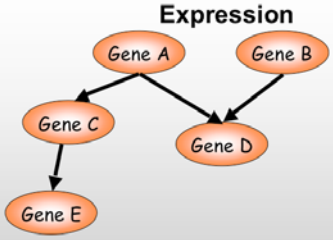
- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
  - General methods
  - Unified models of transcriptional regulation
  - Regulatory networks
- Discussion

Goal: Reconstruct Cellular Networks



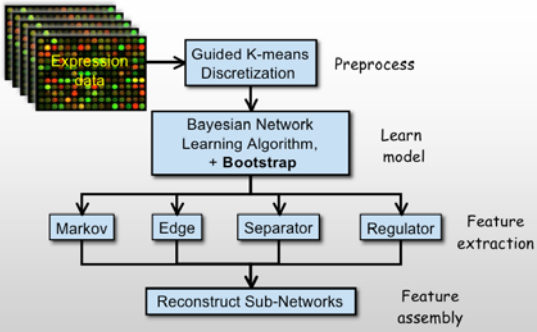
Biocarta. http://www.biocarta.com/

Causal Reconstruction for Gene Expression



- Use language of Bayesian networks to reconstruct causal connections

Friedman et al, JCB 2000

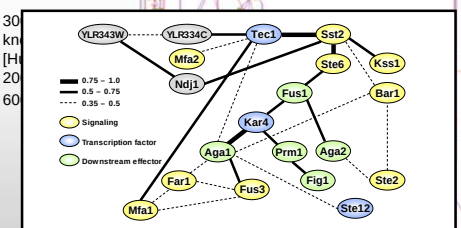


Global network→ Local features → Sub-network

Friedman et al, JCB 2000; Pe'er et al, ISMB 2001

Gene Interaction Maps

- 30
- 60

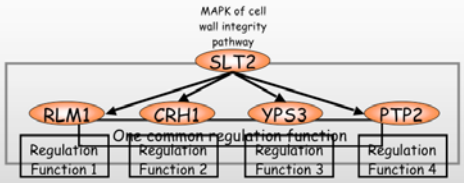


"Mating response" Substructure

STE12 (main TF), Fus3 (Main MAPK) are marginal

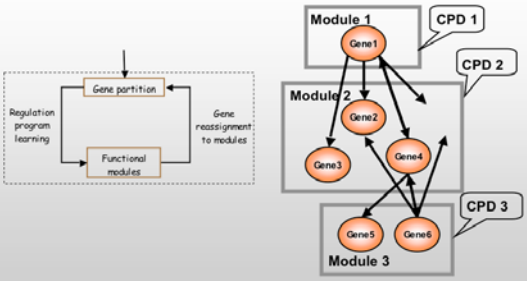
Pe'er et al, ISMB 2001

From Networks to Modules



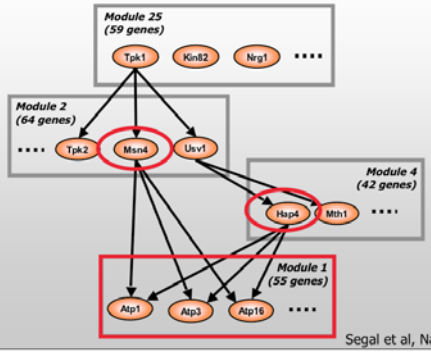
Idea: enforce common regulatory program

- Statistical robustness: Regulation programs are estimated from $m*k$ samples
- Organization of genes into regulatory modules: Concise biological description

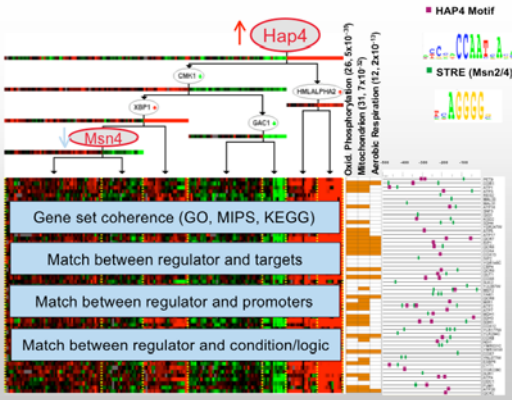


Segal et al, Nat Gen 2003

Learned Network (fragment)

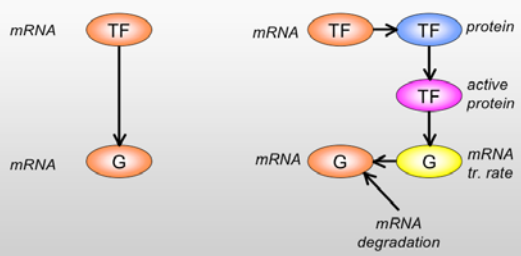


Segal et al, Nat Gen 2003



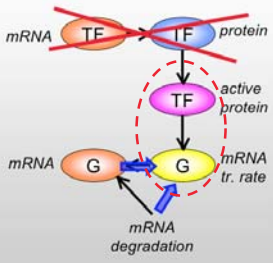
Segal et al, Nat Gen 2003

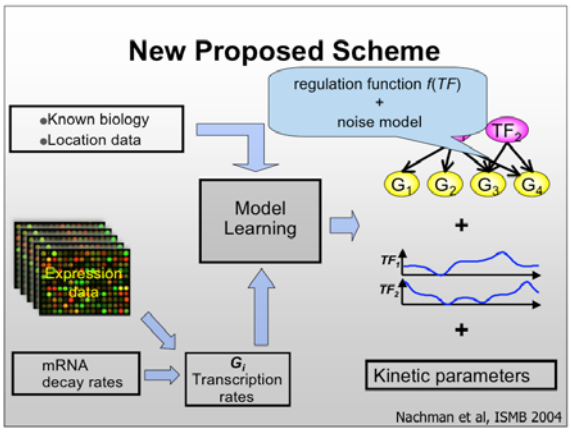
A Major Assumption

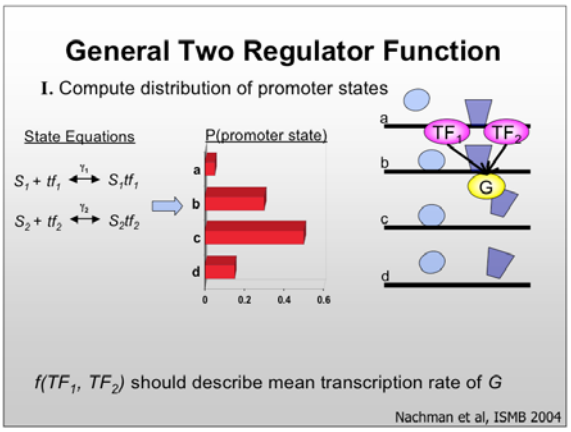


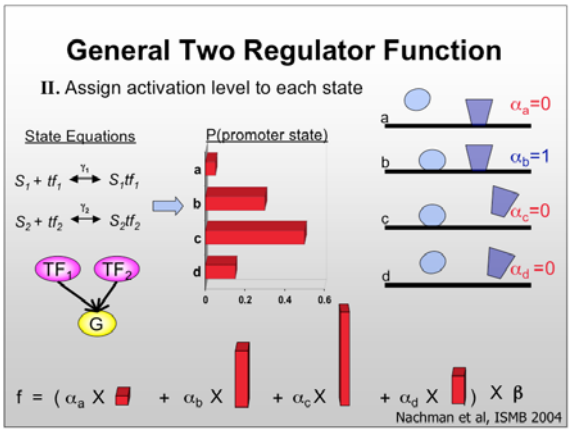
Realistic Regulation Modeling

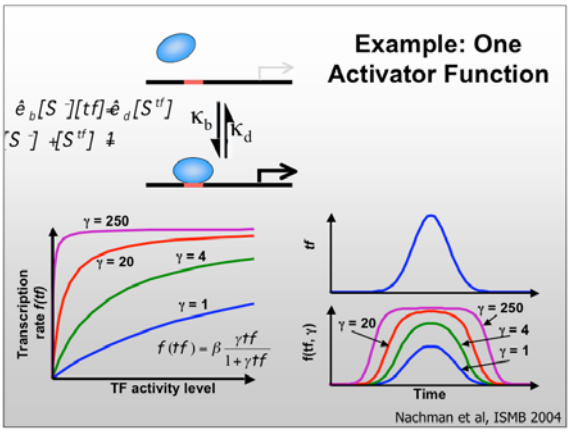
- Model the closest connection
- Active protein levels are not measured
- Transcript rates are computed from expression data and mRNA decay rates

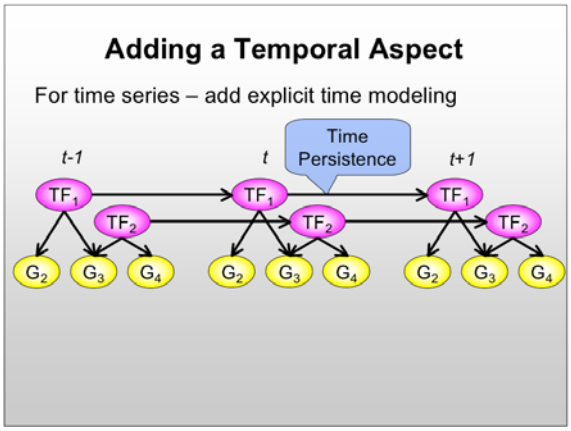


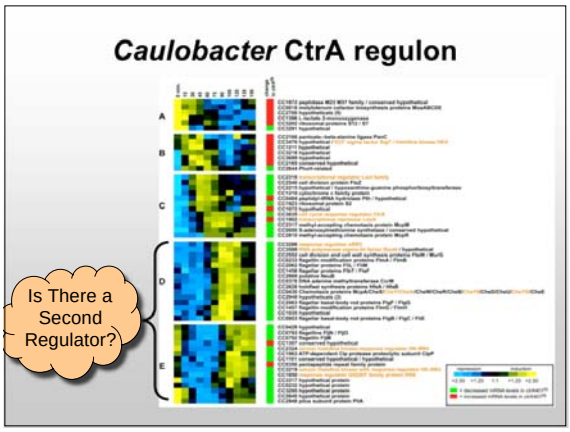

New Proposed Scheme
regulation function f(TF)
+
noise model
•Known biology
•Location data
Model Learning
Expression data
mRNA decay rates
Gi Transcription rates
Kinetic parameters
Nachman et al, ISMB 2004


General Two Regulator Function
I. Compute distribution of promoter states
State Equations
P(promoter state)
f(TF1, TF2) should describe mean transcription rate of G
Nachman et al, ISMB 2004

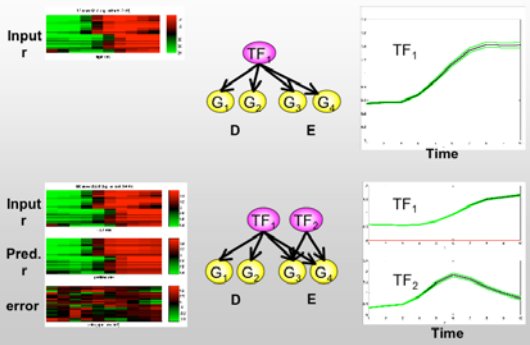

General Two Regulator Function
II. Assign activation level to each state
State Equations
P(promoter state)
Nachman et al, ISMB 2004


Example: One Activator Function
Transcription rate f(tf)
TF activity level
Time
Nachman et al, ISMB 2004


Adding a Temporal Aspect
For time series – add explicit time modeling
Time Persistence
t-1
t
t+1


Caulobacter CtrA regulon
Is There a Second Regulator?

Caulobacter CtrA regulon

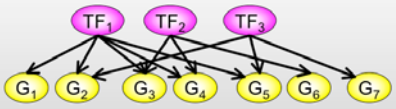


Multiple Regulon Experiments

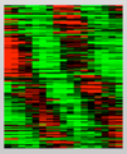
Can we describe the cell transcriptome using a small number of hidden regulators?

few TFs

many genes



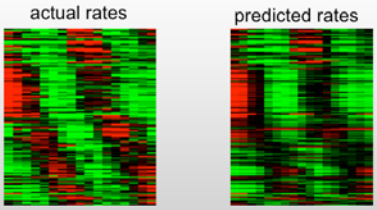
- "Realistic" dimensionality reduction
- Allows prediction of target gene dynamics



Cell Cycle Experiment

actual rates

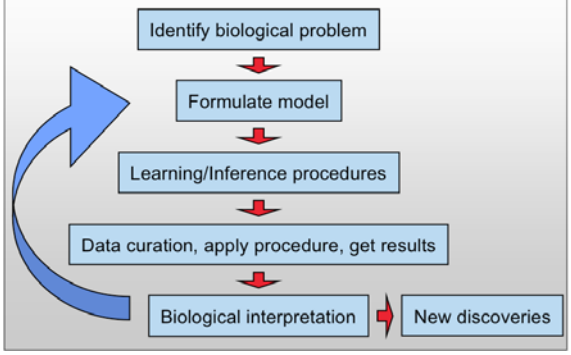
predicted rates



Outline

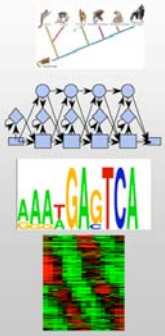
- Sequence evolution
- Protein families
- Transcriptional regulation
  - Transcription factor binding sites
  - Combinatorial regulation
- Gene expression
  - General methods
  - Unified models of transcriptional regulation
  - Regulatory networks
- Discussion

Philosophy



Recap

- Models of evolution
  - Sequence evolution
  - Comparative models
- Sequence Models
  - Profile HMMs
- Transcription Factors
  - Binding sites
  - CIS-regulatory modules
- Gene Expression
  - Clustering, supervised, and beyond
- Combination of subsets of these



Additional Areas

- Gene finding
  - Extended HMMs + evolutionary models
- Analysis of genetic traits and diseases
  - Linkage analysis
  - SNPs, haplotypes, and recombination
- Interaction networks (protein-DNA, protein-protein)
  - Relational models
- Protein structure
  - 2nd-ary and 3rd-ary structure, molecular recognition

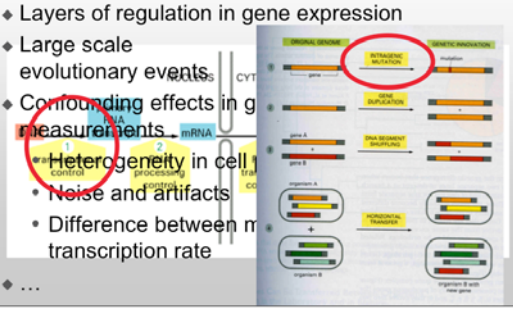
Take Home Message

- **Graphical models as a methodology**
  - Modeling language
  - Foundations & algorithms for learning
  - Allows to incorporate prior knowledge about biological mechanisms
  - Learning can reveal "structure" in data
- **Exploring unified system models**
  - Learning from heterogeneous data
    - Not simply combining conclusions
  - Combine weak evidence from multiple sources
    ⇒ detect subtle signals
  - Get closer to **mechanistic** understanding of the signal

What About Biology?

We skimmed over many details

- Layers of regulation in gene expression
- Large scale evolutionary events
- Confounding effects in g measurements
  - Heterogeneity in cell
  - Noise and artifacts
  - Difference between m transcription rate
- …



The END