



UCSF

Computational Methods in Systems Biology

Nir Friedman Maya Schuldiner

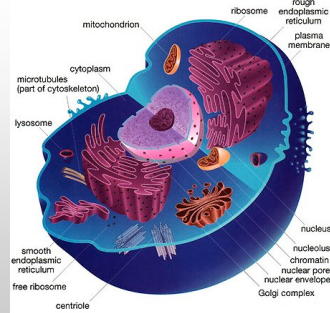
What is Biology?

- ♦ "A branch of knowledge that deals with living organisms and vital processes"
- ♦ The hottest scientific frontier of our times
 - Many great processes have been figured out
 - Much is still unknown
- ♦ Tremendous impact on Medicine
 - Both diagnosis, prognosis, and treatment



2

Bakers Yeast *Saccharomyces Cerevisiae*

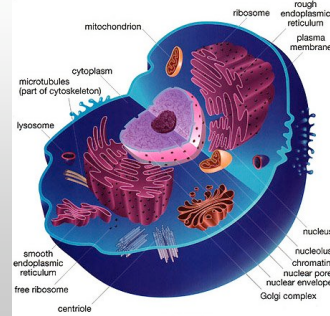


- Used to make bread and beer
- The simplest cell that still resembles human cells

3

Biological Systems are Complex

- The System is NOT just a sum of its parts



4

What is Systems Biology?

"Systems biology is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of that system"

- The last decades lead to revolution on how we can examine and understand biological systems

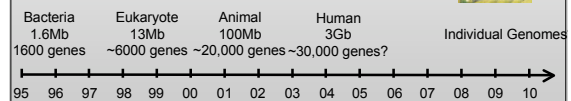
Characterized by

- High-throughput assays
- Integration of multiple forms of experiments & knowledge
- Mathematical modeling

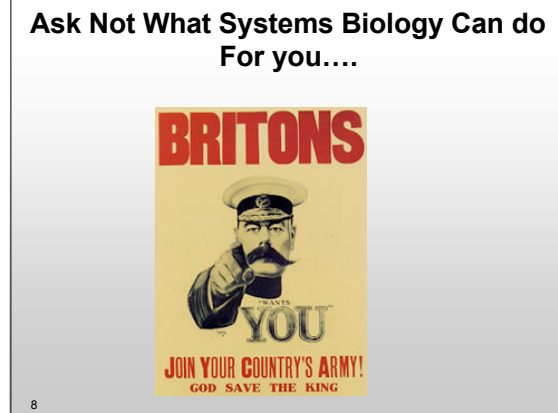
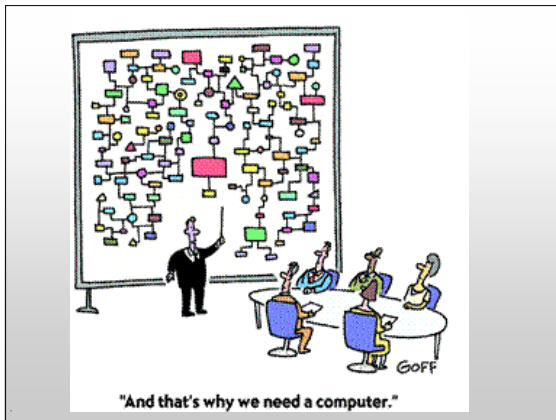
5

The Age of Genomes

404 Complete Microbial Genomes (Thousands in progress)
31 Complete Eukaryotic Genomes (315 in progress!)
3 Complete Plant Genomes (6 in progress)



6

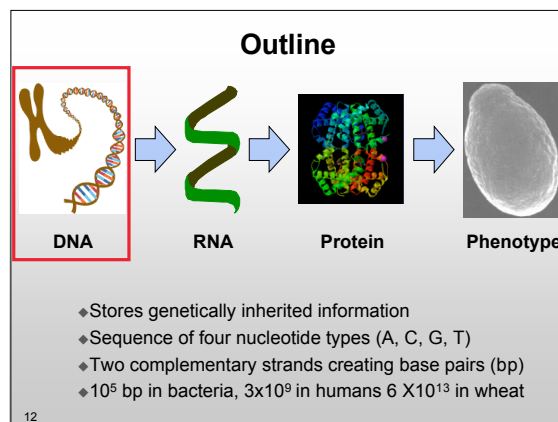
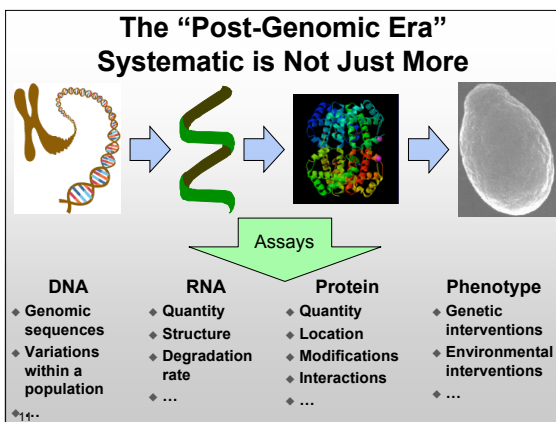
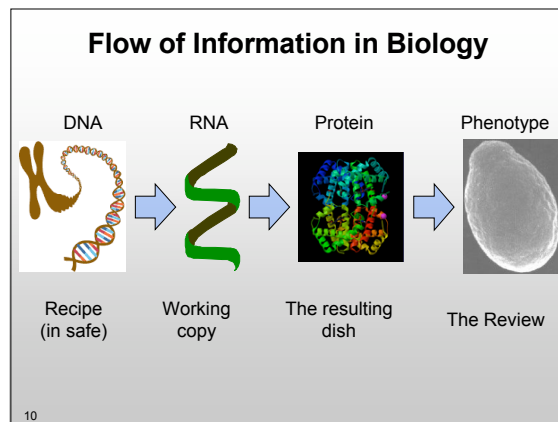


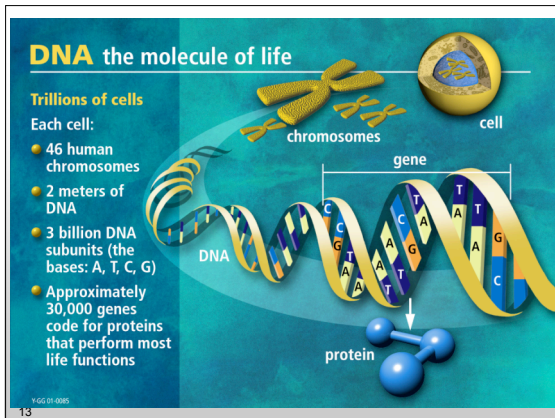
Why Biology for NIPS Crowd?

- ◆ **Quantity**
 - Data-intense discipline: Too vast for manual interpretation
- ◆ **Systematic**
 - Collection of data on **all** genes/proteins/...
- ◆ **Multi-faceted**
 - Measurements of complementary aspects of cellular function, development and disease states
 - Challenge of integration and fusion of multiple data

Has the potential to be medically applicative!

9





Understanding Genome Sequences

~3,289,000,000 characters:

```
aatgtgctctgcgaattatgatagtgatctgtatttactactgcacat
atattgggccaagtgaatttttttaagctaaatagatttttggaacttt
tgacatgactttgtgttaattaaaaacaaaaaagaattgcagaagtgt
tgtaagcttgtaaaaaaattcaacaatgcagacaaatgtgtctgcagt
cttccactcagtatcatctttgtttgtacctatcagaatgtttctatg
tacaagtctttaaatacttgcgaactgtttgtccactgagtatatta
tggaacatcttttcattggcaggacataagattgttttaaggcataaaa
taaaacaaaaaactgattgcgcgggtacgtggctcagcgcctgaatcc
cagcactttgggagatcgaggaggaggtacacctgaagtcaggagttac
agacatggagaacccgcgtctctactaaaaatacaaaattagcctggcgt
ggtggcgatgcctgttaatccagctactcgggaggtgcagggaaggaa
tcgcttgaacccggagcggaggtgcggtgagcgcgagatcgcaacgttg
cactccagcctgggcagcagagcgaactgtctaaacaaacacacaaaa
aaacctgatacatggtatgggaagtacattgtttaaacatgcattgaga
tttaggtgtttccagtttttactgcacagatcggcaatgaatataat
tttatgtatacatcacaataatatacgttggaataatcctagaagtgg
```

Goal:

Identify components encoded in the DNA sequence

14

Open Reading Frame

ATGCTCAGCGTGACCTCA . . . CAGCGTTAA
M L S V T S . . . Q R STP

- ◆ Protein-encoding DNA sequence consists of a sequence of 3 letter **codons**
- ◆ Starts with the START codon (ATG)
- ◆ Ends with a STOP codon (TAA, TAG, or TGA)

15

Finding Open Reading Frames

ATGCTCAGCGTGACCTCA . . . CAGCGTTAA
M L S V T S . . . Q R STP

Try all possible starting points

- ◆ 3 possible offsets
- ◆ 2 possible strands

Simple algorithm finds all ORFs in a genome

- ◆ Many of these are spurious (are not real genes)
- ◆ How do we focus on the real ones?

16

Using Additional Genomes

Basic premise

"What is important is conserved"

Evolution = Variation + Selection

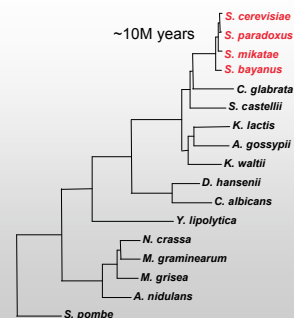
- Variation is random
- Selection reflects function

Idea:

- ◆ Instead of studying a single genome, compare related genomes
- ◆ A real open reading frame will be conserved

17

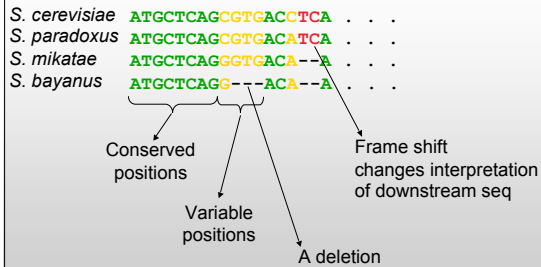
Phylogentic Tree of Yeasts



18

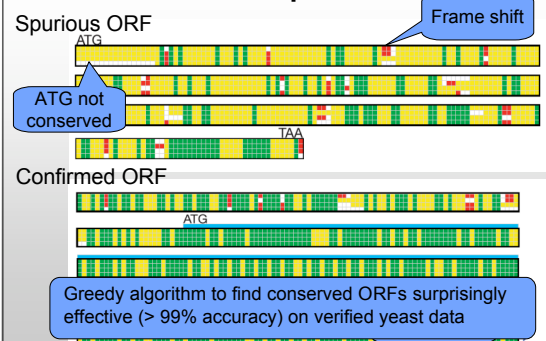
Kellis et al, Nature 2003

Evolution of Open Reading Frame



19

Examples



20

[Kellis et al, Nature 2003]

Defining Conservation

Naïve approach

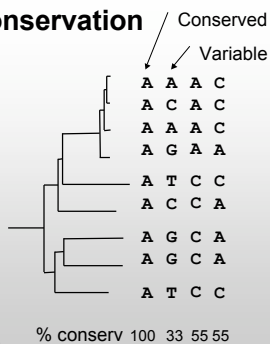
- Consensus between all species

Problem:

- Rough grained
- Ignores distances between species
- Ignores the tree topology

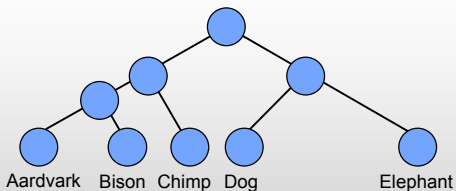
Goal:

- More sensitive and robust methods



21

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

22

Parameterization of Phylogenies

Assumptions:

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

$$P(a \rightarrow c | t + t') = \sum_b P(a \rightarrow b | t) P(b \rightarrow c | t')$$

$$P(a) P(a \rightarrow b | t) = P(b) P(b \rightarrow a | t)$$

governed by a **rate matrix Q**

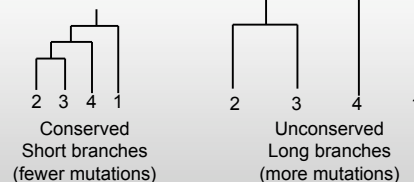
$$Q_{a,b} = \frac{d}{dt} P(a \rightarrow b | t) \Big|_{t=0}$$

$$P(a \rightarrow b | t) = [e^{tQ}]_{a,b}$$

23

Conserved vs. unconstrained

Two hypotheses:



Use $\log \frac{P(\text{position} | \text{unconstrained})}{P(\text{position} | \text{conserved})}$

24

[Boffelli et al, Science 2003]

Genes Are Better Conserved

The figure consists of two side-by-side line graphs sharing a common x-axis labeled 'sequence (bp)' ranging from 0 to 1500.

Left Graph: The y-axis is labeled 'log Fast/Slow' and ranges from -2.0 to 0.0. The green line shows a highly variable pattern with a prominent sharp peak reaching 0.0 at approximately 900 bp. A red horizontal bar is located on the x-axis from 500 to 600 bp.

Right Graph: The y-axis is labeled '% conserved' and ranges from -2.7 to 0.3. The green line shows a sharp peak reaching approximately 0.25 at approximately 200 bp, followed by a decline and then a series of smaller peaks. A red horizontal bar is located on the x-axis from 500 to 600 bp.

Below the x-axis of both graphs is a legend for the red bar, showing three levels of conservation: 100% (dark red), 75% (medium red), and 50% (light red). The red bar in both graphs is solid dark red, indicating 100% conservation.

[Boffelli et al, Science 2003]

Challenges

Other types of genomic elements

- ◆ Small polypeptides (peptohormones, neuropeptides)
- ◆ RNA coding genes
 - rRNA, tRNA, snoRNA...
 - miRNA
- ◆ Regulatory regions

Regulatory Elements

The diagram illustrates the components of a eukaryotic promoter and the process of transcription initiation. A DNA strand is shown with several key regions:

- gene regulatory sequences:** Indicated by a dashed line at the top, encompassing the **spacer DNA** and **gene regulatory proteins** (represented by red and yellow shapes) that bind to specific sites.
- general transcription factors:** Represented by various colored shapes (green, red, blue, yellow) that assemble on the DNA to facilitate transcription.
- RNA polymerase:** A large blue complex that binds to the DNA and initiates transcription.
- TATA box:** A specific DNA sequence (labeled "TATA box") where the transcription machinery assembles.
- start of transcription:** The point where the RNA strand begins to be synthesized, indicated by a red line.
- upstream:** The region of DNA located before the start of transcription.
- promoter:** The region of DNA that includes the TATA box and other regulatory elements.

Three inset diagrams provide detailed views of the DNA-protein interactions:

- The top-left inset shows a DNA double helix with a red protein bound to a specific site.
- The top-right inset shows a DNA double helix with a green protein bound to a specific site.
- The bottom-left inset shows a DNA double helix with a red protein bound to a specific site.

28

[illegible]

Challenges

Other types of genomic elements

- ◆ Small polypeptides (peptohormones, neuropeptides)
- ◆ RNA coding genes
 - rRNA, tRNA, snoRNA...
 - miRNA
- ◆ Regulatory regions

Recognition of elements without comparisons

- ◆ Clearly sequence contains enough information to “parse” it within the living cell

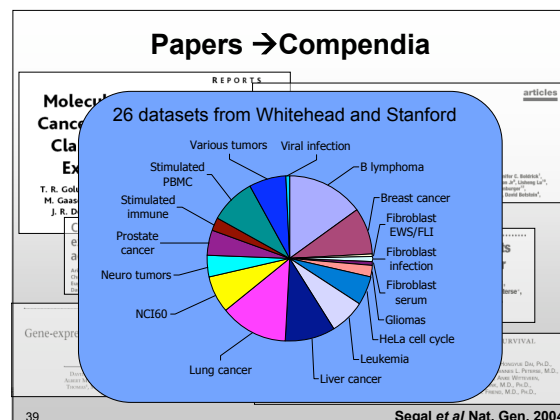
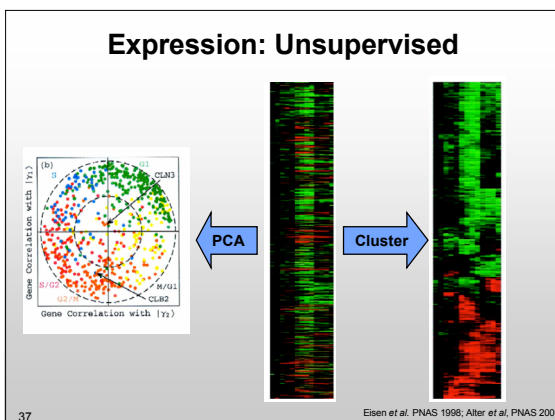
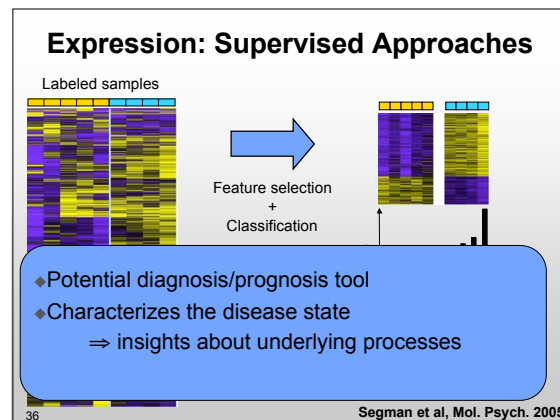
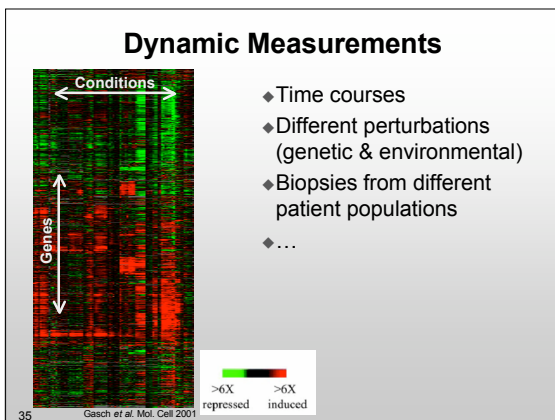
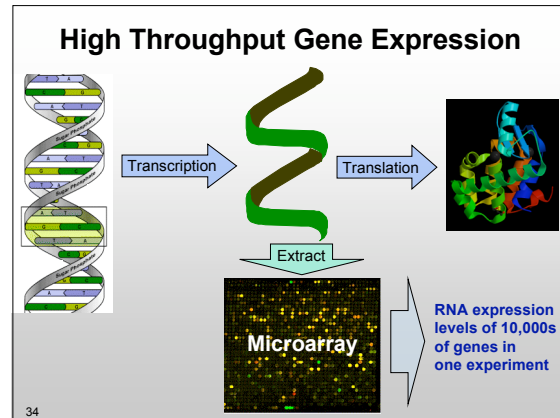
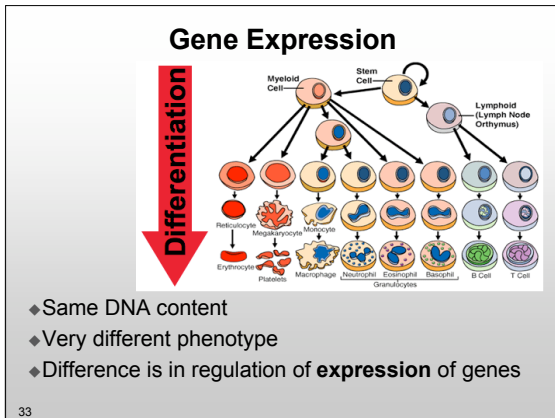
30

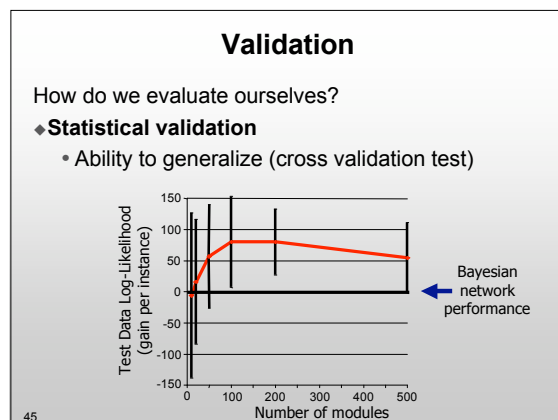
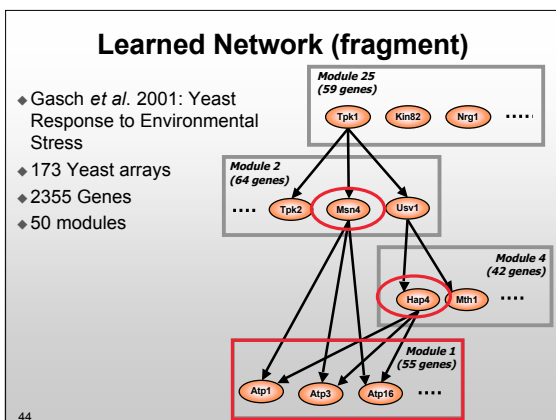
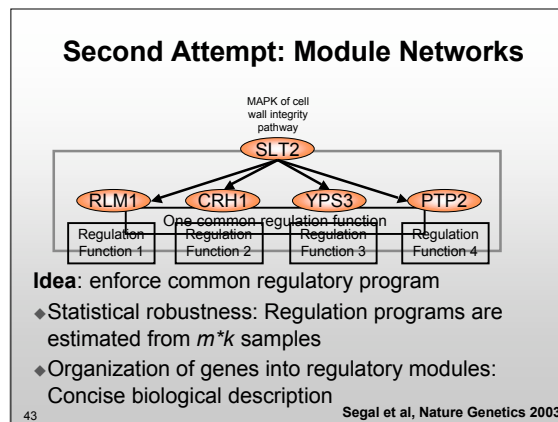
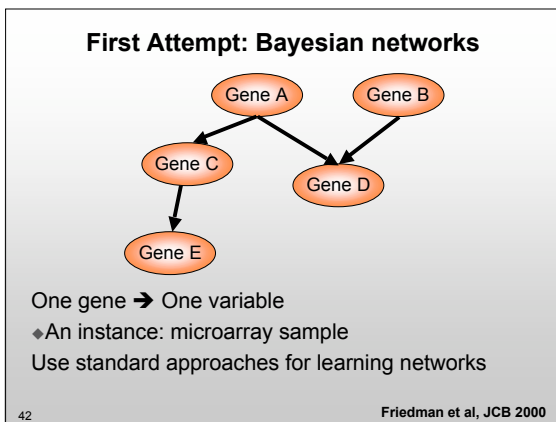
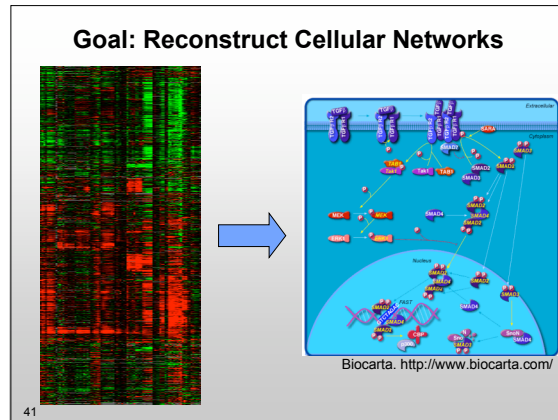
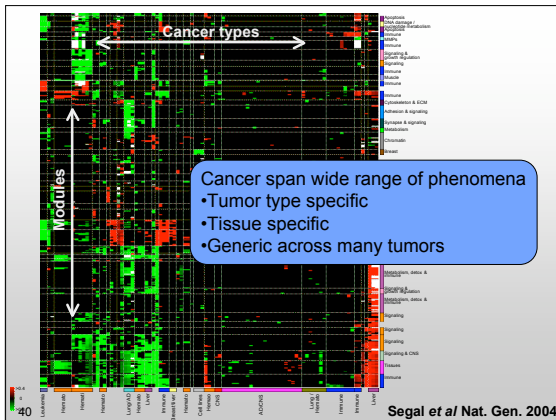
Outline

The diagram illustrates the flow of genetic information. It starts with DNA, represented by a blue double helix and a brown chromosome. A blue arrow points to RNA, shown as a green single helix and highlighted by a red box. Another blue arrow points to Protein, depicted as a 3D molecular model with various colored spheres. A final blue arrow points to Phenotype, shown as a grayscale image of a cell. Labels 'DNA', 'RNA', 'Protein', and 'Phenotype' are placed below their respective images.

- ◆ Copied from DNA template
- ◆ Conveys information (mRNA)
- ◆ Can also perform function (tRNA, rRNA, ...)
- ◆ Single stranded, four nucleotide types (A, C, G, U)
- ◆ For each expressed gene there can be as few as 1 molecule and up to 10,000 molecules per cell.

31





Validation

How do we evaluate ourselves?

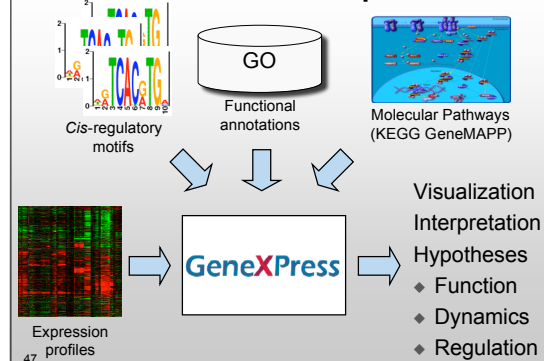
◆ Statistical validation

◆ **Biological interpretation**

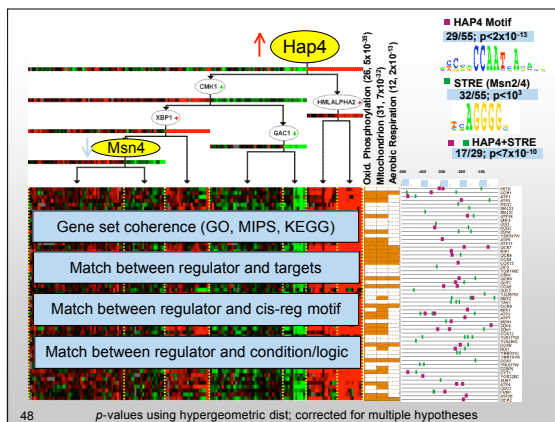
- Annotation database
- Literature reports
- Other experiments, potentially different experiment types

46

Visualization & Interpretation



47



48

Validation

How do we evaluate ourselves?

◆ Statistical validation

◆ Biological interpretation

◆ **Experiments**

- Test causal predictions in the real system
- Lead to additional understanding beyond the prediction
- Experimental validation of three regulators
 - ◆ 3/3 successful results

49

Segal et al, Nature Genetics 2003

Challenges

◆ New methodologies for the huge amount of existing RNA profiles

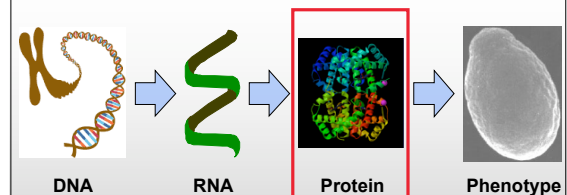
- Meta analysis
- Better mechanistic models
- Contrasting new profiles with existing databases
- Visualization

◆ Other measurements

- Degradation rates
- Localization

50

Outline



- ◆ Proteins are the main executors of cellular function
- ◆ Building blocks are 20 different amino-acid
- ◆ Synthesized from mRNA template
- ◆ Acquires a sequence dependent 3-D conformation
- ◆ Proteomics: Systematic Study of Proteins

51

Why Measure Proteins?

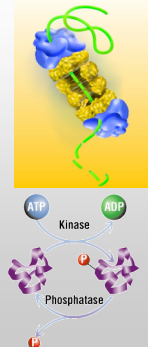
◆ RNA Level ≠ Protein level

Protein quantity is not a direct function of RNA levels

◆ Protein Level ≠ Activity level

Activity of proteins is regulated by many additional mechanisms

- Cellular localization
- Post-translational modifications
- Co-factors (protein, RNA, ...)



52

Challenges in Proteomics

◆ Problematic recognition:

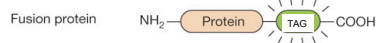
No generic mechanism to detect different protein forms

◆ Thousands of different proteins in the typical cell

◆ Protein abundances vary over several orders of magnitude

53

Making a Protein Generic

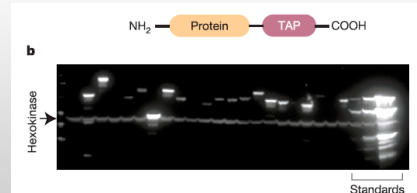


- Tags make a protein generic
- Underlying assumption is that the tag does not change the protein
- All proteins have the same tag
 1. Inability to pool strains
 2. Each experiment is done on a "different" strain

54

TAP-Tag Libraries for Abundance

~4500 Yeast strains have been TAP tagged



- How much is each protein expressed?
- What is the proteome under different conditions?

55

Why Study Protein Complexes?

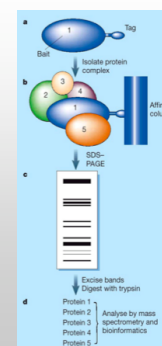
◆ # Most proteins in the cell work in protein complexes or through protein/protein interactions

◆ # To understand how proteins function we must know:

- ◆ - who they interact with
- when do they interact
- where do they interact
- what is the outcome of that interaction

56

Using TAP-Tag to Find Complexes



Large Scale Pull Downs Provide Information on Protein Complexes

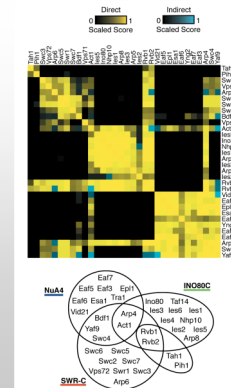
The Venn diagram illustrates the overlap of protein-protein interactions (PPI) between two studies: Gavin et al. (6532) and Krogan et al. (7123). The diagram is set against a light blue background. The top circle, representing the 'Consolidated PPI subset (9074)', is purple and contains 4456 unique interactions. The bottom-left circle, representing 'PPI Gavin et al. (6532)', is green and contains 2963 unique interactions. The bottom-right circle, representing 'PPI Krogan et al. (7123)', is teal and contains 4512 unique interactions. The overlapping regions are: 2036 interactions shared between Gavin et al. and the consolidated subset; 1078 interactions shared between Krogan et al. and the consolidated subset; 1504 interactions shared between both studies; and 29 interactions shared between both studies but not in the consolidated subset.

Region	Count
Consolidated PPI subset (9074) only	4456
PPI Gavin et al. (6532) only	2963
PPI Krogan et al. (7123) only	4512
Overlap: Gavin et al. & Consolidated subset	2036
Overlap: Krogan et al. & Consolidated subset	1078
Overlap: Gavin et al. & Krogan et al.	1504
Overlap: All three studies	29

- Both labs used the same proteins as bait
- Each lab got slightly different results
- The results depended dramatically on analysis method

**Gavin et al. Nature 2006*

**Krogan et al. Nature 2006*




*Gavin et al. Nature 2006

*Krogan et al. Nature 2006

59

We can now define a yeast “interactome”



- Isn't full use of data
- Static picture

- Isn't full use of data
- Static picture

Making a Protein Generic

Fusion protein

NH_2 —Protein—GFP— COOH

1. Fluorescent proteins allow us to visualize the proteins within the cell.
2. Allow us to measure individual cells and the variation/ noise within a population

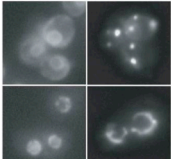


1. Fluorescent proteins allow us to visualize the proteins within the cell.
2. Allow us to measure individual cells and the variation/ noise within a population

61

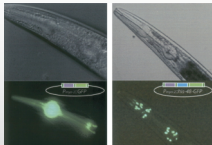
Cellular Localization Using GFP Tags

What can it teach us?



A library of yeast GFP fusion strains has been used to localize nearly all yeast proteins

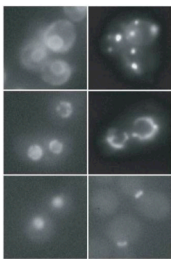
Huh *et al* Nature 2003



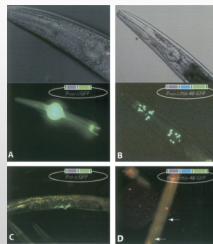
A collection of cloned *C. elegans* promoters is being created for similar purposes

Genome Research 14:2169-2175, 2004

62



A library of yeast GFP fusion strains has been used to localize nearly all yeast proteins


Huh et al *Nature* 2003

A collection of cloned *C. elegans* promoters is being created for similar purposes

Genome Research 14:2169-2175, 2004

Challenges in Fluorescence-based Approaches

- ◆ Better Vision processing will allow to do this in High-Throughput and answer questions like:
 - Changes in localization in response to cellular cues
 - Changes in localization in response to environment cues
 - Changes in localization in various genetic backgrounds
 - Dynamics of localization changes



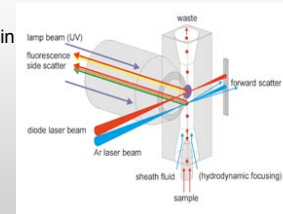
- ◆ Better Vision processing will allow to do this in High-Throughput and answer questions like:
 - Changes in localization in response to cellular cues
 - Changes in localization in response to environment cues
 - Changes in localization in various genetic backgrounds
 - Dynamics of localization changes



THROUGHPUT THE MAJOR BOTTLENECK

Single Cell Measurements: Flow Cytometry

- Cells pass through a flow cell one at a time
- Lasers focused on the flow cell excite fluorescent protein fusions
- Allows multiple measurements (cell size, shape, DNA content)

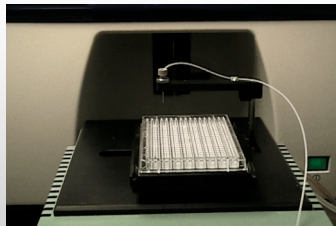


Applications:

- Protein abundance
- Protein-protein interactions
- Single-cell measurements

65

High Throughput Flow Cytometer



- 7 seconds/sample
- ~50,000 counts per sample

66

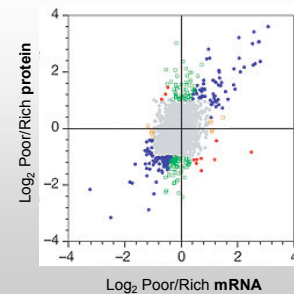
Comparison of mRNA to Protein Levels Allows Identification of Post-transcriptional Regulation

Compare

- Rich media
- Poor media

Observed behaviors

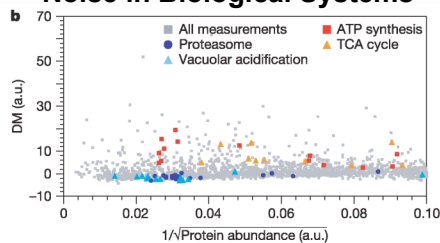
- No change in both
- Coordinated change
- Change in protein, but not mRNA



Newman et al Nature 2006

67

Noise in Biological Systems



- Measurement of 10,000 individual cells allows measurement of variation (noise) in a biological context
- factors that affect levels of noise in gene expression:
 - Abundance, mode of transcriptional regulation, sub-cellular localization

68

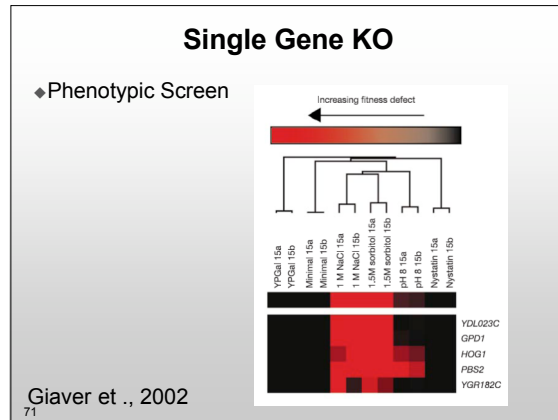
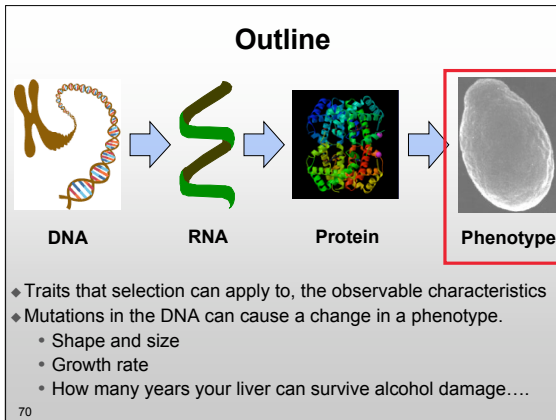
Nature 441, 840-846 (15 June 2006)

Challenges

Proteomics is in its infancy - easier to make an impact

- Integrating this data with other proteomic/genomic data to better predict protein function
- Higher Throughput methods such as flow cytometry will allow generation of varied data: Different growth conditions, Cell cycle, Stress, Mating
- Tagging in mammalian cells becoming more feasible - near future should bring proteomic data on human cells

69



Starting to Probe the Cellular Network

Genetic Interaction

- The effect of a mutation in one gene on the phenotype of a mutation in a second gene
- Different type of interaction - not physical

72

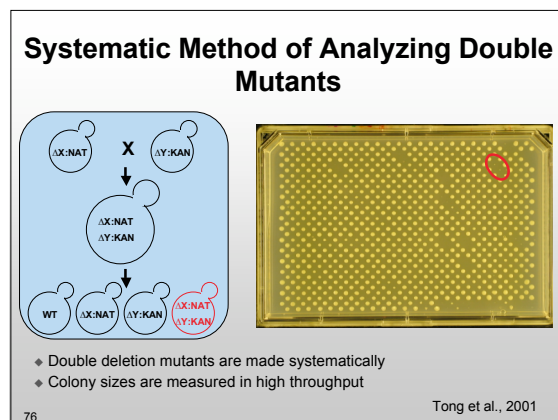
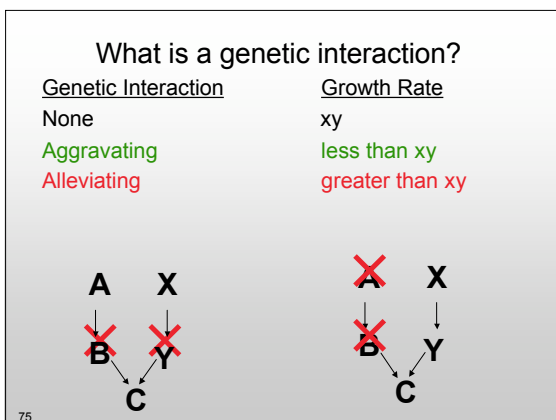
What is a genetic interaction (Epistasis)?

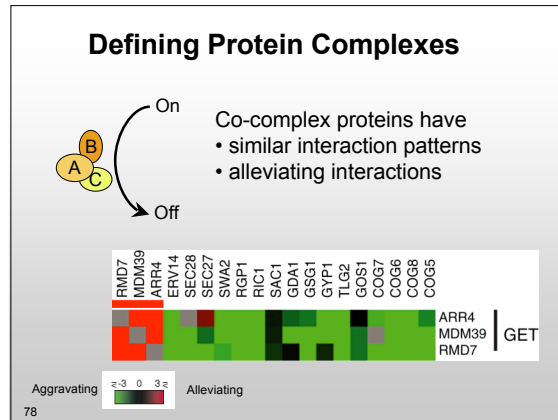
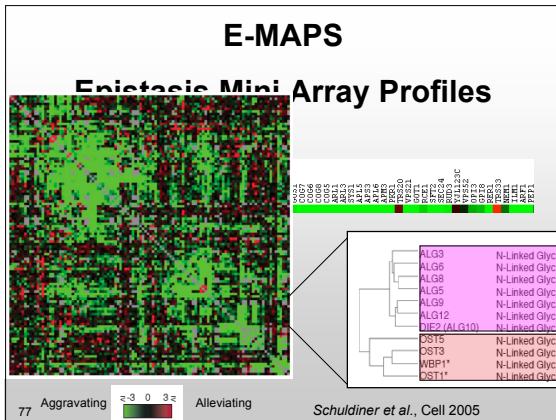
The effect of a mutation in one gene on the phenotype of a mutation in a second gene.

Genotype	Growth Rate
WT	1
$\Delta geneA$	x ($x \leq 1$)
$\Delta geneB$	y ($y \leq 1$)
$\Delta geneA \Delta geneB$	xy (Product)

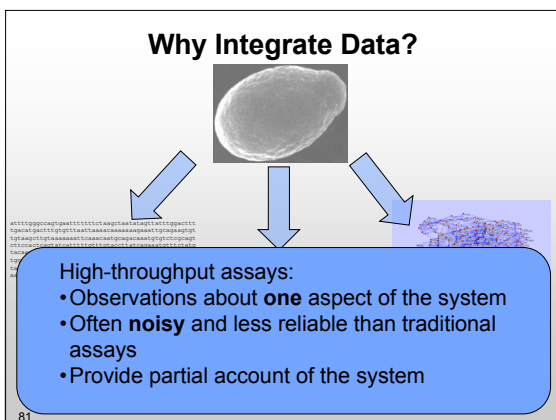
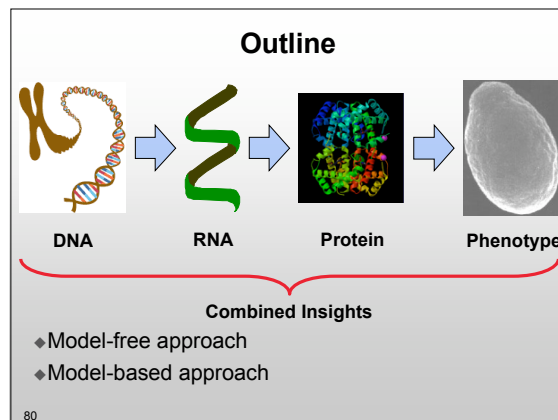
DIFFERENT TYPE OF INTERACTION - NOT PHYSICAL

74





- ### Challenges for the future
- ◆ Only a small fraction of the information has been utilized in E-MAPS made so far
 - ◆ E-MAPS to cover all yeast cellular processes to come out until the end of 2007
 - ◆ Extending this to human cells is now feasible using gene silencing techniques
 - ◆ Amount of data scales exponentially - Higher organisms - more genes

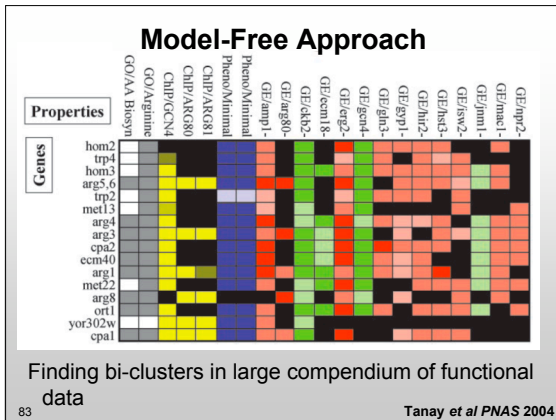


Model-Free Approach

Gene	Location			Expression		Phenotype		Binding sites		
	Nuc	Cyto	Mito	Rich	Poor	Salt	Kan	RAP1	HSF1	GCN4
YAL001C										
YAL002W										
YAL003W										
YAR040W										
YAR041C										

- ◆ Treat different observations about elements as multivariate data
 - Clustering
 - Statistical tests

82



Model-Free Approach

Pros:

- ◆ No assumptions about data
 - Unbiased
 - Can be applied to many data types
- ◆ Can use existing tools to analyze combined data

Cons:

- ◆ No assumptions about data
 - Interpretation is post-analysis
 - No sanity check
- ◆ Cannot deal with data from different modalities (interactions, other types of genetic elements)

84

Model-Based Approach

What is a model?

“A description of a process that could have generated the observed data”

- Idealized, simplified, cartoonish
- Describes the system & how it generates observations

85

Explaining Expression

Key Question:

- ◆ Can we **explain** changes in expression?

General concept:

- ◆ Transcription factor binding sites in promoter region should “explain” changes in transcription

86

Explaining Expression

Relevant data:

- ◆ Expression under environmental perturbations
- ◆ Expression under transcription factors KOs
- ◆ Predicted binding sites of transcription factors
- ◆ Protein-DNA interactions of transcription factors
- ◆ Protein levels/location of transcription factors

87

A Stab at Model-Based Analysis

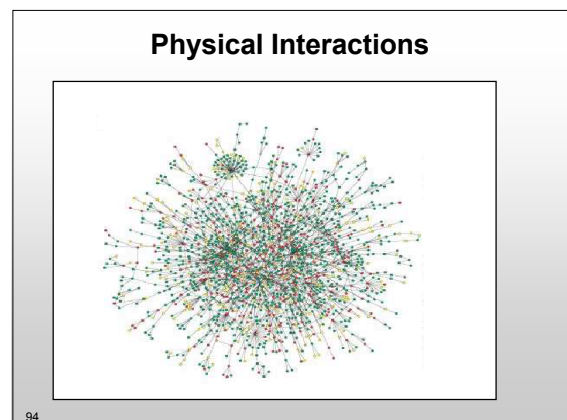
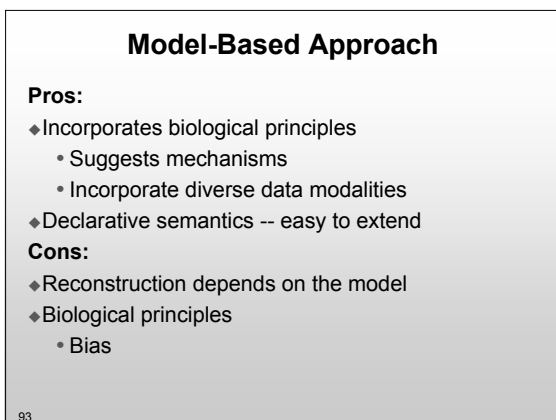
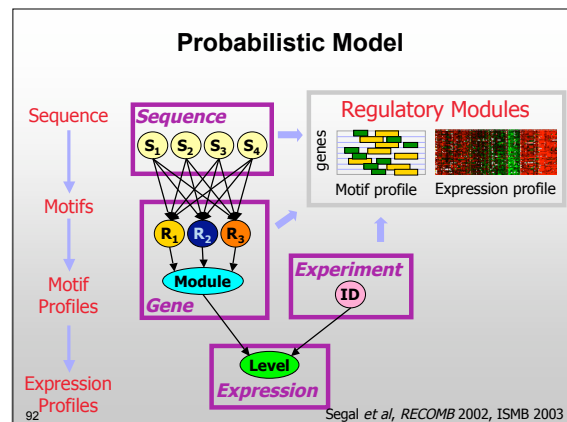
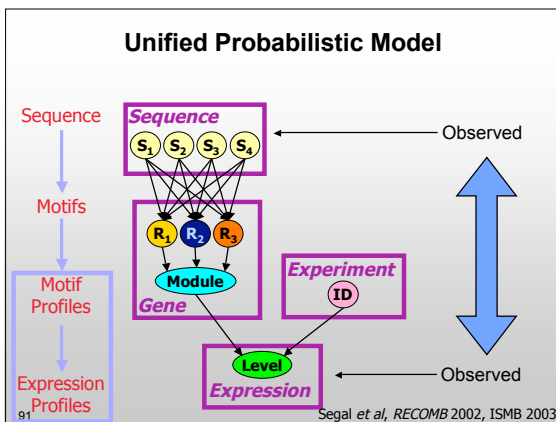
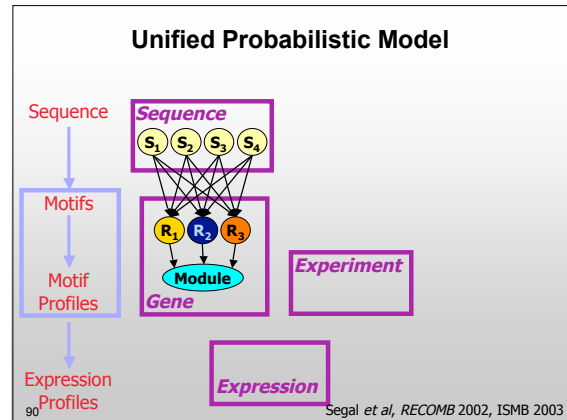
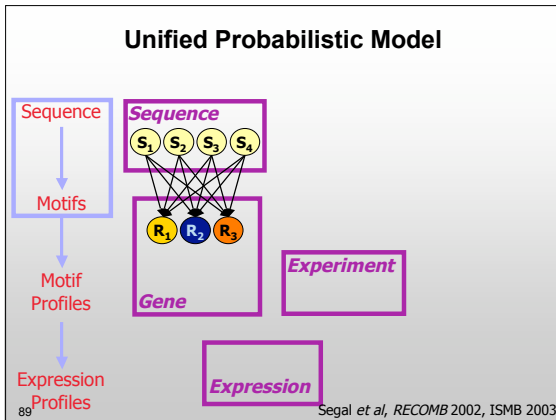
Sequence

Motifs

Motif Profiles

Expression Profiles

88



Physical Interactions

Interaction between two proteins makes it more probable that they

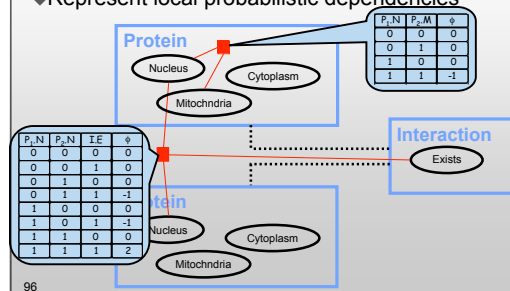
- share a function
- reside in the same cellular localization
- their expression is coordinated
- have similar genetic interactions
- ...

Can we exploit this to make better inference of properties of proteins?

95

Relational Markov Network

- ◆ Probabilistic patterns hold for all groups of objects
- ◆ Represent local probabilistic dependencies



96

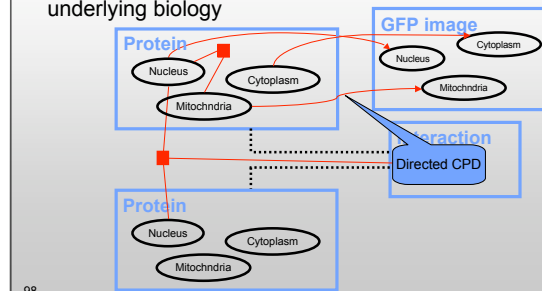
Relational Markov Network

- ◆ Compact model
- ◆ Allows to infer protein attributes by combining
 - Interaction network topology (observed)
 - Observations about neighboring proteins

97

Adding Noisy Observations

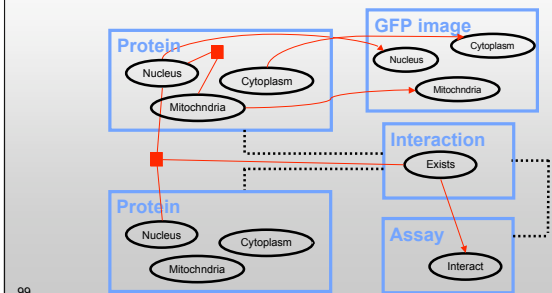
- ◆ Add class for experimental assay
- ◆ View assay result as stochastic function (CPD) of underlying biology



98

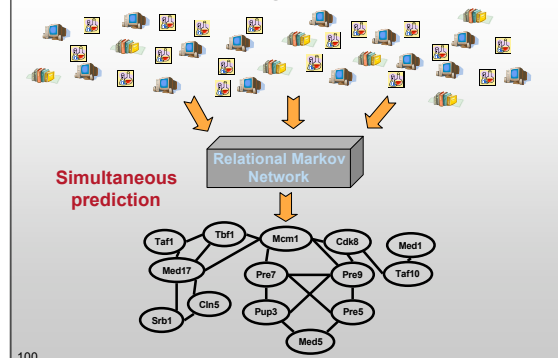
Uncertainty About Interactions

- ◆ Add interaction assays as noisy sensors for interactions



99

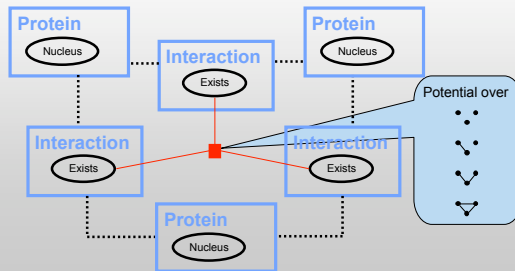
Design Plan



100

Relational Markov Network

- ◆ Add potentials over interactions



101

Relational Markov Models

Combine

- ◆ (Noisy) interaction assays
- ◆ (Noisy) protein attribute assays
- ◆ Preferences over network structures

To find a coherent prediction of the interaction network

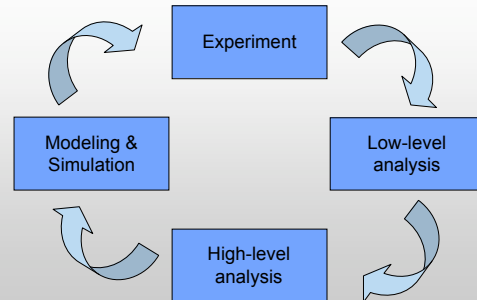
102

Discussion

- ◆ Every day papers are published with high-throughput data that is not analyzed completely or not used in all ways possible
- ◆ The bottlenecks right now are the time and ideas to analyze the data

104

The Need for Computational Methods



106

What are the Options?

- ◆ **Analyze published data**
 - Abundant, easy to obtain
 - Method oriented
 - Don't have to bump into biologists
 - Two million other groups have that data too
- ◆ **Collaborate with an experimental group**
 - Be involved in all stages of project
 - Understand the system and the data better
 - Have priority on the data
 - Involved in generating & testing biological hypotheses
 - Goal oriented
- ◆ **Start your own experimental group...**(yeah, sure)

107

Questions to Keep in Mind

Crucial questions to ask about biological problems

- ◆ **What quantities are measured?**
Which aspects of the biological systems are probed
- ◆ **How are they measured?**
How this measurement represents the underlying system? Bias and noise characteristics of the data
- ◆ **Why are these measurements interesting?**
- ◆ **Which conclusions will make the biggest impact?**

108

Acknowledgements

Slides:

The Computational Bunch

- Yoseph Barash
- Ariel Jaimovich
- Tommy Kaplan
- Daphne Koller
- Noa Novershtern
- Dana Pe'er
- Itsik Pe'er
- Aviv Regev
- Eran Segal

The Biologist Crowd

- David Breslow
- Sean Collins
- Jan Ihmels
- Nevan Krogan
- Jonathan Weissman

Special thanks: Gal Elidan, Ariel Jaimovich

109