ISMB-95
ROBINSON COLLEGE, CAMBRIDGE

Tutorial Programme
Sunday 15 July 1995

TUTORIAL T8

Foundations of Statistical Methods for Multiple Sequence Alignment and Structure Prediction

(Chip Lawrence)

Foundations of statistical methods for multiple sequence alignment and structure prediction

Chip Lawrence

Biometrics Lab, Wadsworth Labs, NYS-DOH, Albany, NY
Chip.lawrence@wadsworth.org

National Center for Biotechnology Information
NLM-NIH Bethesda, MD
Lawrence@ncbi.nlm.nih.gov

Statistical Foundations Tutorial Outline

I. Introduction
   A. Statistical algorithms a breakthrough for multiple sequence alignment.
   B. Conceptual foundations.

II Boltzmann like model of residue frequencies

III Permuted data likelihood
   A. Examples
      1. Gene regulation problem
      2. Coin tossing example
   B. Complete and incomplete data loglikelihoods.
   C. Alternate from of incomplete data loglikelihood

IV. Expectation Maximization algorithm
   A. E step and M step
   B. EM theory

V. Gibbs Sampler
   A. Bayesian background
   B. Predictive inference and predictive update algorithm.

VI. Multiple Elements (gaps)
   A. Gibbs.
   B. Propagation
      1. Forwards/Backwards algorithm
   C. Hidden Markov Models
      1. Missing alignment as using missing triples

VII. Flexibility/Sensitivity tradeoff

VIII. Threading
   A. Pairwise interactions and contacts in 3 space

IX. Stochastic Context Free Grammars
   A. Alignment and Structure (contacts) Missing

Statistical algorithms:
A break through for multiple alignment

1. Previously alignment and statistics as unrelated steps.

These algorithms merge these two steps.

2. Multiple alignment as an NP-hard many to many comparison.

Convert to a series of many to one comparisons

Novel Statistically Based Methods

1. Multiple Sequence Alignment

   Expectation Maximization (EM)

   Gibbs Sampler

   Hidden Markov Models (HMM)

2. Alignment of Sequence to Structures.

   Threading

3. Structural Prediction and Alignment

   Stochastic Context Tree Grammars

Two major conceptual underpinnings

1. **Boltzmann like relationship of residue frequencies and energetic constraints**

   Statistical Mechanics: A states energy and its occupancy
   (Boltzmann & Gibbs)

   Residue frequency analogy
   (Bryant SH and Lawrence CE, (1991), *Proteins*, 9:108-119)

2. **Alignments as Missing data**

   Expectation Maximization
   (Little RJA and Rubin DB, (1987), Statistical Analysis with Missing Data, Wiley, NY NY)

   Data Augmentation
   (Tanner, MA, (1993) Tools for Statistical Inference, Springer-Verlag, NY, NY)

Boltzmann like model of residue frequencies

Maximize entropy subject to energy constraints

$$Max \quad S = \sum p_i \log(p_i)$$

$$St: \quad \sum e_i \, p_i = E$$

Yields

$$p_i = \frac{\exp(-\beta \, e_i)}{Z}$$

$$where \quad Z = \sum \exp(-\beta \, e_i)$$

$$\frac{p_i}{p_0} = \exp(-\beta (e_i - e_o)) = \exp(-\beta \, \Delta e_i)$$

Frequencies of ion pairs

Coulomb`s Law

$$\Delta e = \frac{\epsilon q}{d}$$

Screening and distant dependent dielectric constant

$$\Delta e = \frac{\epsilon(d)\, q}{d}$$

and,

$$\frac{p_i}{p_0} = \exp\left(\frac{-\beta \epsilon(d_i)\, q}{d_i}\right)$$

Alignment as Missing Data

Permuted Data Log Likelihood

Examples:

1. Transcriptional Gene Regulation (Slides 1-4)
2. Coin tossing game

V_n be the vector of correctly aligned residues in the n^{th} sequence.

$V_n \sim f(V_n \mid \Theta)$, where Θ are generic parameters.

In multiple sequence alignment the residues are assumed to be independent:

$$f(V_n \mid \Theta) = \prod_{b=a,c}^{t,g} \prod_{j=1}^{K} p_{b,j}^{I_b(V_{n,j})} \prod_{K+1}^{L} p_{b,0}^{I_b(V_{n,j})}$$

where,

$I_b(z)$ = 1 if z = b and 0 otherwise.

$p_{b,j}$ = the probability of base b at position j in the site.

$p_{b,0}$ = the probability of base b in any non site position.

Ψ be the set of permitted distinguishable permutations, which we will index by r.

Φ_r be the permutation operator for the r^{th} permutation.

Φ_r^{-1} be the inverse permutation operator for the r^{th} permutation.

R_n be a random variable indicating which permutation has been applied to the n sequence.

$Y_{n,r}$ be 1 if $R_n = r$, and 0 otherwise.

X_n be of observed n^{th} sequence.

$$X_n = \Phi_r(V_n)$$

$$V_n = \Phi_r^{-1}(X_n)$$

$$R_n \sim P(R_n \mid \Lambda)$$

For example,

$$P(R_n = r \mid \Lambda) = \lambda_r$$

Complete data loglikelihood {X,Y}

$$L(X,R|\Theta,\Lambda) = P(X,R|\Theta\Lambda)$$

$$\prod_n \prod_{r\in\psi} [f(\phi_r^{-1}(X_n)|\Theta)\, P(R_n|\Lambda)]^{Y_n,r}$$

The incomplete data loglikelihood is obtained by summing over r {X}

$$L(X|\Theta,\Lambda) = P(X|\Theta,\Lambda)$$

$$\prod_n \sum_{r\in\psi} f(\phi_r^{-1}(X_n)|\Theta)\, P(R_n = r|\Lambda)$$

Coins Example with Missing Data: $f(X_n|\theta)$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
T	T	T	H	T	T	T	T	T	H	T	T	H	H	H	T	T	H	T	T	T
T	T	T	H	T	H	T	T	H	T	T	T	T	T	H	T	H	T	T	H	T
T	T	T	T	T	T	T	H	H	T	H	T	T	T	H	T	H	H	H	H	T
T	T	T	T	T	H	H	T	H	T	H	H	H	T	T	T	H	H	T	T	T
T	T	H	T	H	T	T	T	H	H	H	T	T	T	H	T	H	T	T	T	T

Coins Example with Complete Data $P(X_n, Y_n|\theta)$

										1	1	1	1	1	1	1	1	1	1	2
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
T	T	T	H	T	T	T	T	T	H	T	T	H	H	H	T	T	H	T	T	T
												1								
T	T	T	H	T	H	T	T	H	T	T	T	T	T	H	T	H	T	T	H	T
			1																	
T	T	T	T	T	T	T	H	H	T	H	T	T	T	H	T	H	H	H	H	T
																	1			
T	T	T	T	T	H	H	T	H	T	H	H	H	T	T	T	H	H	T	T	T
										1										
T	T	H	T	H	T	T	T	H	H	H	T	T	T	H	T	H	T	T	T	T
								1												

Aligned Coin Sequences $f(V_n|\theta)$

										1	1	1	1	1	1	1	1	1	2
1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
H	H	H	T	T	T	H	T	T	T	T	T	H	T	T	T	T	H	T	T	T
H	T	H	T	T	T	T	T	H	T	T	T	T	T	H	T	H	T	T	H	T
H	H	H	T	T	T	T	T	T	T	H	H	T	H	T	T	T	H	T	H	T
H	H	H	T	T	T	T	T	H	H	T	H	T	T	T	T	H	H	T	T	T
H	H	H	T	T	H	T	H	T	T	T	T	T	T	H	T	H	T	T	T	T

$p_0 = .25, \quad p_1 = .95, \quad p_2 = .8, \quad p_3 = .98$

Multiple Sequence Alignment

Assume positions operate independently

$$f(V_{1,n}, V_{2,n}, ..., V_{L,n}|\Theta) = \prod_j P(V_{n,j}|\Theta)$$

This means the energies are additive

$$\log(f(V_{1,n}, V_{2,n}, ..., V_{L,n}|\Theta)) = \sum_j \log(P(V_{n,j}|\Theta))$$

Block alignment

$$f(V_{1,n}, V_{2,n}, \ldots, V_{L,n}|P) = \prod_{j=1}^{J} P_{j,i}^{I_b(V_{n,j})} \prod_{j \notin S} P_{0,j}^{I_b(V_{n,j})}$$

where $I_b(v) =$ (if v = b, and 0 otherwise).

Complete Data log likelihood {X,Y}

$$\log(P(X,Y|\Theta,\Lambda) = l(X,Y|\Theta,\Lambda) =$$

$$\sum_n \sum_{r\in\Psi} \{Y_{n,r} \log(f(\Phi_r^{-1}(X_n)|\Theta) + Y_{n,r} \log(P(R_n|\Lambda)\}$$

with $f(.|\Theta)$ multinomial with parameters

$$P_{j,b}$$

The maximum likelihood parameter estimates are:

$$\hat{P}_{j,b} = \frac{n_{\phi_r^{-1}(j),b}}{N}$$

EM Algorithm

E Step:

$$Q(\Theta|\Theta^t, X) = E_{y/x}(l(X, Y|\Theta, \Lambda))$$

$$= E_{y/x}(\sum_n \sum_r Y_{n,r} \log(f(\Phi_r^{-1}(X_n|\Theta) + Y_{n,r} \log(P(R_n|\Lambda))$$

$$= (\sum_n \sum_r \zeta_{n,r} \log(f(\Phi_r^{-1}(X_n|\Theta) + \zeta_{n,r} \log(P(R_n|\Lambda))$$

,where

$$\zeta_{n,r} = E_{y/x}(Y_{n,r}) = P(Y_{n,r} = 1|\Theta^t, \Lambda^t, X_n)$$

$$= \frac{P(X_n|Y_{n,r} = 1, \Theta^t)P(R_n|\Lambda^t)}{\sum_r P(X_n|Y_{n,r} = 1, \Theta^t)P(R_n|\Lambda^t)}$$

Coins Example with Missing Data: $F(X_n \mid \theta)$

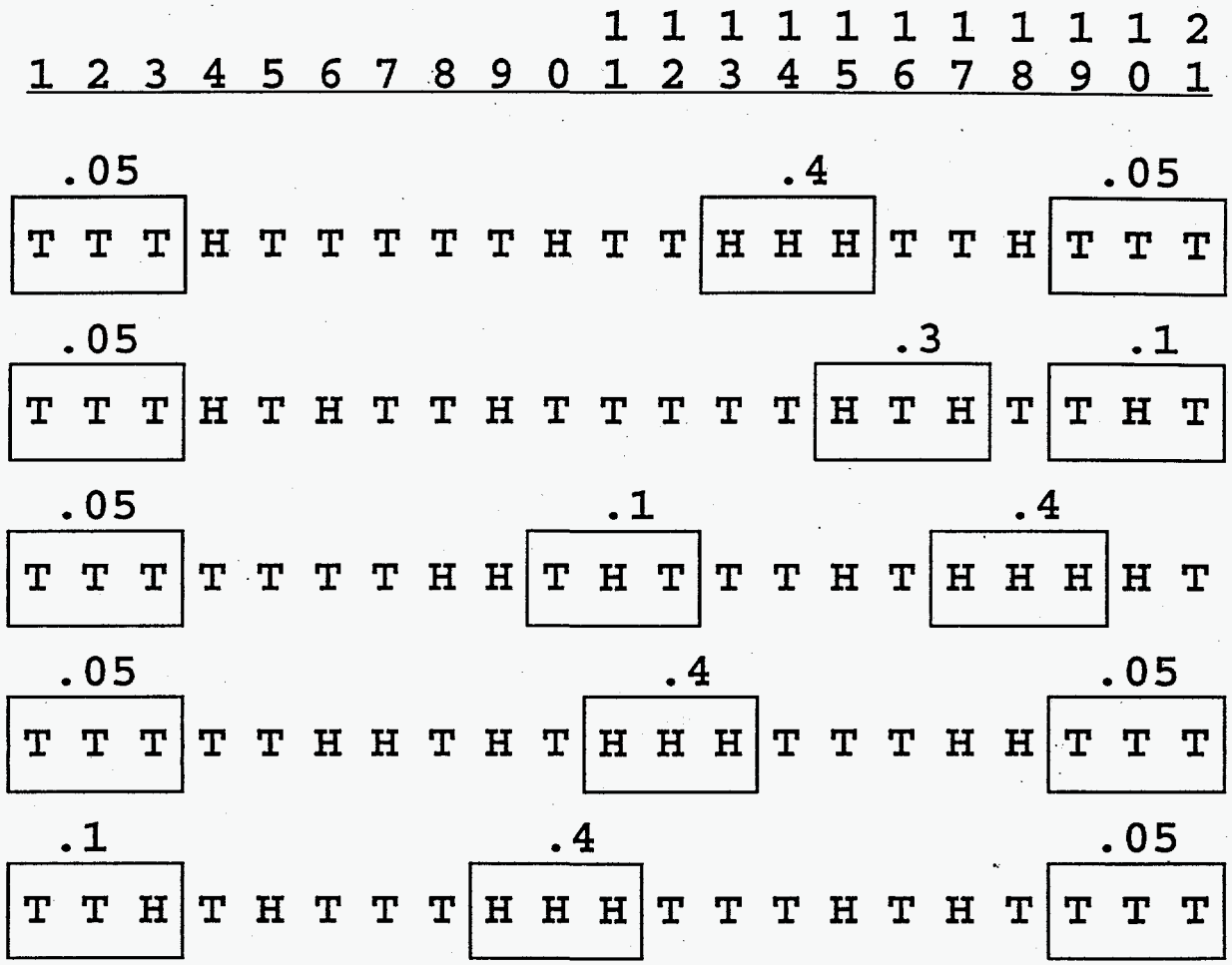
```
                                1 1 1 1 1 1 1 1 1 1 2
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

T T T H T T T T T H T T H H H T T H T T T

T T T H T H T T H T T T T T H T H T T H T

T T T T T T T H H T H T T T H T H H H H T

T T T T T H H T H T H H H T T T H H T T T

T T H T H T T T H H H T T T H T H T T T T
```

$$p(X|Y) = P_{1,X_j} \; p_{2,x_{j+1}} \; p_{3,x_{j+2}} \cdot P_o n_{H,H} \; p_o n_{T,T}$$

Bayes Theorem

P(Y|X)

Coins Example with Missing Data: $F(X_n \mid \theta)$



P(Y|X)

M Step:

$$MaxQ(\Theta|\Theta^t, X) \longrightarrow \Theta^{t+1}$$

Notes:

$$0 \leq \zeta_{n,r} \leq 1$$

replaces

$$Y_{n,r} = (0, 1)$$

and the MLEs are

$$\hat{P}_{j,b} = \frac{\tilde{n}_{j,b}}{N} = \frac{E_{y/x}(n_{j,b})}{N}$$

EM Theory

$$P(X, Y|\Theta, \Lambda) = P(X|\Theta, \Lambda)P(Y|X, \Theta, \Lambda)$$

$$l(X, Y|\Theta, \Lambda) = l(X|\Theta, \Lambda) + \log(P(Y|X, \Theta, \Lambda))$$

$$l(X|\Theta, \Lambda) = l(X, Y|\Theta, \Lambda) - \log(P(Y|X, \Theta, \Lambda))$$

Taking expectations over $P(Y|X, \Theta^t, \Lambda^t)$

$$l(X|\Theta, \Lambda) = Q - H$$

where

$$Q = E_{y/x}(l(X, Y|\Theta, \Lambda))$$

$$H = E_{y/x}(\log(P(y|X, \Theta, \Lambda)$$

$$= \sum_r P(Y|X, \Theta^t, \Lambda^t) \log(P(Y|X, \Theta^t, \Lambda^t)$$

Information Inequality

$$H(\Theta|\Theta^t) \leq H(\Theta^t|\Theta^t)$$

$$l(X|\Theta^{t+1}) - l(X|\Theta^t)$$

$$= \{Q(\Theta^{t+1}|\Theta^t) - Q(\Theta^t|\Theta^t)\} - \{H(\Theta^{t+1}|\Theta^t) - H(\Theta^t|\Theta^t)\}$$

Gibbs Sampler

Bayesian Prospective

$$P(\Theta|X) = \frac{P(X|\Theta)P(\Theta)}{\int P(X|\Theta)P(\Theta)d\Theta}$$

Conjugate prior,

$P(X|\Theta)$ has same form as $P(\Theta)$

For example,

$P(X|\Theta) \sim$ multinomial

$P(\Theta) \sim$ Dirchlet

$P(\Theta|X) \sim$ Dirchlet

Markov Monte Carlo(MMC)

An iterative Markov chain sampling scheme whose equilibrium distribution is the joint distribution of interest.

Gibbs Sampler: MMC using complete set of conditionals

Joint

$$P(R, \Theta, \Lambda | X)$$

Complete set of conditionals

$$P(R | \Theta, \Lambda, X)$$

$$P(\Theta | R, \Lambda, X)$$

$$P(\Lambda | R, \Theta, X)$$

Predictive inference (Liu, 1994, JASA 89:958)

$$\Pi_k = P(R_k|R_1, ..., R_{k-1}, R_{k+1}, ...R_L) = P(R_k|R_{[k]})$$

$$= \int P(R_k|\Theta)P(\Theta|R_{[k]})d\Theta$$

Predictive Update Multiple Alignment

Under the condition that $P(R_k|\Lambda)$ is a constant (Lawrence et.al., 1994, HICSS 27th 5:245-255)

$$\Pi_n = P(R_k|R_{[k]}) \propto E(P(\Theta|R_{[k]})$$

$$= \frac{\prod_j(\tilde{n}_{j,b} + \beta_b)}{N + B}$$

where is a defined above, $\bar{\beta}$ is vector of pseudo count priors, $N = \sum \tilde{n}_{j,b}$, and $B = \beta_b$.

Multiple Elements (of arbitrary size) & gaps

For example,

1) Multiple ligand binding motifs

2) Multiple structurally conserved regions

Gibbs:

$$P(R_k^{(1)}, R_k^{(2)}, ..., R_k^{(q)} | \bar{R}_{[k]})$$

Complete set of conditionals

$$P(R_k^{(1)} | R_k^{(2)}, ..., R_k^{(q)}, \bar{R}_{[k]})$$

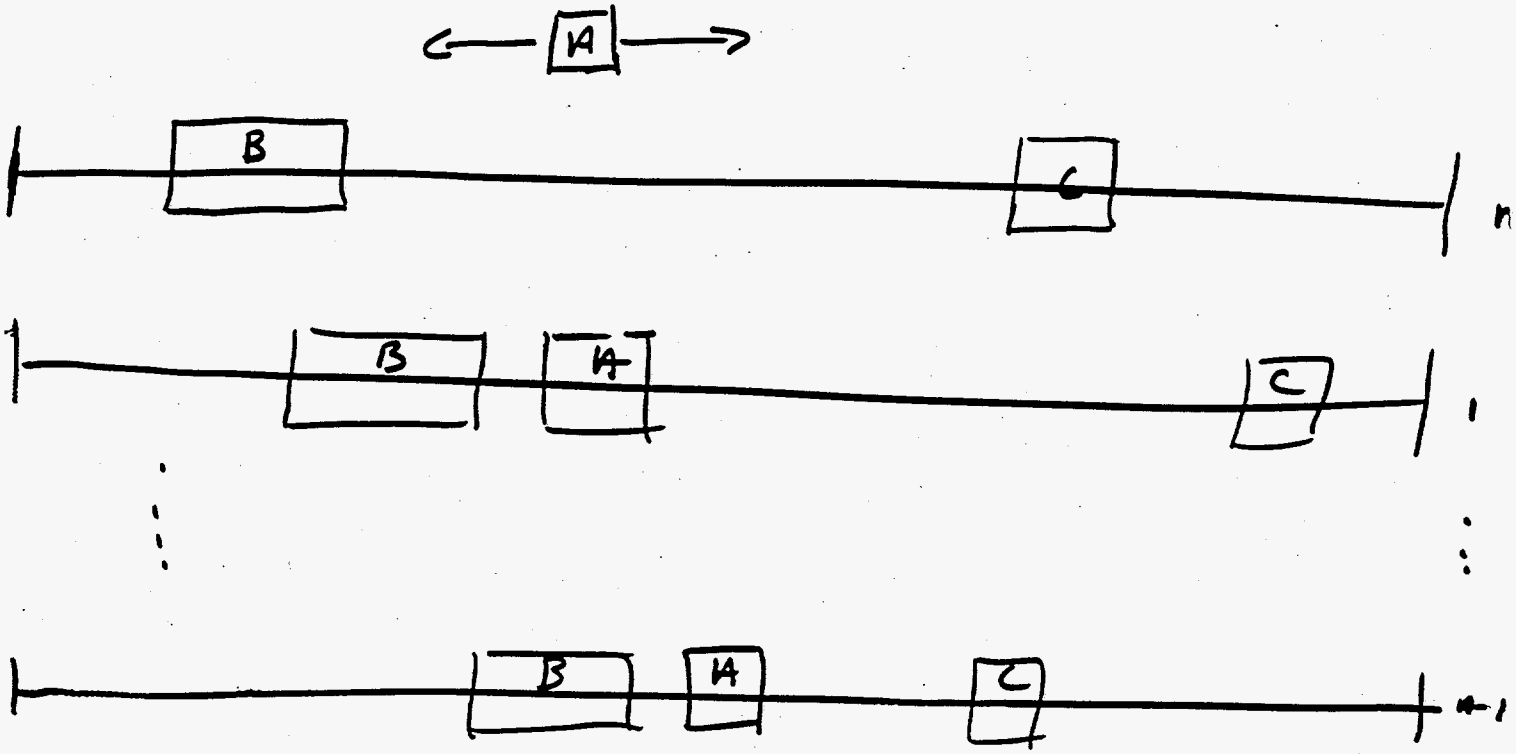
$$P(R_k^{(2)} | R_k^{(1)}, ..., R_k^{(q)}, \bar{R}_{[k]})$$

$$P(R_k^{(q)} | R_k^{(1)}, ..., R_k^{(q-1)}, \bar{R}_{[k]})$$

Three element example

Residue frequency model from $\bar{R}_{[k]}$, and conditioning on the alignments of $R_k^{(b)}, R_k^{(c)}$.

$$P(R_k^{(a)} | R_k^{(b)}, R_k^{(c)}, \bar{R}_{[k]})$$



$\bar{R}_1, \ldots \bar{R}_{n-1}$ - ① residue frequency model

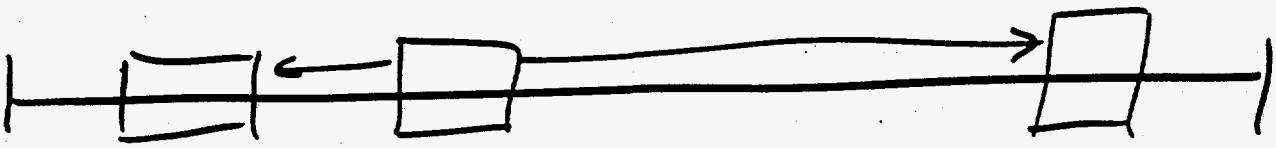
② Joint alignment information of A, B, C.

$$P(R_n^{(a)} \mid \bar{R}_1, \bar{R}_2 \ldots \bar{R}_{n-1}, R_n^b, R_n^c)$$

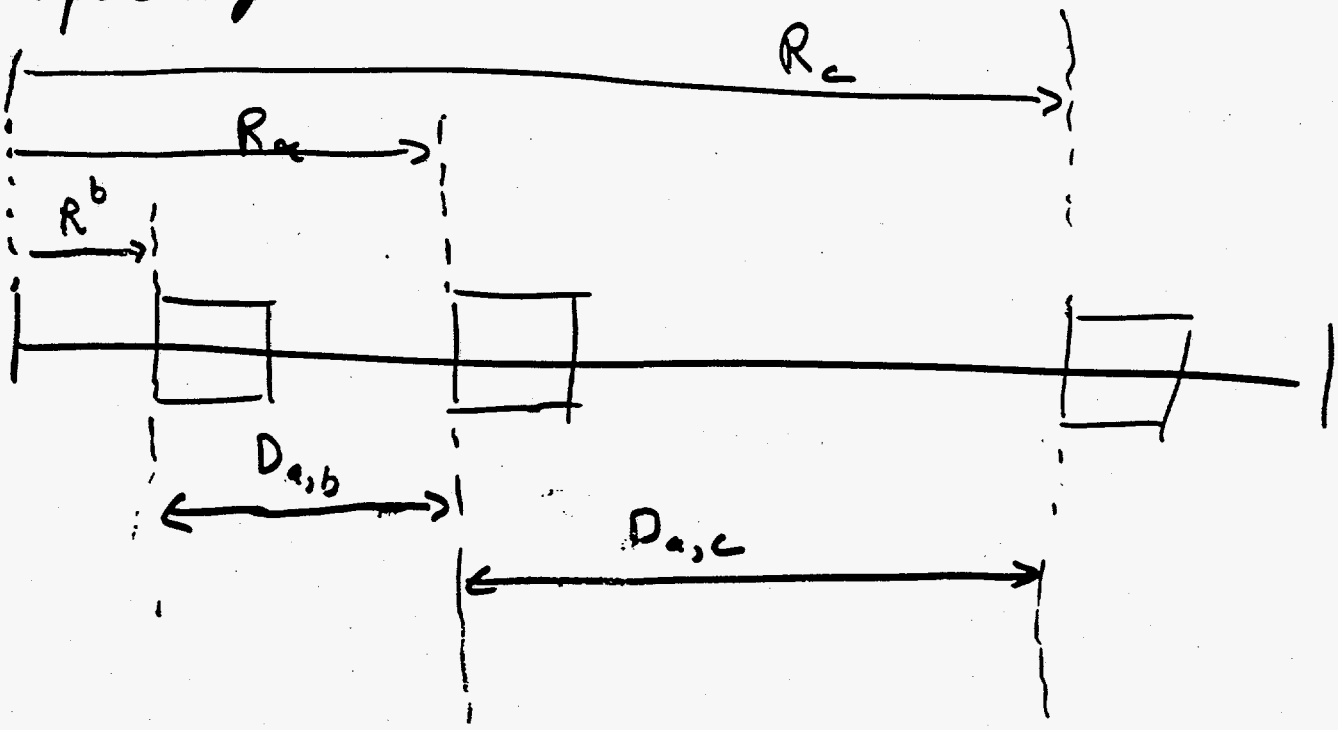
Joint Alignment Information

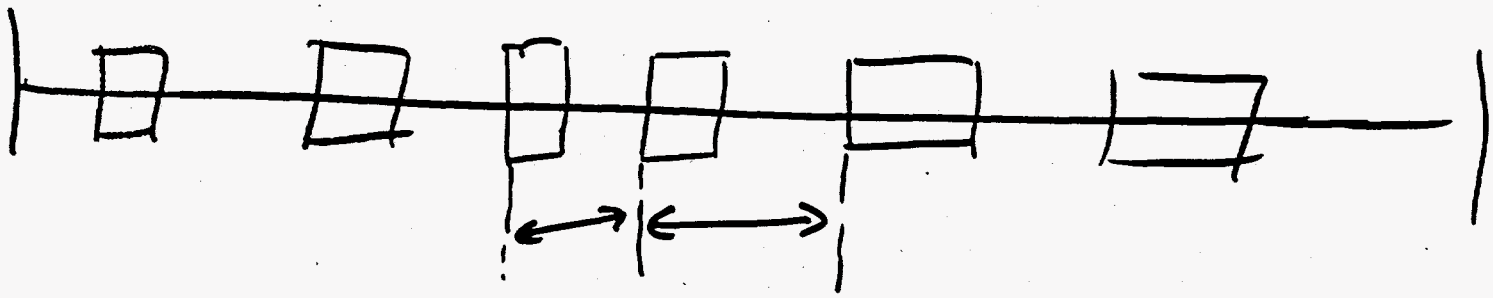
1) Overlaps (& deletions)

2) Element order



3) Spacing





Colinearity: Markov Dependence

$$P(R_k^v | R_k^{[v]}, \bar{R}_{[k]}) = P(R_k^v | R_k^{v-1}, R_k^{v+1}, \bar{R}_{[k]})$$

1) Dependence only on two nearest neighbors

2) How can we capture dependence on both?

Forward Backward Algorithm

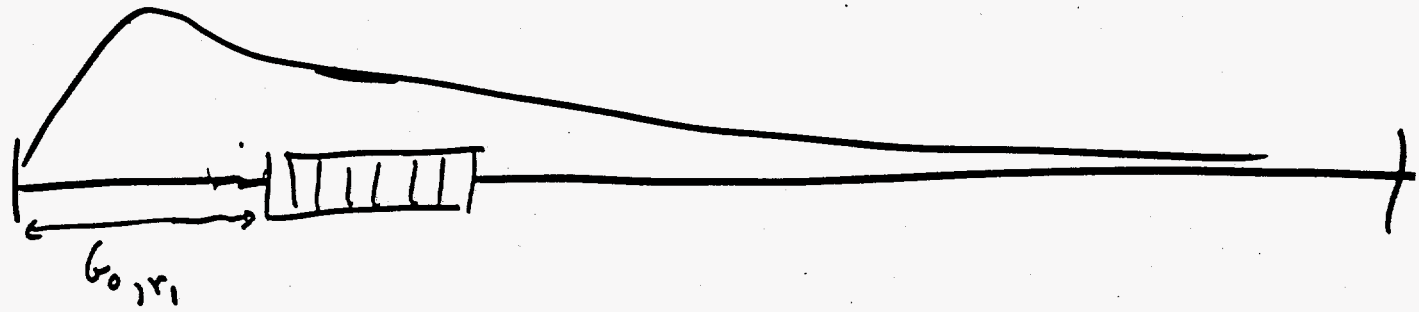
Forward

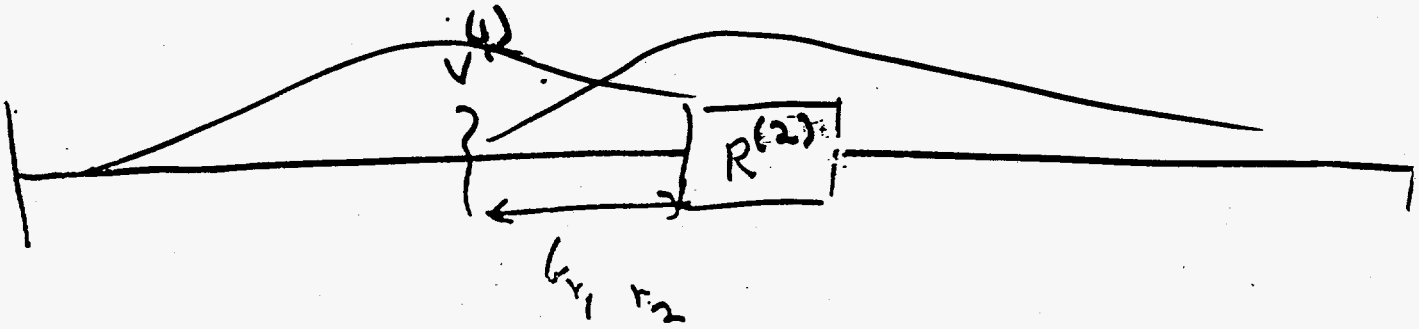
$$\nu^1 = P(R^1|.) \propto \Pi_1 G_{0,1}$$

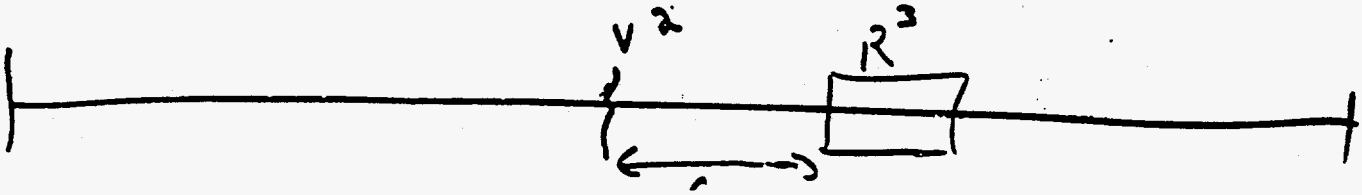
$$\nu^2 = P(R^2|.R^1)\nu^1 \propto \Pi_2 G_{1,2}\nu^1$$

$$\nu^3 = P(R^3|.R^2, R^1) = P(R^3|.R^2)P(R^1|.) \propto \Pi_2 G_{1,2}\nu^2$$

$$\nu^q \propto \Pi_q G_{q,q-1}\nu^{q-1}$$


G_{0,r_1}

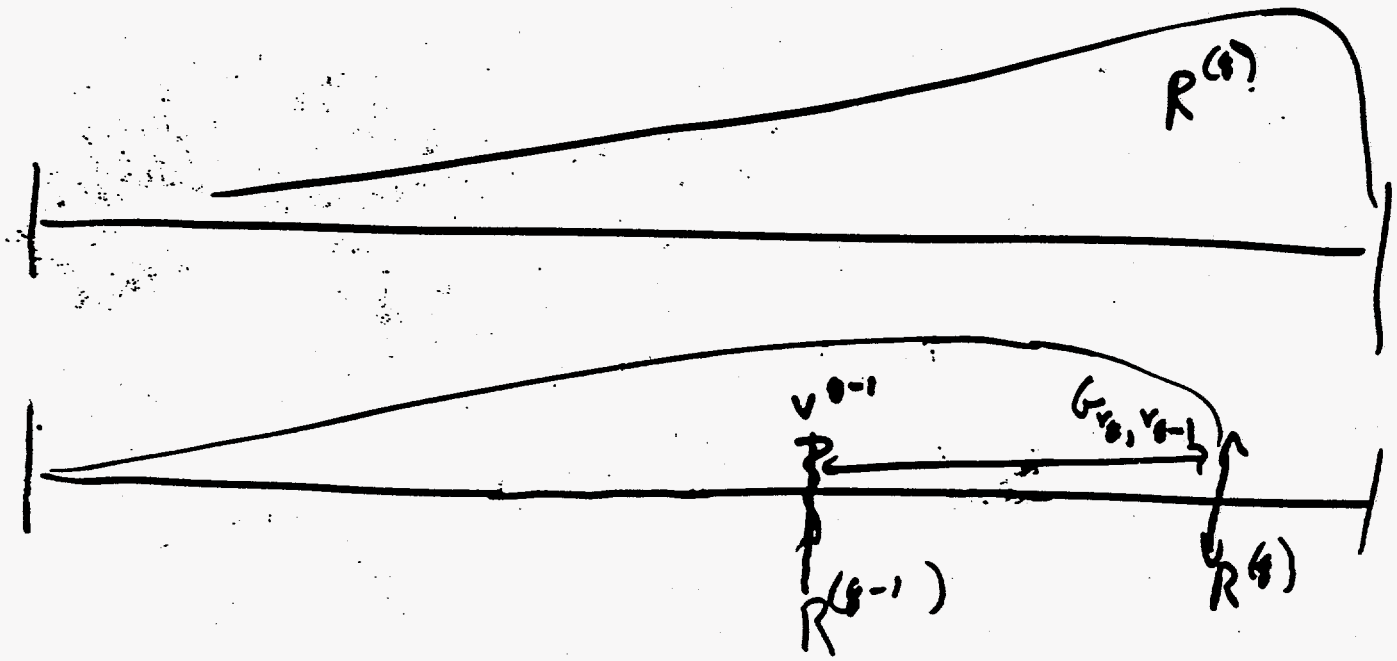

$V^{(1)}$
$R^{(2)}$
G_{r_1,r_2}


V^2
R^3

Backward

1) Draw R_l^q from ν^q , true marginal distribution of the last element

2) Draw R_l^{q-1} conditional on location of R_l^q including the effect of $G_{q,q-1}$



HMMs for Multiple Sequence Alignment

(Krogh, et.al. 1994, JMB 235:1501)(Baldi, et.al, 1994, PNAS 91:1059)

EM

Use Markov dependence

Forward/Backward algorithm

Product multinomial model of residue frequency

Missing Alignment Data

Set of 3 state hidden variables

Match

Insert

Delete

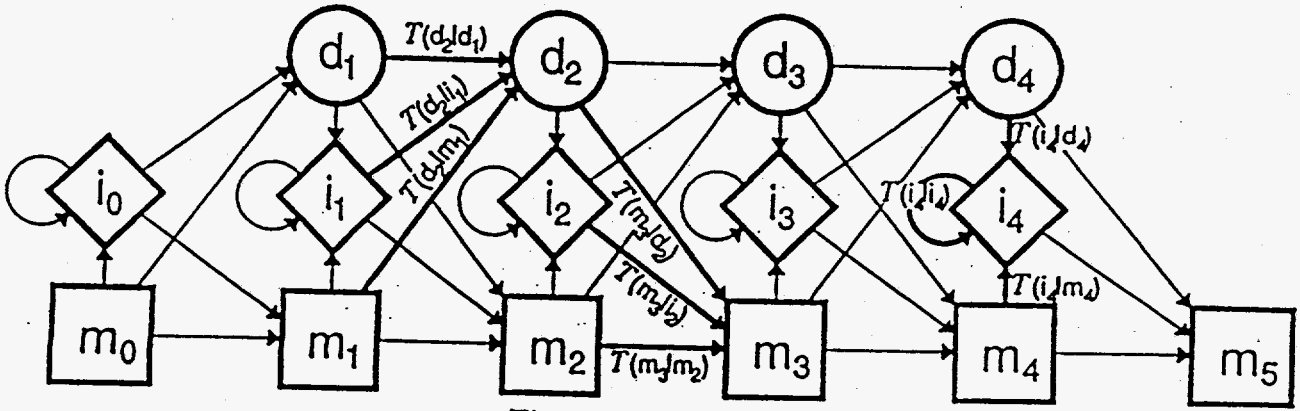
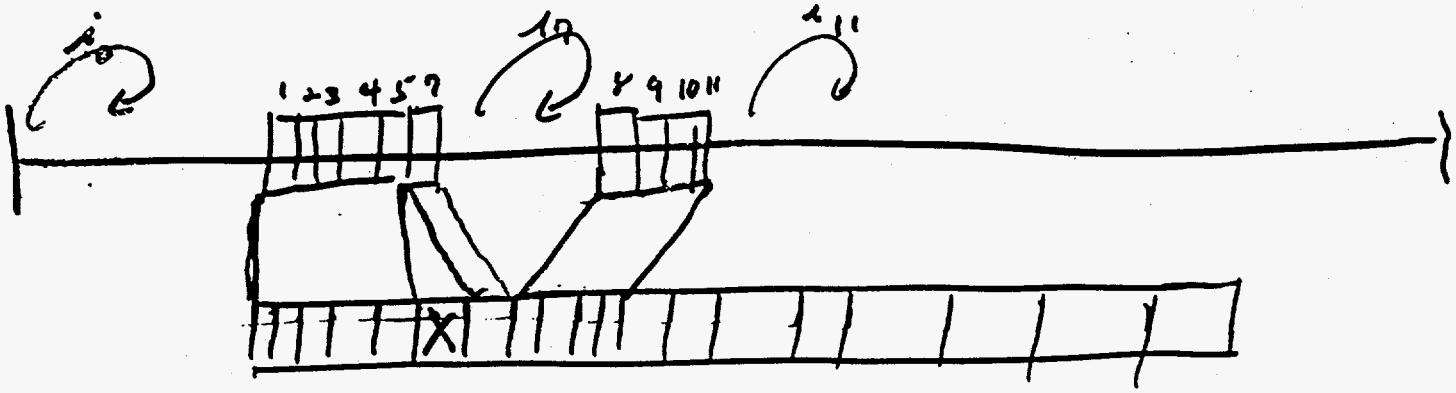


Figure 1. The model.



The missing data model

$P(R|\Lambda) =$ Markov Chain Model of transitions between hid

Learning of parameters of missing data model

As before, even though we observe no of the missing variables the algorithm can nevertheless learn parameter estimates about these models. In this way position specific gap penalties are learned.

Flexibility Sensitivity Trade-off

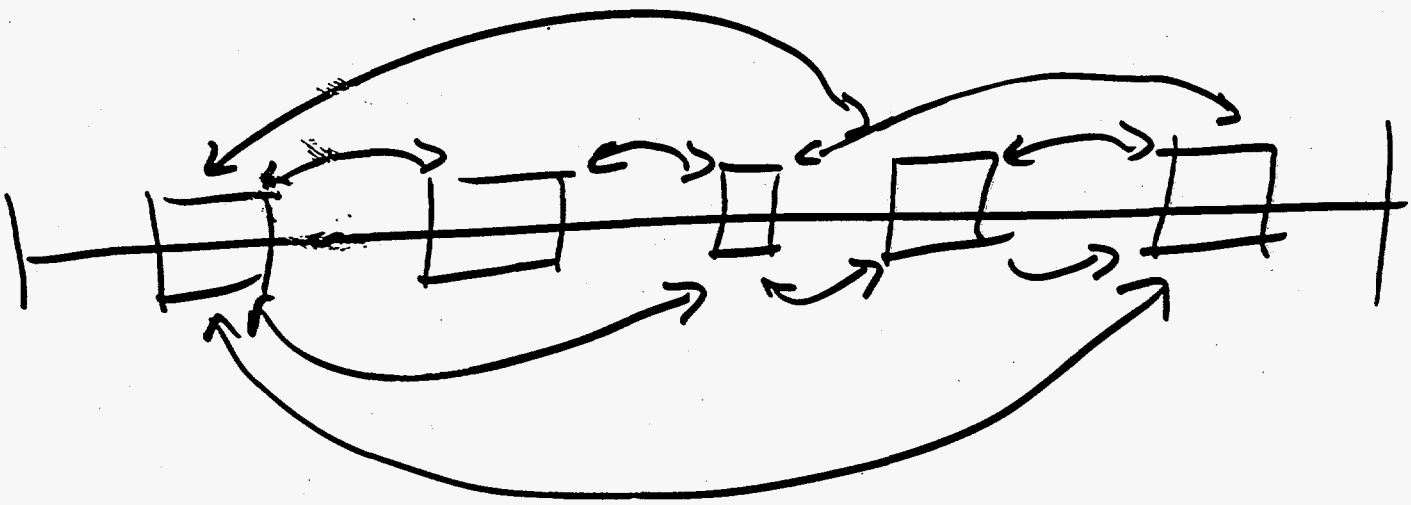
$$l(X|\Theta,\Lambda) = Q - H$$

$$= \sum\sum \zeta_{n,r} \log(f(\Phi_r^{-1}(X_n)|\Theta)$$

$$+ \sum\sum \zeta_{n,r} \log(P(R_n|\Lambda)) - \sum\sum \zeta_{n,r} \log(\zeta_{n,r})$$

$$= \sum\sum \zeta_{n,r} \log(f(\Phi_r^{-1}(X_n)|\Theta)$$

$$- \sum\sum P(R_n|\Lambda,\Theta,X) \log(\frac{P(R_n|\Lambda,\Theta,X)}{P(R_n|\Lambda)})$$



Threading

$$P(R_1, R_{2,...,}R_n|\Theta) = \prod_{pairs} P(R_i, R_j|\Theta)$$

$$\log(P(R_1, R_{2,...,}R_n|\Theta)) = \sum_j \mu_{R_j} + \sum_j \sum_{i>j} \mu_{R_i,R_j,d_{i,j}}$$

Alignment of sequence to structurally conserved regions

Markov dependence is gone

PBD derived empirical potentials provide a set of known parameters

Gibbs Sampling Algorithm

$$P(R^v | R^{[v]}, \bar{\mu})$$

Statistical Foundations Tutorial Outline

I. Introduction
   A. Statistical algorithms a breakthrough for multiple sequence alignment.
   B. Conceptual foundations.

II Boltzmann like model of residue frequencies

III Permuted data likelihood
   A. Examples
      1. Gene regulation problem
      2. Coin tossing example
   B. Complete and incomplete data loglikelihoods.
   C. Alternate from of incomplete data loglikelihood

IV. Expectation Maximization algorithm
   A. E step and M step
   B. EM theory

V. Gibbs Sampler
   A. Bayesian background
   B. Predictive inference and predictive update algorithm.

VI. Multiple Elements (gaps)
   A. Gibbs.
   B. Propagation
      1. Forwards/Backwards algorithm
   C. Hidden Markov Models
      1. Missing alignment as using missing triples

VII. Flexibility/Sensitivity tradeoff

VIII. Threading
   A. Pairwise interactions and contacts in 3 space

IX. Stochastic Context Free Grammars
   A. Alignment and Structure (contacts) Missing

APPENDIX

Foundations of Statistical Algorithms for Multiple Sequence Alignment and Structure Prediction

Chip Lawrence

Biometrics Lab, Wadsworth Labs, Albany, NY 12201
Internet: Chip.Lawrence@wadsworth.org
Phone (518) 473-3382 FAX (518) 474-8590

and

National Center for Biotechnology information
NLM, NIH, Bethesda, MD
Internet: lawrence@ncbi.nlm.nih.gov
Phone: (301) 496-2475 FAX: (301) 480-9241

Abstract

Recently, statistical algorithms have proved to be useful for several problems in computational molecular biology. These included the following: an EM algorithm for identification and characterization of DNA regulatory binding sites (Lawrence and Reilly, 1990); a Gibbs sampling algorithm for local multiple alignment of subtle sequence signals in protein sequence (Lawrence et al. , 1993); alignment of large families of protein sequence using hidden Markov models (HMM) (Baldi et.al, 1994 & Haussler et al. 1994); the threading of sequence through structural motifs (Bryant and Lawrence, 1993) and the prediction of common RNA secondary structures using context-free grammars (Sakakibara et al., 1994) (Eddy and Durbin 1994). In each of these cases, critical alignment and/or structural data are missing.

In the 1970s it became widely recognized that many statistical problems are most easily addressed by pretending that critical missing data are available. In fact, for some problems, statistical inference is facilitated by creating a set of latent variables, none of whose values are observed. The key observation was that conditional probabilities for the values of the missing data could be inferred by application of Bayes theorem to the observed data. Statistical inference based on this concept was called the "missing information principle" (Orchard and Woodbury, 1972). Its application became widely known through a deterministic maximum likelihood algorithm, expectation maximization (EM) algorithm (Dempster et al., 1977). The use of sampling methods, known as data augmentation methods, for problems involving missing data were developed in the late 1980s (Tanner and Wong, 1987 & Li, 1988). This common statistical framework forms the basis of all of the above algorithms.

I. Introduction

These methods derive their power for multiple sequences problems from two key characteristics. 1) They reduce a many-to-many comparison problem to a many-to-one comparison: each sequence to a common evolving statistical model. 2) They effectively capture both the characteristics common to the set of sequences and the variability across its members in a pair of stochastic models. These methods employ two stochastics models: the first models the residue probabilities in observed sequence data given an alignment, while the second models the missing alignment data given the residue probabilities. Clearly the problem would be easy if we knew the alignment, for then the residue probability parameters could be estimated via a simple tabulation. The reverse is also true if we know the residue probabilities for the common elements, then the probability of the alterative alignment can readily determined. These problems are made challenging by the fact that at the outset neither the alignment nor the residue probabilities are known. These methods iteratively cycle between these two models to adaptively "learn" both the alignment and the residue probabilities.

II. Boltzmann like models

The success of these methods is to a large extent dependent on how well the chosen models represent the underlying biology. Mutations in biopolymer sequences may be classified into four categories: point mutations, insertions/deletions, transpositions, and duplications. The products of point mutations will be accepted if they satisfy the functional and structural constraints of the biopolymer. At the molecular level, these constraints often take the form of energetic requirements on the interactions of the residues of the biopolymer with one another or with their environment. The relationship between energetic constraints and frequencies forms the basis of statistical mechanics, pioneered by Gibbs and Boltzmann. There is an analogous relationship for residue frequencies subject to random point mutations (Berg and von Hipple 1987, Bryant and Lawrence 1991), which forms the foundation of the models used here. From a statistical modeling prospective this relationship is quite valuable since it allows for the translation from the language of physics and chemistry that governs molecular behavior to the language of statistics, to yield a stochastic model that represents the underlying science.

Multinomial models have been used successfully to capture both the variability and the limitations shared by common elements in a set of sequences. Because the most important interactions of the residues of a biopolymer are frequently with the environment as opposed to with one another (Bryant and Lawrence, 1993), multiple sequence alignment models that assume independence of the residue positions have enjoyed considerable success.

III. Permuted data likelihood

The other three classes of mutations, insertions/deletions, transpositions, and duplications, result in changes in the length of the sequence or in reordering of the sequence. These events result in

a permutation of the indices of the data, and since the effects of these events are not directly observable in biopolymer sequences, these data are missing. It thus falls in the class of statistical problems concerned with the analysis of data with unobserved index permutations (Lawrence and Reilly, 1996). The fundamental feature of all of these statistical methods is use of stochastic models to "impute" this missing data given the data that is observed, the sequences. Let us consider in more detail the effect of deletions with specific attention on the conserved segments. A deletion mutation will remove a segment of a protein sequence, and the resulting two adjacent fragments of the protein will be shifted to form a continuous chain. If the deletion is ``upstream" of the conserved segment, then this segment will be shifted to the left, and thus misaligned with respect to its predecessor. Insertions operate in an analogous manner but add sequence segments. Transpositions move a segment to a new location in the sequence. Duplications replicate segments and then insert them in new locations. To account for these unobserved events, Bayes theorem is employed to find the conditional distribution of the alignment variables given the sequences and a residue frequency model of the form described in the previous paragraph.

To help fix the ideas, consider the following coin tossing analogy. The game is played with L coins, say 50. (L-J) of these coins, say 40 (plain coins), all have the same probability of heads not necessarily a half. The remaining J =10 special coins have probabilities of heads different from the plain coins and different from one another. On each of K independent trials, corresponding to K observed sequences in alignment problems, the coins are shaken in a tumbler and laid out in a row. Consequently, not only have the coins been flipped but also their order has been permutted. The player is challenged to estimate the probabilities of the 11 types of coins, the parameters of the residue frequency model, and to specify the locations, the alignment, of the special coins in each trial. Various restrictions of the permitted permutations tend to simplify the problem. If the special coins are required to remain in order, sequences are said to be collinear. If further, the 10 special coins are positioned in a contiguous block, then only a single motif must be aligned.

IV & V. EM and Gibbs sampler algorithms

We used this last and simplest case to explain the concepts behind the two major algorithm classes of algorithms that have been employed to find solutions to these alignment problems: expectation maximization (EM) algorithms and Gibbs sampling algorithms. When the permutation constraints of this simplest case are relaxed, we must deal with the joint distribution of the multiple alignment variables for each sequence. Because the maximization step of EM algorithm required direct access to the joint distribution, it is not (straightforward) to extend these algorithms to more complex alignment problems. On the other hand, Gibbs sampler provides a means to access this complex distribution through the conditional distribution of one alignment variable given the others; thus, extension to arbitrarily complex joint distributions is conceptually straightforward and practical for many real alignment problems. However, if the number of free alignment variables becomes large, the convergence of the Gibbs sampling algorithm becomes problematic.

VI. Multiple elements

If attention is restricted to collinear alignments, then the recursive relationship which forms the basis of the well known dynamic programming algorithms for the alignment of pairs of sequences can be brought to bare. When this recursive relationship is used with Gibbs sampling, a "propagation" algorithm which samples from the conditional distribution of all alignment variables in one sequence, given the current alignment in the remaining sequences, makes it practical to analyze arbitrarily complex collinear multiple alignment data sets. Hidden Markov models (HMM) combine the use of this recursive relationship with the EM algorithm to provide another means to analyze collinear multiple sequence data sets which have large joint alignment spaces for each sequence. Because HMM and dynamic programming algorithms for the alignment of pairs of sequences both characterize the missing alignment data through the use of a trichotomy (insert, delete or match) for each residue, the recursive relationship is almost identical. In contrast, since propagation treats the missing alignment data as permutations of the indices of the data, as do the other methods described above, the recursive relationship used by propagation departs from the others.

VII. Flexibility/Sensitivity tradeoff

HMMs and propagation allow for full flexibility in collinear alignments, in that gaps may be placed between any pair of residues, and position-specific gap distributions may be employed. However, this flexibility comes at a price in sensitivity. Quantitation of this flexibility/sensitivity tradeoff can be accessible from a fundamental relationship of EM theory.

VIII & IX. Threading and stochastic context free grammars

When the assumption that residue positions act independently is relaxed to account for non-local interactions of residues, higher order biopolymer structure must be considered. Threading methods take structural motifs as given, and employ higher order models implied by these structures to align sequences to structural motifs. Methods for the structural prediction of multiple RNA sequences treat structural variables, as well as alignment variables, as missing data and consequently employ a third component model to impute the missing structural data. Residue interactions are described by the higher order interaction implied by these imputed structures.

This tutorial reviews the common statistical framework behind all of these algorithms, illustrates its application to each of them, and describes the biological underpinnings of the models they use.

Annotated bibliography

The bibliography is not intended to be exhaustive. Rather, it is intended to give the reader entry into this literature.

Blotzmann like models

Bryant SN, and Lawrence CE, 1991, The frequencies of ion pair substructures in proteins is quantitatively related to electrostatic potentials: a statistical model for non bonded interactions, Proteins 9: 92-112. (Statistical model for and empirical evidence of Boltzmann like model of residues in proteins)

Berg OG, and von Hipple PH, 1987, Selection of DNA binding sites by regulatory proteins: statistical mechanical theory and application to operators and promoters, J. Mol Biol, 193: 723-750. (Statistical mechanical theory and empirical evidence in DNA for Bailsmann like model)

General Missing data references:

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B 39: 1-38. (Gave EM its name and a major boost)

Li, K.H. 1988. Imputation using Markov chains. J. Stat. Comp. 30: 57-79. (Early data augmentation reference)

Little, RJA and Rubin, DB, 1987, Statistical analysis with missing data, J. Wiley and Son, New, York. (Reference for EM methods)

Orchard, T. and M.A. Woodbury. 1972. A missing information principle: theory and applications. Proc. of the 6th Berkeley Symposium on Math. Stat. and Prob. (First missing data reference)

Tanner, MA, 1990, Tools for statistical analysis: Observed data and data augmentation methods, Springer Verlag, Berlin. (Reference for most missing data statistical methods)

Tanner, M.A. and W.A. Wong. 1987. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 82: 528-540. (Early data augmentation reference)

Permuted data statistical methods

Lawrence, CE and Reilly, AA, 1996, Likelihood inference for permuted data with application to

gene regulation, J Amer Stat Assoc, to appear. (Provides statistical background for permute data analysis and gives a gene regulation application)

EM alignment

Cardon LR, and Stormo GD, 1992, Expectation maximization algorithms for identifying protein binding sites with variable lengths from unaligned DNA fragments, J. Mol. Biol. 223: 159-170 (Extends Lawrence & Reilly to permit one gap in the binding site)

Lawrence, C.E. and A.A. Reilly. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, Proteins Struc. Func. Genet. 7: 41-51. (First use of missing data concepts for multiple sequence alignment)

Gibbs sampler for multiple alignment

Lawrence, C.E., S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262:208-214. (First Gibbs sampling for multiple sequence alignment reference)

Liu JS, Neuwald AF, and Lawrence CE, 1995, Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, J Amer Statis Assoc, to appear 12/95. (Gives rigorous statistical basis for Gibbs sampling for multiple sequence alignment)

Neuwald AF, Liu JS, and Lawrence CE, 1995, Gibbs motif sampling: detection of bacterial outer membrane protein repeats, Protein Science, to appear fall 95. (Gives latest Gibbs methods for multiple sequence alignment and applications)

HMM

Baldi, P. Chauvin, Y. Hunkapiller T. McClure, M.A. 1993. Hidden Markov models of biological primary sequence information, Proc Natl Acad Sci 91:1059-1063. (One of two first HMM for multiple alignment reference)

Krogh A, Brown M, Mian IS, and Sjolander K, Haussler DA, 1994. Hidden Markov Models in computational biology: applications to protein modeling, J. Mol. Biol. 235:1501-1531. (one of two first HMM for multiple alignment reference)

Threading

Bryant SH and Lawrence CE, 1993. An empirical energy function for threading protein sequence through the folding motif, Proteins 16:92-112. (Threading reference that is most easily seen from the missing data prospective)

Rost B, and Sander C, 1994, Structural prediction of proteins - where are we now? Curr Opin Biotechnol 5:372-380 (review)

Wodak SJ, and Rooman MJ, 1993, Generating and testing protein folds, Curr Opin Struct Biol 3:247-259 (review)

Stochastic context free grammars

Eddy SR; Durbin R, 1994, RNA sequence analysis using covariance models, Nucleic Acids Res 22: 2079-88(one of two first references on stochastic context free grammars for RNA alignment and structural prediction)

Sakakibara Y; Brown M; Hughey R; Mian IS; Sjolander K; Underwood RC; Haussler D, 1994, Stochastic context-free grammars for tRNA modeling, Nucleic Acids Res 22: 5112-20 (one of two first references on stochastic context free grammars for RNA alignment and structural prediction)