Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

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


Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819931008%293%262%26C208%3ADSSAG%3E2.0.CO%3B2-A

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Science is published by The American Association for the Advancement of Science. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/aaas.html.

Science
©1993 The American Association for the Advancement of Science

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2001 JSTOR
REFERENCES AND NOTES


9. Diol 1b was neutralized by preparing the correponding carboxylic acid suspended in PBS (sodium phosphate buffer) with one equivalent of sodium hydroxide.

10. All new adducts were characterized by NMR and mass spectrometry.

11. In the crude mixture of the cycladdition, we could not detect any trace of the metac regioiso- mer by 'H NMR (>>80% ortho regioselective).

12. M. J. Frisch et al., Gaussian 92 (Gaussian, Pittsburgh, PA, 1992).

13. R. J. Loncharich, F. K. Brown, K. N. Houk, J. Org. Chem. 54, 1129 (1989). All structures were fully optimized using the 3-21G basis set. Reactants were also fully optimized with the 6-31G* basis set. Vibrational frequencies for all transition structure were calculated using the HF/3-21G level and shown to have one imaginary frequency. In addtion, the energies were evaluated by single-point calculations with the 6-31G* basis set on 3-21G geometries.


18. This strategy to avoid product inhibition was first described by Braisted and Schultz in (4).

19. The stereochemistry was established on the basis of 1H NMR analysis. For example: 5-exo, 6δ F = 2.65 ppm and 6δ H = 3.5 Hz; 5-exo, 6δ H = 4.2 ppm, and 6δ H = 0 Hz.

20. Goehler and C. Mitschke, Nature 266, 496 (1977); the KLH conjugate was prepared by slowly add- ing 2 mg of the hapten dissolved in 100 μl of dimethylformamide to 2 mg of KLH in 900 μl of 0.01 M sodium phosphate buffer, pH 7.4. Four 8-week-old 129GIXC mice each received 2 intraperitoneal (IP) injections of 100 μg of the hap- ten conjugated to KLH in 0.2 ml 0.9% saline solution (MPL and TDM emulsion) 2 weeks apart. A 50 μg IP

Patterns shared by multiple protein or nucleic acid sequences shed light on molecular structure, function, and evolution. The recognition of such patterns generally relies upon aligning many sequences, a complex, multifaceted research process whose difficulty has long been appreciated. This problem may be divided into "global multiple alignment" (1, 2), whose goal is to align complete sequences, and "local multiple alignment" (2-11), whose aim is to locate relatively short patterns shared by otherwise dissimilar sequences. We report a new algorithm for local multiple alignment that assumes no prior information on the patterns or their locations within the se- quences; it determines these locations from only the information intrinsic to the sequences themselves. We focus on subtle amino acid sequence patterns that may vary greatly among different proteins. Much research on the alignment of such patterns uses additional information to sup- plement algorithmic analyses of the actual sequences, including data on three-dimen- sional structure, chemical interactions of residues, effects of mutations, and interpret- ation of sequence database search results. However, such research, which has led to many discoveries of sequence relations and structure and function predictions [see (12) for a recent example], is laborious and requires frequent input of expert knowledge.

A wealth of protein and DNA sequence data is being generated by genome projects and other sequencing efforts. A crucial barrier to deciphering these sequences and understanding the relations among them is the difficulty of detecting subtle local residue patterns common to multiple sequences. Such patterns frequently reflect similar molecular structures and biological properties. A mathematical definition of this "local multiple alignment" problem suitable for full computer automation has been used to develop a new and sensitive algorithm, based on the statistical method of iterative sampling. This algorithm finds an optimized local alignment model for N sequences in N-linear time, requiring only seconds on current workstations, and allows the simultaneous detection and optimization of multiple patterns and pattern repeats. The method is illustrated as applied to helix-turn-helix proteins, lipocalins, and prenyltransferases.

Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment

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

A wealth of protein and DNA sequence data is being generated by genome projects and other sequencing efforts. A crucial barrier to deciphering these sequences and understanding the relations among them is the difficulty of detecting subtle local residue patterns common to multiple sequences. Such patterns frequently reflect similar molecular structures and biological properties. A mathematical definition of this "local multiple alignment" problem suitable for full computer automation has been used to develop a new and sensitive algorithm, based on the statistical method of iterative sampling. This algorithm finds an optimized local alignment model for N sequences in N-linear time, requiring only seconds on current workstations, and allows the simultaneous detection and optimization of multiple patterns and pattern repeats. The method is illustrated as applied to helix-turn-helix proteins, lipocalins, and prenyltransferases.

Patterns shared by multiple protein or nucleic acid sequences shed light on molecular structure, function, and evolution. The recognition of such patterns generally relies upon aligning many sequences, a complex, multifaceted research process whose difficulty has long been appreciated. This problem may be divided into “global multiple alignment” (1, 2), whose goal is to align complete sequences, and “local multiple alignment” (2-11), whose aim is to locate relatively short patterns shared by otherwise dissimilar sequences. We report a new algorithm for local multiple alignment that assumes no prior information on the patterns or their locations within the sequences; it determines these locations from only the information intrinsic to the sequences themselves. We focus on subtle amino acid sequence patterns that may vary greatly among different proteins. Much research on the alignment of such patterns uses additional information to supplement algorithmic analyses of the actual sequences, including data on three-dimensional structure, chemical interactions of residues, effects of mutations, and interpretation of sequence database search results. However, such research, which has led to many discoveries of sequence relations and structure and function predictions [see (12) for a recent example], is laborious and requires frequent input of expert knowledge. These approaches are becoming increasingly overwhelmed by the quantity of sequence data.

A number of automated local multiple alignment algorithms have been developed (2-11), and some have proved valuable as part of integrated software workbenches. Unfortunately, rigorous algorithms for finding optimal solutions have been so computationally expensive as to limit their applicability to a very small number of sequences, and heuristic approaches have gained speed by sacrificing sensitivity to
Our method is both fast and sensitive and generally finds an optimized local alignment model for $N$ sequences in $N$-linear time. This advantage is achieved by incorporating some recent developments in statistics and by using a formulation of the problem that models well the underlying biology but avoids the explicit treatment of gaps. We illustrate the application of this method with a diverse set of difficult but well understood test cases.

**Problem and methods.** Our problem is to locate and describe a pattern thought to be contained within a set of biopolymer sequences. The model we use has three fundamental characteristics. First, we seek a relatively small number of sequence elements or patterns, each consisting of one ungapped segment from each of the input sequences. Second, a single pattern is described by a probabilistic model of residue frequencies at each position. Third, the location of the pattern within the sequences is described by a set of probabilistically inferred position variables. These features are derived from well established principles of protein structure and knowledge of the sources of sequence pattern variation (13). These principles are valid in general for globular protein families, although a few interesting counterexamples are known.

First, homologous proteins or protein domains typically are characterized by a core of common secondary structure elements separated by intervening loops (13). Gaps in sequence alignments stem primarily from variations in loop length, and loops that participate in active sites are constrained to maintain their geometry, and thus frequently retain their length as well.

Common sequence patterns therefore can generally be described by a relatively small number of ungapped elements.

Second, physicochemical constraints influence which particular residues may occur at each position in a sequence element. The similarities of closely related sequences stem largely from recent common ancestry and are relatively easy to locate by various methods (2-11), including ours. In contrast, our primary concern is to locate the common features of sequences that differ greatly. Here, similar local residue patterns reflect structural and functional constraints that arise from the energetic interactions among residues or between residue and ligand, irrespective of evolutionary history. The relation between a state's energy and frequency forms the basis of statistical mechanics, and an analogous relation governs the frequencies of residues subject to random point mutations (14). Residue frequency models are therefore natural in the present context.

Third, genomic rearrangements, as well as insertions, deletions, and duplications of sequence segments, result in the occurrence of a common pattern at different positions within sequences. However, these mutational events are "unobserved" because no data directly specify their effects on the positions of the patterns (6). As recognized by statisticians since the 1970s (15), many problems with unobserved data are most easily addressed by pretending that critical missing data are available. The key "missing information principle" (15) is that the probabilities for the unobserved positions may be inferred through the application of Bayes theorem to the observed sequence data.

The optimization procedure we use is the predictive update version (16) of the Gibbs sampler (17). Strategies based on iterative sampling have been of great interest in statistics (18). The algorithm can be understood as a stochastic analog of expectation maximization (EM) methods previously used for local multiple alignment (6, 7). It yields a more robust optimization procedure and permits the integration of information from multiple patterns. In addition, a procedure for the automatic determination of pattern width has been developed. For clarity, we first describe the identification of a single pattern of fixed width within each input sequence and then generalize to variable widths and multiple patterns.

**The basic algorithm.** We assume that we are given a set of $N$ sequences $S_1, \ldots, S_N$ and that we seek within each sequence mutually similar segments of specified width $W$. The algorithm maintains two evolving data structures. The first is the pattern description, in the form of a probabilistic model of residue frequencies for each position $i$ from 1 to $W$, and consisting of the variables $q_{i,j}, \ldots, q_{i,20}$. This pattern description is accompanied by an analogous probabilistic description of the "background frequencies" $p_{i1}, \ldots, p_{i20}$ with which residues occur in sites not described by the pattern. The second data structure, constituting the alignment, is a set of positions $a_{kj}$ for $k$ from 1 to $N$, for the common pattern within the sequences. Our objective will be to identify the "best," defined as the most probable, common pattern. This pattern is obtained by locating the alignment that maximizes the ratio of the corresponding pattern probability to background probability.

The algorithm is initialized by choosing random starting positions within the various sequences. It then proceeds through many iterations to execute the following two steps of the Gibbs sampler:

1) Predictive update step. One of the $N$ sequences, $z$, is chosen either at random or in specified order. The pattern description $q_{i,j}$ and background frequencies $p_i$ are then calculated, as described in Eq. 1 below, from the current positions $a_{kj}$ in all sequences excluding $z$.

2) Sampling step. Every possible segment of width $W$ within sequence $z$ is considered as a possible instance of the pattern. The probabilities $Q_{i,j}$ of generating each segment $x$ according to the current pattern probabilities $q_{i,j}$ are calculated, as are the probabilities $P_i$ of generating these segments by the background probabilities $p_i$. The weight $A_i = Q_i/p_i$ is assigned to segment $x$, and with each segment so weighted, a random one is selected (19). Its position then becomes the new $a_{ij}$.

This simple iterative procedure constitutes the basic algorithm. The central idea is that the more accurate the pattern description constructed in step 1, the more accurate the determination of its location in step 2, and vice versa. Given random positions $a_{ij}$ in step 2 the pattern description $q_{ij}$ will tend to favor no particular segment. Once some correct $a_{ij}$ have been selected by chance, however, the $q_{ij}$ begin to reflect, albeit imperfectly, a pattern existent within other sequences. This process tends to recruit further correct $a_{ij}$, which in turn improve the discriminating power of the evolving pattern.

An aspect of the algorithm alluded to in step 1 above concerns the calculation of the $q_{ij}$ from the current set of $a_{ij}$. For the $i$th position of the pattern we have $N - 1$ observed amino acids, because sequence $z$ has been excluded; let $c_{ij}$ be the count of amino acid $j$ in this position. Bayesian statistical analysis suggests that, for the purpose of pattern estimation, these $c_{ij}$ should be supplemented with residue-dependent "pseudocounts" $b_j$ to yield pattern probabilities

$$q_{ij} = \frac{c_{ij} + b_j}{N - 1 + B}$$

where $B$ is the sum of the $b_j$. The $p_i$ are calculated analogously, with the corresponding counts taken over all nonpattern positions (20).

After normalization, $A_i$, gives the probability that the pattern in sequence $z$ belongs at position $x$. The algorithm finds the most probable alignment by selecting a set of $a_{ij}$'s that maximizes the product of these ratios. Equivalently, one may maximize $F$, the sum of the logarithms of these ratios. In the notation developed above, $F$ is given by the formula

$$F = \sum_{i=1}^{N} \sum_{j=1}^{W} c_{ij} \log \frac{q_{ij}}{p_j}$$

where the $c_{ij}$ and $q_{ij}$ are calculated from the complete alignment (Fig. 1).

**Research Articles**

**Science** • Vol. 262 • 8 October 1993
Phase shifts. One defect of the algorithm as just described is the "phase" problem. The strongest pattern may begin, for example, at positions 7, 19, 8, and 23, and so forth within the various sequences. However, if the algorithm happens to choose $a_1 = 9$ and $a_2 = 21$ in an early iteration, it will then most likely proceed to choose $a_3 = 10$ and $a_4 = 25$. In other words, the algorithm can get locked into a nonoptimal "local maximum" that is a shifted form of the optimal pattern. This situation can be avoided by inserting another step into the algorithm (16). After every $M$th iteration, for example, one may compare the current set of $a_i$ with sets shifted left and right by up to a certain number of letters. Probability ratios may be calculated, as above, for all possibilities, and a random selection is made among them with appropriate corresponding weights.

Pattern width. The algorithm as so far described requires the pattern width to be input. It is possible, of course, to execute the algorithm with a range of plausible widths and then select the best result according to some criterion. One difficulty is that the function $F$ is not immediately useful for this purpose, as its optimal value always increases with increasing width $W$.

The problem here corresponds to the well-known issue of model selection encountered in statistics. The difficulty stems from the change in the dimensionality with the additional freely adjustable parameters. Several criteria that incorporate the effects of variable dimension have been useful in other applications (21). Unfortunately, these criteria did not perform well at selecting those pattern widths that identified correct alignments in data sets with known solutions.

A superior criterion proved to be one based on the incomplete-data log-probability ratio $G(22)$, which subtracts from the function $F$ the information required to determine the location of the pattern in each of the input sequences. We found that dividing $G$ by the number of free parameters needed to specify the pattern ($19W$ in the case of proteins) produced a statistic useful for choosing pattern widths. We call this quantity the information per parameter. The use of this empirical criterion is discussed in the examples section below and is illustrated in Figs. 2 and 3.

Multiple patterns. As described above, a pattern within a set of sequences can be described as consisting of several distinct elements separated by gaps. The Gibbs sampler may easily maintain several distinct patterns rather than a single one. Seeking several patterns simultaneously rather than sequentially allows information gained about one to aid the alignment of others. The relative positions of elements within the sequences can be used to improve their simultaneous alignment. Because only one element in sequence $i$ is altered at a time, the combinatorial problem of joint positioning is circumvented. Nevertheless, because no element's position is permanently fixed, the best joint location of all elements may be identified.

Incorporating models of element location that favor consistent ordering (colinearity) and of element spacing that favor close packing accommodates insertions and deletions. Our implementation of a multi-element version of the Gibbs sampler (23) includes ordering probabilities (24). As illustrated below, this joint information improves the prediction of the correct alignment of colinear elements. Constraints on loop length variation result in similarities in the spacing of the elements of homologous proteins. Thus, inclusion of an element spacing component in the model should improve alignment. However, we have not yet found it necessary to incorporate spacing effects into the algorithm (25).

Examples. To examine the algorithm, we have chosen three examples that present different classes of difficulties for automated multiple alignment. First is the helix-turn-helix (HTH) motif, which represents a large class of sequence-specific DNA binding structures involved in numerous cases of gene regulation. Such HTH motifs generally occur singly as local, isolated structures in different sequence contexts. Detection and alignment of HTH motifs is a well-known problem because of the great sequence variation compatible with the same basic structure. Second are the lipocalins, a family of proteins that bind small, hydrophobic ligands for a wide range of biological purposes. These proteins show widely spaced sequence motifs within highly variable sequences but share the same

---

**Fig. 1.** Alignment and probability ratio model for the helix-turn-helix pattern common to 30 proteins (45). (A) The alignment. Columns from left to right are: sequence name; locations $a_i$ of the left end of the common pattern in each sequence; locations $a_{i+17}$ of the right end of the common pattern; NBRF/PRI accession number; and NBRF/PRI code name, if available. Asterisks (*) below the alignment indicate the 20-residue segment previously described on the basis of structural superpositions (26, 27). Almost equal values of information per parameter were given by pattern widths of 18 to 21 residues (Fig. 2); the longer extensions added to the right the 18-residue pattern shown. (B) Probability ratios $(100 \times q_i/p_i)$ for each amino acid at each position in the model pattern.
structural topology throughout the polypeptide chain. We chose the five most divergent lipocalins with known 3D structure for analysis because the correct alignment of their sequence motifs has previously depended on structural superposition. Third are isoprenyl-protein transferases, essential components of the cytoplasmic signal transduction network. The B subunits of these enzymes contain multiple copies of multiple motifs that have not previously been satisfactorily characterized by automated alignment methods.

**Single site: HTH proteins.** The widespread DNA binding HTH structure comprises 20 contiguous amino acids (26). In our test set of 30 proteins (Fig. 1), the correct location of the motif is known (26, 27) from x-ray and nuclear magnetic resonance structures, or from substitution mutation experiments, or both. The rest of the 3D structure of these proteins, apart from the HTH motif itself, is quite different in different subfamilies. Furthermore, the element is found at positions throughout the polypeptide chain. Our test set represents a typically diverse cross section of HTH sequences. Close homologs have been excluded. The difficulty of detection and alignment of the HTH motif from such sequences is well recognized. There have been several attempts to develop position-specific weight matrices and other empirical pattern discriminators diagnostic for this structure (28). These have achieved some success in making several predictions that were later confirmed and that have also aroused controversy (29).

We used this example to develop two important features of the algorithm. First, the empirical criterion of information per parameter allowed for the automated determination of element width (Fig. 2). Second, heuristic convergence criteria substantially shortened the time required to find the best model (Fig. 3, legend). These two features enabled the algorithm to identify and align all 30 HTH motifs quickly and consistently. The structure itself, is completely obtained with five pattern widths in the range from 17 to 22 residues (Fig. 2), of which 21 residues had the highest converged value of information per parameter. These results compare favorably with the 20-residue view based previously on structural superpositions (26, 27). The criteria developed empirically with the HTH example have worked consistently well in all of our subsequent applications.

**Multiple sites: Lipocalins.** The majority of protein sequence families contain multiple collinear elements separated by variable-length gaps (13). We have successfully aligned distantly related sequences for several problems in this class, including protein kinases, aspartyl proteinases, 3'RNA ligases and mammalian helix-loop-helix proteins. We report here on one of the most difficult of these test cases: in lipocalins (30, 31), two weak sequence motifs, centered on the generally conserved residues -Gly-X-Trp- and -Thr-Asp-, are recognized from structural comparisons (31, 32). 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 [such as (33)], fail

---

**Fig. 2.** Information per parameter as the criterion of pattern width for helix-turn-helix (HTH) proteins. The points indicate the maximum values of information per parameter found by the algorithm. The upper points (▲ and +) used the complete sequences of the 30 HTH proteins listed in Fig. 1A. (▲) All of the sequences in the data set were aligned in the correct order (as in Fig. 1A). (+) One or more of the sequences in the data set were incorrectly aligned. All completely correct alignments in the width range from 17 to 22 residues gave greater values of information per parameter than any incorrect alignments outside this width range. (●) The "nonsites" sequence data of the 30 HTH proteins, constructed by deleting the 18 residues of the HTH pattern itself (Fig. 1A) from each of the sequences. (●) A shuffled data set (46) of the 30 HTH sequences. The alignments from the nonsites background of the HTH proteins give values slightly greater than random expectation.

**Fig. 3.** Convergence behavior of the Gibbs sampling algorithm. Because the Gibbs sampler, when run for finite time, is a heuristic rather than a rigorous optimization procedure, one cannot guarantee the optimality of the results it produces. Therefore, the best solution found in a series of runs will be called "maximal." A single pattern of width 18 residues was sought in the data set of 30 HTH proteins shown in Fig. 1A. Solid lines show the course of three independent runs with different random seeds. Evolving models in such runs rapidly reach intermediate "background" information values (1.0 to 1.2 bits per parameter) and then sample different models in this plateau region for a widely variable number of iterations before converging rapidly. Curve 1 is typical in showing a very short lag time on the plateau; longer lags as in curves 2 and 3 are less common. Curves 1 and 3 illustrate the stochastic behavior of the Gibbs sampler; once "converged," the model stays predominantly at the maximal value of 1.84 bits per parameter but is never permanently in this solution. In the infinite limit, the sampler will spend the plurality of its time on the pattern that maximizes F and therefore the information per parameter (22). Curve 2 demonstrates persistence (after escape from the background plateau) in a submaximal state (1.60 bits per parameter), which is a "phase-shifted" version of the best model. When sufficiently large stochastic phase shifts are allowed (see text), such states do not normally trap the evolving model for many iterations. Curve 3 reaches exactly the same maximal value as curve 1, suggesting one possible strategy for detecting convergence, namely, recurrence of exactly the same pattern with different seeds. The following heuristic approach was found to greatly reduce the time required to find the apparently optimal alignment: (i) For a given random seed, repeat the basic algorithm a fixed number of times (typically 10 times for each input sequence) beyond the last iteration in which the best pattern observed (with this seed) improved; and (ii) try at most some fixed number of seeds (usually 10 for a single element), but stop when the best pattern is reproduced by a specified number of different seeds (usually 2). The rationale underlying this approach is that it is unlikely for the identical suboptimal solution to be found on several independent trials before the optimal solution is found once. The dotted line represents a run on the nonsite data set (Fig. 2). Such runs never exceed 1.2 bits per parameter and thus are stuck permanently in states resembling the background plateaus from which the models of the HTH motif alignments eventually escape. Repeated runs in which shuffled sequences as input data are used can provide criteria for how strong a pattern is required to be considered significant (38).

---

**SCIENCE • VOL. 262 • 8 OCTOBER 1993**
to align these motifs for the full spectrum of lipocalin sequences. Challenged with five such diverse sequences of known crystal structure, our algorithm correctly aligned these two regions and extended the width of both to 16 residues (Fig. 4), in agreement with the structural evidence (31, 32).

Multiple copies of multiple sites: Prenyltransferases. Internal repeats in protein sequences underlie many important structural functions and are more common than is generally recognized (34). These repeats are often obscured by sequence divergence following duplication, rendering their detection and characterization a challenging problem. The analysis of repeats is often labor-intensive, relying in part on visual inspection of “dot plots” (10, 34)—a procedure that limits searches and surveys of large databases.

An example of recent interest involves sequence repeats in the subunits of the heterodimeric protein-isoprenyltransferases (10, 35). These enzymes are responsible for targeting and anchoring members of the ras superfamily of small guanosine triphosphates to their sites of action on various cellular membranes (36). The β subunits of prenyltransferases contain a subtel internal repeat of possible function significance (34). Although no direct structural information is yet available for these proteins, previous sequence analysis suggested that the β subunit repeat consists of three motifs separated by variable-length gaps and that this entire tripartite structure is repeated three to five times in each of four proteins (10, 35).

The challenge here is therefore to identify a relatively large number of weak patterns covering up to 80 percent of the length of the sequences. The resulting crowding of elements increases interelement dependencies and the complexity of the joint probability surface over which the algorithm must find the most probable alignment.

The previous analysis was subjective and time-consuming, relying on the combined use of several different multiple alignment methods. In contrast, the Gibbs sampling algorithm quickly and objectively reproduced and extended the previous results (Fig. 5).

Evaluation and comparison. The main difficulties of automated local multiple alignment stem from the high dimensionality of the search space and the existence of many local optima. Here, the large search space is explored one dimension at a time by comparing each sequence to an evolving residue frequency model. Stochastic sampling permits the algorithm to escape local optima in which deterministic approaches may get trapped. Including a phase shift step expedites convergence by permitting the sampler to explore related local optima.

Tests showed the algorithm to be relatively insensitive to various numbers of negative examples included among the input sequences. To cope with large numbers of input negative examples, we have extended the algorithm to seek a pattern in only a specified number of input sequences.

The use of an appropriate model for interelelemet spacing would improve the algorithm’s sensitivity, but this feature has
not been needed to identify even the subtle patterns described above. The problem of highly correlated input sequences can be addressed by various weighting schemes (37), but we have yet to implement such a feature. Choosing an optimal number of elements requires further study. We have found that an input with elements is not warranted when multiple random seeds lead to many different alignments and when the resulting information per parameter consistently fails to exceed that obtained from shuffled sequences (38). Prior knowledge concerning amino acid ratios (39) has been used profitably in pairwise protein sequence alignment as well as in pattern construction methods (8, 40). We have modified the Gibbs sampler to use such prior information, but in practice, for even moderate numbers of sequences (≥5), we have not found it to yield any improvement. However, an interesting new approach to incorporating prior information has been described (41), and there is much room for further experimentation.

Some basic similarities between our method and several earlier ones should be noted. Stormo and Hartzell and Hertz et al. (5) seek the pattern that maximizes a measure similar to F. Their approach differs mainly in the heuristic optimization procedure used, which is an adaptation of an algorithm first proposed by Bacon and Anderson (4). We have implemented the method of (5) and tested it on a variety of examples. This approach uses only a small subset of the data for the early sequences examined, and thus is easily misled. As a result, the solution found was rarely as good as that produced by the sampler. Furthermore, the need to construct an alignment for each possible segment in the initial sequence requires on average more passes through the input data than does the sampler (see below), resulting in greater execution times. Both EM methods (42) and the Gibbs sampler are built on a common statistical framework. Two EM approaches for multiple alignment have been described, block-based methods (6, 7) and gap-based methods in the form of hidden Markov models (43). For multielement problems, the Gibbs sampler outperforms block-based EM methods. Because EM methods are forced to sum over all possibilities, the time complexity grows exponentially with additional elements. In contrast, the Gibbs sampler never needs to consider more than one element at a time. The speed of the sampler stems partly from the fact that it always deals with a specific model alignment rather than a weighted average. Also, because EM methods are deterministic, they tend to get trapped by local optima which are avoided by the sampler. Hidden Markov models, because they permit arbitrary gaps, have great flexibility in modeling patterns, but suffer the penalties of this added complexity discussed above.

Several other approaches to the local multiple alignment problem bear a brief review. Methods that seek a “consensus” word with the highest aggregate score among segments within the output sequences have been described (3). Their space requirements effectively limit them to protein patterns of six residues, and their time requirements effectively allow only closely related words to contribute to a consensus. These constraints greatly decrease the sensitivity of these methods to weak patterns.

Algorithms that compare all input sequences with one another and then coalesce consistent pairwise local alignments have been described (8–10). The MACAW algorithm (8) has comparable speed to the Gibbs sampler for a relatively small number of input sequences and can locate many distinct patterns in a single run. Its time complexity, however, is at least quadratic in the aggregate length of the input sequences, and it tends to be less sensitive to weak sequence patterns. The performance of methods that must compare all input sequences with one another may degrade as the number of sequences increases. In contrast, the power of the Gibbs sampler and EM methods increases with additional sequences because the pattern model is improved by more data. As illustrated above, the Gibbs sampler is successful even with a relatively small number of input sequences. A version of the Gibbs sampling algorithm has been added to the MACAW program (8), and the updated program is available upon request.

The memory requirements for the Gibbs sampler are negligible; storing the input sequences is usually the dominant space demand. When flexible halting criteria, such as those described in Fig. 3, are used, it is difficult to analyze the worst-case time complexity of the method. However, for typical protein sequence data sets, we have found that, for a single pattern width, each input sequence needs to be sampled on average fewer than T = 100 times before convergence. In the more time-consuming step 2 of the basic algorithm, approximately LW 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 7N/2, where L is the average length of the N input sequences (44). 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.
23. A version of the Gibbs sampling procedure for locating patterns within nucleic acids has been implemented in the C programming language, and is available from the authors upon request.

24. During each iteration, the orders of all elements in $N - 1$ of the input sequences are available. Counts of observed orders are combined with prior probabilities (taken as uniform in our applications) to calculate "model" order probabilities, analogously to Eq. 1. When choosing an element's new position, the weights $A_i$ of all candidate segments may be adjusted naturally by multiplication with these posterior order probabilities. A probabilistic model of element location based jointly on the observed order and residue frequency is produced. In our applications we have set the total number of order pseudocounts to NSTA, where $N$ is the number of sequences, $S$ is the number of sites per sequence, $T$ is the number of types of element, and $A$ has been chosen so that each of the ordering model will be described elsewhere (J. S. Liu et al., in preparation).

25. We have developed and initially tested a stochastic multiple alignment procedure which permits complete flexibility in gaps. It uses the same Markov characteristic that is used in dynamic programming and perfect matching. To date we found that the increase in gap flexibility permitted by this algorithm is not worth the cost of the increase in noise from chance optimal alignments.


38. When sufficient data are available, it is possible to derive an asymptotic test for the statistical significance of any pattern found. In practice, sufficient data are not usually available for this asymptotic test, especially for protein sequences. Given the speed of the algorithm, Monte Carlo tests based on these sequences (S. F. Altschul and B. W. Erickson, Mol. Biol. Evol. 2, 526 (1985) provide a viable alternative. In any case, Monte Carlo tests are superior under many circumstances to asymptotic tests [P. Hall and D. M. Titterington, J. R. Stat. Soc. B 51, 459 (1989)].


45. For a representative data set, such as that from the HTH problem studied above, $T = 50$, $N = 30$, $I = 250$, and $W = 20$, yielding $8,000,000$ iterations of the innermost loop. The number of possible solutions for such a problem is $10^{28}$.


49. J. S. L. is supported in part by NIH grant MH-31-164.

14 May 1993; accepted 3 September 1993