Markov Chain Monte Carlo in Bayesian Statistics, Phylogenetic Reconstruction and Protein Structure Prediction

Biomath Seminar

September 28, 2010

Outline

The Bayesian Paradigm
Conditional Probablity
Bayes Formula
Markov Chains
Transition Probabilities
Stationary Measures
Reversibility
Ergodic Theorem
Monte Carlo
Simple Monte Carlo
Markov Chain Monte Carlo
Metropolis Hastings Algorithm
Gibbs Sampling
Applications
Bayesian Posteriors
Phylogenetic Reconstruction
Protein Structure Prediction - ROSETTA

Conditional Probability

For a *probability* P and an event c with $P(C) > 0$, define the *conditional probability*

$$P(A|C) = \frac{P(A \cap C)}{P(C)}.$$

The defining formula for conditional probability can be rewritten to obtain *the multiplication principle*

$$P(A \cap C) = P(A|C)P(C).$$

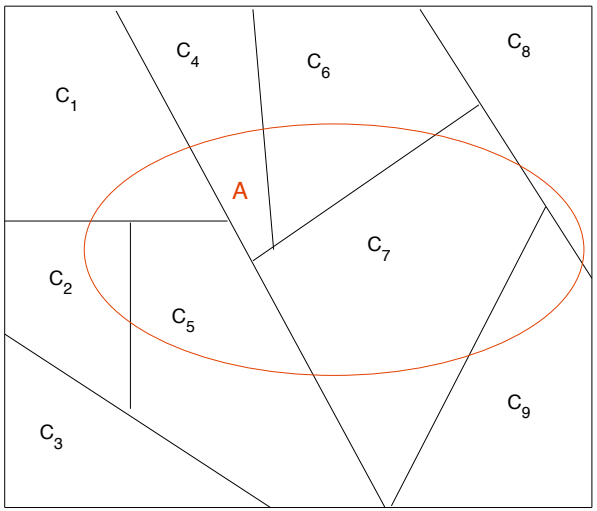
Using the multiplication formula **twice**

$$P(A \cap C) = \begin{cases} P(A|C)P(C) \\ \\ P(C|A)P(A) \end{cases}$$

$$P(C|A)P(A) = P(A|C)P(C) \quad \text{or} \quad P(C|A) = \frac{P(A|C)P(C)}{P(A)}.$$

Law of Total Probability

A *partition* of the sample space Ω is a finite collection of pairwise disjoint events $\{C_1, C_2, \ldots, C_n\}$ whose union is Ω.



Thus, every point belongs to *exactly* one of the C_i. In particular, distinct members of the partition are mutually exclusive. ($C_i \cap C_j = \emptyset$, if $i \neq j$.)

Theorem (Law of total probability)

Let P be a probability on Ω. and let $\{C_1, C_2, \ldots, C_n\}$ be a partition of Ω chosen so that $P(C_i) > 0$ for all i. Then, for any event $A \subset \Omega$

$$P(A) = \sum_{i=1}^{n} P(A|C_i)P(C_i).$$

Theorem (Bayes formula)

Let P be a probability on S. and let $\{C_1, C_2, \ldots, C_n\}$ be a partition of Ω chosen so that $P(C_i) > 0$ for all i. Then, for any event $A \subset \Omega$ and any j

$$P(C_j|A) = \frac{P(A|C_j)P(C_j)}{\sum_{i=1}^{n} P(A|C_i)P(C_i)}.$$

Bayesian Statistics

We begin with a *parameter space* Ψ. Our goal is to use data to estimate parameter values.

The *Bayesian approach to statistics* takes into account external information to determine a *prior density* π for the value of the parameter θ. Thus, in this approach, both the parameter and the data are random.

Estimation and hypothesis testing are based on Bayes formula.

Bayesian Statistics

Let $\tilde{\Theta}$ be a random variable having the given prior density π. In the case where both $\tilde{\Theta}$ and the data take on only a finite set of values, then Bayes formula is

$$\begin{aligned} f_{\Theta|X}(\theta|\mathbf{x}) &= P\{\tilde{\Theta} = \theta | X = \mathbf{x}\} \\ &= \frac{P\{X = \mathbf{x}|\tilde{\Theta} = \theta\}P\{\tilde{\Theta} = \theta\}}{\sum_{\psi} P\{X = \mathbf{x}|\tilde{\Theta} = \psi\}P\{\tilde{\Theta} = \psi\}} \\ &= \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi\{\theta\}}{\sum_{\psi} \mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi\{\psi\}} \end{aligned}$$

Given data $\mathbf{x}$, the function of θ, $f_{\Theta|X}(\theta|\mathbf{x})$ is called the *posterior density*.

For a continuous distribution on the parameter space, π is now a density function and the sums in Bayes formula become integrals.

$$f_{\Theta|X}(\theta|\mathbf{x}) = \frac{\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta)\pi(\theta)}{\int \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\, d\psi}$$

Example

Suppose that the prior density is a normal random variable with mean θ_0 and variance $1/\lambda$. Data X are independent normal random variables with unknown mean θ, variance 1.

The posterior density is also normally distributed, mean

$$\theta_1(\mathbf{x}) = \frac{\lambda}{\lambda + n}\theta_0 + \frac{n}{\lambda + n}\bar{x}.$$

and variance $1/(\lambda + n)$.

Bayesian Statistics

We often call $\mathbf{f}_{X|\Theta}(\mathbf{x}|\theta) = L(\theta|\mathbf{x})$ the *likelihood function* and

$$\log L(\theta|\mathbf{x})$$

the *score function*. In this way we write

$$f_{\Theta|X}(\theta|\mathbf{x}) \propto L(\theta|\mathbf{x})\pi(\theta).$$

The posterior distribution is proportional to the product of the likelihood function and the prior distribution.
The constant of proportionality

$$\zeta(\mathbf{x}) = \int \mathbf{f}_{X|\Theta}(\mathbf{x}|\psi)\pi(\psi)\ d\psi$$

can be difficult to evaluate.

Markov Chains

Definition

A process X is called a *Markov chain* with values in a state space S if

$$\begin{aligned} P\{X_{n+m} \in A | X_1, X_2, \ldots X_n\} &= P\{X_{n+m} \in A | X_n\} \\ &= \phi(m, n, X_n, A) \end{aligned}$$

for all $m, n \geq 0$ and sets A.

In words, given the entire history of the process up to time n, the only part that is useful in predicting the future is X_n, the position of the process at time n.

If ϕ does not depend on n, we call X *time homogenuous*.

Markov Chains - Transition Probabilities

Let the state space S be countable, then we can define the *transition probabilities*

$$T(x, y) = P\{X_{n+1} = y | X_n = x\}.$$

Then the probability of any event can be determined using T and the *initial distribution*

$$\alpha(A) = P\{X_0 \in A\}.$$

For example, the n time step transition

$$P\{X_n = y | X_0 = x\} = T^n(x, y).$$

Markov Chains - Classification of States

Definition

A state y is *accessible* from a state x (written $x \longrightarrow y$) if for some $n \geq 0$, $T^n(x, y) > 0$. If $x \longrightarrow y$ and $y \longrightarrow x$, then we write $x \longleftrightarrow y$ and say that x and y *communicate*.

$\longleftrightarrow$ is an *equivalence relation*. Thus communication partitions the state space by its equivalence classes.

1. Call a set of states C *closed* if for all $x \in C$ and all $n \geq 0$, $P_x\{X_n \in C\} = 1$.
2. A Markov chain is called *irreducible* if all states communicate.

Markov Chains - Classification of States

Set $\tau_y = \min\{n > 0 : X_n = y\}$. The state y is

1. *recurrent* if $P_y\{\tau_y < \infty\} = 1$,
2. *transient* if $P_y\{\tau_y = \infty\} > 0$,
3. *positive recurrent* if y is recurrent and $E_y\tau_y < \infty$,

   The *period* of a state y, $\ell(y)$ if τ_y is distributed on the lattice $L_{\ell(y)}$ given $X_0 = y$. The state y is
4. *periodic* if $\ell(y) > 2$,
5. *aperiodic* if $\ell(y) = 1$,
6. *ergodic* if it is positive recurrent and aperiodic.

Theorem

Communicating states have the same period.

Markov Chains - Recurrence and Stationary Measures

Definition

A measure σ for a transition probability T is called *stationary* if

$$\sigma\{y\} = \sum_{x \in S} \sigma\{x\} T(x, y),$$

or in matrix form $\sigma = \sigma T$. Note that we are *not* requiring that σ be a *probability* measure.

To explain the term stationary measure, note that if σ is a probability measure and is the initial distribution, then the identity above becomes

$$P_\sigma\{X_0 \in A\} = P_\sigma\{X_1 \in A\} = \cdots = P\{X_n \in A\}.$$

Call a Markov chain *stationary* if X_n has the same distribution for all n

Markov Chains - Reversibility

For a stationary Markov chain, Bayes' formula allows us to look at the process in reverse time.

$$\begin{aligned} T_-(x,y) &= P_\sigma\{X_n = y | X_{n+1} = x\} \\ &= \frac{P_\sigma\{X_{n+1} = x | X_n = y\} P_\sigma\{X_n = y\}}{P_\sigma\{X_{n+1} = x\}} \\ &= \frac{\sigma\{y\}}{\sigma\{x\}} T(y,x). \end{aligned}$$

Definition

The Markov chain X^- in reverse time is called the *dual Markov process*. If $T = T_-$, then the Markov chain is called *reversible* and the stationary distribution satisfies *detailed balance*

$$\sigma\{y\} T(y,x) = \sigma\{x\} T(x,y). \tag{1}$$

Sum this equation over y to see that it is a stronger condition than stationarity.

Markov Chains - Ergodic Theorem

Theorem

If T is the transition matrix for an ergodic Markov chain with stationary distribution σ, then

$$\lim_{n\to\infty} T^n(x,y) = \frac{1}{E_y \tau_y} = \sigma\{y\}.$$

Theorem

(Ergodic theorem for Markov chains) Assume X is an ergodic Markov chain and that f is bounded, then for any initial distribution α,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to^{a.s.} \int_S f(y)\ \sigma(dy).$$

Simple Monte Carlo

If the goal is to compute an integral

$$\int g(x)\ \sigma(dx),$$

then, in circumstances in which the probability measure σ is easy to simulate, *simple Monte Carlo* suggests creating independent samples

$$X_0(\omega), X_1(\omega), \ldots$$

having distribution σ. Then, by the *law of large numbers*,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} g(X_j(\omega)) = \int g(x)\ \sigma(dx) \text{ with probability } 1.$$

The error is determined by the *central limit theorem*. If that does not get the job done ...

Markov Chain Monte Carlo

Given a probability distribution σ, one way to sample from this distribution is to

- construct an ergodic Markov chain whose stationary distribution is σ, and
- use the ergodic theorem.

This strategy is used to

- Find the posterior distribution in a Bayesian statistics problem
- The distribution of phylogenetic trees consistent with the data.
- Find the (relative) positions of the atoms in a protein.

We will now explore the similarities and differences in these two questions.

Markov Chain Monte Carlo

We construct an irreducible Markov chain

$$\tilde{X}_0, \tilde{X}_1, \ldots$$

having stationary distribution σ. The most commonly used strategy to define this sequence is the method developed by Metropolis and extended by Hastings.
Assume

- a countable state space S, (to avoid technical issues on the first pass)
- σ is not trivial,
- T be a Markov transition matrix on S
- so that the the chain immediately enter states that have positive σ probability.

In other words, if $T(x, y) > 0$ then $\sigma\{y\} > 0$.

Metropolis Hastings Algorithm

Define

$$\alpha(x,y) = \begin{cases} \min\left\{\frac{\sigma\{y\}T(y,x)}{\sigma\{x\}T(x,y)}, 1\right\}, & \text{if } \sigma\{x\}T(x,y) > 0, \\ 1, & \text{if } \sigma\{x\}T(x,y) = 0. \end{cases}$$

- If $\tilde{X}_n = x$, generate a *candidate value* y with probability $T(x,y)$.
- With probability $\alpha(x,y)$, this candidate is accepted and $\tilde{X}_{n+1} = y$.
- Otherwise, the candidate is rejected and $\tilde{X}_{n+1} = x$.

Consequently, the transition matrix for this Markov chain is

$$\tilde{T}(x,y) = \alpha(x,y)T(x,y) + (1 - \alpha(x,y))\delta_x\{y\}.$$

Metropolis Hastings Algorithm

Note that

- this algorithm only requires that we know the ratios $\sigma\{y\}/\sigma\{x\}$ and thus we do not need to normalize σ.
- if $\sigma\{x\}T(x,y) > 0$ and if $\sigma\{y\} = 0$, then $\alpha(x,y) = 0$ and thus the chain cannot visit states with $\sigma\{y\} = 0$.

Claim. $\tilde{T}$ is the transition matrix for a reversible Markov chain with stationary distribution σ.

We must show that σ satisfies the detailed balance equation (1). Consequently, we can limit ourselves to the case $x \neq y$.

Metropolis Hastings Algorithm

Case 1. $\sigma\{x\}T(x,y) = 0$.

In this case $\alpha(x,y) = 1$ and $T(x,y) = \tilde{T}(x,y)$.

- If $\sigma\{y\} = 0$, then, $\sigma\{y\}\tilde{T}(y,x) = 0$,

$$\sigma\{x\}\tilde{T}(x,y) = \sigma\{x\}T(x,y) = 0,$$

  and (1) holds.
- If $\sigma\{y\} > 0$ and $T(y,x) > 0$, then $\alpha(y,x) = 0$, $\tilde{T}(y,x) = 0$ and (1) holds.
- If $\sigma\{y\} > 0$ and $T(y,x) = 0$, then $\alpha(y,x) = 1$, $\tilde{T}(y,x) = T(y,x) = 0$ and (1) holds.

Metropolis Hastings Algorithm

Case 2. $\sigma\{x\}T(x,y) > 0$ and $\alpha(x,y) = 1$

In this case,

$$\sigma\{x\}\tilde{T}(x,y) = \sigma\{x\}T(x,y).$$

In addition, $\alpha(y,x) \leq 1$ and

$$\sigma\{y\}\tilde{T}(y,x) = \sigma\{y\}\frac{\sigma\{x\}T(x,y)}{\sigma\{y\}T(y,x)}T(y,x) = \sigma\{x\}T(x,y).$$

Case 3. $\sigma\{x\}T(x,y) > 0$ and $\alpha(x,y) < 1$.

$$\sigma\{x\}\tilde{T}(x,y) = \sigma\{x\}\frac{\sigma\{y\}T(y,x)}{\sigma\{x\}T(x,y)}T(x,y) = \sigma\{y\}T(y,x).$$

In addition, $\alpha(y,x) = 1$ and

$$\sigma\{y\}\tilde{T}(y,x) = \sigma\{y\}T(y,x).$$

Thus, the claim holds.

Metropolis Hastings Algorithm

Example

The original Metropolis algorithm had $T(x, y) = T(y, x)$ and thus

$$\alpha(x, y) = \min\left\{\frac{\sigma\{y\}}{\sigma\{x\}}, 1\right\}.$$

Example (Independent Chains)

Let $\{X_n; n \geq 0\}$ be independent discrete random variable with distribution function $f(x) = P\{X_0 = x\}$. Then

$$\alpha(x, y) = \min\left\{\frac{w(y)}{w(x)}, 1\right\},$$

where $w(x) = f(x)/\sigma\{x\}$ is the *importance weight* function that would be used in importance sampling if the observations if observations were generated from f.

Metropolis Hastings Algorithm

Example

Take T to be the transition matrix for a random walk on a graph and let σ to be uniform measure. Then

$$T(x,y) = \begin{cases} \frac{1}{\deg(x)}, & \text{if } x \text{ and } y \text{ are adjacent} \\ 0, & \text{if not.} \end{cases}$$

$$\alpha(x,y) = \begin{cases} \min\left\{\frac{\deg(x)}{\deg(y)}, 1\right\}, & \text{if } x \text{ and } y \text{ are adjacent} \\ 1, & \text{if not.} \end{cases}$$

Thus, to visit each node of the graph with equal probability, always move to a point with lower degree and move to a point with a higher degree according to the ratio of degrees.

Gibbs Sampling

The Gibbs sampler is a Metropolis Hastings algorithm attuned to the case that the state space S is a high dimensional vector space. Call this dimension N.

Let x_k indicate the k-th coordinate of $\mathbf{x}$ and let $\mathbf{x}^{-k}$ be $\mathbf{x}$ with the k-th coordinate removed. Thus, for successive values of $k = 1, 2, \ldots N$, we find transition probabilities

$$T(x_k, \mathbf{x}^{-k})$$

that change only one coordinate.

Implementation Issues

- Any appropriate T will produce a sample from σ, but some choices are much better than others.
- One long run is better than many short.
- Determinging run length - what is the *burn-in time*?
- Variance reduction
- Monitoring the output.
- Numerical stability

Bayesian Posterior Distributions

The state space S is now the parameter space Ψ. The probability distribution is the posterior distribution with density

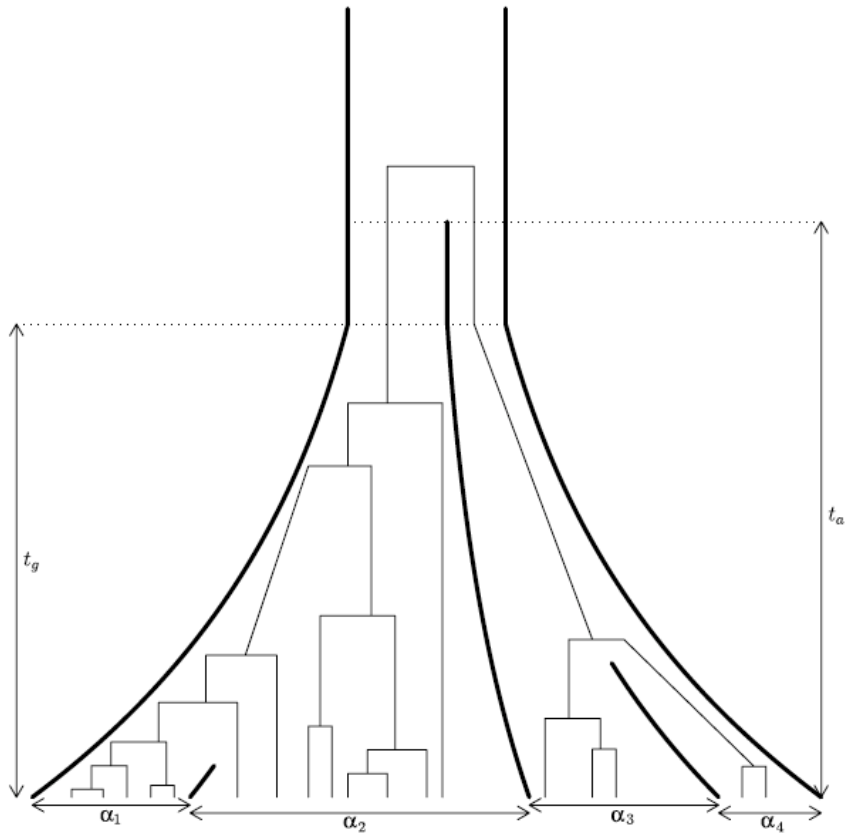
$$\sigma(\theta) = f_{\Theta|X}(\theta|\mathbf{x}) = \zeta(\mathbf{x})L(\theta|\mathbf{x}).\pi(\theta)$$

Thus, the Metropolis-Hastings algorithm has

$$\alpha(\theta,\psi) = \begin{cases} \min\left\{\frac{f_{\Theta|X}(\psi|\mathbf{x})T(\psi,\theta)}{f_{\Theta|X}(\theta|\mathbf{x})T(\theta,\psi)}, 1\right\}, \text{ if } f_{\Theta|X}(\theta|\mathbf{x})T(\theta,\psi) > 0, \\ 1, \text{ if } f_{\Theta|X}(\theta|\mathbf{x})T(\theta,\psi) = 0. \end{cases}$$

Phylogenetic Reconstruction

The goal: Find the distribution of phylogenetic trees consistent with the data.



Phylogenetic Reconstruction - Model Parameters

For a Splitting Plus Growth Model Parameters

- N_a - ancestral population size
- t_a - time of subpopulation split
- t_g - time of population growth
- α_i - subpopulation growth rates
- π_i - subpopulation split proportions

Nuisance Parameters - from mutation model

- mutation rates
- mutation probabilities

Phylogenetic Reconstruction - State Space

A genealogical tree can be described using three attributes.

- tree topology
- genotype of nodes
- branch lengths

With n terminal nodes, we have

$$(2n-3) \times (2n-5) \times \cdots \times 3 \times 1$$

topologies.

nodes	5	10	20	30
topologies	105	3.446×10^7	8.201×10^{21}	4.952×10^{38}

Phylogenetic Reconstruction - Markov Chain

Choose parameter values according a prior distribution π
The Markov chain T allows three types of moves

- Change an interior node genotype
- Change a branch length
- Change a tree topology
  - Pick a node and detach a subtree from the parent of the node.
  - Attach with a probablility that depends on the genetic similarity of the new parent.

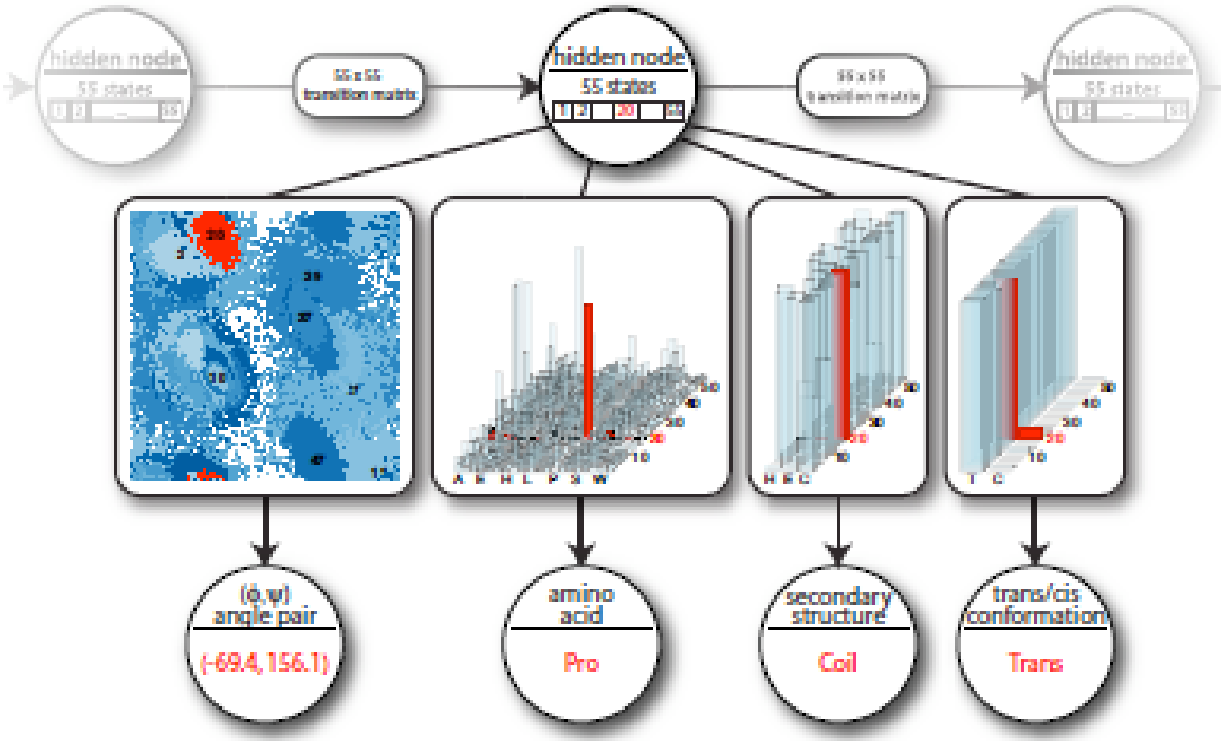
Changing topology is a rare event compared to the other two changes. Likelihood computed using a *pruning algorithm*.

Protein Structure Prediction

- *Prior distribution*
  - experimental parameters the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{\alpha}$, and $^{1}H^{N}$ *nuclear magnetic resonance chemical shifts* of the polypeptide backbone atoms
  - protein fragments from the *Protein Data Bank*
  - interproton distance restraints from *multidimensional nuclear Overhauser enhancement spectra*. (Side chain assignments are the most time consuming step.)
  - ROSETTA selects two hundred fragments from the crystallographic structural database that are similar in amino acid sequence
  - A Monte Carlo based assembly searches for compact, low energy folds - *the negative of a penalized score function*

Protein Structure Prediction- Bayesian network

Using a hidden Markov model to give a better Gibbs sampler.



hidden state $\longrightarrow$ dihedral angles amino acids (a), secondary structure (s), and the*cis* or *trans* conformation of the peptide bond (c).