Graphical Models and Bayesian Methods in Bioinformatics: From Structural to Systems Biology

David L. Wild

Keck Graduate Institute of Applied Life Sciences, Claremont, CA, USA

October 3, 2005

Outline

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Motivation

- Goal of this talk: to demonstrate how Graphical Models and Bayesian Methods may be used for a variety of modeling problems in Bioinformatics
- Inferring Gene Regulatory Networks from Microarray Data
- Protein Structure Prediction
- Biomarker Discovery in Microarray Data
- Identifying Protein Complexes in High-Throughput Protein Interaction Screens
- Clustering Protein Sequences and Structures

Basic Rules of Probability

$P(x)$	probability of x
$P(x\|\theta)$	conditional probability of x given θ
$P(x,\theta)$	joint probability of x and θ

$$P(x,\theta) = P(x)P(\theta|x) = P(\theta)P(x|\theta)$$

Bayes Rule:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

Marginalization

$$P(x) = \int P(x,\theta)\, d\theta$$

Bayes Rule Applied to Machine Learning

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	likelihood of θ
$P(\theta)$	prior probability of θ
$P(\theta\|\mathcal{D})$	posterior of θ given $\mathcal{D}$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

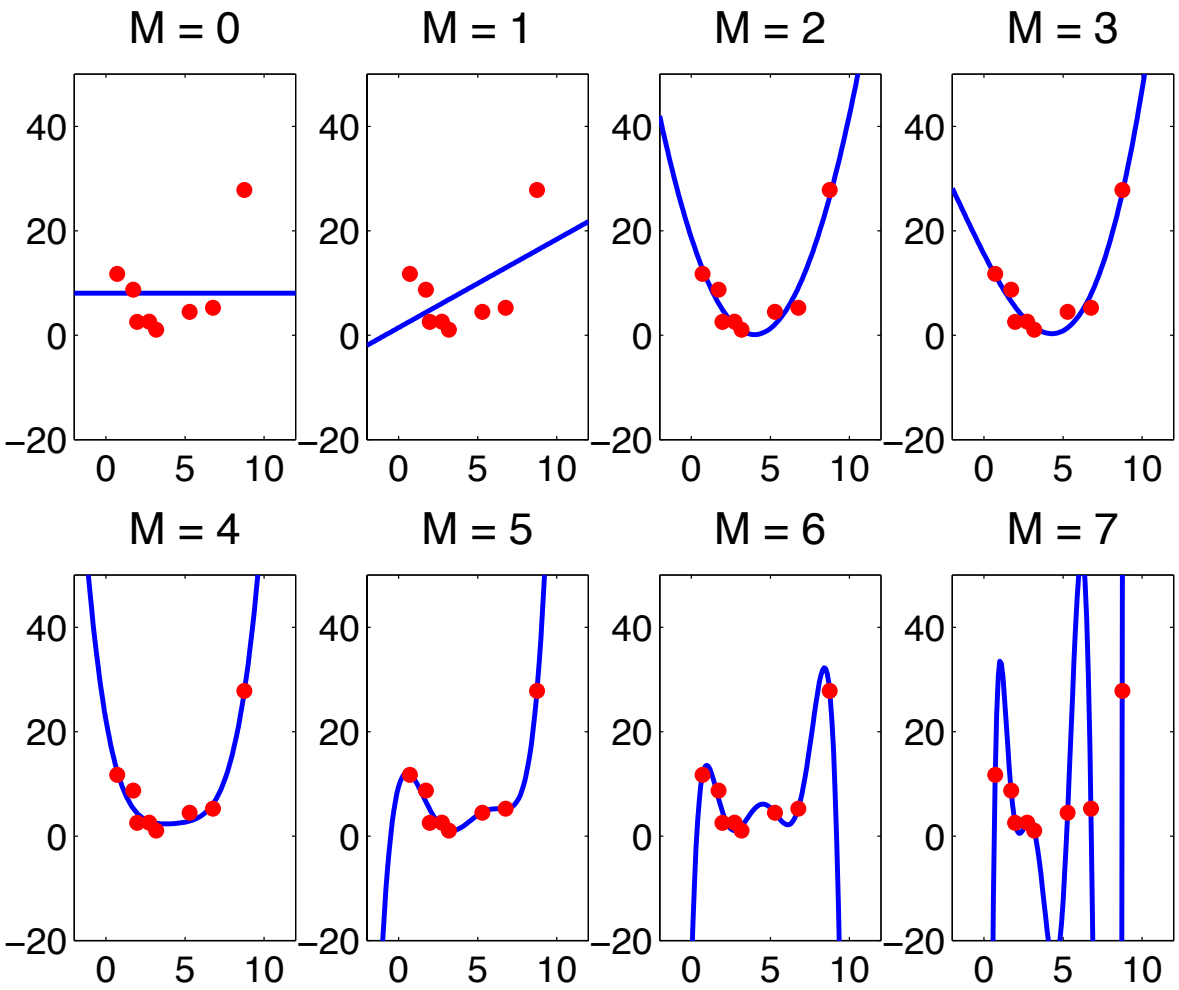
$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

$$P(x|\mathcal{D}, m) = \int P(x|\theta)P(\theta|\mathcal{D}, m)d\theta \qquad \text{(for many models)}$$

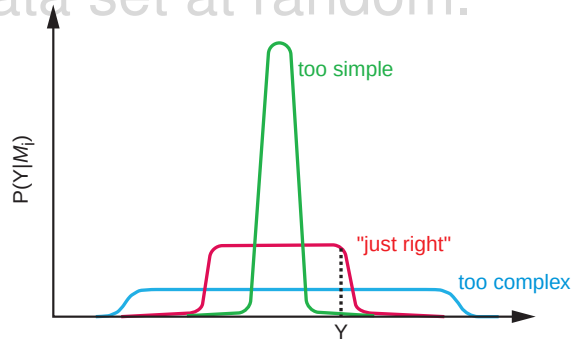
Model structure and overfitting: a simple example



Using Bayesian Occam's Razor to Learn Model Structure

Select the model class m_i with the highest probability given the data by computing the Marginal Likelihood ("evidence"):
Interpretation: The probability that *randomly selected* parameters from the prior would generate the data set.

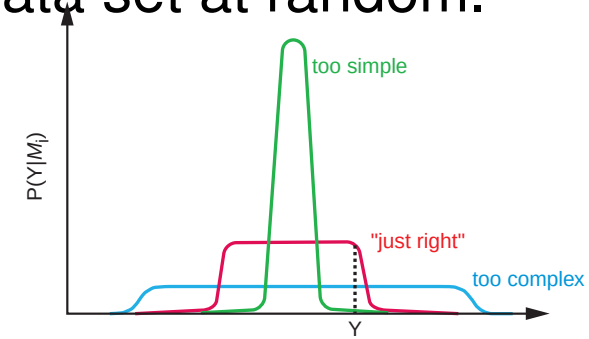
- Model classes that are too simple are unlikely to generate the data set.
- Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



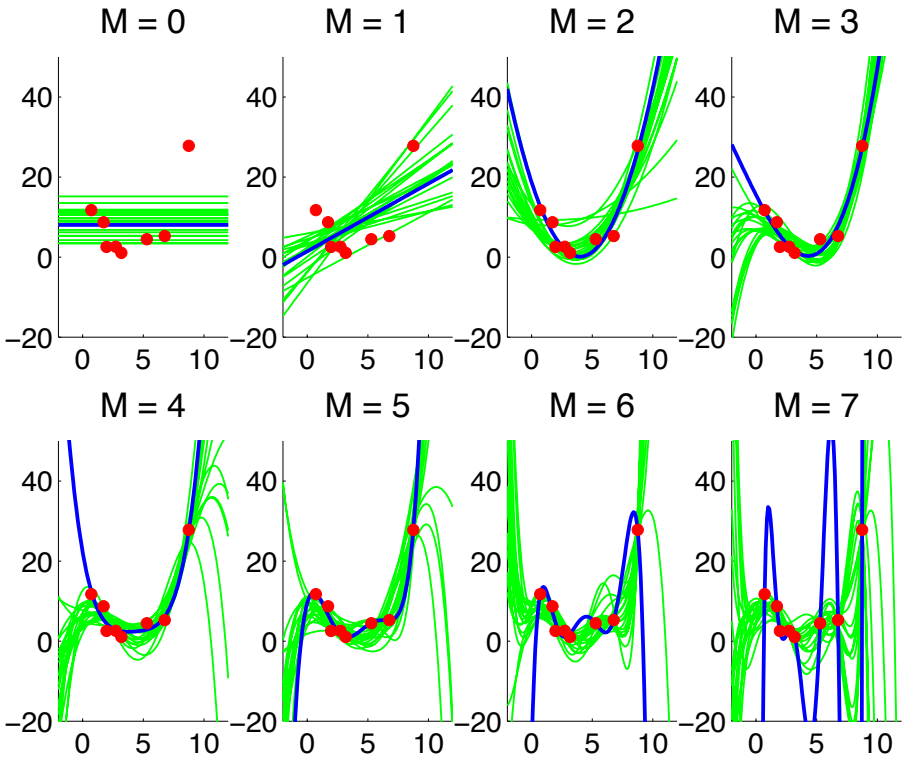
Using Bayesian Occam's Razor to Learn Model Structure

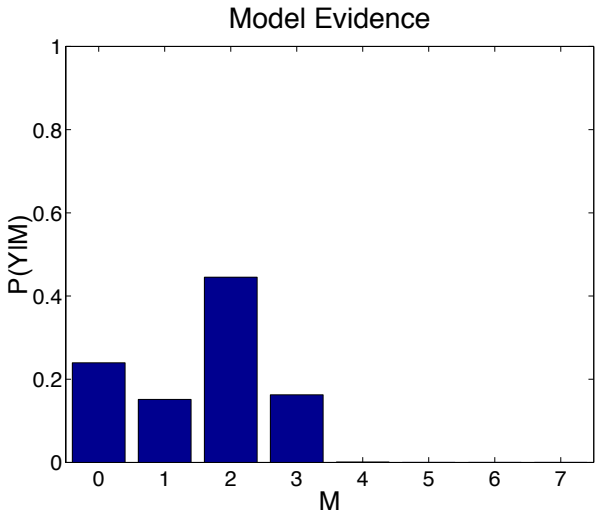
Select the model class m_i with the highest probability given the data by computing the Marginal Likelihood ("evidence"):
Interpretation: The probability that *randomly selected* parameters from the prior would generate the data set.

- Model classes that are too simple are unlikely to generate the data set.
- Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Bayesian Model Selection: Occam's Razor at Work

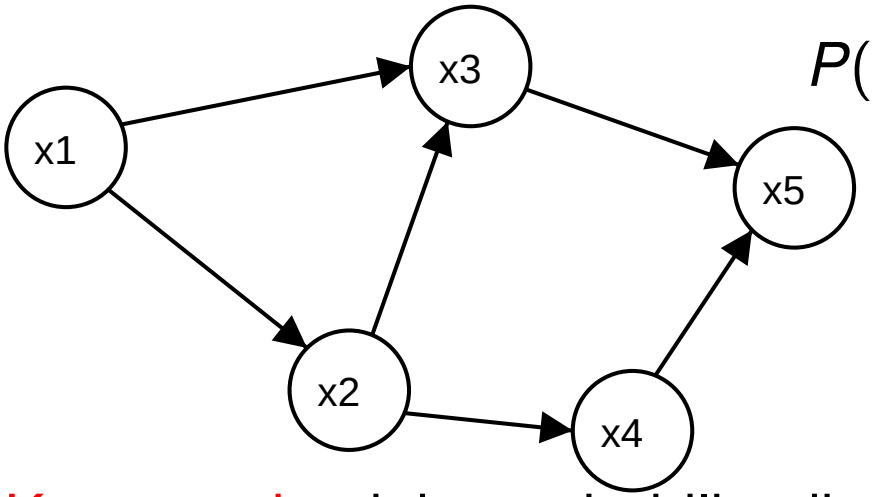




e.g. for quadratic (M=2): $y = a_0 + a_1 x + a_2 x^2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$ and $\theta_2 = [a_0\ a_1\ a_2\ \tau]$

Graphical Models

Directed acyclic graph where each node corresponds to a random variable.



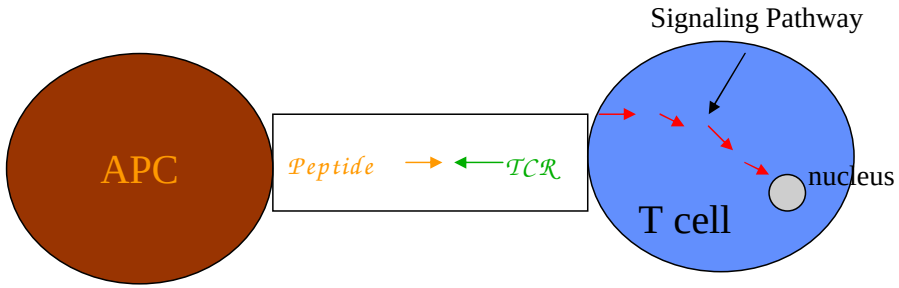
$$P(\mathbf{x}) = P(\mathbf{x}_1)P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_3|\mathbf{x}_1,\mathbf{x}_2) P(\mathbf{x}_4|\mathbf{x}_2)P(\mathbf{x}_5|\mathbf{x}_3,\mathbf{x}_4)$$

Key quantity: joint probability distribution over nodes:
$P(\mathbf{x}) = P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
The graph specifies a factorization of this joint probability distribution.
Also known as Bayesian Networks, Belief Nets and Probabilistic Independence Nets.

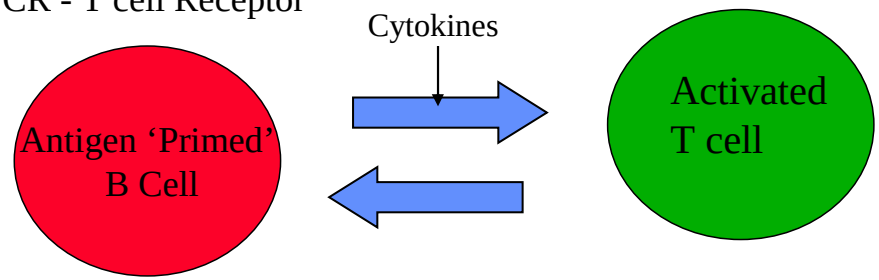
T cell activation

The Biological System

- The central event in the generation of an immune response is the activation of T cells.



APC - Antigen Presenting Cell
TCR - T cell Receptor

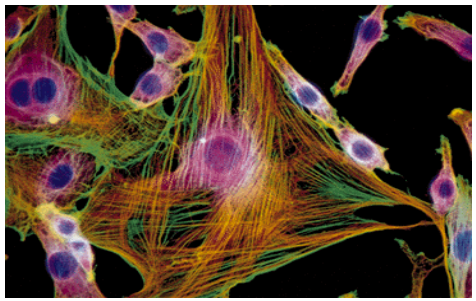


A Model of T cell Activation

***In Vitro* model of T-cell activation for analysis of transcriptional pathways.**

Stimulus

PMA
Ionomycin



Human Jurkat cells

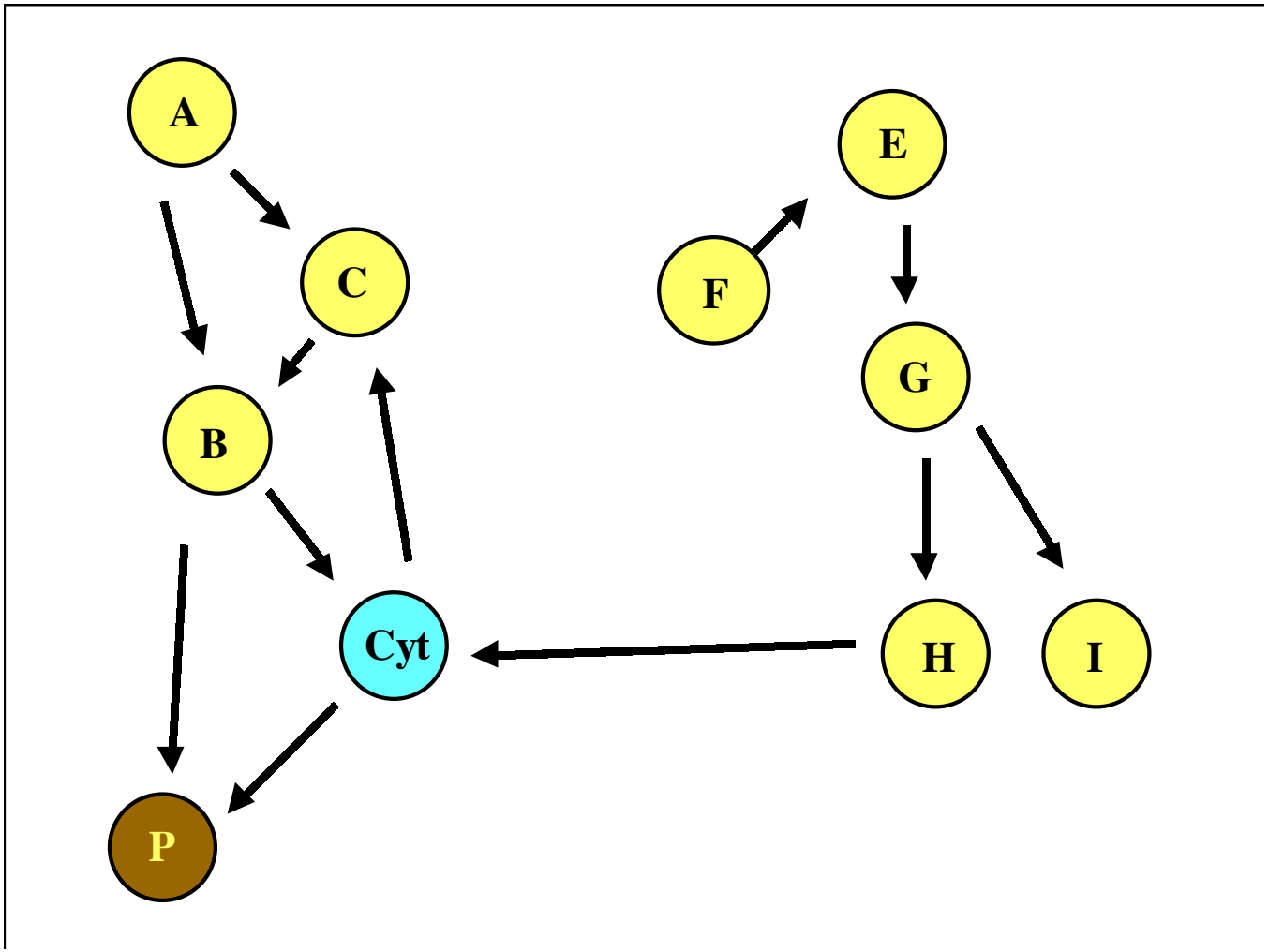
Response

Apoptosis:
Jun B, Caspase 8

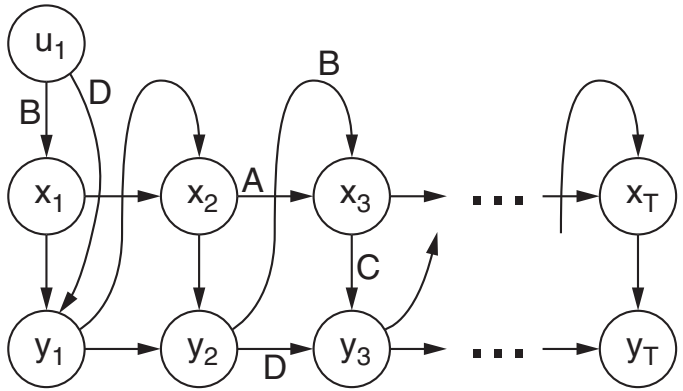
Inflammation:
FYB, IL3R α

Cell cycle/ Adhesion:
Cyclin A2, Integrin α

Hypothetical Networks Involved in T-cell Activation



A Gaussian State-Space Model with Feedback



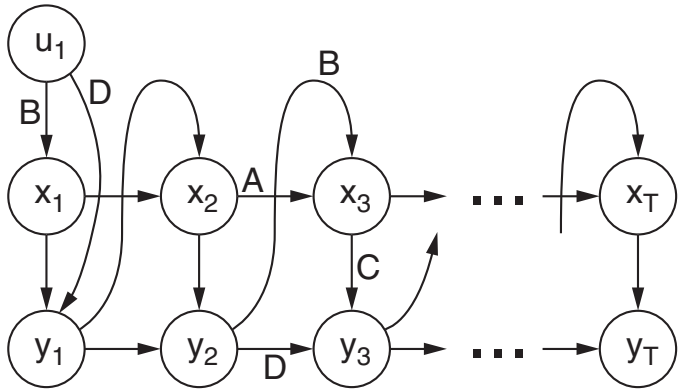
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



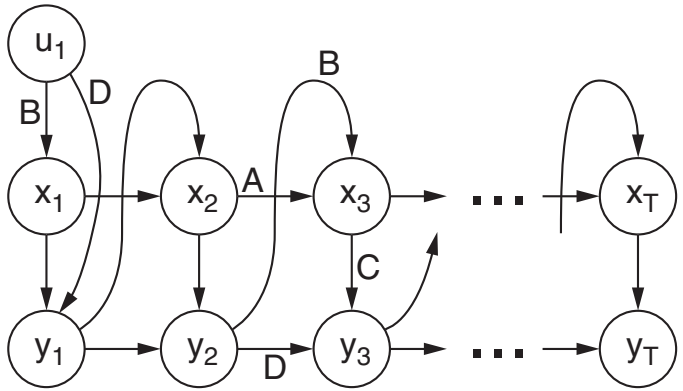
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



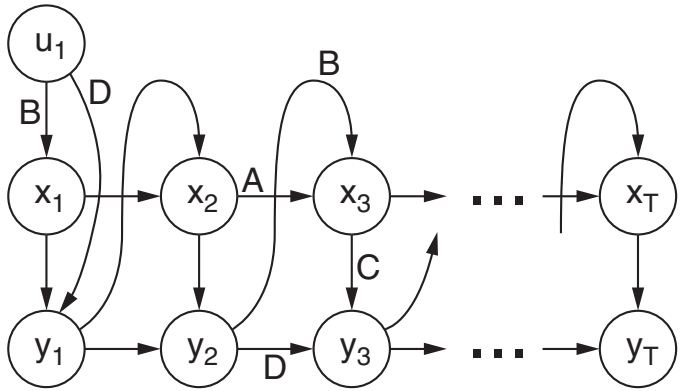
Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

A Gaussian State-Space Model with Feedback



Output equation: $\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation: $\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

Our Approach

- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Classical statistical approach uses cross-validation and bootstrapping (Rangel et al., *Bioinformatics*, 2004).
- Can also use variational approximations to perform approximate Bayesian inference in state-space models (Beal et al., *Bioinformatics*, 2005).

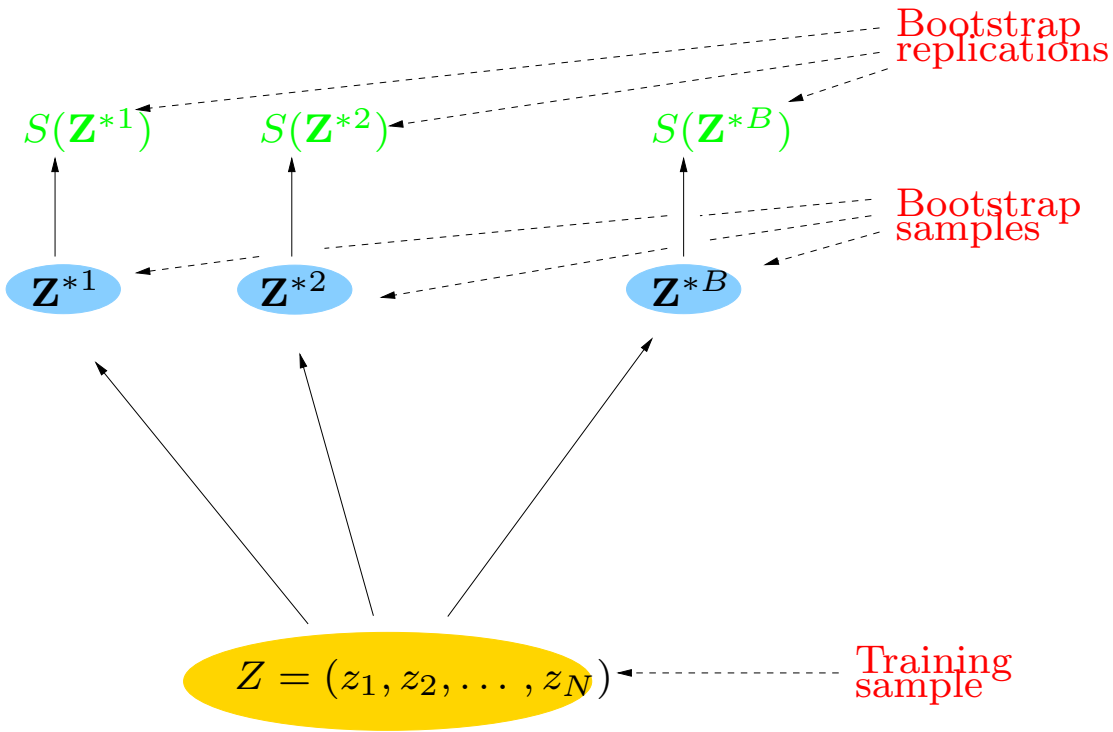
Our Approach

- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Classical statistical approach uses cross-validation and bootstrapping (Rangel et al., *Bioinformatics*, 2004).
- Can also use variational approximations to perform approximate Bayesian inference in state-space models (Beal et al., *Bioinformatics*, 2005).

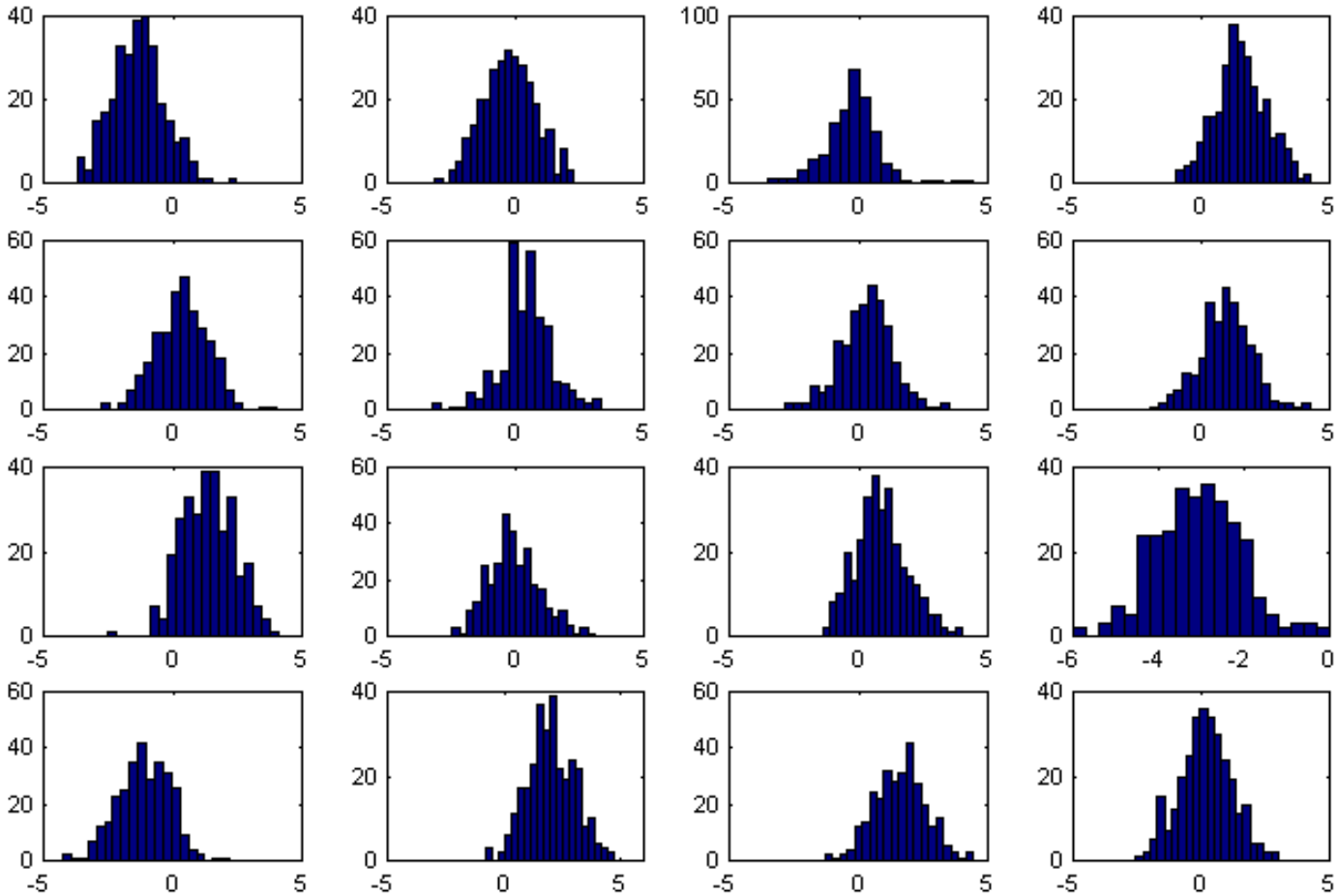
Our Approach

- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Classical statistical approach uses cross-validation and bootstrapping (Rangel et al., *Bioinformatics*, 2004).
- Can also use variational approximations to perform approximate Bayesian inference in state-space models (Beal et al., *Bioinformatics*, 2005).

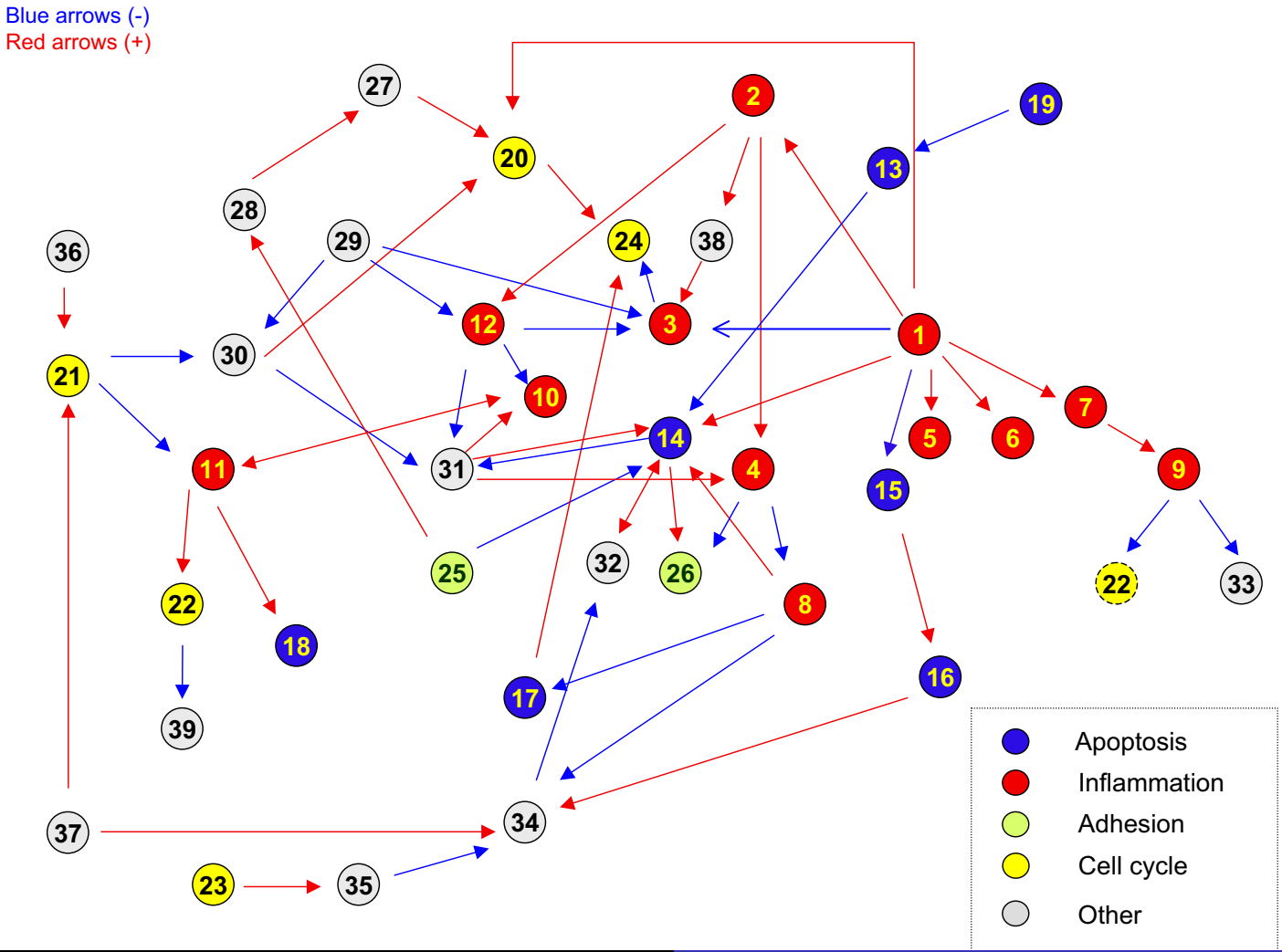
Bootstrap Procedure for Parameter Confidence Intervals (1)



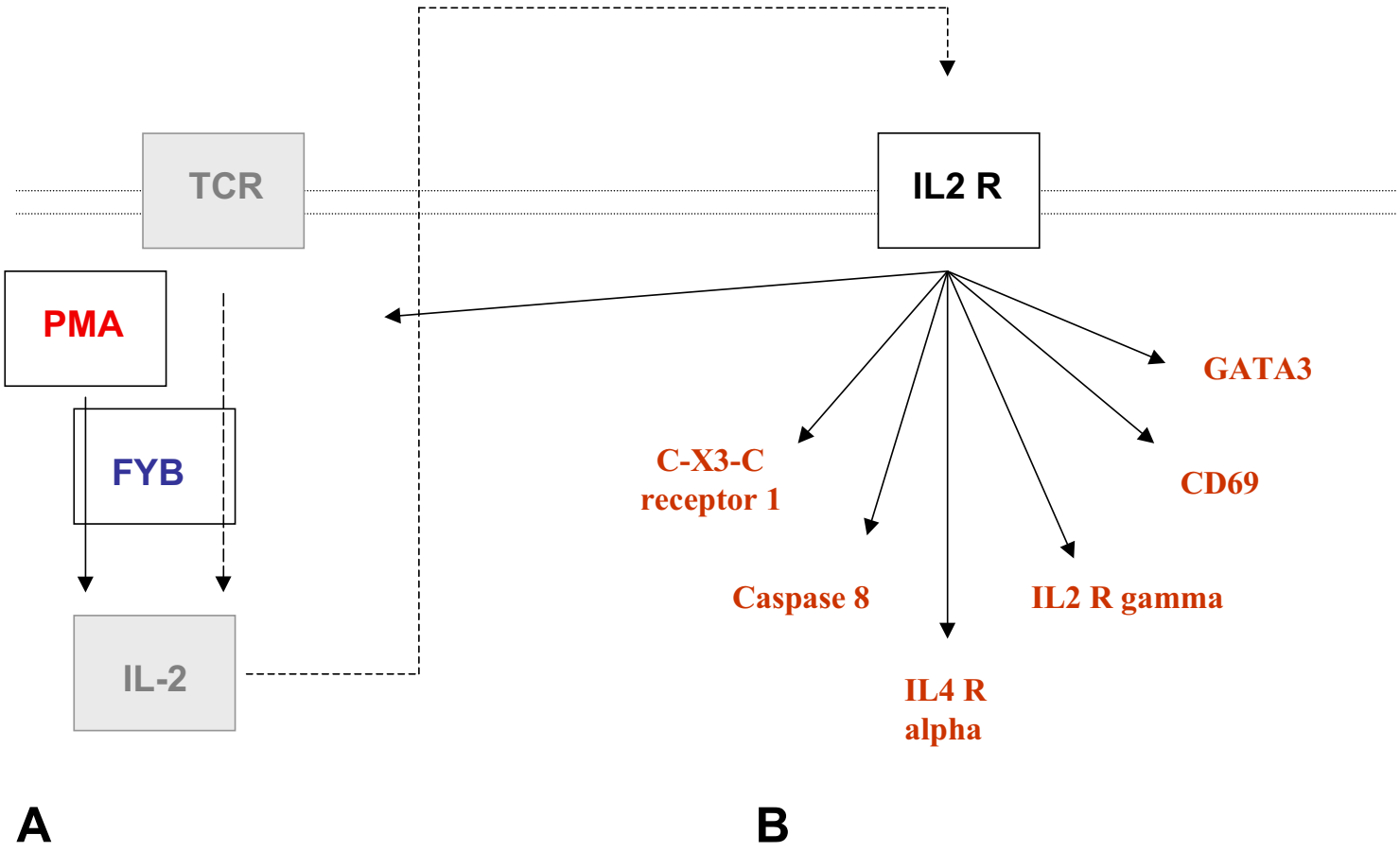
Bootstrap Procedure for Parameter Confidence Intervals (2)



Inferring Regulatory Networks

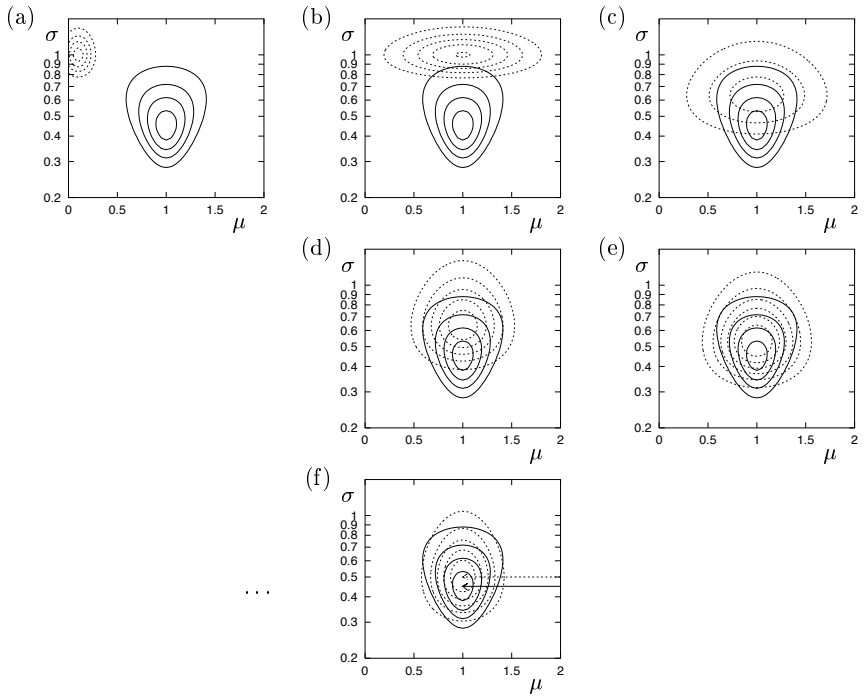


In-Silico Hypotheses



Variational Bayesian Approach

Variational free energy minimization is a method of approximating a complex distribution $p(\mathbf{x})$ by a simpler distribution $q(\mathbf{x};\theta)$. We adust the parameters θ so as to get q to best approximate p in some sense.



From David J.C. MacKay “Information Theory, Inference and Learning Algorithms”

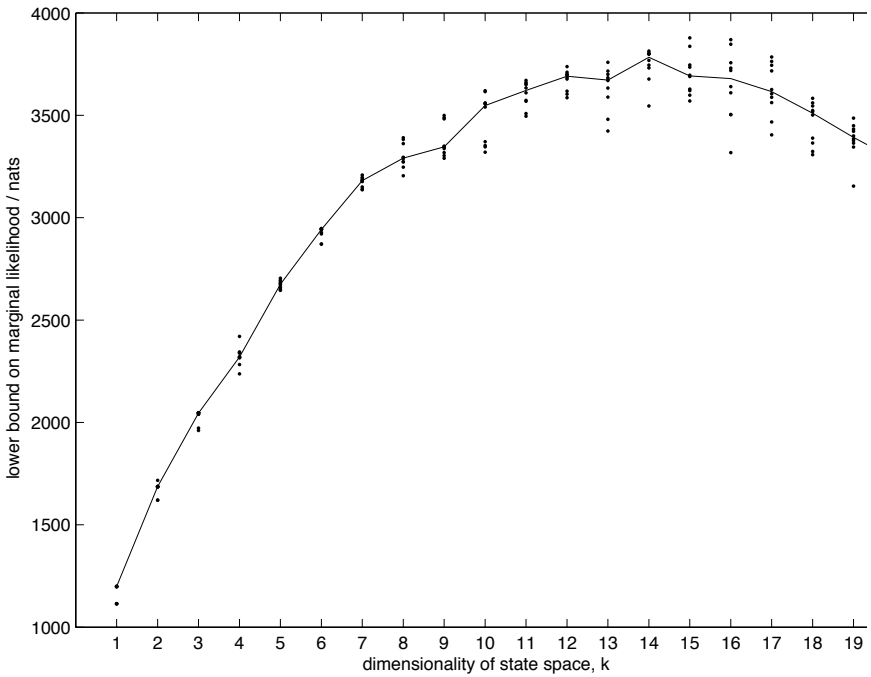
Lower Bounding the Marginal Likelihood

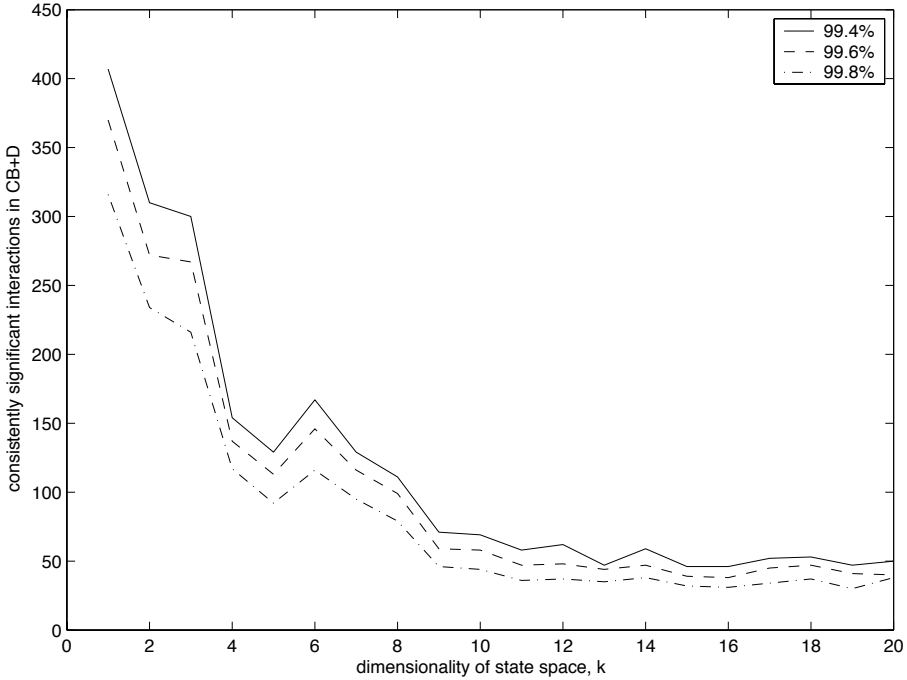
We can also lower bound the marginal likelihood:
Using a simpler, factorised approximation to $q(\mathbf{x}, \theta) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\theta}(\theta)$:

$$\ln p(\mathbf{y}|m) \quad = \quad \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\theta}(\theta), \mathbf{y}).$$

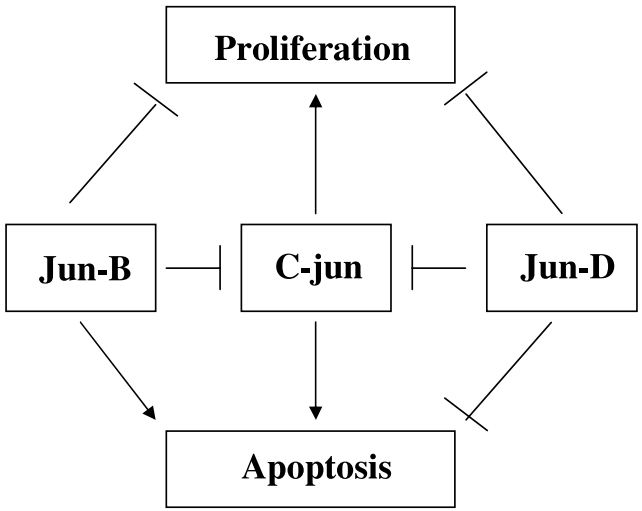
Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** iterative updates. $-\mathcal{F}_m$ is a variational free energy

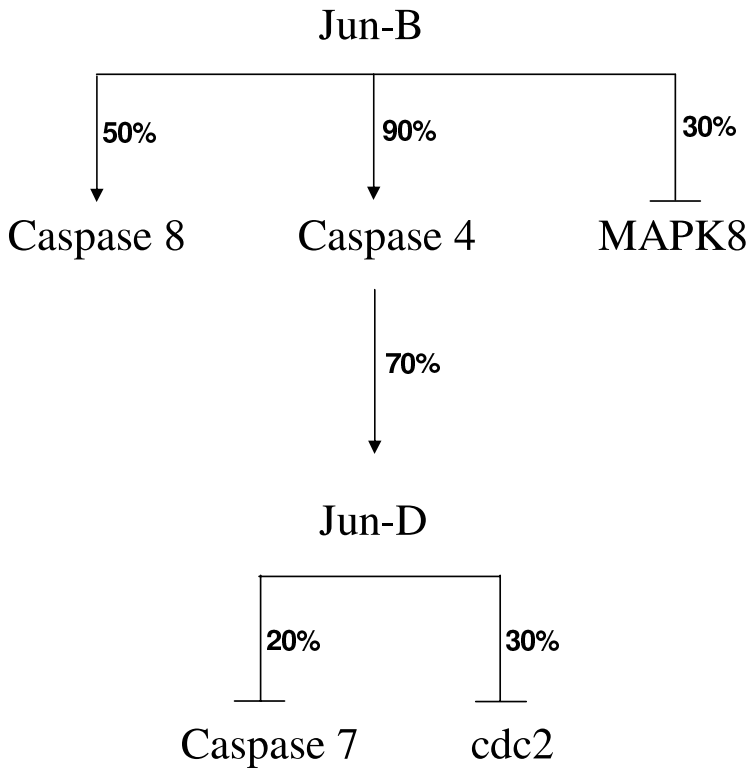
Results from the Variational Bayesian Approach





In-Silico Hypotheses (2)





Future Work

A framework to build on with future work:

- incorporating biologically plausible nonlinearities
- adding prior knowledge (especially in the form of constraints on positive and negative interactions)
- making and testing gene silencing and overexpresson predictions
- combining gene and protein expression data with metabolomic data

Future Work

A framework to build on with future work:

- incorporating biologically plausible nonlinearities
- adding prior knowledge (especially in the form of constraints on positive and negative interactions)
- making and testing gene silencing and overexpresson predictions
- combining gene and protein expression data with metabolomic data

Future Work

A framework to build on with future work:

- incorporating biologically plausible nonlinearities
- adding prior knowledge (especially in the form of constraints on positive and negative interactions)
- making and testing gene silencing and overexpresson predictions
- combining gene and protein expression data with metabolomic data

Future Work

A framework to build on with future work:

- incorporating biologically plausible nonlinearities
- adding prior knowledge (especially in the form of constraints on positive and negative interactions)
- making and testing gene silencing and overexpresson predictions
- combining gene and protein expression data with metabolomic data

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Protein Secondary Structure Prediction

- Discriminant approach with neural networks,
- Seminal work by Qian and Sejnowski (1988)
- PHD (Rost and Sander, 1993) - evolutionary information from multiple sequence alignment
- Jones (1999) - position-specific scoring matrices (PSSM)
- Cuff and Barton (2000) evaluated different types of multiple sequence alignment profiles
- Generative model (Schmidler, 2002) using primary structure only with lower prediction accuracy

Our Approach

- Proteins as collection of local structural segments which may be shared by unrelated proteins
- Build a probabilistic generative graphical model that describes the relationship between protein primary structure and its secondary structure
- Incorporate biological constraints (residue propensities, long range interactions)
- Learn model parameters from data sets of proteins with known structure
- Predict structure of novel proteins using Bayesian inference

Our Approach

- Proteins as collection of local structural segments which may be shared by unrelated proteins
- Build a probabilistic generative graphical model that describes the relationship between protein primary structure and its secondary structure
- Incorporate biological constraints (residue propensities, long range interactions)
- Learn model parameters from data sets of proteins with known structure
- Predict structure of novel proteins using Bayesian inference

Our Approach

- Proteins as collection of local structural segments which may be shared by unrelated proteins
- Build a probabilistic generative graphical model that describes the relationship between protein primary structure and its secondary structure
- Incorporate biological constraints (residue propensities, long range interactions)
- Learn model parameters from data sets of proteins with known structure
- Predict structure of novel proteins using Bayesian inference

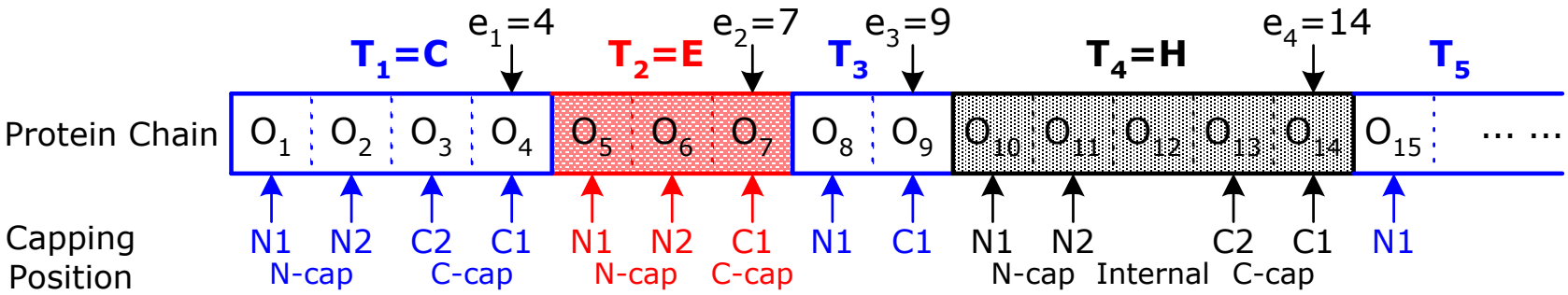
Our Approach

- Proteins as collection of local structural segments which may be shared by unrelated proteins
- Build a probabilistic generative graphical model that describes the relationship between protein primary structure and its secondary structure
- Incorporate biological constraints (residue propensities, long range interactions)
- Learn model parameters from data sets of proteins with known structure
- Predict structure of novel proteins using Bayesian inference

Our Approach

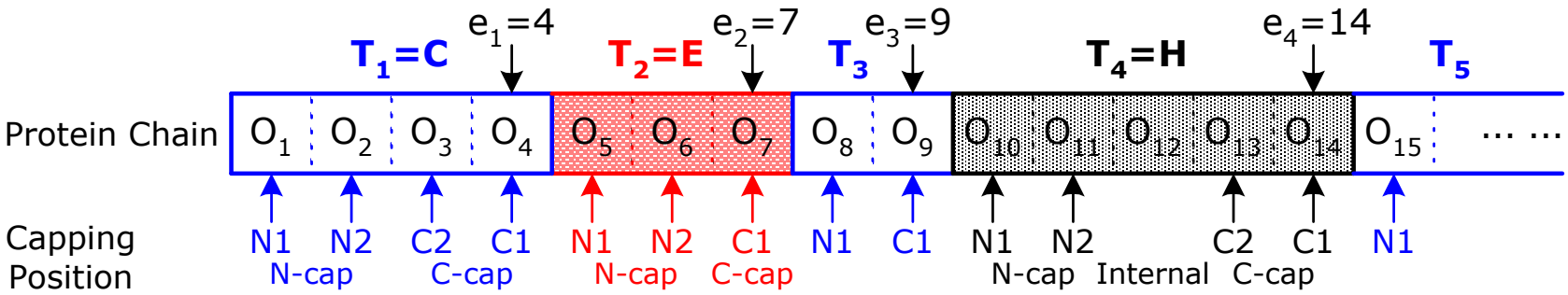
- Proteins as collection of local structural segments which may be shared by unrelated proteins
- Build a probabilistic generative graphical model that describes the relationship between protein primary structure and its secondary structure
- Incorporate biological constraints (residue propensities, long range interactions)
- Learn model parameters from data sets of proteins with known structure
- Predict structure of novel proteins using Bayesian inference

Segmental Model



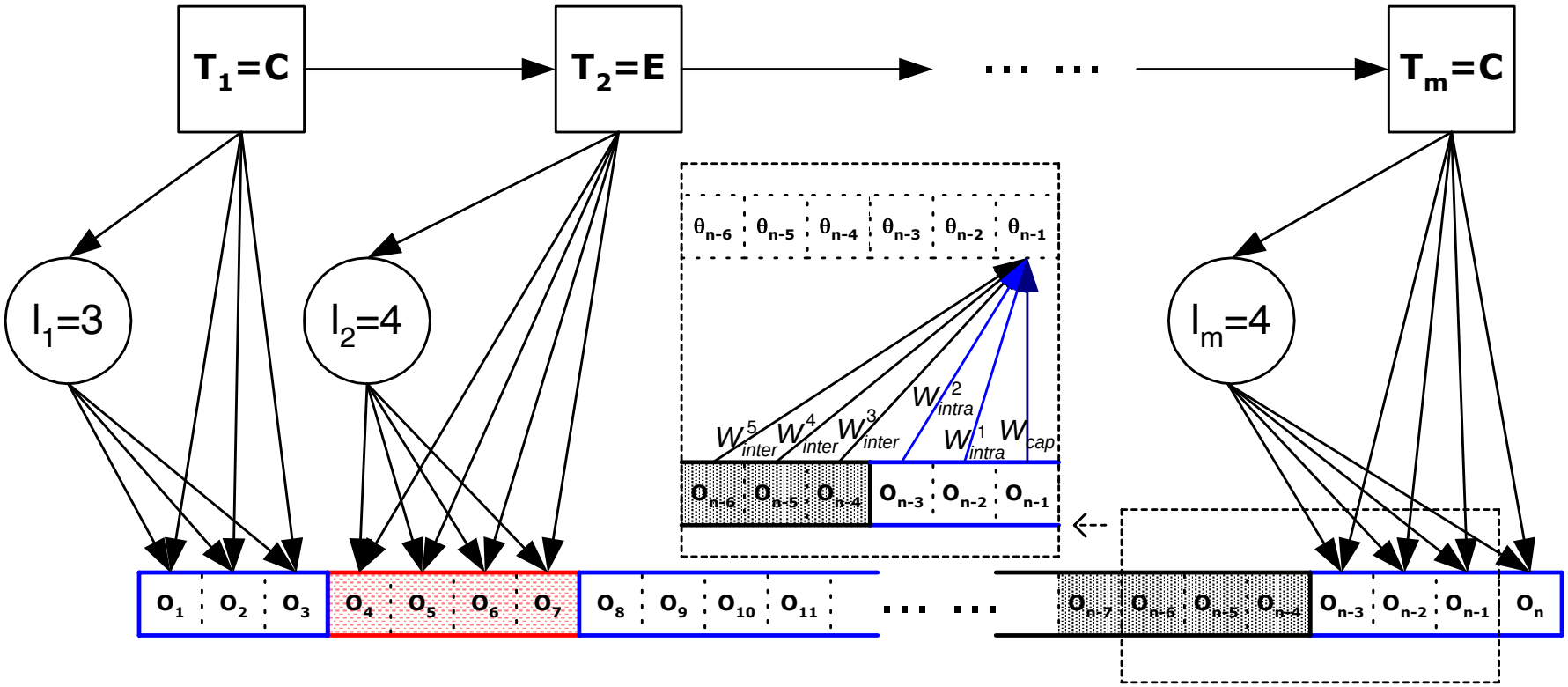
1. A sequence of observations on n amino acid residues $O = [O_1, O_2, \ldots, O_n]$
2. A set of segmental variables, (m, e, T), where m is the number of segments, the segmental endpoints $e = [e_1, e_2, \ldots, e_m]$ and the segment types $T = [T_1, T_2, \ldots, T_m]$.

Segmental Model



1. A sequence of observations on n amino acid residues $O = [O_1, O_2, \ldots, O_n]$
2. A set of segmental variables, (m, e, T), where m is the number of segments, the segmental endpoints $e = [e_1, e_2, \ldots, e_m]$ and the segment types $T = [T_1, T_2, \ldots, T_m]$.

Segmental Semi-Markov Models



Chu et al. *ICML*, 2004

Individual Likelihood

This is a Dirichlet-Multinomial distribution.

$$\mathcal{P}(O_k|O_{[1:k-1]}, T_i) = \int_{\theta_k} \mathcal{P}(O_k|\theta_k, T_i)\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i)\, d\theta_k$$

- Multinomial: $\mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a\in\mathcal{A}} (\theta_k^a)^{O_k^a}$
- Dirichlet Prior: $\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a\in\mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1}$
- Weights:
  $\gamma_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j}$.

Individual Likelihood

This is a Dirichlet-Multinomial distribution.

$$\mathcal{P}(O_k|O_{[1:k-1]}, T_i) = \int_{\theta_k} \mathcal{P}(O_k|\theta_k, T_i)\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i)\, d\theta_k$$

- Multinomial: $\mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a\in\mathcal{A}} (\theta_k^a)^{O_k^a}$
- Dirichlet Prior: $\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a\in\mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1}$
- Weights:
  $\gamma_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j}.$

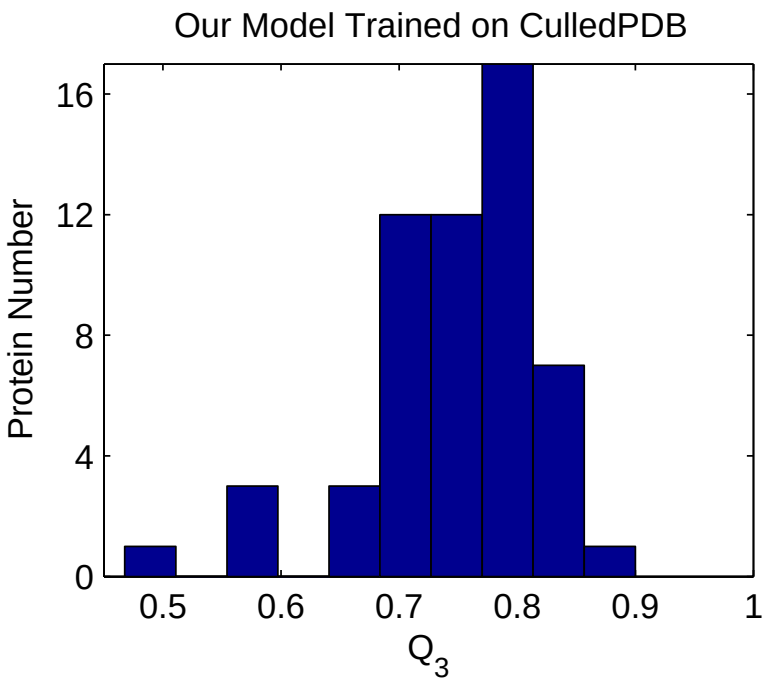
Individual Likelihood

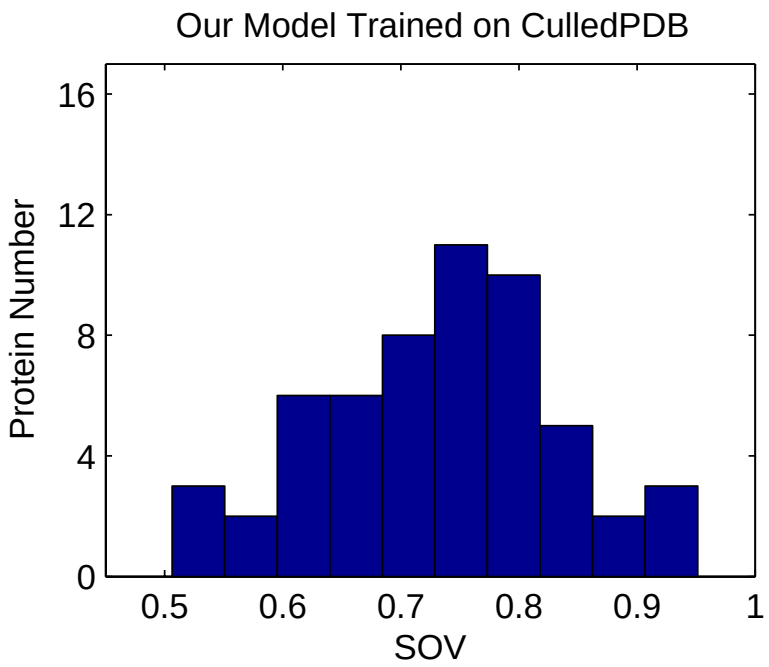
This is a Dirichlet-Multinomial distribution.

$$\mathcal{P}(O_k|O_{[1:k-1]}, T_i) = \int_{\theta_k} \mathcal{P}(O_k|\theta_k, T_i)\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i)\, d\theta_k$$

- Multinomial: $\mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a\in\mathcal{A}} (\theta_k^a)^{O_k^a}$
- Dirichlet Prior: $\mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a\in\mathcal{A}} (\theta_k^a)^{\gamma_k^a - 1}$
- Weights:
  $\gamma_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j}.$

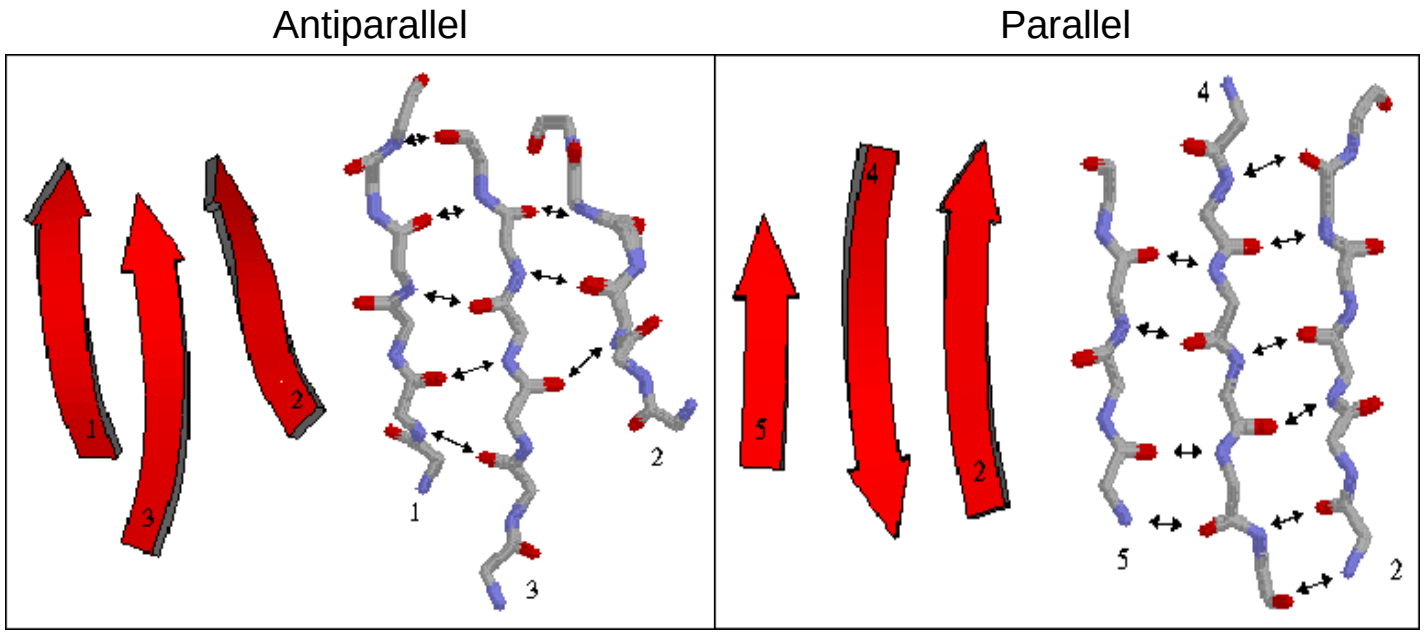
CASP5 Results





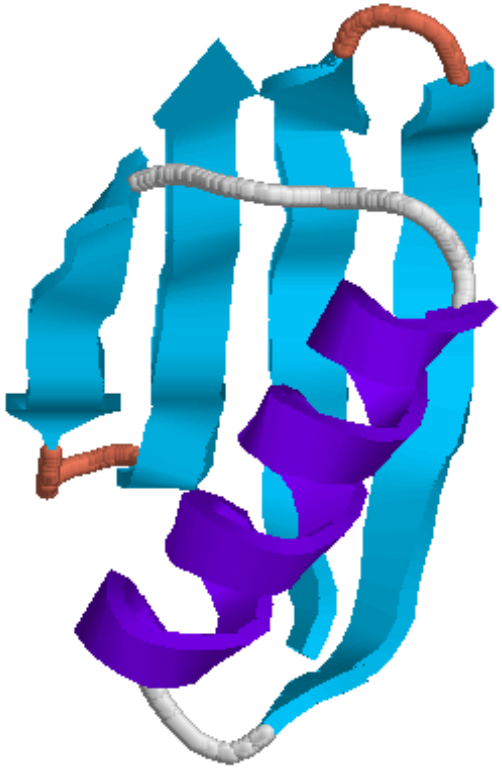
	Chain Length	Q_3^{casp5}	SOV^{casp5}	Q_3^{culled}	SOV^{culled}
Average	215.75	74.6±10.3 %	73.4±12.3 %	74.9±7.5%	73.1±10.3%

Long-range Interactions in β-sheets

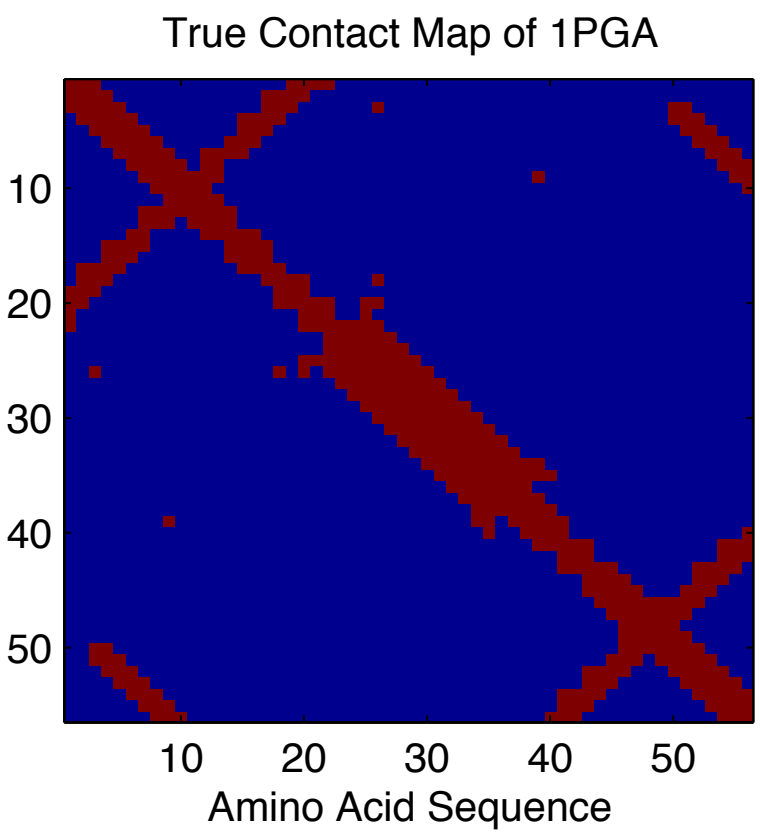


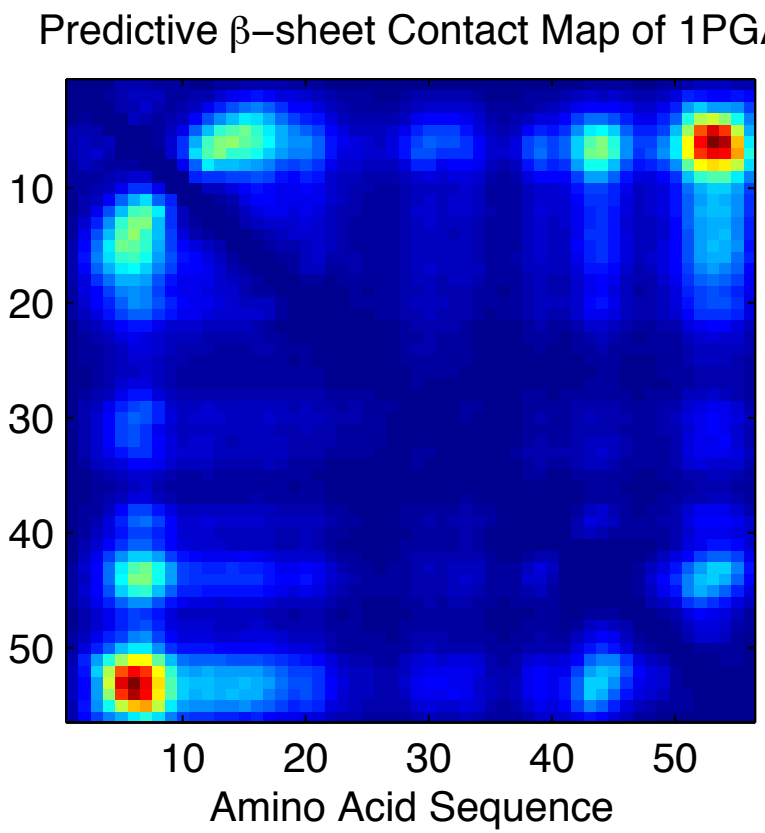
The β-sheet space is the set of all the possible combinations of β-sheets;
A set of interaction variables, $\mathcal{I}$, to describe one possible case.

1PGA - PROTEIN G



1PGA - PROTEIN G

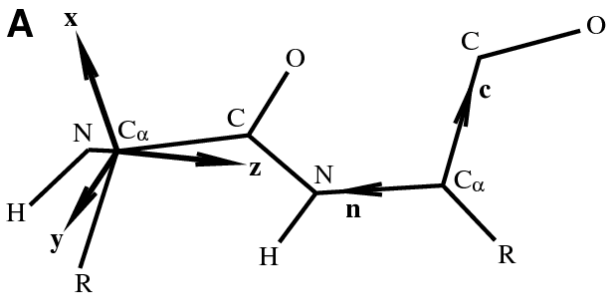




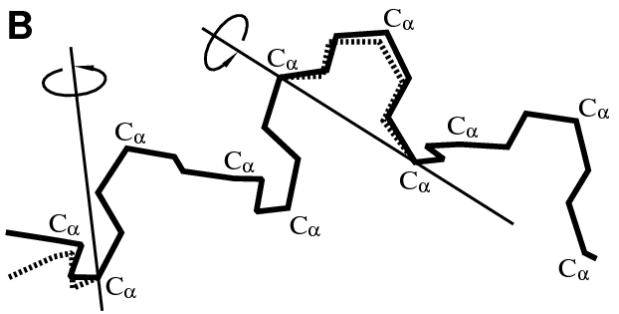
Combining the Probabilistic Model with Steric Constraints

Model and moves

- Planar rigid peptide bonds
- Elastic C_α valence geometry



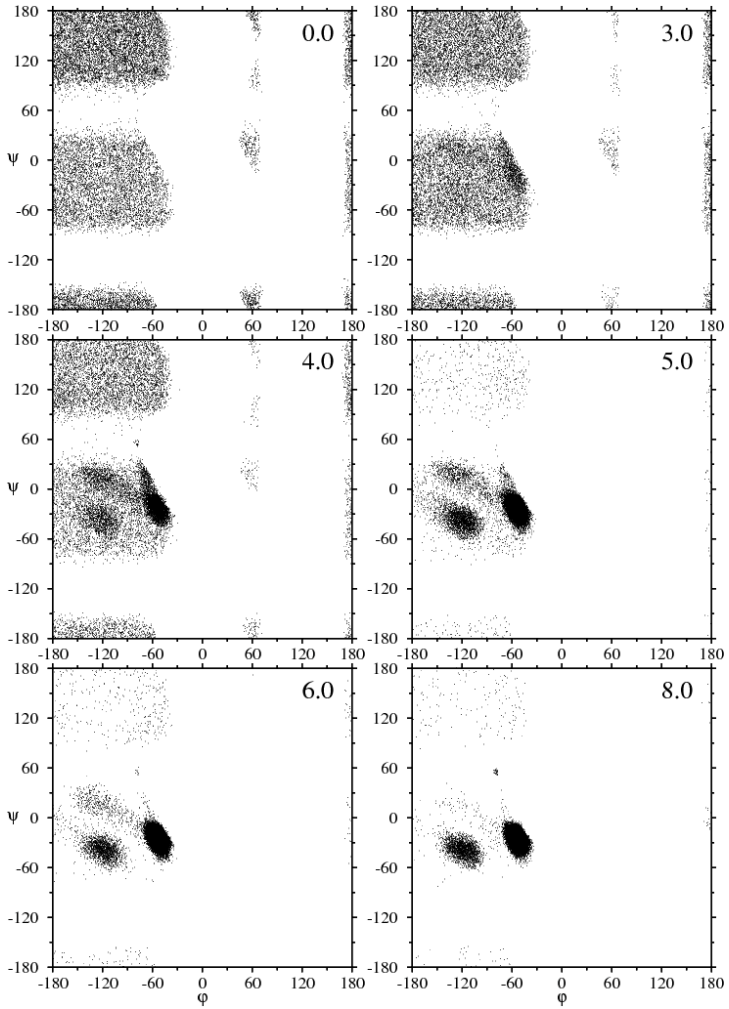
- Random pivotal rotations
- Random crankshaft rotations



Ramachrandran Plots for Polyalanine

Simulated Ramachandran plots

- When $H/RT = 0$, extended conformations are 70% more likely compact helical ones
- At high H/RT values, three distinctive compact conformations dominate the distribution



Podtelezhnikov and Wild, *Proteins*, 2005.

Ramachrandran Plots from PDB

Ho et al. (2003) *Protein Science* 12:2508–2522

- 500 nonhomologous proteins from the PDB
- C-capping (panel D) contains $\varphi = -120°$ and $\psi = -40°$

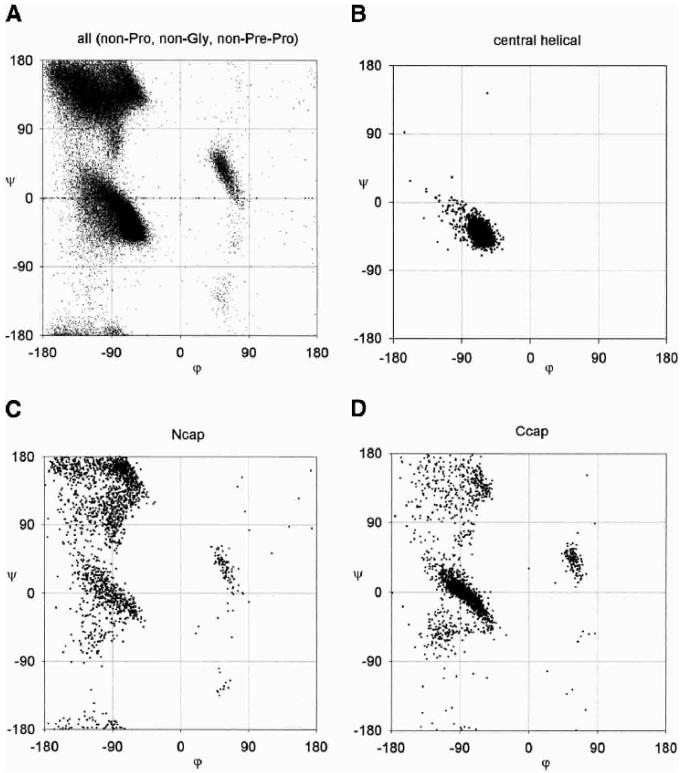
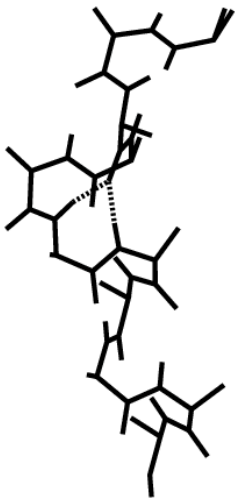


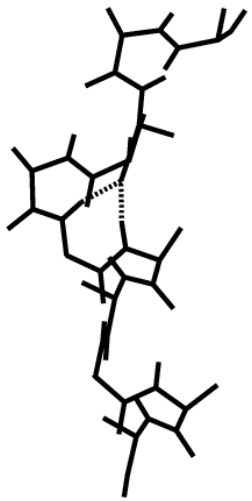
Figure 2. Ramachandran plots. (*A*) All residues excluding Pro, Gly, and pre-Pro; (*B*) residues in the center of the α-helix, which are more constrained than for all residues; (*C*) the Ncap residue; and (*D*) the Ccap residue in the α-helix, which are scattered throughout the entire allowed region.

Hydrogen Bonding Patterns

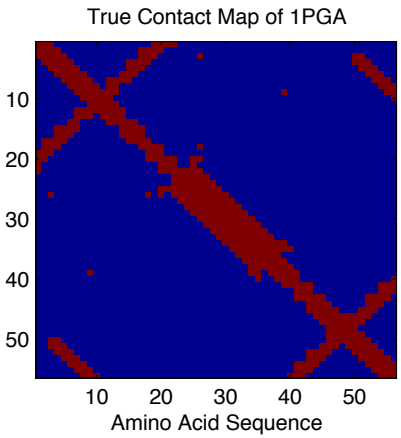
H-bonding patterns

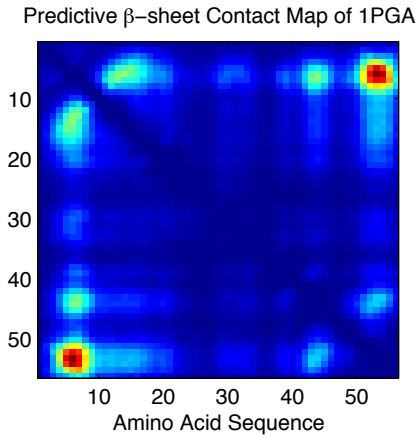
- 3_{10}-helices are 3 times more likely than α-helices
- Double H-bonds with a common acceptor are responsible for $\varphi = -120°$ and $\psi = -40°$

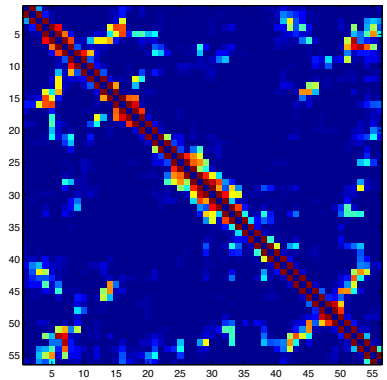




Contacts Sampled in Monte Carlo Procedure







Future Work

- Combining probabilistic model and steric constraints
- De-novo protein design

Future Work

- Combining probabilistic model and steric constraints
- De-novo protein design

Conclusions

- Graphical models and Bayesian methods can be used for a variety of modeling problems in Bioinformatics.
- They allow robust statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Automatic model selection via Bayesian “Occam’s Razor”
- We have looked at two problem domains: inferring genetic regulatory networks and protein structure prediction
- Models produce plausible biological hypotheses which can be experimentally validated

Conclusions

- Graphical models and Bayesian methods can be used for a variety of modeling problems in Bioinformatics.
- They allow robust statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Automatic model selection via Bayesian “Occam’s Razor”
- We have looked at two problem domains: inferring genetic regulatory networks and protein structure prediction
- Models produce plausible biological hypotheses which can be experimentally validated

Conclusions

- Graphical models and Bayesian methods can be used for a variety of modeling problems in Bioinformatics.
- They allow robust statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Automatic model selection via Bayesian “Occam’s Razor”
- We have looked at two problem domains: inferring genetic regulatory networks and protein structure prediction
- Models produce plausible biological hypotheses which can be experimentally validated

Conclusions

- Graphical models and Bayesian methods can be used for a variety of modeling problems in Bioinformatics.
- They allow robust statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Automatic model selection via Bayesian “Occam’s Razor”
- We have looked at two problem domains: inferring genetic regulatory networks and protein structure prediction
- Models produce plausible biological hypotheses which can be experimentally validated

Conclusions

- Graphical models and Bayesian methods can be used for a variety of modeling problems in Bioinformatics.
- They allow robust statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- Automatic model selection via Bayesian “Occam’s Razor”
- We have looked at two problem domains: inferring genetic regulatory networks and protein structure prediction
- Models produce plausible biological hypotheses which can be experimentally validated

Acknowledgements

- Claudia Rangel (University of Southern California)
- Matthew Beal (SUNY Buffalo)
- Alexei Podtelezhnikov (KGI)
- Wei Chu (University College London)
- Zoubin Ghahramani (University College London)
- Francesco Falciani (Univeristy of Birmingham, UK)
- This work is supported by NIH Grant Number 1 P01 GM63208 (Tools and Data Resources in Support of Structural Genomics) and NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)

Acknowledgements

- Claudia Rangel (University of Southern California)
- Matthew Beal (SUNY Buffalo)
- Alexei Podtelezhnikov (KGI)
- Wei Chu (University College London)
- Zoubin Ghahramani (University College London)
- Francesco Falciani (Univeristy of Birmingham, UK)
- This work is supported by NIH Grant Number 1 P01 GM63208 (Tools and Data Resources in Support of Structural Genomics) and NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)

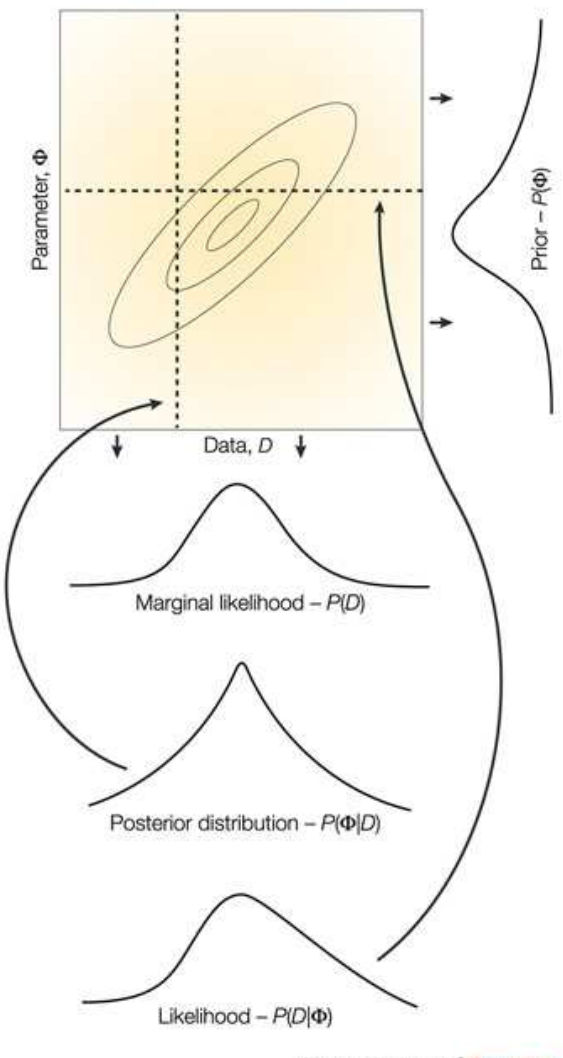
Acknowledgements

- Claudia Rangel (University of Southern California)
- Matthew Beal (SUNY Buffalo)
- Alexei Podtelezhnikov (KGI)
- Wei Chu (University College London)
- Zoubin Ghahramani (University College London)
- Francesco Falciani (Univeristy of Birmingham, UK)
- This work is supported by NIH Grant Number 1 P01 GM63208 (Tools and Data Resources in Support of Structural Genomics) and NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)

Acknowledgements

- Claudia Rangel (University of Southern California)
- Matthew Beal (SUNY Buffalo)
- Alexei Podtelezhnikov (KGI)
- Wei Chu (University College London)
- Zoubin Ghahramani (University College London)
- Francesco Falciani (Univeristy of Birmingham, UK)
- This work is supported by NIH Grant Number 1 P01 GM63208 (Tools and Data Resources in Support of Structural Genomics) and NSF Grant Number CCF-0524331 (Reconstructing Metabolic and Transcriptional Networks using Bayesian State Space Models)

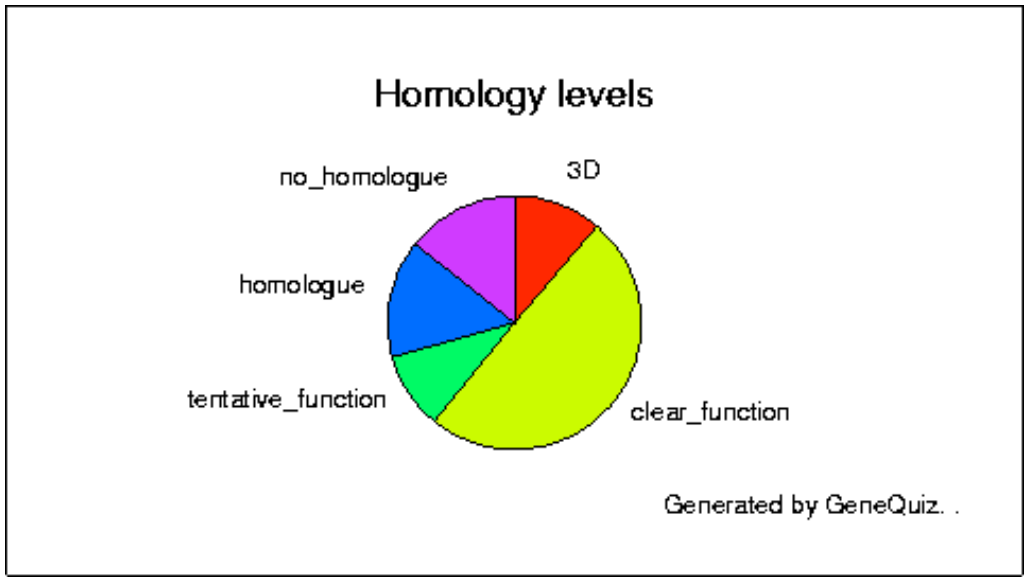
The basic features that underlie Bayesian Inference



From M.A. Beaumont and B. Rannala “The Bayesian Revolution in Genetics”

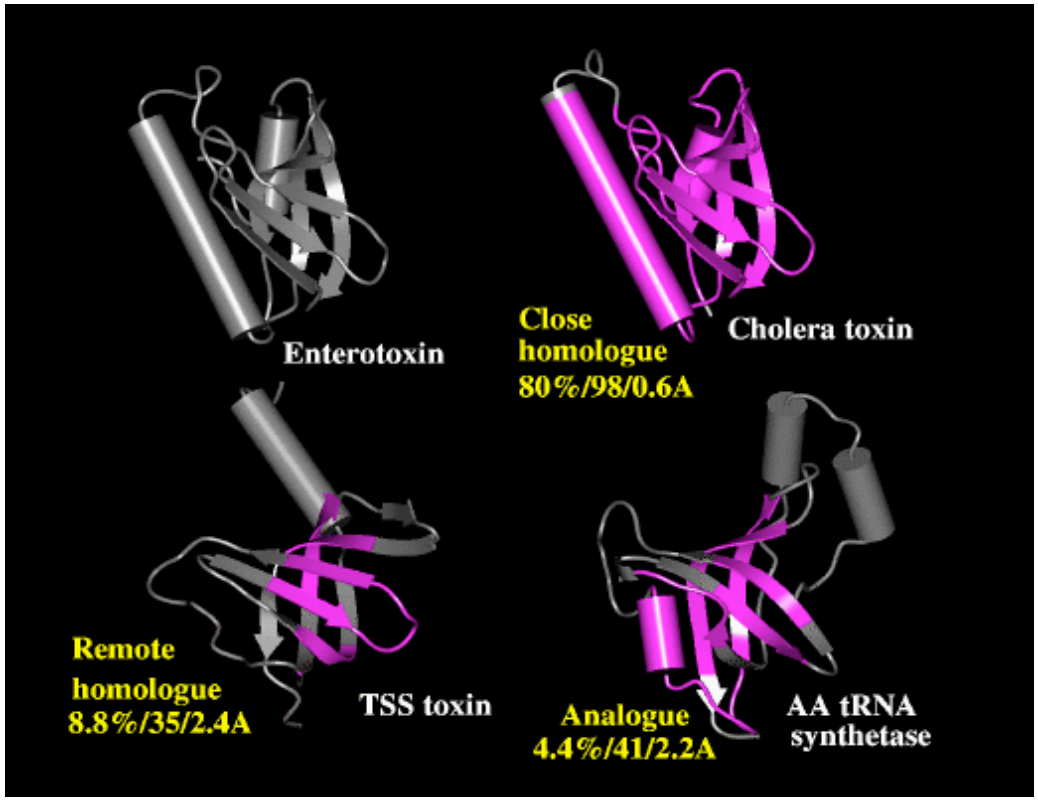
The Function-Homology Gap

Functional assignment by homology: the function-homology gap



yeast data analyzed by GeneQuiz

Structural Homologs and Analogs (1)



Russell et al. J. Mol. Biol (1997) 269, 423-439

Comparison to Cuff and Barton (2000)

Method Description	Q_3
Networks using frequency profile from CLUSTALW	71.6%
Networks using BLOSUM62 profile from CLUSTALW	70.8%
Networks using PSIBLAST alignment profiles	72.1%
Arithmetic sum based on the above three networks	73.4%
Networks using PSIBLAST PSSM	75.2%
Our algorithm with MSAP	71.3%
Our algorithm with PSIBLAST PSSM	72.2%

The Helix-Coil Transition in Polyalanine

Helix-coil transition

- Hydrogen bonds are formed and broken cooperatively
- Zimm-Bragg parameters of the helix-coil transition: $s = 0.013e^{-H/RT}$ and $\sigma = 0.3$