A Gibbs Sampler for the Detection of Subtle Motifs in Multiple Sequences

Charles E. Lawrence, Stephen F. Altschul, John C. Wootton, Mark S. Boguski,
Andrew F. Neuwald, and Jun S. Liu

Wadsworth Labs, Albany, NY; National Center for Biotechnology
Information, Bethesda, MD; Harvard University, Cambridge, MA

Abstract

We describe a new statistically based algorithm that aligns sequences by means of predictive inference. Using residue frequencies, this Gibbs sampling algorithm iteratively selects alignments in accordance with their conditional probabilities. The newly formed alignments in turn update an evolving residue frequency model. When equilibrium is reached the most probable alignment can be identified. If a detectable pattern is present, generally convergence is rapid. Effectively, the algorithm finds optimal local multiple alignments in linear time (seconds on current workstations). Its use is illustrated on test sets of lipocalins and prenyltranferases.

I. Introduction

Multiple sequence alignment has proved to be a remarkably successful means of representing and organizing much of the present deluge of inferred protein sequence data. It is crucial to research on the structure and function of proteins, promoting the detection and description of sequence motifs and aiding efforts at protein modeling, structure prediction and engineering. In addition, by organizing information on mutational variation, multiple sequence alignment can elucidate molecular evolution and serve as the input for phylogenetic reconstruction.

The importance of local multiple sequence alignment has long been appreciated and has been the subject of extensive study (30, 2, 29, 34, 35, 19, 22, 32, 31, 37, 5). The goal of automated methods is to produce optimized alignments, using only the information intrinsic to the sequences themselves. Unfortunately, rigorous algorithms for finding optimal solutions are so computationally expensive as to limit their application to a very small number of sequences. On the other hand, many heuristic approaches gain speed at the sacrifice of sensitivity to subtle patterns. We recently presented a relatively non-technical description of a new local multiple sequence alignment algorithm (24) based on Gibbs sampling (33). Here we develop the statistical foundation and mathematical model on which the algorithm is based.

When biopolymers have been subjected to a limited amount of evolutionary change, their commonality stems primarily from their mutational history. Such closely related sequences are relatively easy to align. The focus here will be on the more difficult case that arises when the sequences have been subjected to extensive change, and any common patterns that remain are subtle. Among such distantly related sequences, common features stem primarily from structural or functional constraints. These constraints arise from the energetic interactions among residues or between residues and ligand. 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 (28, 8). This relationship suggests that residue frequency models can be a valuable tool for representing the structural and functional constraints common to a set of proteins or protein domains. Accordingly, a residue frequency model is at the core of the methods described here.

1060-3425/94 $3.00 © 1994 IEEE

Structurally related proteins share a common core composed largely of secondary structural elements. Energetic interactions among the residues of these core elements are the primary determinants of protein structure and stability. Sequence length variations, corresponding to gaps in alignments, stem primarily from variations in the lengths of loops that connect these core elements. Furthermore, in order to maintain the proper interactions between loop residues and a ligand, loop geometry often must be maintained. Length variation in such loops is thus uncommon, and corresponding loop residues are also subject to energetic constraints. These aspects of molecular structure suggest that sequences of protein domains sharing similar structure and/or function will contain ungapped blocks of residues subject to common energetic constraints.

Among the classes of variation in macromolecular sequences one, point mutation, alters the identity of a residue at a given position in the sequence. In contrast, transpositions, insertions or deletions, and sequence duplications result in the misalignment of sequences. The need for sequence alignment algorithms stems from the fact that direct data on the effects of these latter unobserved events 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 (17). 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 first described by Orchard and Woodbury (27) and called the "missing information principle". Its application became widely known through a deterministic maximum likelihood algorithm, expectation maximization (EM) algorithm (13).

Gelfand and Smith (15) developed a sampling based approach, which they named the Gibbs sampler. It was developed for the case in which the posterior distribution is complicated, and thus difficult or impossible to obtain by direct integration. They employed this sampling algorithm both to develop a Bayesian description of the complete posterior distribution, and to find maximum a posteriori (MAP) estimates. They chose the name "Gibbs sampler" because a key required theorem from statistical physics, the Hammersley Clifford theorem, employs Gibbs/Boltzmann potentials to model joint probabilities from a complete set of conditionals. The use of sampling methods for problems involving missing data was first undertaken by Tanner and Wong (36) and Li (25). In the last few years, this sampling approach and its extensions have become a topic of great interest in statistics (15, 33). Most statistical applications have little connection with statistical mechanics, and thus the name Gibbs sampling has fallen into disfavor among some statisticians. Because of the connections of this work with statistical physics, the name Gibbs sampler here is entirely appropriate.

The missing information principle was first used for sequence alignment to develop a block based expectation maximization (EM) algorithm for the identification and characterization of common motifs in biopolymer sequences (22). This work subsequently was extended to permit small variations in the spacing of pairs of blocks (10). More recently, EM algorithms for gap-based alignment methods, in the form of Hidden Markov Models (HMM), have been described (18). A more complete description of statistical aspects of the use of these ideas for misaligned data is given by Lawrence and Reilly (23). Following this tradition, the algorithm presented here assumes the existence of missing alignment data and imputes probabilities to them. A Gibbs sampling approach is used to exploit these inferred probabilities.

In section II we define local multiple sequence alignment as a missing data problem, show how it may be simplified to a problem of predictive inference, and describe a Gibbs sampling algorithm to obtain maximum a posteriori (MAP) estimates for the possible solutions of this optimization problem. Applications to lipocalins and prenyltransferases are presented in section III.

II. A Gibbs Sampling Algorithm for Local Multiple Sequence Alignment

Let $\mathbf{R}_n = \{R_{1,n},R_{2,n},\ldots,R_{J,n}\}$ be the vector of residues observed in sequence n. Assume that the sequences share K blocks containing w_k (k = 1,...,K) positions with common residue propensities, and let $W = \sum_k w_k$. Also, let $\mathbf{A}_n = \{A_{1,n},\ldots,A_{K,n}\}$ be the vector of starting positions of the K elements in sequence n, and $\mathbf{A}_k = \{A_{k,1},A_{k,2},\ldots,A_{k,N}\}$ be a vector of the starting positions of element k in each of the N sequences. Our goal is to find the most probable alignment of all elements given the all of the sequence data, i.e. max $\{P(\mathbf{A}|\mathbf{R})\}$.

Given an alignment, the joint distribution of the residue types at the W positions in the common core may be represented by a multinomial residue frequency model. Interactions of residues with ligand, backbone atoms and water are essential to protein structure and function, and impose first order constraints on residue frequencies. Forces between pairs of residues are also key determinants of protein structure, and impose pairwise interaction constraints on residue frequencies. Since the multinomial model is a member of the exponential family (20) and we consider at most pairwise interactions, the log joint distribution of residue frequencies may be described as a sum over first order terms plus the sum of pairwise terms over the set that mutually interact (4). In other words,

$$\log(P(R_n\,|A_n) = \sum_{m=1}^{W} \mu_{m,r} + \sum_{C} \mu_{i,r,j,s}\,, \qquad (4)$$

where C is the set of residue pairs that make contact, $\mu_{m,r}$ is a parameter for the log probability of observing residue type r at position m in the structure, and $\mu_{i,r,j,s}$ is the parameter for the log probability of observing a pair of residues of types r and s at positions i and j in the structure. The two summations represent respectively the first and second order "free energy" parameters. Such a model has been successfully employed for "threading" sequences through folding motifs (9).

Even when ligand specific effects are ignored, over 65% of the information concerning residue pair frequencies in proteins of known structure is captured by first order "hydrophobicity" terms (9). Consequently, in spite of the important contribution of pairwise interactions to protein stability, much of the information contained in protein motifs is captured by first order terms alone. Accordingly, in what follows we will restrict attention to first order residue frequency models, i.e. product multinomials.

Since the common core of a structure is composed of ungapped blocks of sequence positons in the structure, a specific column, say w, in a specific block, say k, maps to a specific residue position, say m, in the structure. Using this mapping, let $\theta_{k,w,r} = \exp(\mu_{m,r})$ be the residue probabilities. If the position of the various elements within sequence n is $\mathbf{A}_n$, then the probability of observing the constituent residues is

$$P(R_n|\,A_n,\theta) =$$

$$\Big(\prod_{k=1}^{K}\prod_{w=0}^{w_k\,1} \theta_{k,w,r(A_{k,n}+w)}\Big)\ \Big(\prod_{l\in N} \theta_{0,r(l)}\Big), \qquad (2)$$

where N is the set of sequence positions not in any element, $\theta_{O,r}$ are the residue frequency parameters for non site positions. Application of Bayes theorem yields

$$P(A_n|R_n,\theta) = \frac{P(R_n|A_n,\theta)\ P(A_n|\theta)}{\sum P(R_n|A_n,\theta)\ P(A_n|\theta)}. \qquad (3.a)$$

Notice that under the non-informative prior assumption that all alignments are equally likely,

$$P(A_n|R_n,\theta) \propto P(R_n|A_n,\theta). \qquad (3.b)$$

In fact, alignment data is missing and the residue frequency parameters are unknown. In Bayesian statistics both of these are treated as random variables whose distribution, given the observed data, is the object of analysis. Thus it appears our interest should focus on $P(A_n,\theta|R_n)$. However, as we shall see, θ integrates out of the problem. We begin by considering the conditional posterior distributions of θ given the alignment and the data, which is obtained as follows:

$$P(\theta|A_n,R_n) = \frac{P(R_n|A_n,\theta)\ P(\theta)}{\int P(R_n|A_n,\theta)\ P(\theta)\ d\theta}. \qquad (4)$$

We choose as the prior for θ the Dirchlet distribution, D(θ), since it is conjugate to the multinomial. Because the posteriors will also be Dirichlet this choice facilitates analysis. Because this is a flexible distribution for θ, little is lost by assuming this from. We restrict attention to non-informative priors, and thus employ a single set of priors for all positions in both the element and non-element positions. Thus the priors are distributed as $D(\beta_1,\beta_2,...,\beta_{20})$ where β_j is a prior hyperparameter for residue type j. If the position of the kth element is known in all but the nth sequence the posterior distribution for the parameters of the corresponding residue frequency model is the product of independent 20 parameter Dirichlet distributions. Specifically,

$$P(\theta|R,A_{k,[n]})=$$
$$(\prod_{w=0}^{w_k 1} D(c_{k,w,1}[n]+\beta_1, c_{k,w,2}[n]+\beta_2, ... ,c_{k,w,20}[n]+\beta_{20})$$
$$(D(c_{0,1}[n]+\beta_1, c_{0,2}[n]+\beta_2, ... ,c_{0,20}[n]+\beta_{20})), \qquad (5)$$

where $c_{1,k,r}$[-n] is the count of the number of residues of type r in all of the sequences except n at position w of the kth element, $A_{k,[-n]}$ is the alignment of the kth element in every sequence except the nth, and $c_{0,j}$ [-n] is the count of the number of residues of type j in the non-element positions.

The Gibbs algorithm permits us to sample the joint distribution P(θ,A|R) by iteratively sampling from the complete set of conditionals. Equation (5) provides the conditional distribution of θ, while equations (2) and (3) provide the conditional distribution of $A_{k,n}$. However, sampling from the Dirichlet distribution is unnecessary and time consuming (26).

When the location of the kth element is given in all but the nth sequence, probabilities for the missing location of this element in the remaining sequence can be attained from the predictive distribution,

$$P(A_{k,n}|R,A_{k,[n]})=\int P(A_{k,n}|\theta,R)P(\theta|A_{k,[n]},R)\,d\theta \quad (6)$$

Furthermore, this distribution can be estimated taking account of other elements in the nth sequence. In this case the overlap of elements within a sequence must be excluded, by setting $P(A_{k,n}|\theta,R,A_{[-k],n})$ to zero if the kth element overlaps another. Normalization is required to attain a proper distribution.

Integration of the right hand side of equation (6) yields an expression in the form of a multinomial Dirichlet distribution. Fortunately, the expression reduces to the following simple form

$$P(A_{k,n}|R,A_{k,[n]}) \propto$$
$$\prod_{l=0}^{w_k 1}(c_{l,r(a+k)}+\beta_{r(a+k)}) \ \prod_{l\in N}(c_{l,r(l)}+\beta_{r(l)}), \qquad (7)$$

(26). Sampling using equation (7) greatly simplifies the computation, for we need only maintain the residue counts in all element positions, and sampling from the Dirichlet distribution is eliminated.

The Gibbs sampling algorithm is now straightforward. The process begins with a random selection of element locations in all sequences, from which the associated residue counts are calculated. Then, in a systematic fashion, elements are removed for relocation, and the associated counts are decremented. After normalization, equation (7) provides the distribution for sampling the new location of the element in question, after which the appropriate counts are incremented. The procedure is repeated until equilibrium is achieved, from which the most probable alignment may be identified. The Gibbs sampling algorithm defined by these distributions yields an irreducible Markov chain with the stable distribution P(A|R). A geometric convergence rate to the true alignment distribution can be guaranteed (21). While it is not assured, we have observed rapid convergence in numerous applications.

Phase Shifts

The most important obstacle to quick convergence is the non-convexity of the probability surface from which the algorithm samples. This creates local optima, or energy traps. There are two classes of such optima: chance optima and shifted optima. Assume that the true element locations occur at positions {23,66,189,53,... }. If the sampler happens to draw positions 21,64 and 187 within the first three sequences, then position 51 becomes a likely location within the 4th sequence, and position 53 an unlikely one. Thus, the sampler tends to become trapped in this locally optimal alignment, and substantial further sampling may be required to escape.

Notice that simultaneously shifting the alignment of all elements two positions to the right places them correctly. A shift of s positions to the right corresponds to removing the s leftmost columns of the element model, and adding s columns to the model's right end. Shifts left are analogous. Changes in the product multinomial associated with each possible shift are easily calculated. When shifts of up to S positions in either direction are considered, the relative probabilities for the 2S+1 possibilities are proportional to the corresponding product multinomials. The algorithm explores these alternatives by sampling.

Element Order and Colinearity

Equation (7), which gives the predictive distribution for the location of the site in the nth sequence, carries no information concerning element order. Thus when exon shuffling results in element transpositions, the sampler may identify common elements in spite of their reordering. However, because biopolymers frequently evolve without transposition, colinearity is a safe assumption for many multiple alignment problems. Notice that at each sampling iteration, the orders of the elements in the N-1 sequences not under consideration are known. This ordering information may be employed as follows to improve the sensitivity of the sampler. Observations on order can be combined with residue frequency observations and their priors in equation (6) to yield a predictive distribution which jointly incorporates these two kinds of information. In the lipocalins example given below we employ this joint information to correctly identify a pair of element locations that could not be identified independently.

Repeating Sequence Elements

Biopolymer duplications which result in sequence repetition are much more common an event than is often assumed. Because the sampler conditions on the locations of all elements except the current one, the residue frequencies from repeated elements within the current sequence can be included in the predictive distribution given in equation (7). In this way, information from repeated elements can be employed to aid alignment. This information may be used to identify repeats within a single sequence, or to construct a local alignment from multiple sequences containing repeats. The prenyltransferase example below illustrates the use of the sampler in such a context.

III. Applications

The algorithm has been developed and tested using several examples of protein classes that present different types of difficulty for automated multiple alignment methods. As previously described (24), DNA binding proteins of the helix-turn-helix class were employed to examine the performance of the method in detecting and aligning a single, highly variable motif. This example was employed to develop convergence heuristics and an empirical method for automated determination of pattern width. Applications of the algorithm to more complex test cases, as described in (24), are illustrated here with lipocalins and prenyltransferases.

Lipocalins

As discussed above, proteins, protein domains, and even most protein motifs, are composed of multiple core segments and ligand binding segments, with intervening loops of variable length. As a consequence the sequences of these families contain multiple blocks of ungapped elements which are separated by dissimilar sequences of highly variable length. We have successfully aligned several such protein families, including protein kinases, aspartyl proteinases, and aminoacyl-tRNA ligases. We report here on the most difficult of these test cases, the lipocalins. Lipocalins bind a wide variety of hydrophobic ligands and share a common polypeptide fold, but have extremely diverse sequences. These proteins have two weak sequence motifs, centered on the generally conserved residues -GXW- and -TD-, which are recognized from structural comparisons (12, 14). The rest of the topologically conserved lipocalin folds have very different sequences. Conventional automated sequence alignment methods, although successful for selected subsets of the data, fail to align these motifs for the full spectrum of lipocalin sequences such as the five aligned here.

For many multiple element problem including the lipocalins, colinearity of the elements is a reasonable prior assumption. We thus incorporated the ordering component of the sequence model described in section II. As indicated in Figure 1 (taken from 24) , challenged with five very diverse lipocalin sequences of known crystal structure, the sampler correctly aligned these two regions at width 16 residues of both elements, in agreement with the structural evidence (12, 14). In this case only five sequences were made available and the patterns common to all of these sequences are subtle. As a consequence we found that inclusion of the ordering information described in section II was required to identify these ten sequence elements correctly.

```
                                 Motif A                                    Motif B

                            17                32                       104               119
YA_MANSE ..     GYCPDVKPVN  DFDLSAFAGAWHEIAK  LPLENENQGK...FGQRVVNLVP  WVLATDYKNYAINYNC  DYHPDKKAHS
                            25                40                       109               124
LACB_BOVIN..    QALIVTQTMK  GLDIQKVAGTWYSLAM  AASDISLLDA...KIDALNENKV  LVLDTDYKKYLLFCME  NSAEPEQSLA
                            16                31                       100               115
BBP_PIEBR ..    GACPEVKPVD  NFDWSNYHGKWWEVAK  YPNSVEKYGK...YGGVTKENVF  NVLSTDNKNYIIGYYC  KYDEDKKGHQ
                            14                29                       105               120
RETB_BOVIN..    CRVSSFRVKE  NFDKARFAGTWYAMAK  KDPEGLFLQD...SFLQKGNDDH  WIIDTDYETFAVQYSC  RLLNLDGTCA
                            27                42                       109               124
MUP2_MOUSE..    HAEEASSTGR  NFNVEKINGEWHTIIL  ASDKREKIED...SVTYDGFNTF  TIPKTDYDNFLMAHLI  NEKDGETFQL
                                     * **                                     ***  *
```

Two motifs located automatically in five lipocalins of known crystal structure. The sequences, defined by SwissProt database codes, are, from top to bottom, *Manduca sexta* insecticyanin; Bovine β-lactoglobulin; *Pieris brassicae* bilin-binding protein; bovine plasma retinol-binding protein; mouse major urinary protein 2. Asterisks (***) below the alignment denote generally conserved residues recognized from structural comparisons (12).

Figure 1: Lipocalins Multiple Alignment

Prenyltransferases

Repeated elements underlie a broad spectrum of biological problems. Because these elements often play somewhat different roles their sequences frequently diverge substantially after duplication, rendering their detection and characterization challenging. The analysis of repeats is often labor-intensive, relying in part on visual inspection of "dot plots" (6) -- a procedure that limits searches and surveys of large databases.

Repeats in the subunits of the heterodimeric protein-isoprenyltransferases provides an example of subtle internal repeats (6). These enzymes are responsible for targeting and anchoring members of the ras superfamily of small GTPases to their sites of action on various cellular membranes (11). A subtle internal repeat of possible functional significance is contained in prenyltransferases (6). Structural information is not yet available for these

```
                A                                   L                B
Ram1
    109                                                              MLYWIANSLKVM DRDWLSDD--
    129 TKRKIVVKLFTI SPSG-------------------- GPFGGGPGQLSH LA- STYAAINALSLC DNIDGCWDRID
    181 DRKGIYQWLISL KEPN-------------------- GGFKTCLEVGEV DTR GIYCALSIATLL NILTEEL----
    230 LTEGVLNYLKNC QNYE-------------------- GGFGSCPHVDEA HGG YTFCATASLAIL RSMDQIN----
    279 NVEKLLEWSSAR QLQEE------------------- RGFCGRSNKLVD GC- YSFWVGGSAAIL EAFGYGQCF--
    331 NKHALRDYILYC CQEKEQ------------------ PGLRDKPGAHSD FY- HTNYCLLGLAVA E----------
                                          SSYSCTPNDSPH                   ...27 aa...
    415 NVRKIIHYFKSN LSSPS
FT-β
     74 QREKHFHYLKRG LRQLTDAYECLDAS
     99 SRPWLCYWILHS LELLDEPIPQIV
    122 VATDVCQFLELC QSPD-------------------- GGFGGGPGQYPH LA  PTYAAVNALCII GTEEAYNVIN
    173 NREKLLQYLYSL KQPD-------------------- GSFLMHVGGEVD VR  SAYCAASVASLT NIITPDL---
    221 LFEGTAEWIARC QNWE-------------------- GGIGGVPGMEAH GG  YTFCGLAALVIL KKERSLN---
    269 NLKSLLQWVTSR QMRFE------------------- GGFQGRCNKLVD GC  YSFWQAGLLPLL ...20 aa...
    331 HQQALQEYILMC CQCPA------------------- GGLLDKPGKSRD FY  HTCYCLSGLSIA ...
Bet2
      8 LKEKHIRYIESL DTKKHNFEYWLTEHLRLN---------------------   GIYWGLTALCVL DSPETFV---
     56 LKEEVISFVLSC WDDKY------------------- GAFAPFPRHDAH LL  TTLSAVQILATY DALDVLGKDR
    108 RKVRLISFIRGN QLED-------------------- GSFQGDRFGEVD TR  FVYTALSALSIL GELTSEV---
    156 VVDPAVDFVLKC YNFD-------------------- GGFGLCPNAESH AA  QAFTCLGALAIA NKLDMLSDDQ
    207 QLEEIGWWLCER QLPE-------------------- GGLNGRPSKLPD VC  YSWWVLSSLAII GRLDWIN---
    255 NYEKLTEFILKC QDEKK------------------- GGISDRPENEVD VF  HTVFGVAGLSLM ...
GGT-β
     19 LLEKHADYIASY GSKKDDYEYCMSEYLRMS----------------------  GVYWGLTVMDLM GQLHRM
     67 NKEEILVFIKSC QHEC-------------------- GGVSASIGHDPH LL  YTLSAVQILTLY DSIHVI
    115 NVDKVVAYVQSL QKED-------------------- GSFAGDIWGEID TR  FSFCAVATLALL GKLDAI
    163 NVEKAIEFVLSC MNFD-------------------- GGFGCRPGSESH AG  QIYCCTGFLAIT SQLHQV
    211 NSDLLGWWLCER QLPS-------------------- GGLNGRPEKLPD VC  YSWWVLASLKII GRLHWI
    259 DREKLRSFILAC QDEET------------------- GGFADRPGDMVD PF  HTLFGIAGLSLL ...
Cdc43
     12 VTKKHRKFFERH  ...103 aa...
    127 DKRSLARFVSKC Q ...52 aa...
    191 DTEKLLGYIMSQ QCYN-------------------   GAFGAHNEPHSG --  YTSCALSTLALL SSLEKLSDKF
    240 FKEDTITWLLHR QVSSHGCMKFESELNASYDQSDD   GGFQGRENKFAD TC  YAFWCLNSLHLL TKDWKMLC--
    309 QTELVTNYLLDR TQKTLT-----------------   GGFSKNDEEDAD LY  HSCLGSAALALI ...
```

Repeating motifs in prenyltransferase subunits. Ram1, Bet2 and Cdc43 are yeast gene products with NBRF/PIR accession numbers S07864, S15399 and A40875, respectively. FT-β is the β subunit of farnesyltransferase from rat brain (acc. no. A40037) and GGT-β is the β subunit of rab geranylgeranyl transferase from rat brain. The primary structures of these proteins have been shown to contain a variable number of tripartite internal repeats each of which is composed of "A" and "B" subdomains separated by a "linker region" containing multiple glycine and proline residues (6) . When subjected to analysis by the Gibbs sampler, these previously-defined motifs were identified, and additional copies were also observed (cf. Fig. 1 in 6). Dashes indicate the locations and extents of gaps between motifs; ellipses (...) accompanied by a number and the abbreviation "aa" indicate the locations and extents of larger intervening subsequences expressed as the number of amino acid residues. The spacing between motifs L and B is only 2-3 residues whereas that between motifs A and L is greater and more variable.

Figure 2: Prenyltransferase Beta Subunits

proteins. However, previous sequence analysis indicates that the β subunit contains a tripartite motif that is repeated 3-5 times in each of four proteins.

This example contains four characteristics that make it challenging. 1) The repeats are subtle. 2) Because downstream information must be utilized to exploit data from internal repeats, the Markovian character of dynamic programming and other methods that use colinearity to advantage is not available. 3) The many elements per sequence result in a great number of possible alignments of the elements within each sequence. 4) The crowding of elements increases inter-element dependency and the complexity of the joint probability surface which the algorithm must explore.

The previous analysis of this problem was subjective and time-consuming, relying on the combined use of several different multiple alignment methods. In contrast, as illustrated in Fig. 2 (taken from 24), the Gibbs sampling algorithm quickly and objectively reproduced and extended the previous results.

IV. Conclusions

We view distantly related proteins or protein sub-structures as a set of ungapped core elements subjected to random point mutations. These point mutations tend to maximize sequence entropy, but are subject to energetic constraints associated with protein structure and function. Consequently, we use a residue frequency model to describe the common character of distantly related proteins. This is a key factor in the sampler's speed. It permits the high-dimensional search space of these problems to be explored one dimension at a time, by means of comparing each sequence to a common evolving residue frequency model.

EM methods use a similiar strategy. Two classes of EM algorithms have been described: block based methods (22) and gap based methods, i.e. HMMs (18). The sampler goes a step further then these, conditioning on the alignment of all but the current element. This avoids the need to exploit the Markov property that underlies collinear models of sequence similarity and is required by dynamic programming and gap based EM methods. This relaxation permits the sampler to identify non-collinear similarities such as those arising from transpositions. Furthermore, downstream residue frequencies of similar elements such as those arising from sequence duplications may be exploited. The sampler outperforms block based EM methods on multiple element problems because of this conditioning. Forced to sum over all possibilities, block based EM methods have a time complexity that grows exponentially with additional elements. In contrast the sampler never needs to consider more than one element at a time.

Choosing an appropriate number of elements of appropriate width remains an important problem. An empirical method for determining element width is described elsewhere (24). Choosing an optimal number of elements requires further study. However, we have found that an additional element is not warranted when multiple random seeds lead to many different alignments, and when the maximum posterior probabilities consistently fail to exceed those obtained from shuffled sequences. Employing an appropriate model for inter-element spacing will improve the algorithm's sensitivity, but this feature has not been needed to identify even the subtle patterns described above.

Inherent to a Bayesian approach is the concept that all unknowns are treated as random variables. Thus, we postulate a Dirichlet model for the residue frequency parameters. Through the use of predictive inference, computation is greatly reduced, avoiding sampling from posterior Dirichlet distributions, albeit at the cost of failing to obtain sampling estimates from the the residue frequency distribution.

While prior information concerning the locations of the elements in the sequence and the frequency of residues in the common elements can be incorporated in our formulation, we have employed only noninformative priors. In no test case to date have we found it necessary to relax this assumption.

The existence of multiple local optima presents a major challenge to multiple sequence alignment methods. As described above, the inclusion of phase shifts permits the sampler to avoid shifted optimum through direct explorations. Furthermore, its inherent stochastic character permits the sampler to escape chance local optima that cause difficulties for deterministic methods.

The memory requirements for the method are negligible; storing the input sequences is usually the dominant space demand. It is difficult to analyze the worst case time complexity of this algorithm. However, for typical protein sequence data sets, we have found that each input sequence needs to be sampled on average fewer than T = 100 times before convergence. In the more time consuming sampling step of the basic algorithm, approximately Lw_k multiplications are performed, where L is the length of the sequence that has been removed from the model. Therefore the total number of multiplications needed to execute the Gibbs sampler is approximately $TNWL^*$, where L^* is the average length of the N input sequences. The factor T is expected to grow with increasing L^*. However, experimentation suggests that T tends to decrease slowly with increasing N when the common pattern exists at roughly equal strength within the input sequences. Thus, linear time complexity has been observed in applications. In practice all of the examples we have examined to date have been solved in under two minutes on current workstations.

Through the combination of a mathematical model that represents basic properties of protein structure and change, and a randomized optimization procedure, the Gibbs sampler has objectively solved difficult multiple sequence alignment problems in a matter of seconds in the absence of any expert knowledge, or ancillary information derived from three-dimensional structures or other sources.

References

1. Akrigg, D., T.K. Attwood, A.J. Bleasby, J.B. Findlay, A.C. North, N.A. Maughan, D.J. Parry-Smith, D.N. Perkins, and J.C. Wootton. 1992. SERPENT--an information storage and analysis resource for protein sequences. Comput. Appl. Biosci. 8: 295-296.

2. Bacon, D.J. and W.F. Anderson. 1986. Multiple sequence alignment. J. Mol. Biol. 191: 153-161.

3. Berg, O.G. and P.H. von Hippel. 1987. Selection of DNA binding sites by regulatory protiens. Statistical mechanical theory and application to operators and promoters. J. Mol. Biol. 193: 723-750.

4. Besag, J. 1974. Spatial interactions and the statistical analysis of lattice systems. J. Roy. Stat. Soc. B 35: 192-236.

5. Boguski, M.S., R.C. Hardison, S. Schwartz, and W. Miller. 1992. Analysis of conserved domains and sequence motifs in cellular regulatory proteins and locus control regions using new software tools for multiple alignment and visualization. New Biol. 4: 247-260.

6. Boguski, M.S., A.W. Murray, and S. Powers. 1992. Novel repetitive sequence motifs in the α and β subunits of prenyl-protein transferases and homology of the α subunit to the MAD2 gene product of yeast. New Biologist 4: 408-411.

7. Boguski, M.S., J. Ostell, and D.J. States. 1992. Molecular sequence databases and their uses. In Protein Engineering: A Practical Approach. A.R. Rees, M.J.E. Sternberg, and R. Wetzel, Editors. Oxford: IRL Press, pps. 57-88.

8. Bryant, S.H. and C.E. Lawrence. 1991. The frequencies of ion pair substructures in proteins is quantitatively related to electrostatic potentials: a statistical model for nonbonded interactions. Proteins Struct. Func. Genet. 9: 108-119.

9. Bryant, S.H. and C.E. Lawrence. 1993. An empirical energy function for threading protein sequence through the folding motif. Proteins Struc. Func. Genet. 16: 92-112.

10. Cardon, L.R. and G.D. Stormo. 1992. Expectation maximization algorithm for identifying protein binding sites with variable lenghts from unaligned DNA fragments. J. Mol. Biol. 223: 159-170.

11. Clarke, S. 1992. Protein isoprenylation and methylation at carboxyl-terminal cysteine residues. Annu. Rev. Biochem. 61: 355-386.

12. Cowan, S.W., M.E. Newcomer, and T.A. Jones. 1990. Crystallographic refinement of human serum retinol binding

protein at 2A resolution. Proteins Struct. Func. Genet. 8: 44-61.

13. 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.

14. Flower, D.R., A.C.T. North, and T.K. Attwood. 1993. Structure and sequence relationships in the lipocalins and related proteins. Protein Science 2: 753-761.

15. Gelfand, A.E. and A.F.M. Smith. 1990. Sampling based approaches to calculating marginal probabilities. J. Am. Stat. Assoc. 85: 389-409.

16. Geman, D. and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transaction in Pattern Analysis and Machine Intelligence 6: 721-741.

17. Goodman, L.A. 1974. Exploratory latent structural analysis using both identified and unidentified models. Biometrika 61: 215-231.

18. Haussler, D., A. Krogh, S. Mian, and K. Sjolander. 1992. Protein modeling using hidden Markov models. Computer and Information Sciences. Tech. Rep. No. UCSC-CRL-92-23, University of California at Santa Cruz.

19. Hertz, G.Z., G.W. Hartzell III, and G.D. Stormo. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput. Appl. Biosci. 6: 81-92.

20. Kendall, M. and A. Stuart. 1979. The Advanced Theory of Statistics. New York: Macmillan.

21. Kalos, M.H. and P.A. Whitlock. 1986. Monte Carlo Methods, Vol I: Basics. New York: Wiley & Sons.

22. 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.

23. Lawrence, C.E. and A.A. Reilly. 1992. Likelihood inference with uncertain indices with application to gene regulation. Tech. Rep. No. 121. Biometrics Laboratory, Wadsworth Center for Laboratories and Research, Albany, N.Y.

24. 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 (in press).

25. Li, K.H. 1988. Imputation using Markov chains. J. Stat. Comp. 30: 57-79.

26. Liu, J. The collapsed Gibbs sampler and other issues with application to a protein binding problem. J. Am. Stat. Assoc. (in press).

27. 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.

28. Pohl, F.M. 1971. Emperical protein energy maps. Nature New Biol. 234: 277-279.

29. Posfai, J., A.S. Bhagwat, G. Posfai, and R.J. Roberts. 1989. Predictive motifs derived from cytosine methyltransferases. Nucl. Acids Res. 17: 2421-2435.

30. Queen, C.M., N. Wegman, and L.J. Korn. 1982. Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. Nucl. Acids Res. 10: 449-456.

31. Schuler, G.D., S.F. Altschul, and D.J. Lipman. 1991. A workbench for multiple alignment construction and analysis. Proteins Struc. Func. Genet. 9: 180-190.

32. Smith, H.O., T.M. Annau, and S. Chandrasegaran. 1990. Finding sequence motifs in groups of functionally related proteins. Proc. Natl. Acad. Sci. USA 87: 826-830.

33. Smith, A.F.M. and G.O. Roberts. 1993. Baysian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. Roy. Stat. Soc. B 55: 3-23.

34. Staden, R. 1989. Methods for discovering novel motifs in nucleic acid sequences. Comput. Appl. Biosci. 5: 293-298.

35. Stormo, G.D. and G.W. Hartzell III. 1989. Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. USA. 86:1183-1187.

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

37. Vingron, M. and P. Argos. 1991. Motif recognition and alignment for many sequences by comparison of dot matrices. J. Mol. Biol. 218:33-43.

38. Yudkin, M.D. 1987. The prediction of helix-turn-helix DNA-binding regions in proteins. Protein Eng. 1: 371-372.