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

1. Motivation and Background
2. Inferring Gene Regulatory Networks from Microarray Data
3. Protein Structure Prediction
4. Conclusions
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 $\theta$
- $P(x, \theta)$: joint probability of $x$ and $\theta$

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

**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

\[ P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} \]

- \( P(D | \theta) \) : likelihood of \( \theta \)
- \( P(\theta) \) : prior probability of \( \theta \)
- \( P(\theta | D) \) : posterior of \( \theta \) given \( D \)

Model Comparison:

\[ P(m | D) = \frac{P(D | m)P(m)}{P(D)} \]

\[ P(D | m) = \int P(D | \theta, m)P(\theta | m) \, d\theta \]

Prediction:

\[ P(x | D, m) = \int P(x | \theta, D, m)P(\theta | D, m) \, d\theta \]

\[ P(x | D, m) = \int P(x | \theta)P(\theta | D, m) \, d\theta \quad \text{(for many models)} \]
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.

\[
P(Y| M_i)
\]

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

\[
P(x) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)P(x_4|x_2)P(x_5|x_3, x_4)
\]

Key quantity: joint probability distribution over nodes:
\[
P(x) = P(x_1, x_2, \ldots, x_n)
\]

The graph specifies a factorization of this joint probability distribution.

Also known as Bayesian Networks, Belief Nets and Probabilistic Independence Nets.
The central event in the generation of an immune response is the activation of T cells.
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
A Gaussian State-Space Model with Feedback

Output equation:
\[ y_t = Cx_t + Dy_{t-1} + v_t \]

State dynamics equation:
\[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

Key Concept: \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) 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.
Output equation: $y_t = Cx_t + Dy_{t-1} + v_t$

State dynamics equation: $x_t = Ax_{t-1} + By_{t-1} + w_t$

Key Concept: $y_t$ represents the measured gene expression level at time step $t$ and $x_t$ 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: \[ y_t = Cx_t + Dy_{t-1} + v_t \]
State dynamics equation: \[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

Key Concept: \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) models the many unmeasured (hidden) factors such as
- genes that have not been 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:
\[ y_t = Cx_t + Dy_{t-1} + v_t \]

State dynamics equation:
\[ x_t = Ax_{t-1} + By_{t-1} + w_t \]

Key Concept: \( y_t \) represents the measured gene expression level at time step \( t \) and \( x_t \) 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)

\[ Z = (z_1, z_2, \ldots, z_N) \]

\[ Z^*_{1} \quad Z^*_{2} \quad Z^*_{B} \]

S(\(Z^*_{1}\)) \quad S(\(Z^*_{2}\)) \quad S(\(Z^*_{B}\))

Bootstrap replications
Bootstrap samples
Training sample

From Hastie, Tibshirani and Friedman “The Elements of Statistical Learning”
Bootstrap Procedure for Parameter Confidence Intervals (2)
Inferring Regulatory Networks

Some key genes:
- FYB (1), IL3Rα (2), CD 69 (3), TRAF5 (4), IL4Rα (5), GATA binding protein 3 (6), IL-2Rγ (7), chemokine receptor CX3CR1 (9), interleukin-16 (11), Jun B (13), Caspase 8 (14), Clusterin (15), Caspase 7 (18), survival of motor neuron 1 (19), Cyclin A2 (20), CDC2 (21), PCNA (22), Integrin alpha-M (26), MCL-1 (31)
In-Silico Hypotheses

**A**
- TCR
- PMA
- FYB
- IL-2

**B**
- IL2 R
- C-X3-C receptor 1
- Caspase 8
- IL2 R gamma
- IL4 R alpha
- GATA3
- CD69
Variational free energy minimization is a method of approximating a complex distribution $p(x)$ by a simpler distribution $q(x; \theta)$. We adjust the parameters $\theta$ so as to get $q$ to best approximate $p$ in some sense.

From David J.C. MacKay “Information Theory, Inference and Learning Algorithms”
We can also **lower bound** the marginal likelihood: Using a simpler, factorised approximation to 
\( q(x, \theta) \approx q_x(x)q_\theta(\theta) \):

\[
\ln p(y|m) = \mathcal{F}_m(q_x(x), q_\theta(\theta), y).
\]

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)

Motivation and Background
Inferring Gene Regulatory Networks from Microarray Data
Protein Structure Prediction
Conclusions
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 overexpression 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 overexpression predictions
- combining gene and protein expression data with metabolomic data
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 overexpression 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 overexpression 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
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 $T = [T_1, T_2, \ldots, T_m]$. 
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
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:** \( \mathcal{P}(O_k|\theta_k, T_i) = \frac{(\sum_a O_{ak}^a)!}{\prod_a O_{ak}^a!} \prod_{a \in A} (\theta_k^a)^{O_{ak}^a} \)
- **Dirichlet Prior:** \( \mathcal{P}(\theta_k|O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a \in A} (\theta_k^a)^{\gamma_k^a - 1} \)
- **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.

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

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

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

- **Multinomial:** \( P(O_k | \theta_k, T_i) = \frac{(\sum_a O_k^a)!}{\prod_a O_k^a!} \prod_{a \in A} (\theta_k^a)^{O_k^a} \)
- **Dirichlet Prior:** \( P(\theta_k | O_{[1:k-1]}, T_i) = \frac{\Gamma(\sum_a \gamma_k^a)}{\prod_a \Gamma(\gamma_k^a)} \prod_{a \in A} (\theta_k^a)^{\gamma_k^a-1} \)
- **Weights:**
  \[ \gamma_k = W_{cap} + \sum_{j=1}^{\ell_k} W_{intra}^j \cdot O_{k-j} + \sum_{j=\ell_k+1}^{\ell} W_{inter}^j \cdot O_{k-j}. \]
Motivation and Background
Inferring Gene Regulatory Networks from Microarray Data
Protein Structure Prediction
Conclusions

CASP5 Results

<table>
<thead>
<tr>
<th>Chain Length</th>
<th>$Q_3^{casp5}$</th>
<th>$SOV^{casp5}$</th>
<th>$Q_3^{culled}$</th>
<th>$SOV^{culled}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>215.75</td>
<td>74.6±10.3 %</td>
<td>73.4±12.3 %</td>
<td>74.9±7.5%</td>
</tr>
</tbody>
</table>
The $\beta$-sheet space is the set of all the possible combinations of $\beta$-sheets;
A set of interaction variables, $\mathcal{I}$, to describe one possible case.
1PGA - PROTEIN G

True Contact Map of 1PGA

Predictive β-sheet Contact Map of 1PGA
Combining the Probabilistic Model with Steric Constraints

Model and moves

- Planar rigid peptide bonds
- Elastic $C_\alpha$ valence geometry
- Random pivotal rotations
- Random crankshaft rotations

![Diagram of protein structure with $C_\alpha$ bonds and rotational movements]
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

Motivation and Background
Inferring Gene Regulatory Networks from Microarray Data
Protein Structure Prediction
Conclusions

Ramachandran Plots from PDB


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

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

- $3_{10}$-helices are 3 times more likely than $\alpha$-helices.
- Double H-bonds with a common acceptor are responsible for $\phi = -120^\circ$ and $\psi = -40^\circ$.
Contacts Sampled in Monte Carlo Procedure

True Contact Map of 1PGA

Predictive β-sheet Contact Map of 1PGA
Future Work

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

- Combining probabilistic model and steric constraints
- De-novo protein design
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.
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 (University 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 (University 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 (University 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 (University 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”
Functional assignment by homology:
the function-homology gap

yeast data analyzed by GeneQuiz
Structural Homologs and Analogs (1)

## Comparison to Cuff and Barton (2000)

<table>
<thead>
<tr>
<th>Method Description</th>
<th>$Q_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Networks using frequency profile from CLUSTALW</td>
<td>71.6%</td>
</tr>
<tr>
<td>Networks using BLOSUM62 profile from CLUSTALW</td>
<td>70.8%</td>
</tr>
<tr>
<td>Networks using PSIBLAST alignment profiles</td>
<td>72.1%</td>
</tr>
<tr>
<td>Arithmetic sum based on the above three networks</td>
<td>73.4%</td>
</tr>
<tr>
<td>Networks using PSIBLAST PSSM</td>
<td>75.2%</td>
</tr>
<tr>
<td>Our algorithm with MSAP</td>
<td>71.3%</td>
</tr>
<tr>
<td>Our algorithm with PSIBLAST PSSM</td>
<td>72.2%</td>
</tr>
</tbody>
</table>
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 \( \sigma = 0.3 \)