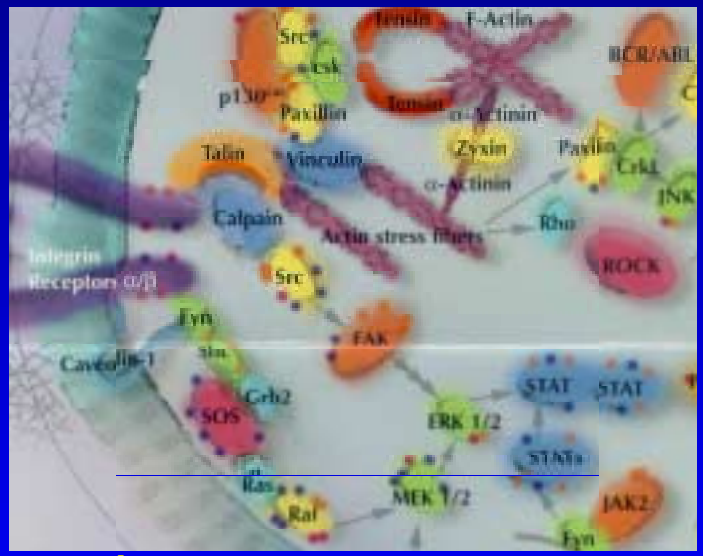
Bayesian Decomposition Expression to Pathways

Michael Ochs

Cancer Biology



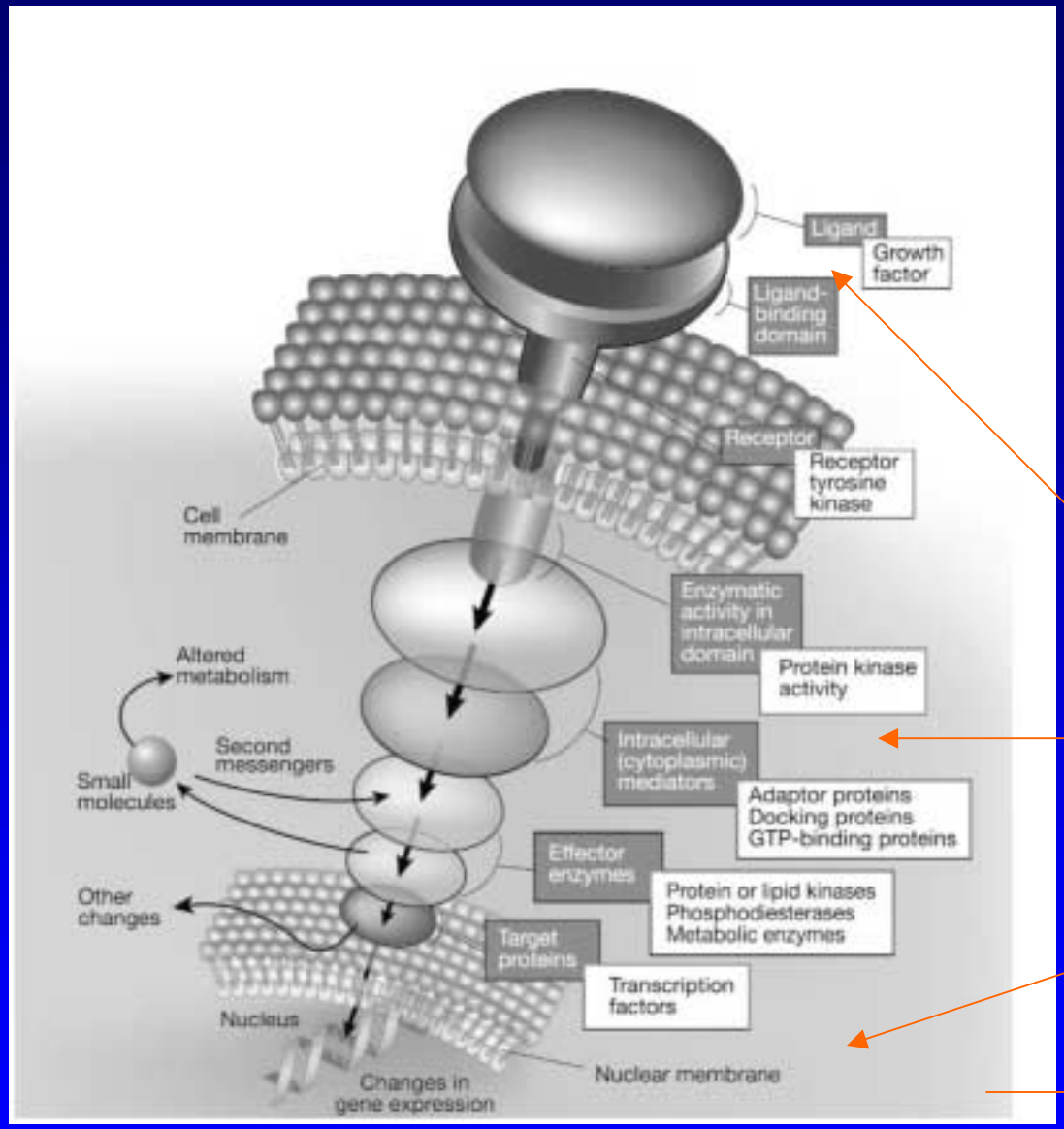
www.biosource.com

Cancer is many Diseases but with a Single Theme

- a cell becomes immortal
- a cell becomes mobile

Signalling and Metabolic Pathways Hold the Key

Signalling Pathways



Stimulus

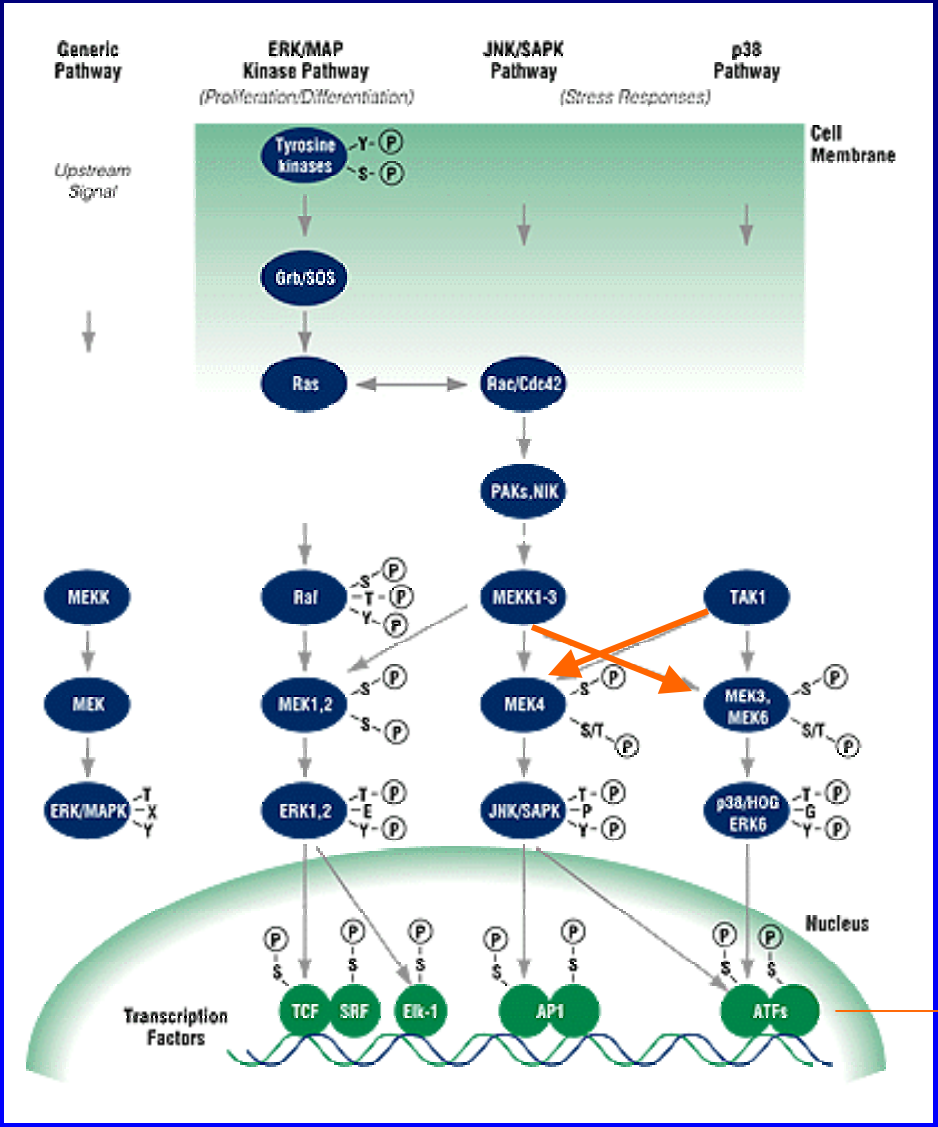
Signal Transduction

Transcription

mRNA

Downward, *Nature*, 411, 759, 2001

Identifying Pathways



Interacting Pathways Lead to Confusion if All Genes Need to Lie in a Single Cluster

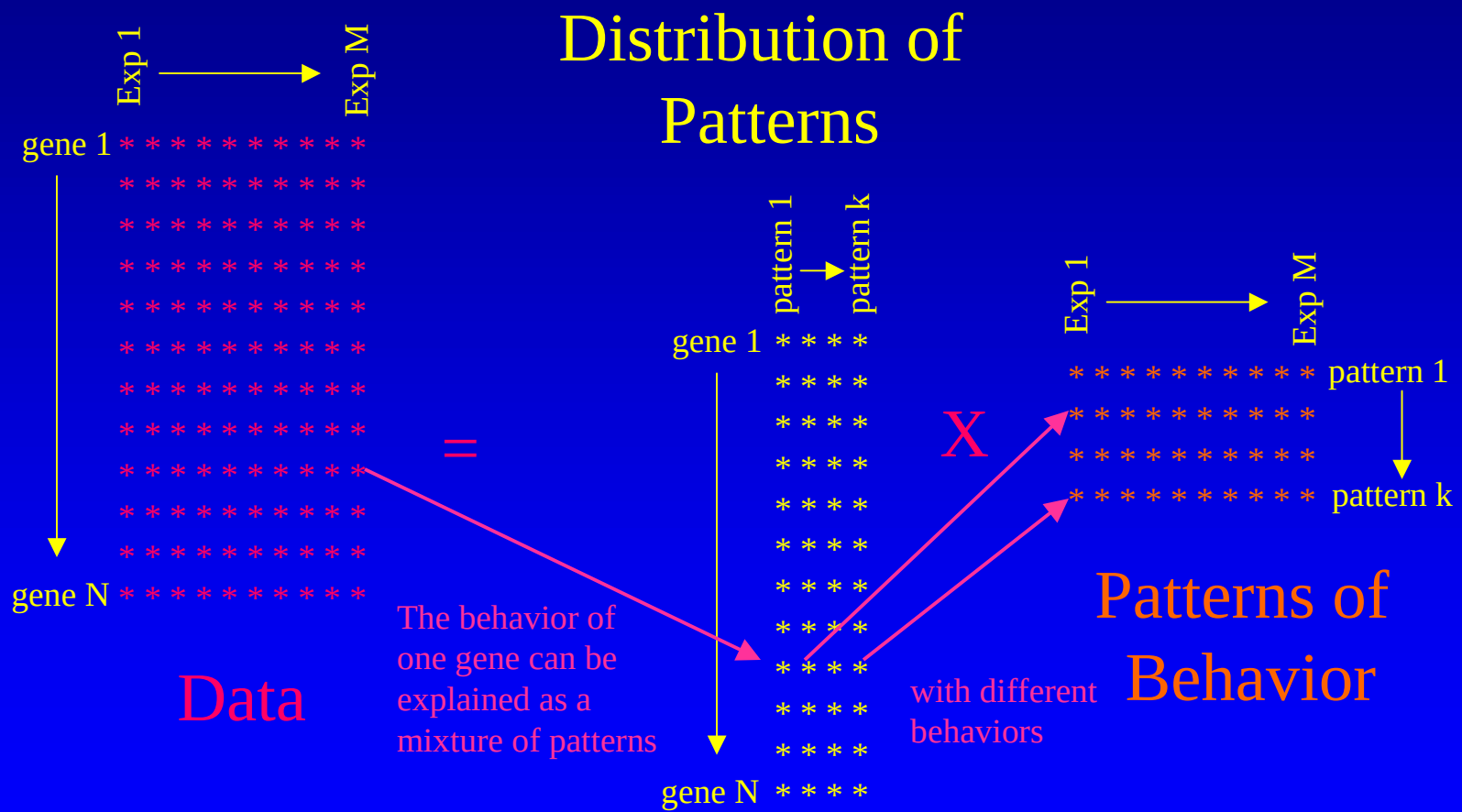
mRNA

www.promega.com

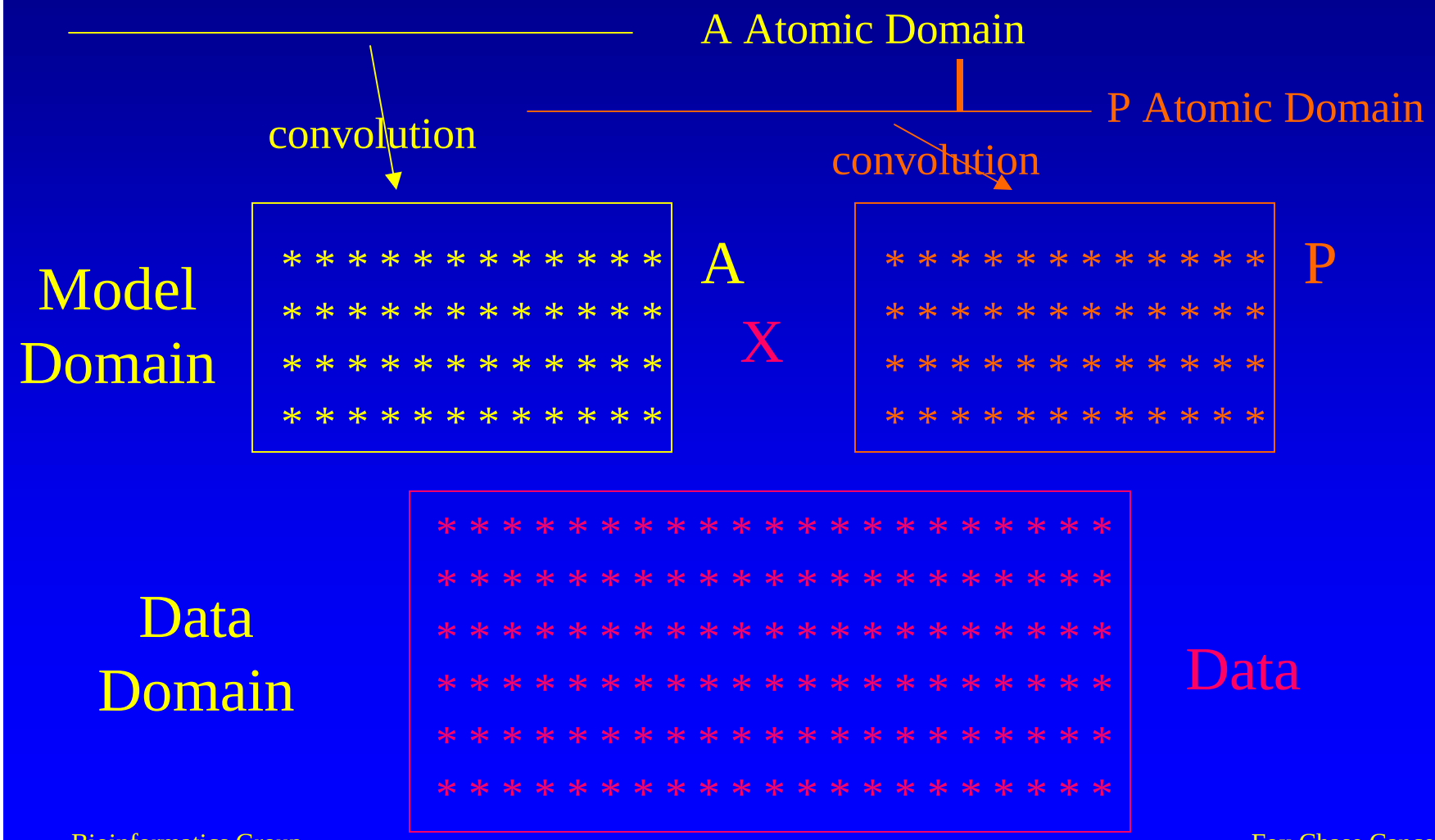
Bayesian Decomposition

- Data Mining/Pattern Recognition Algorithm
  - Unsupervised Method
  - Create Multiple, Overlapping “Clusters”
    - Each Gene can be in Multiple Patterns
    - Get to Pathways: Key for Cancer Development
- Methodology
  - Markov Chain Monte Carlo Algorithm
  - Simulated Annealing
  - Integration of Prior Knowledge

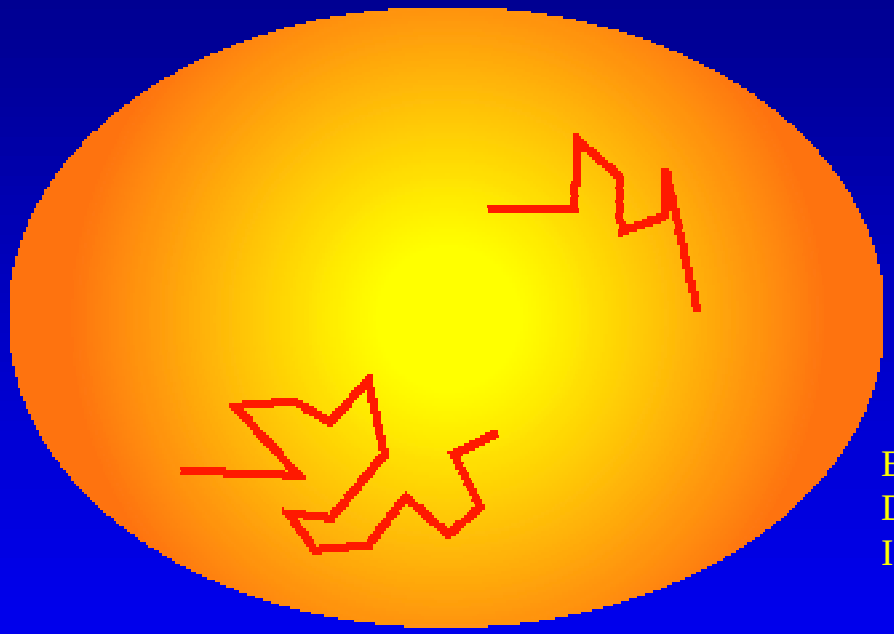
BD: Matrix Decomposition



BD: Domains


A Atomic Domain
P Atomic Domain
convolution
convolution
Model Domain
A
X
P
Data Domain
Data

BD: Markov Chain MC



Based on Maximum Entropy
Data Consultants Massive
Inference Sampler

Cloud in N-Dimensional Space: Probability Density
for the Model Results from Atomic Domain Prior,
Model Functions (Prior), and the Likelihood

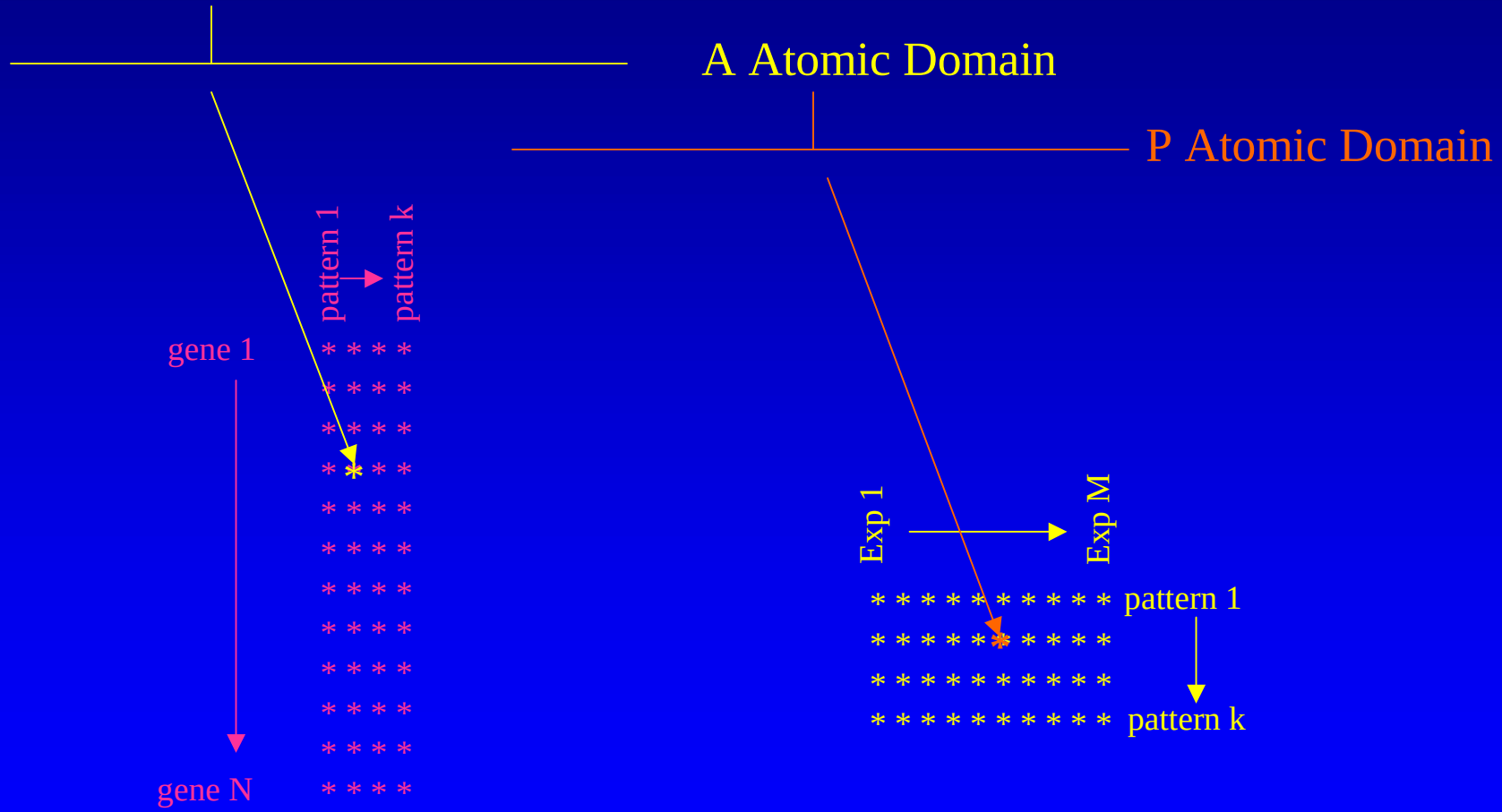
BD Requirements

- Data Points > (A + P) Points
- Atomic Domains (Sibisi and Skilling, *J R Stat Soc B*, **59**, 217, 1997)
  - Positive Additive Distributions
  - Infinitely Divisible Process
- Model Domains
  - Linked to Atomic Domains by Model Function
  - Correlations between Parameters are Introduced by Model Functions (Atomic > Model)

BD Features

- Basis Vectors (Patterns) are Nonorthogonal
  - Physically Meaningful if Good Model
  - Artifacts Removed if Do Not Fit Model
- Noise is Treated
  - Noise is Integral Part of Fitting Process
  - Artifacts Often Appear in Residuals (i.e. noise)
- Markov Chain Sampling Yields
  - Mean of Probable Distributions and Patterns
  - Uncertainties for Distributions and Patterns

BD: Gene Expression



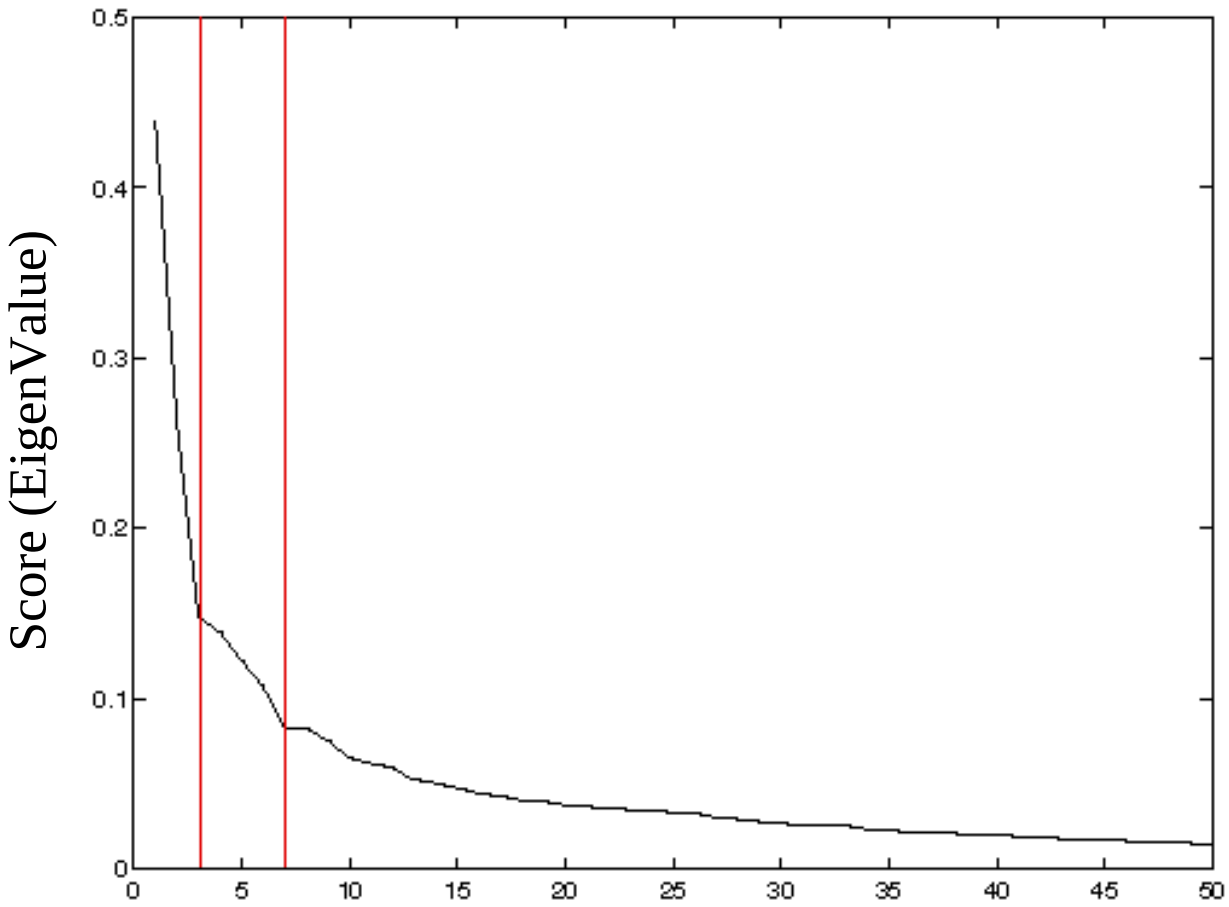
Rosetta Data Set

- Filtering
  - Eliminate Genes
    - >25% Data Missing in Ratios or Uncertainties
    - < 2 Experiments with 3 Fold Change
  - Eliminate Experiments
    - < 2 Genes Changing by 3 Fold
- Uncertainties
  - Used Values from Rosetta Error Model
  - Missing Data Log Ratio =1, Log Unc = 100

Analysis

- Analyzed Full Experimental Data with PCA
  - Estimate of Dimensionality of Data
- Bayesian Decomposition
  - Filtered Data: 764 Genes, 228 Experiments
  - Ran Multiple Seeds, Multiple Pattern Number
  - Focus on Dimensions Suggested by PCA
- Data Driven
  - Let Analysis Determine Where to Look

PCA Results


Score (EigenValue)
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
30
35
40
45
50
Principal Component

Bayesian Decomposition

- Distributions
  - Assignment of Genes to Patterns
- Patterns
  - Each Pattern Defines Behavior Across Experiments
- Experimental Patterns
  - Experiments explained by a single pattern
  - Correlations between experiments
- Genes in Patterns
  - Identify biological processes
  - Identify correlations in genes

Experiments High in One Pattern

- Pattern 1
  - YHR034C 56%
- Pattern 2
  - rpd3 89%
- Pattern 3
  - ssn6 (cyc8) 76%
  - YER024W 56%
  - tup1 54%
- Pattern 5
  - YJL107C 53%
  - yap3 51%

Genes in Patterns

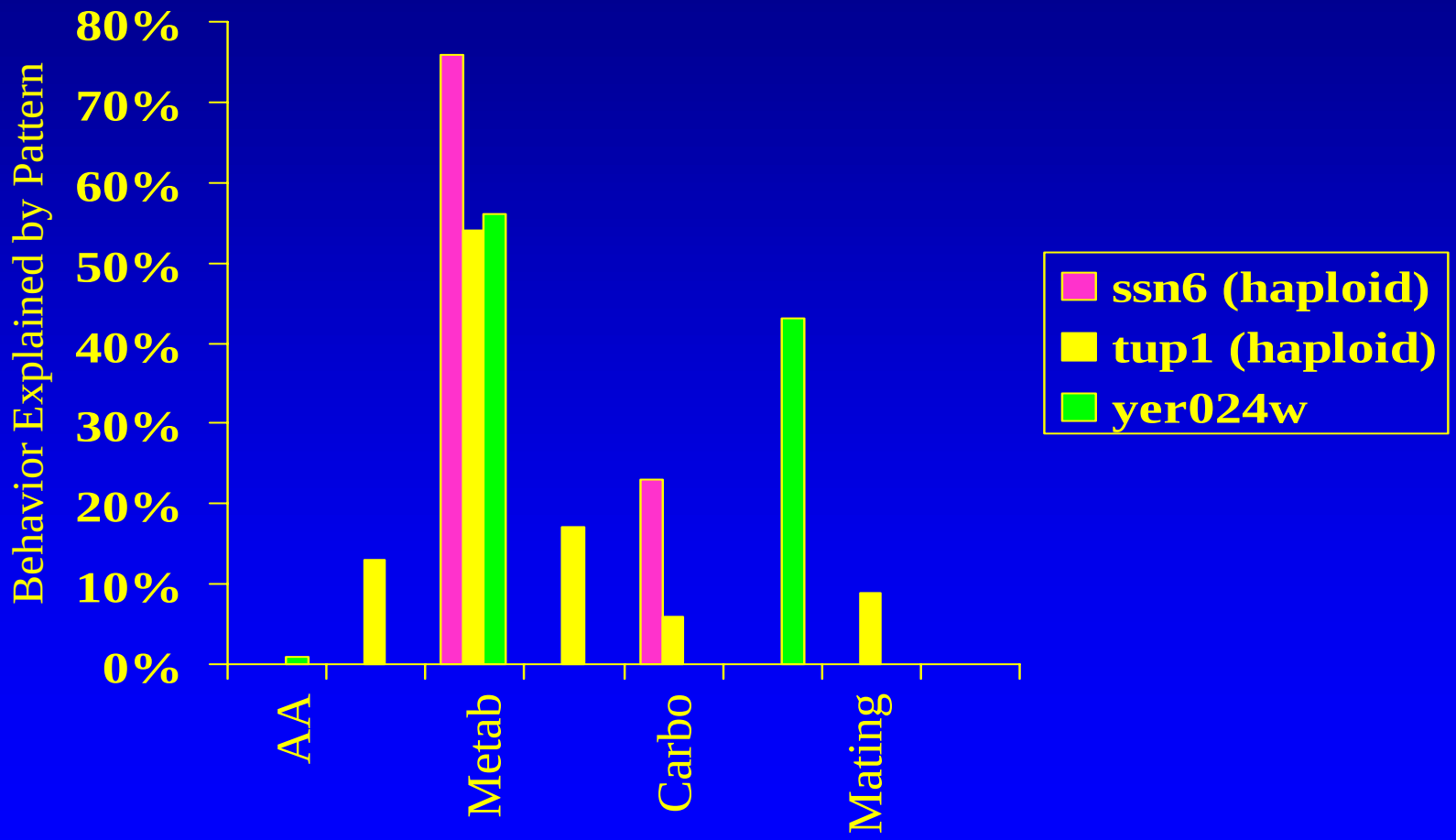
(Proteome Database Cellular Role)

- Pattern 1 AA Pattern
  - 403 Genes
  - 22/36 AA metabolism
  - 9 additional metabolism
- Pattern 2
  - 410 Genes
  - 7/27 metabolism
  - 7/27 DNA/RNA processing
  - 6 transport
- Pattern 3 Metabolic Pattern
  - 390 Genes
  - 13/26 metabolism
  - 6 transport, 4 Pol II
- Pattern 4
  - 276 Genes, 30/50 Unknown
- Pattern 5 Carbo Pattern
  - 355 Genes
  - 14/37 carbohydrate metabolism
  - 7/37 cell stress
  - 6 transport
- Pattern 6
  - 297 Genes, 30/50 unknown
- Pattern 7 Mating Pattern
  - 223 Genes
  - 13/23 mating response
  - 5/23 meiosis

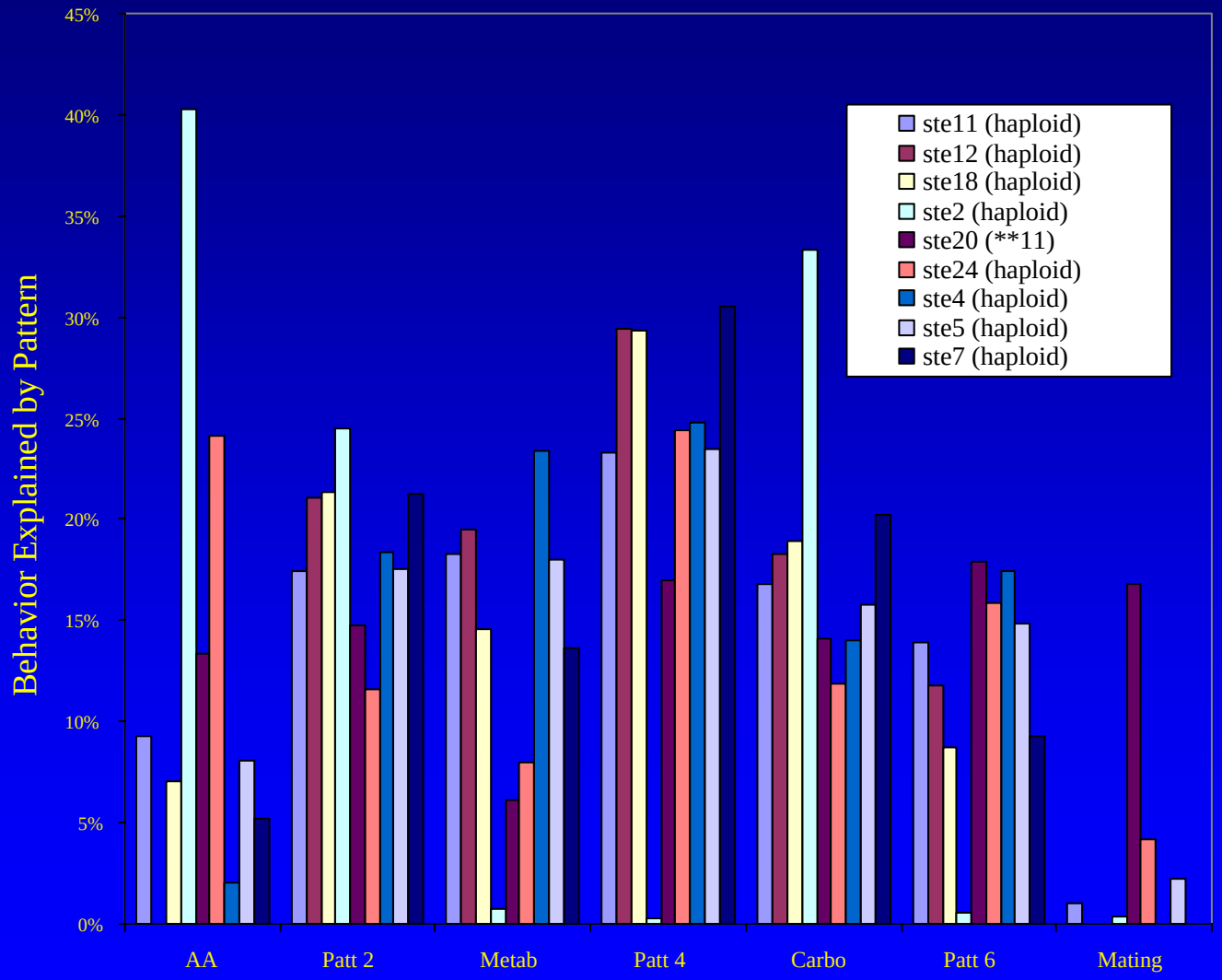
Metabolic Patterns

- Patterns 1 and 5
  - yap 3 98%
  - YJL107C 98%
  - YHR034C 98%
  - FR901,228 98%
- Patterns 1, 3, and 5
  - ssn6 100%
  - swi6 99%
  - yap 3 98%
  - YJL107C 98%
  - YHR034C 98%
  - FR901,228 98%

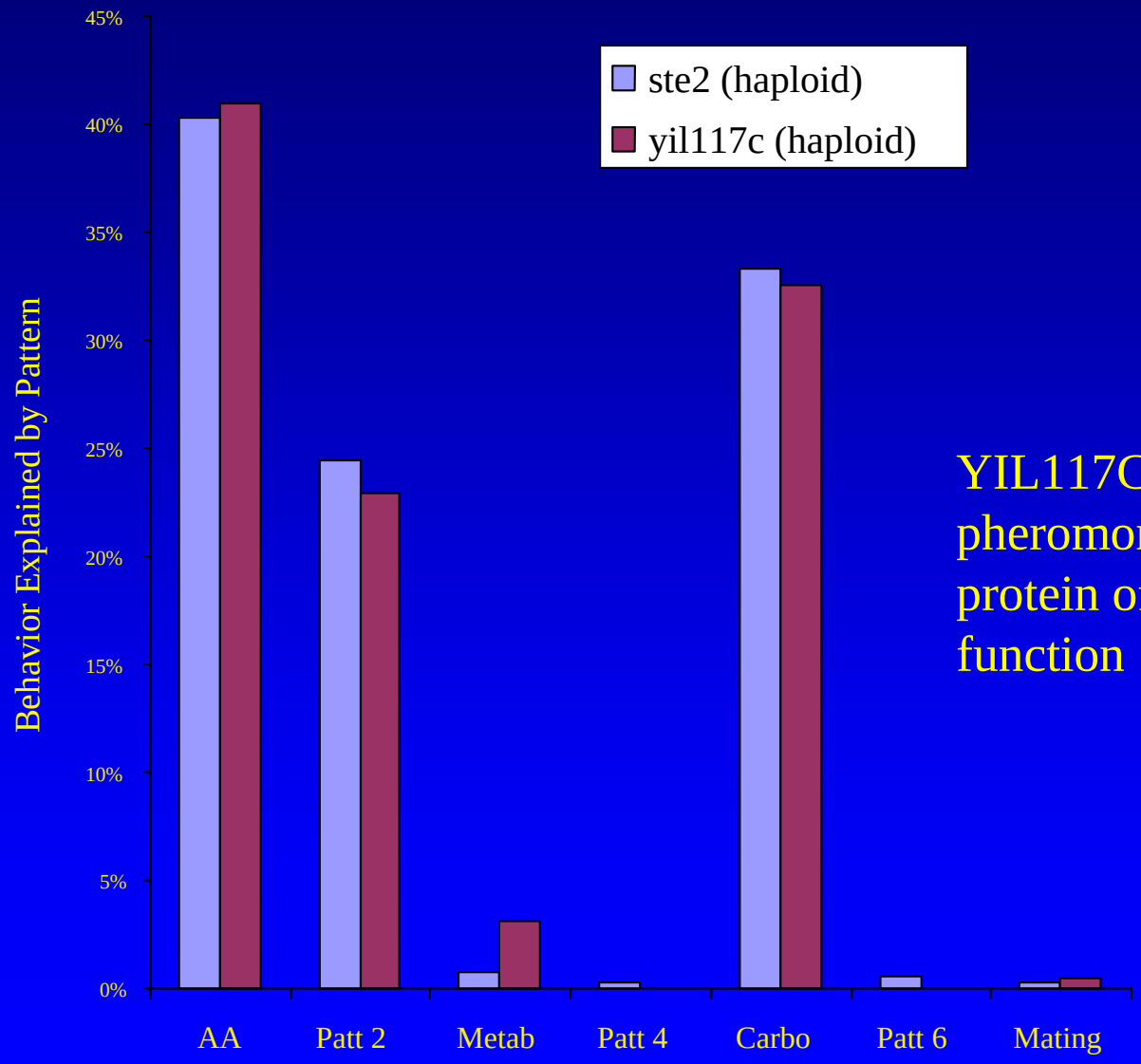
Metabolic Pattern



Sterile Family Proteins

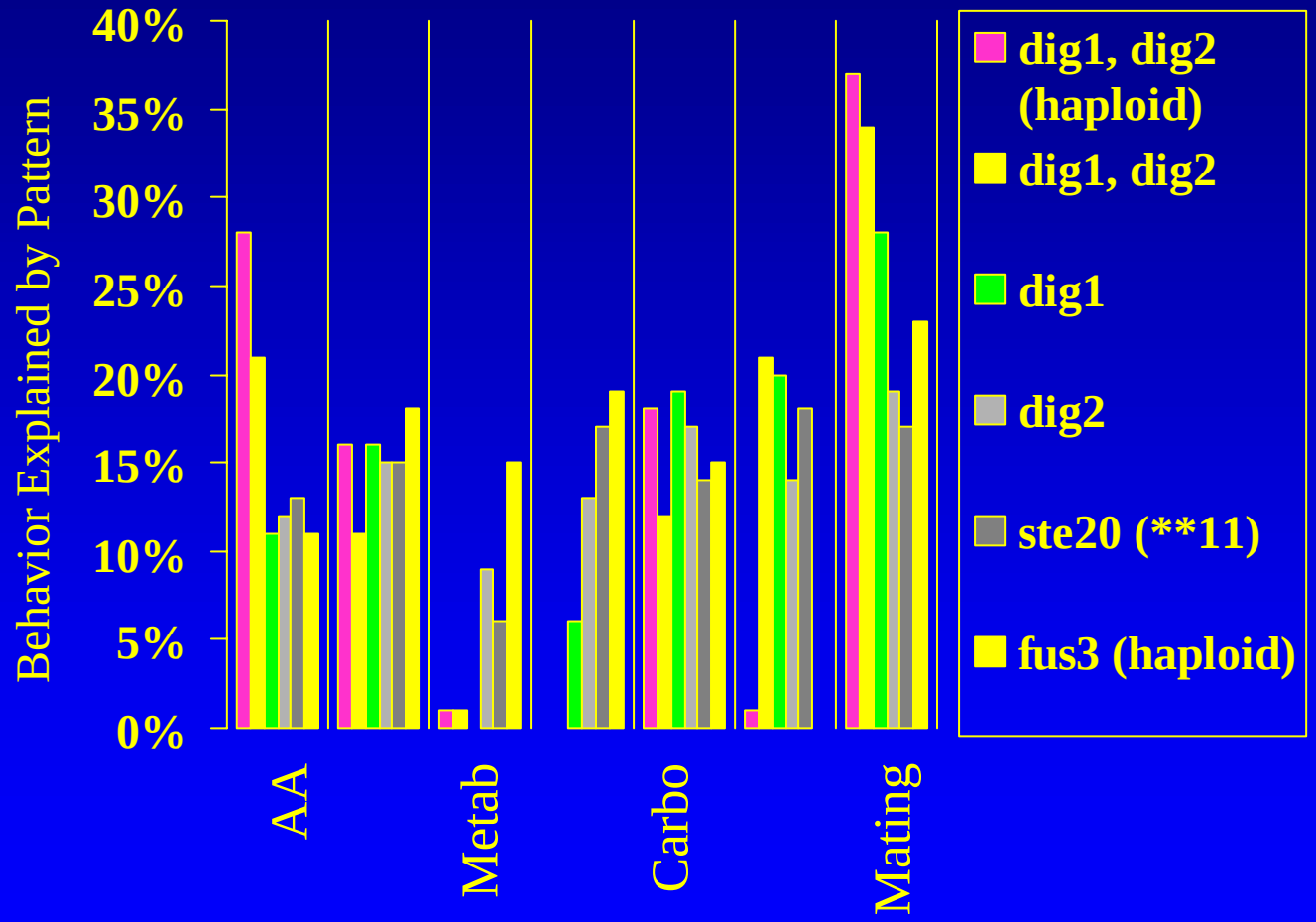


Ste2

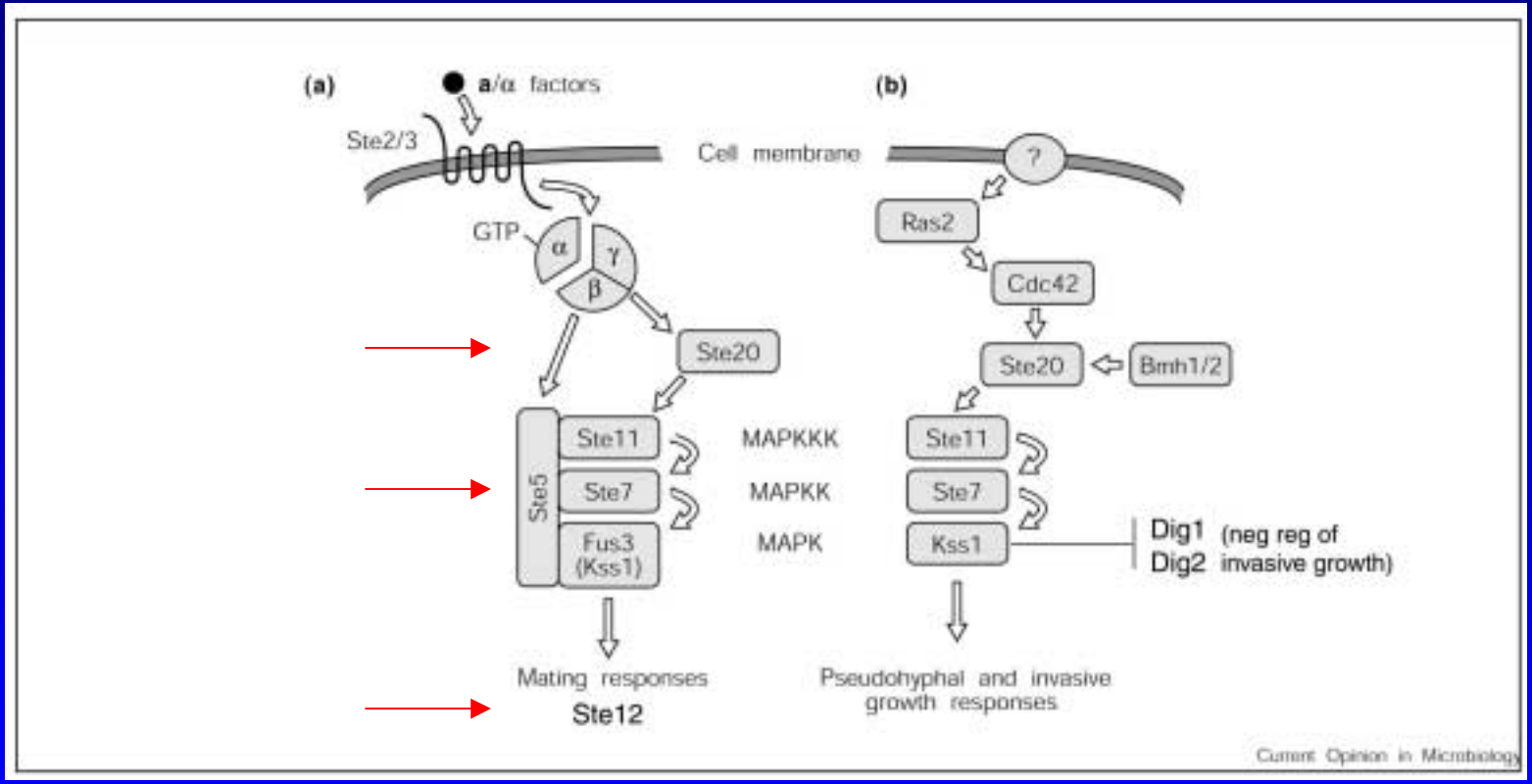


YIL117C is prm5, a pheromone regulated protein of unknown function

Mating Pattern

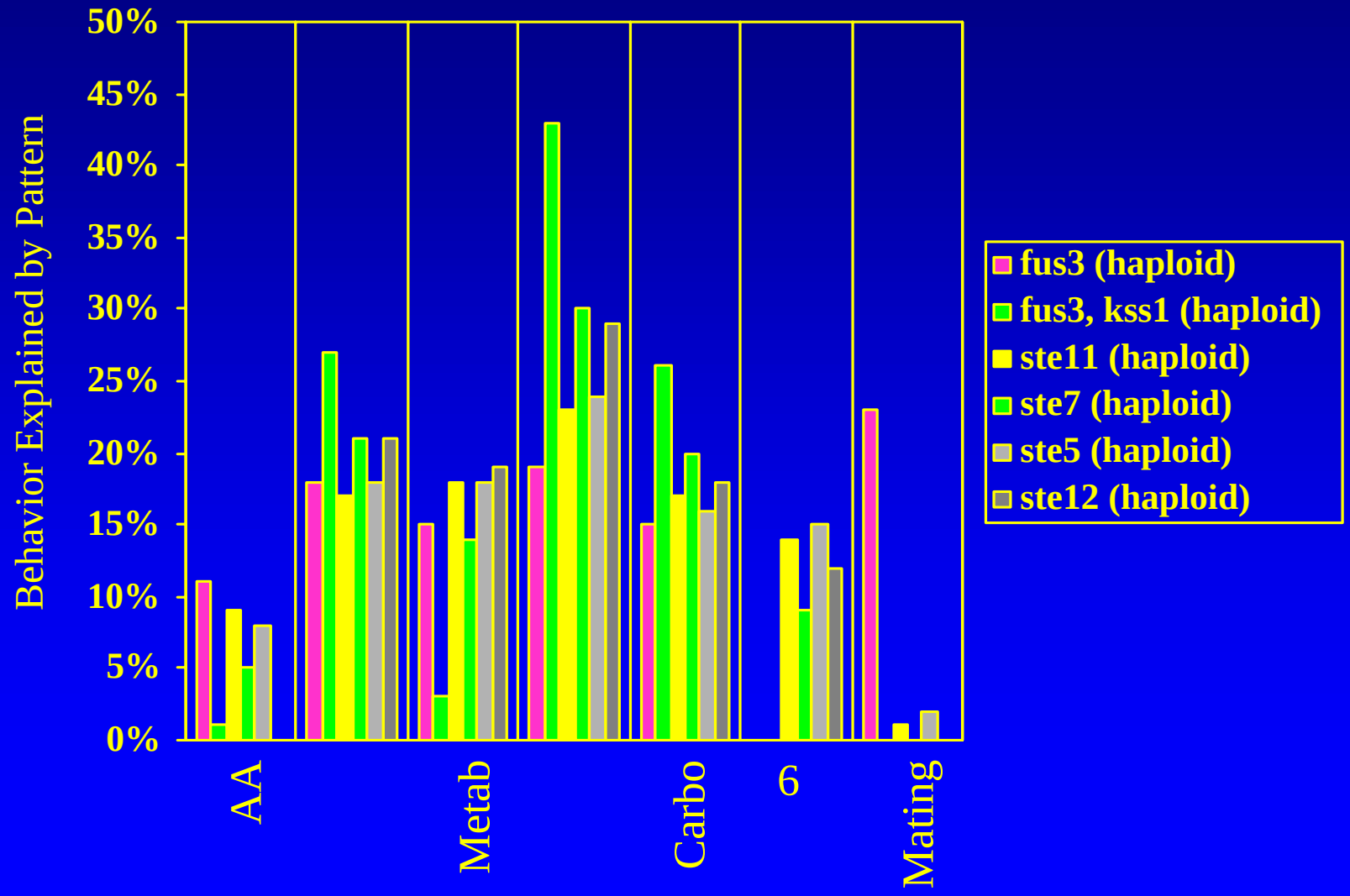


Mating Pathway

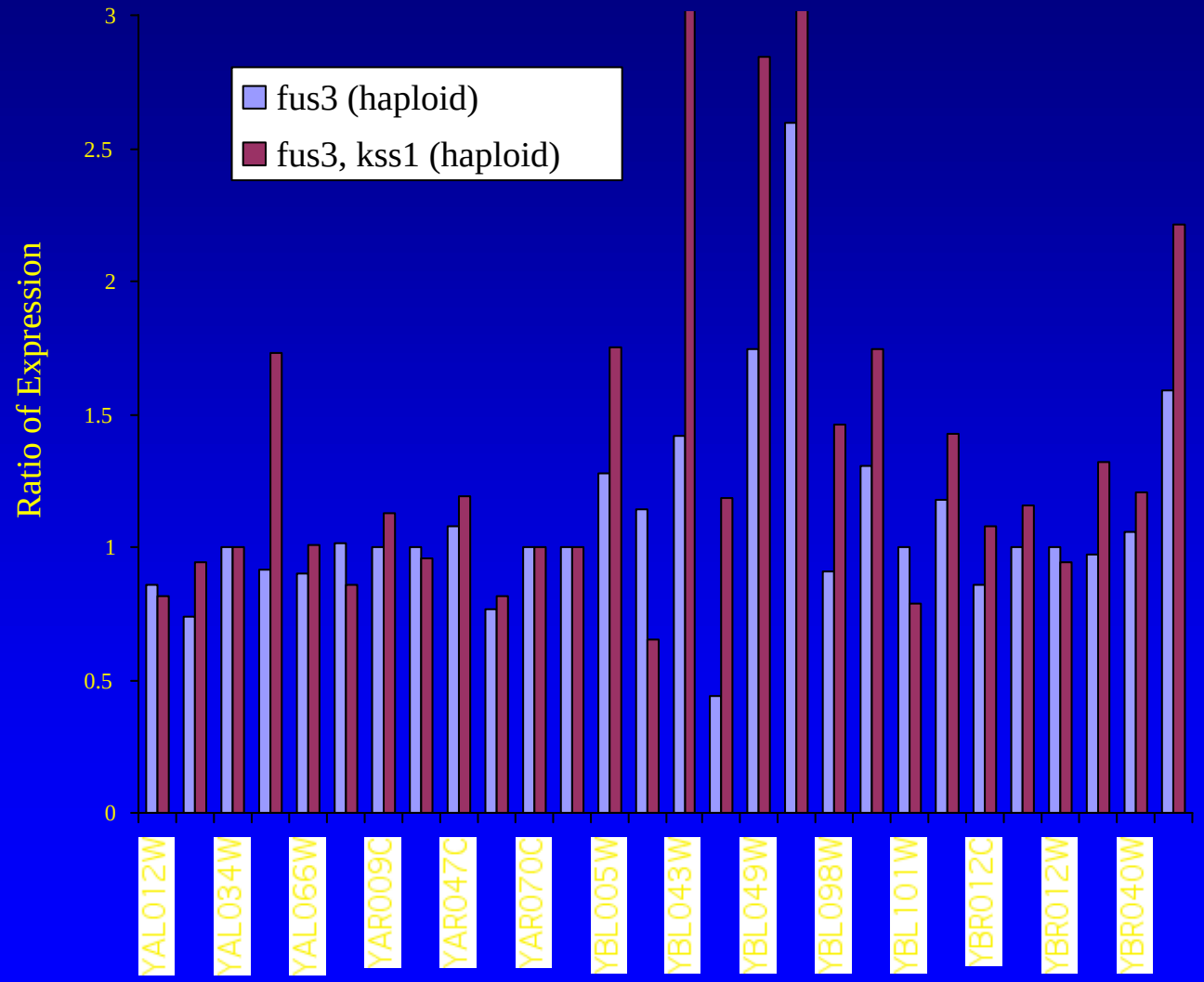


Posas, et al, Curr Opin Microbiology, 1, 175, 1998

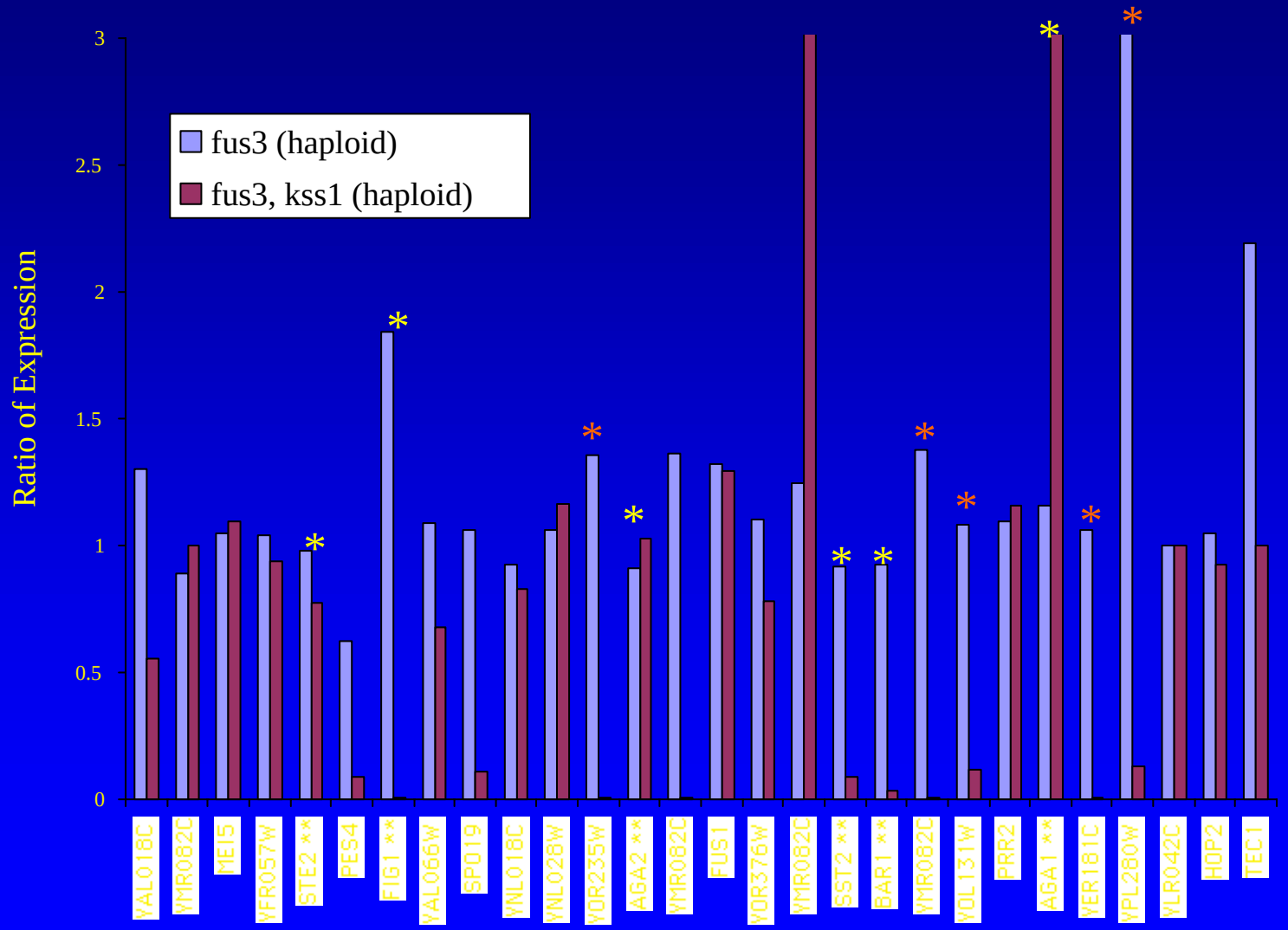
Mating Pattern



Correlations (Fus3/Kss1)



Mating Pattern Correlations



Pattern 6

- Mating Pattern -> Pattern 6 for Ste11, Ste7, Ste5, and Ste12
- PSI-BLAST and SMART pick up 5 matches among unknown ORFs to transposon and retroposon proteins

Conclusions

- Life is Very Complex
  - Multiple Pathways and Interactions for Each Protein with Transcription/Translation
  - Natural Stochastic Variations
- Analysis Tools Must
  - Isolate Areas of Interest without Loss of Knowledge Discovery
  - Incorporate Maximal Prior Knowledge to Reduce “Search Space”

Bayesian Decomposition

- Ability to Identify Overlapping Coexpression Can Lead to Identification of Pathways Affected in an Experiment
- Capability to Encode Prior Knowledge Can Improve Results
  - Ochs et al, *J Magn Res*, **137**, 161, 1999
  - Ochs et al, *Magn Res Med*, in press

Future Directions

- Application to Clinical Trials
  - Study of GIST Tumor Response to Gleevec
  - Study of Drug Response of HOSE Lines
- Bayesian Decomposition Development
  - Incorporation of Time Domain Modelling
  - Incorporation of Knowledge of Coexpression

Credits (Definitely Due)

- Fox Chase
  - Frank Manion
  - Thomas Moloshok
  - Jeffrey Grant
  - Yue Zhang
  - Ghislain Bidaut
  - Burt Eisenberg
  - Andy Godwin
- City of Hope
  - Bob Klevecz
- Johns Hopkins
  - Giovanni Parmigiani
- Fox Chase (NMR)
  - Truman Brown (Columbia)