



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

Science, New Series, Vol. 262, No. 5131. (Oct. 8, 1993), pp. 208-214.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819931008%293%3A262%3A5131%3C208%3ADSSSAG%3E2.0.CO%3B2-0>

Science is currently published by American Association for the Advancement of Science.

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.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aaas.html>.

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.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

REFERENCES AND NOTES

- O. Diels and K. Alder, *Justus Liebigs Ann. Chem.* **460**, 98 (1928); O. Diels, J. H. Blom, W. Koll, *ibid.* **443**, 242 (1925); B. M. Trost, I. Fleming, L. A. Paquette, *Comprehensive Organic Synthesis* (Pergamon, Oxford, 1991), vol. 5, pp. 316; F. Friguelli and A. Taticchi, *Dienes in the Diels-Alder Reaction* (Wiley, New York, 1990), and references therein; W. Carruthers, *Cycloaddition Reactions in Organic Synthesis* (Pergamon, Oxford, 1990), and references therein; G. Desimoni, G. Tacconi, A. Barco, G. P. Pollini, *ACS Monograph No. 180* (American Chemical Society, Washington, DC, 1983), and references cited therein.
- For recent reviews, see U. Pindur, G. Lutz, C. Otto, *Chem. Rev.* **93**, 741 (1993); H. B. Kagan and O. Riant, *ibid.* **92**, 1007 (1992), and references therein.
- D. Hilvert, K. W. Hill, K. D. Nared, M.-T. M. Auditor, *J. Am. Chem. Soc.* **111**, 9261 (1989).
- A. C. Braisted and P. G. Schultz, *ibid.* **112**, 7430 (1990).
- C. J. Suckling, M. C. Tedford, L. M. Bence, J. I. Irvine, W. H. Stimson, *Biorg. Med. Chem. Lett.* **2**, 49 (1992).
- K. D. Janda, C. G. Shevlin, R. A. Lerner, *Science* **259**, 490 (1993).
- L. E. Overman, G. F. Taylor, C. B. Petty, P. J. Jessup, *J. Org. Chem.* **43**, 2164 (1978).
- For example: *dl*-pumiliotoxin, C. L. E. Overman and P. J. Jessup, *Tetrahedron Lett.* 1253 (1977); *J. Am. Chem. Soc.* **100**, 5179 (1978); *dl*-perhydrogephyrotoxin, L. E. Overman and C. Fukaya, *ibid.* **102**, 1454 (1980); *dl*-isogabaculine, S. Danishefsky and F. M. Hershenov, *J. Org. Chem.* **44**, 1180 (1979); tylidine, L. E. Overman, C. B. Petty, R. J. Doedens, *ibid.*, p. 4183; amaryllidaceae alkaloids, L. E. Overman *et al.*, *J. Am. Chem. Soc.* **103**, 2816 (1981); gephyrotoxine, L. E. Overman, D. Lesuisse, M. Hashimoto, *ibid.* **105**, 16, 5373 (1983).
- Diene **1b** was prepared by neutralizing the corresponding carboxylic acid suspended in PBS (sodium phosphate buffer) with one equivalent of NaOH.
- All new adducts were characterized by NMR and mass spectrometry.
- In the crude mixture of the cycloaddition, we could not detect any trace of the *meta* regioisomer by ¹H NMR (>98% *ortho* regioselective).
- M. J. Frisch *et al.*, *Gaussian 92* (Gaussian, Pittsburgh, PA, 1992).
- R. J. Loncharich, F. K. Brown, K. N. Houk, *J. Org. Chem.* **54**, 1129 (1989). All structures were fully optimized with the 3-21G basis set. Reactants were also fully optimized with the 6-31G* basis set. Vibrational frequencies for all transition structures were calculated at the RHF/3-21G level and shown to have one imaginary frequency. In addition, the energies were evaluated by single-point calculations with the 6-31G* basis set on 3-21G geometries.
- K. N. Houk, R. J. Loncharich, J. Blake, W. Jorgensen, *J. Am. Chem. Soc.* **111**, 9172 (1989).
- L. E. Overman, G. F. Taylor, K. N. Houk, I. N. Domelsmith, *ibid.* **100**, 3182 (1978).
- M. I. Page and W. P. Jencks, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 1678 (1971).
- F. Bernardi *et al.*, *J. Am. Chem. Soc.* **110**, 3050 (1989); F. K. Brown and K. N. Houk, *Tetrahedron Lett.* **25**, 4609 (1984).
- This strategy to avoid product inhibition was first described by Braisted and Schultz in (4).
- The stereochemistry was established on the basis of ¹H NMR analysis. For example: **5-exo**, $\delta H_6 = 2.65$ ppm and $^4J_{6,7a} = 3.5$ Hz; **5-endo**, $\delta H_6 = 3.12$ ppm, and $^4J_{6,7a} = 0$ Hz.
- G. Kohler and C. Milstein, *Nature* **256**, 495 (1975); the KLH conjugate was prepared by slowly adding 2 mg of the hapten dissolved in 100 μ l of dimethylformamide to 2 mg of KLH in 900 μ l of 0.01 M sodium phosphate buffer, pH 7.4. Four 8-week-old 129GIX+ mice each received 2 intraperitoneal (IP) injections of 100 μ g of the hapten conjugated to KLH and RIBI adjuvant (MPL and TDM emulsion) 2 weeks apart. A 50 μ g IP fluoroacetic acid/acetoneitrile, at 1.5 ml/min, 240 nm. The retention time for the *trans* isomer is 6.00 min and for the *cis* isomer is 6.60 min.
- We thank J. Ashley for technical assistance in the kinetic analysis; L. Ghosez, K. B. Sharpless, K. C. Nicolaou, E. Keinan, and D. Boger for helpful discussions and comments on the manuscript; and R. K. Chadha for the x-ray crystallographic study. Supported in part by the National Science Foundation (CHE-9116377) and the A. P. Sloan Foundation (K.D.J.). We also thank Fundación Ramón-Areces (Spain) for a fellowship to B.P.T.

16 July 1993; accepted 3 September 1993

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

C. E. Lawrence, S. F. Altschul, M. S. Boguski, A. F. Neuwald, and J. C. Wootton are with the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894. J. S. Liu is with the Statistics Department, Harvard University, Cambridge, MA 02138. C. E. Lawrence is also affiliated with the Biometrics Laboratory, Wadsworth Center for Labs and Research, New York State Department of Health, Albany, NY 12201.

highly variable patterns.

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, \dots, 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 $q_{i,1}, \dots, q_{i,20}$. This pattern description is accompanied by an analogous probabilistic description of the "background frequencies" p_1, \dots, p_{20} with which residues occur in sites not described by the pattern. The second data structure, constituting the alignment, is a set of positions a_k , 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_j are then calculated, as described in Eq. 1 below, from the current positions a_k 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_x of generating each segment x according to the current pattern probabilities $q_{i,j}$ are calculated, as are the probabilities P_x of generating these segments by the background probabilities p_j . The weight $A_x = Q_x/P_x$ is assigned to segment x , and with each segment so weighted, a random one is selected (19). Its position then becomes the new a_z .

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_k , in step 2 the pattern description $q_{i,j}$ will tend to favor no particular segment. Once some correct a_k have been selected by chance, however, the $q_{i,j}$ begin to reflect, albeit imperfectly, a pattern extant within other sequences. This process tends to recruit further correct a_k , 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 $q_{i,j}$ from the current set of a_k . For the i th position of the pattern we have $N - 1$ observed amino acids, because sequence z has been excluded; let $c_{i,j}$ be the count of amino acid j in this position. Bayesian statistical analysis suggests that, for the purpose of pattern estimation, these $c_{i,j}$ should be supplemented with residue-dependent "pseudocounts" b_j to yield pattern probabilities

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B} \quad (1)$$

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

After normalization, A_x 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_k '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

$$F = \sum_{i=1}^W \sum_{j=1}^{20} c_{i,j} \log \frac{q_{i,j}}{p_j} \quad (2)$$

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

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, 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_k 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 (19 W in the case of proteins) produced a statistic useful for choosing pattern width. 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 z 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-recognized 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

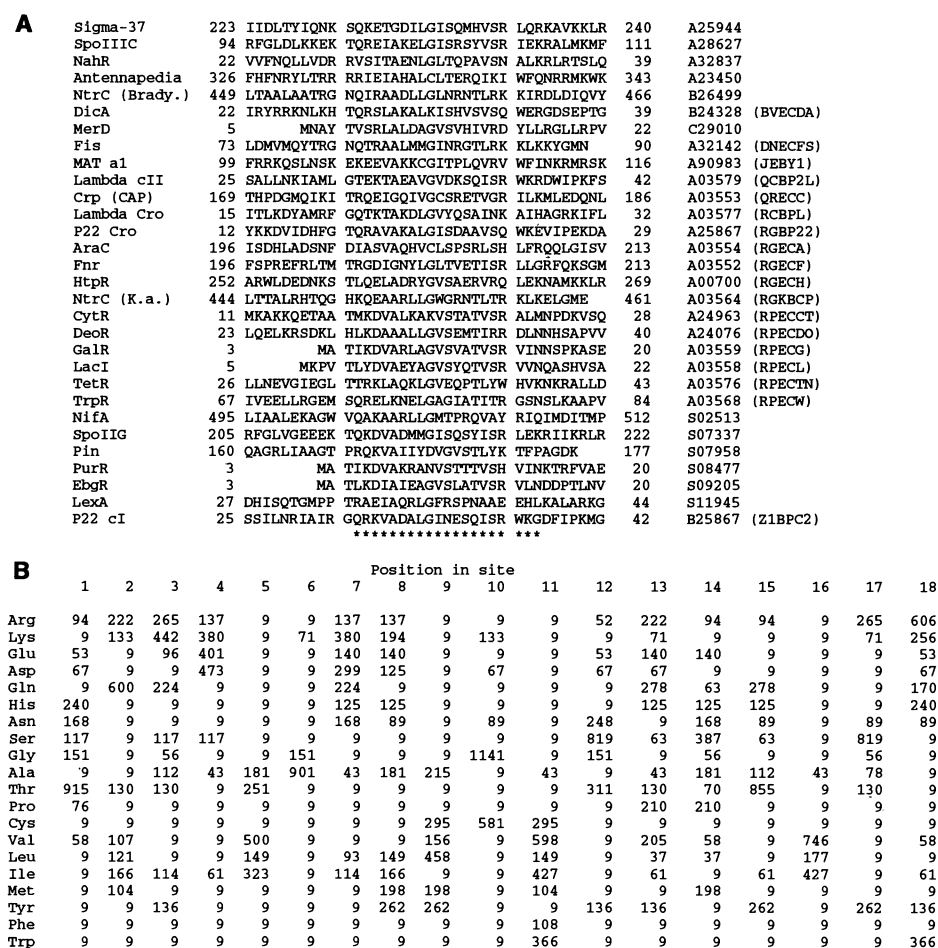


Fig. 1. Alignment and probability ratio model for the helix-turn-helix pattern common to 30 proteins (45). (A) The alignment. Columns from left to right are: sequence name; locations a_k of the left end of the common pattern in each sequence; aligned sequences, including residues flanking the 18-residue common pattern; right-end positions ($a_k + 17$) of the common pattern; NBRF/PIR accession number; and NBRF/PIR 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 widths extended to the right the 18-residue pattern shown. (B) Probability ratios ($100 \times a_{ij}/p$) for each amino acid at each position in the pattern model.

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 β 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 structure itself, is completely 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. Correct alignments were obtained with six 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 colinear 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, aminoacyl-

tRNA 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

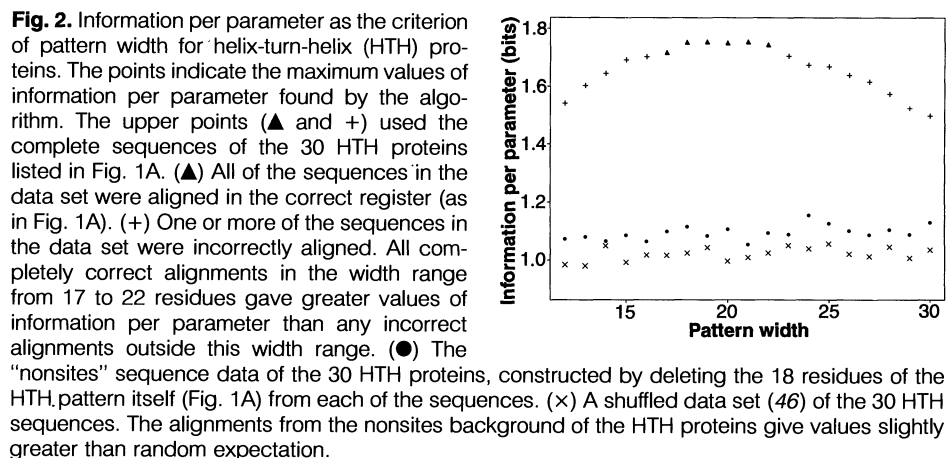


Fig. 2. Information per parameter as the criterion of pattern width for helix-turn-helix (HTH) proteins. The points indicate the maximum values of information per parameter found by the algorithm. The upper points (▲ and +) used the complete sequences of the 30 HTH proteins listed in Fig. 1A. (▲) All of the sequences in the data set were aligned in the correct register (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. (x) 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.

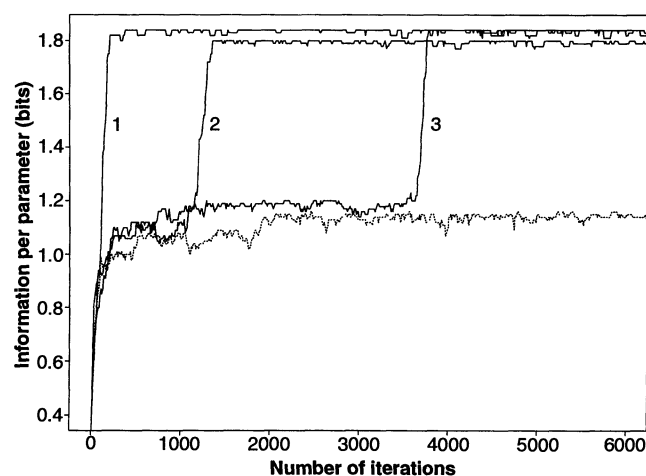
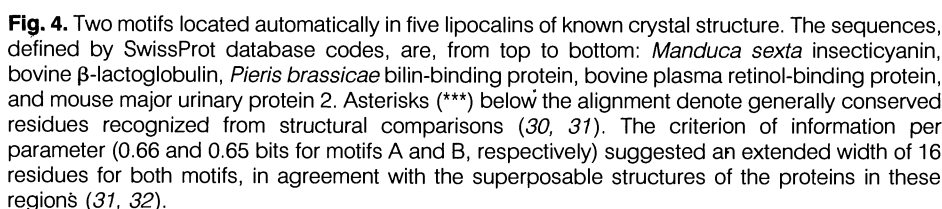


Fig. 3. Convergence behavior of the Gibbs sampling algorithm. Because the Gibbs sampler, when run for finite time, is a heuristic rather than a rigorous optimization procedure, one cannot guarantee the optimality of the results it produces. Therefore, the best solution found in a series of runs will 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 (1.80 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).

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.

The use of an appropriate model for interelement spacing would improve the algorithm's sensitivity, but this feature has



A		L		B	
Ram1					
109					
129	TKRKIVVKKFTI	SPSG-----	GPFGGGPGQLSH	LA-	MLYWIANSKLKVM
181	DRKGIYQWLISL	KEPN-----	GGFKTCLEVGVEV	DTR	STYAAINALSLK
230	LTEGVLYNLKNC	QNYE-----	GGFGSGCPHVDEA	HGG	GYICALSITALL
279	NVEKLLEWSSAR	QLQEE-----	RGFCGRSNKLVD	HGG	YTFCATASLAIL
331	NKHALRDYILYC	CQKEQ-----	PGLRDKPGAHS	GC-	YSFWVGWGSAAIL
			SSYSTCPNDSPH	FY-	HTNYCLLGLAVA
					...27 aa...
415	NVRKIIHYFKSN	LSSPS			
FT-β					
74	QREKHFHYLKRG	LRQLTDAYECLDAS			
99	SRPWLVCYWLHS	LELLDEPIQIV			
122	VATDVCQFLELC	QSPD-----	GGFGGGPGQYPH	LA	PTYAAVNALCII
173	NREKLLQYLYSL	KQPD-----	GGFLMHVGGVEVD	VR	SAYCAASVASLT
221	LFEGTAEWIARC	QNWFE-----	GGIGGVPGMEAH	GG	YTFCGLAALVIL
269	NLKSLLQWVTSR	QMRFE-----	GGFGQGRCNKLV	GC	YSFWQAGLLPLL
331	HQQAQLQEIYLMC	CQCPA-----	GGLLDKPGKSRD	FY	HTCYCLSGLSIA
					...20 aa...
					...
Bet2					
8	LKEKHIRYIESL	DTKKHNFEYWLTEHLRLN			
56	LKEEIVSFVLSC	WDDKY-----	GAFAPFPRHDAH	LL	TIYVGLTALCVL
108	RKVRLLISFIRGN	QLED-----	GSFQGDPRFGEVD	TR	TTLSAVQILATY
156	VVDPAVDVLKRC	YNFD-----	GGFGLCPNAESH	AA	FVYTALSALSIL
207	QLEEIQWVLCER	QLPFE-----	GGLNCRPSKLPD	VC	QAFTCLGALATA
255	NYEKLTEFILKC	QDEKK-----	GGISDRPENEVD	VF	YSWVWVLSLAI
					HTVFGVAGLSLM
					...20 aa...
					...
GGT-β					
19	LLEKHADYIASY	GSKKDDYECMSEYLRMS			
67	NKEELTVFIKSC	QHEC-----	GGVSASITGDHPH	LL	GVYWGLTVMDML
115	NVDVAVVAVQSL	QKED-----	GSFAGDITWGEID	TR	YTLTSAVQILLTYL
163	NVEKALIEFVLSC	MNFD-----	GGFGCRPGSSESH	AG	FSFCAVATLALL
211	NSDLLGWVLCER	QLPS-----	GGLNCRPEKLPD	VC	QYVCTGTFLATK
259	DREKLRSFILAC	QDEET-----	GGFADRPGDMVD	PF	YSWVWVLSLAKII
					HTLFGIAGLSLL
					...20 aa...
					...
Cdc43					
12	VTKKHKRFFERH	...103 aa...			
127	DKRSLARFVSKC	Q ...52 aa...			
191	DTEKLLGYIMSG	QCYN-----	GAFGAHNEPHSG	--	YTSCALSTLALL
240	FKEDTITWLLHR	QVSSHGCMKFESLNASYDQSD	GGFQGRNKFAD	TC	YAFWCLNSLHLL
309	QTELVTNYLLDR	TQKTLT-----	GGFSKNDEEDAD	LY	HSCLSGAALALI
					...20 aa...
					...

Fig. 5. Repeating motifs in prenyltransferase subunits. Ram1 (Swiss-Prot, accession number P22007) and FT- β (Swiss-Prot, Q02293) are the β subunits of farnesyltransferase from the yeast, *Saccharomyces cerevisiae*, and rat brain, respectively. Bet2 (PIR International, S22843) is the β subunit of type I geranylgeranyltransferase from *S. cerevisiae*. GGT- β (GenBank, L10416) and Cdc43 (Swiss-Prot, P18898) are the β subunits of type II geranylgeranyltransferase from *S. cerevisiae* and rat brain, respectively. 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 Gly and Pro residues (10, 35). When analyzed by the Gibbs sampler, these previously defined motifs were identified and additional copies were also observed [compare with figure 1 in (35)]. The information per parameter for motifs A, L, and B was 2.3, 2.3, and 2.4 bits, respectively. 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 gaps expressed as the number of amino acid residues. The spacing between motifs L and B is only two or three residues, whereas that between motifs A and L is greater and more variable.

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 additional element 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 relations (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 foundation. 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 against segments within the input 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 $T \approx 100$ 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 $TNLW$, where \bar{L} is the average length of the N input sequences (44). The factor T is expected to grow with increasing \bar{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 conclusion, as illustrated by our examples, the Gibbs sampler objectively solves 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. By adopting a randomized optimization procedure in the place of deterministic approaches, it is able to retain both speed and sensitivity to weak but biologically significant patterns.

REFERENCES AND NOTES

1. S. F. Altschul and D. J. Lipman, *SIAM J. Appl. Math.* **49**, 197 (1989); W. Bains, *Nucleic Acids Res.* **14**, 159 (1986); G. J. Barton and M. J. Sternberg, *J. Mol. Biol.* **198**, 327 (1987); M. P. Berger and P. J. Munson, *Comput. Appl. Biosci.* **7**, 479 (1991); H. Carrillo and D. Lipman, *SIAM J. Appl. Math.* **48**, 1073 (1988); D. Feng and R. F. Doolittle, *J. Mol. Evol.* **25**, 351 (1987); D. Gusfield, *Bull. Math. Biol.* **155**, 141 (1993); C. M. Henneke, *Comput. Appl. Biosci.* **5**, 141 (1989); D. G. Higgins, A. J. Bleasby, R. Fuchs, *ibid.* **8**, 189 (1992); M. S. Johnson and R. F. Doolittle, *J. Mol. Evol.* **123**, 267 (1986); D. J. Lipman, S. F. Altschul, J. D. Kececioglu, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 4412 (1989); M. Murata, J. S. Richardson, J. L. Sussman, *ibid.* **82**, 3073 (1985); R. B. Russell and G. J. Barton, *Proteins* **14**, 309 (1992); D. Sankoff, *SIAM J. Appl. Math.* **28**, 35 (1975); S. Subbiah and S. C. Harrison, *J. Mol. Biol.* **209**, 539 (1989); W. R. Taylor, *Methods Enzymol.* **183**, 456 (1990); M. Vingron and P. Argos, *Comput. Appl. Biosci.* **5**, 115 (1989); M. S. Waterman and M. D. Perlwitz, *Bull. Math. Biol.* **46**, 567 (1984).
2. G. J. Barton and M. J. Sternberg, *J. Mol. Biol.* **212**, 389 (1990); C. Chappey, A. Dancckaert, P. Dessen, S. Hazout, *Comput. Appl. Biosci.* **7**, 195 (1991); E. Depiereux and E. Feytmans, *ibid.* **8**, 501 (1992); D. J. Parry-Smith and T. K. Attwood, *ibid.* **7**, 233 (1991); *ibid.* **8**, 451 (1992); E. Sobel and H. Martinez, *Nucleic Acids Res.* **14**, 363 (1986).
3. J. Postfal, A. S. Bhagwat, G. Postfal, R. J. Roberts, *Nucleic Acids Res.* **17**, 2421 (1989); C. M. Queen, N. Wegman, L. J. Korn, *ibid.* **10**, 449 (1982); H. O. Smith, T. M. Annau, S. Chandrasegaran, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 826 (1990); R. Staden, *Comput. Appl. Biosci.* **5**, 293 (1989); M. S. Waterman, R. Arratia, D. J. Galas, *Bull. Math. Biol.* **46**, 515 (1984).
4. D. J. Bacon and W. F. Anderson, *J. Mol. Biol.* **191**, 153 (1986).
5. G. D. Stormo and G. W. Hartzell III, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 1183 (1989); G. Z. Hertz, G. W. Hartzell III, G. D. Stormo, *Comput. Appl. Biosci.* **6**, 81 (1990).
6. C. E. Lawrence and A. A. Reilly, *Proteins* **7**, 41 (1990).
7. L. R. Cardon and G. D. Stormo, *J. Mol. Biol.* **223**, 159 (1992).
8. G. D. Schuler, S. F. Altschul, D. J. Lipman, *Proteins* **9**, 180 (1991).
9. M. Vingron and P. Argos, *J. Mol. Biol.* **218**, 33 (1991).
10. M. S. Boguski, R. C. Hardison, S. Schwartz, W. Miller, *New Biol.* **4**, 247 (1992).
11. N. N. Alexandrov, *Comput. Appl. Biosci.* **8**, 339 (1992); S. Henikoff and J. G. Henikoff, *Nucleic Acids Res.* **19**, 6565 (1991); S. Henikoff, *New Biol.* **3**, 1148 (1991); M. Y. Leung, B. E. Blaisdell, C. Burge, S. Karlin, *J. Mol. Biol.* **221**, 1367 (1991); M. A. Roytberg, *Comput. Appl. Biosci.* **8**, 57 (1992).
12. P. Bork, C. Sander, A. Valencia, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7290 (1992).
13. J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981); A. M. Lesk and C. Chothia, *J. Mol. Biol.* **136**, 225 (1980); C. Chothia, *Annu. Rev. Biochem.* **53**, 537 (1984); A. Sali and T. L. Blundell, *J. Mol. Biol.* **212**, 403 (1990); R. F. Doolittle, *Protein Sci.* **1**, 191 (1992).

14. F. M. Pohl, *Nature New Biol.* **234**, 277 (1971); O. G. Berg and P. H. von Hippel, *J. Mol. Biol.* **193**, 723 (1987); S. H. Bryant and C. E. Lawrence, *Proteins* **16**, 92 (1993).
15. T. Orchard and M. A. Woodbury, *Proceedings of the Sixth Berkeley Symposium on Mathematics, Statistics and Probability* (Univ. of California Press, Berkeley, 1972), vol. 1, pp. 697–715; L. A. Goodman, *Biometrika* **61**, 215 (1974).
16. J. Liu, *Department of Statistics Research Report No. R-426* (Harvard Univ. Press, Cambridge, MA, 1992).
17. In the context of simulated annealing [N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, *J. Chem. Phys.* **21**, 1087 (1953); S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **220**, 671 (1983)], D. Geman and D. Geman [IEEE *Trans. Pattern Anal. Mach. Intell.* **6**, 721 (1984)] introduced Gibbs sampling as a particular stochastic relaxation step. They chose its name because a key theorem from statistical physics (the Hammersley Clifford Theorem) uses Gibbs/Boltzmann potentials. Because we use residue frequencies based on Gibbs/Boltzmann-like free energies, the name is more appropriate here than in many other applications.
18. M. Tanner and W. H. Wong, *J. Am. Stat. Assoc.* **82**, 528 (1987); A. E. Gelfand and A. F. M. Smith, *ibid.* **85**, 389 (1990).
19. Segment x is chosen with probability $A_x/\sum A_i$, where the sum is taken over all possible segments.
20. One could choose $q_{i,j}$ simply proportional to $c_{i,j}$, but this would imply a zero probability for any amino acid not actually observed. This difficulty may be surmounted through the use of Bayesian predictive inference (18). Bayesian analysis makes use of subjective "prior probabilities" for the values of the parameters to be estimated. A common choice for such priors when multinomial models are involved is the Dirichlet distribution [J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis* (Cambridge Univ. Press, New York, 1972)], which results in the simple addition of "pseudocounts" to the observed counts, as given in Eq. 1. These pseudocounts should capture a priori expectations concerning the occurrence of the various letters in different pattern positions and may vary from one application to another. We have found that letting $b_j = Bp_j$, where p_j is the frequency of residue j in the complete data set, is effective. We used this noninformative prior probability in all of our applications. The total number of pseudocounts B is also part of the prior specification, for which there is no "correct" choice. The smaller B is taken, the greater the reliance placed on current observation vis-a-vis a priori expectation. We have found that choosing $B \approx \sqrt{N}$ generally works well, yielding pseudocounts b_j applicable to the calculation of both $q_{i,j}$ and p_j .
21. H. Akaike, *IEEE Trans. Autom. Control* **19**, 716 (1974); G. Schwarz, *Ann. Stat.* **6**, 461 (1978).
22. G is equal to

$$F - \sum_{i=1}^N \left(\log L'_i + \sum_{j=1}^{L'_i} Y_{i,j} \log Y_{i,j} \right)$$
 where L'_i is the number of possible positions for the pattern within sequence i , and $Y_{i,j}$ is the normalized weight of position j , that is the weight Q_j/P_j divided by the sum of these weights within sequence i . This adjustment accounts for the fact that the position of the pattern within each sequence is not known [see (6) and R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data* (Wiley, New York, 1987)]. It can be shown that G increases monotonically with increasing F , so an optimization algorithm for F remains appropriate.
23. A version of the Gibbs sampling procedure for locating patterns within multiple sequences 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_x 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 NST/k , where N is the number of sequences, S is the number of sites per sequence, T is the number of types of element, and k has been chosen as 20. Details 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 to align two sequences. 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 local optima.
26. R. G. Brennan and B. W. Matthews, *J. Biol. Chem.* **264**, 1903 (1989); C. O. Pabo and R. T. Sauer, *Annu. Rev. Biochem.* **61**, 1053 (1992); J. Treisman, E. Harris, D. Wilson, C. Desplan, *Bioessays* **14**, 145 (1992).
27. D. Kostrewa *et al.*, *J. Mol. Biol.* **226**, 209 (1992); R. M. Lamerichs *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6863 (1989); M. A. Schell, P. H. Brown, S. Raju, *J. Biol. Chem.* **265**, 3844 (1990); A. Contreras and M. Drummond, *Nucleic Acids Res.* **16**, 4025 (1988); M. H. Drummond, A. Contreras, L. A. Mitchenall, *Mol. Microbiol.* **4**, 29 (1990); L. M. Shewchuk *et al.*, *Biochemistry* **28**, 2340 (1989); H. S. Yuan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9558 (1991); A. Brunelle and R. Schleif, *J. Mol. Biol.* **209**, 607 (1989); S. Spiro *et al.*, *Mol. Microbiol.* **4**, 1831 (1990); C. S. Barbier and S. A. Short, *J. Bacteriol.* **174**, 2881 (1992); M. D. Yudkin, J. H. Millonig, L. Appleby, *Mol. Microbiol.* **3**, 257 (1989); C. Berens, L. Altschmied, W. Hillen, *J. Biol. Chem.* **267**, 1945 (1992); R. J. Rolfes and H. Zalkin, *ibid.* **263**, 19653 (1988).
28. M. Drummond, P. Whitty, J. C. Wootton, *EMBO J.* **5**, 441 (1986); I. B. Dodd and J. B. Egan, *Nucleic Acids Res.* **18**, 5019 (1990); D. Akrigg *et al.*, *Comput. Appl. Biosci.* **8**, 295 (1992).
29. M. D. Yudkin, *Protein Eng.* **1**, 371 (1987); I. B. Dodd and J. B. Egan, *ibid.* **2**, 174 (1988).
30. A. Nagata *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 4020 (1991); M. C. Peitsch and M. S. Boguski, *New Biol.* **2**, 197 (1990).
31. S. W. Cowan, M. E. Newcomer, T. A. Jones, *Proteins* **8**, 44 (1990).
32. L. Sawyer, *Nature* **327**, 659 (1987); A. C. T. North, *Int. J. Biol. Macromol.* **11**, 56 (1989); Z. Bocskei *et al.*, *Nature* **360**, 186 (1992); H. M. Holden, W. R. Rypniewski, J. H. Law, I. Rayment, *EMBO J.* **6**, 1565 (1987); R. Huber *et al.*, *J. Mol. Biol.* **198**, 499 (1987); H. L. Monaco and G. Zanotti, *Biopolymers* **32**, 457 (1992); M. Z. Papiz *et al.*, *Nature* **344**, 383 (1986); D. R. Flower, A. C. T. North, T. K. Attwood, *Protein Sci.* **2**, 753 (1993).
33. M. S. Boguski, J. Ostell, D. J. States, in *Protein Engineering: A Practical Approach*, A. R. Rees, M. J. E. Sternberg, R. Wetzel, Eds. (IRL Press, Oxford, 1992), pp. 57–88.
34. D. J. States and M. S. Boguski, in *Sequence Analysis Primer*, M. Gribskov and J. Devereux, Eds. (Freeman, New York, 1991), pp. 141–148.
35. M. S. Boguski, A. W. Murray, S. Powers, *New Biol.* **4**, 408 (1992).
36. S. Clarke, *Annu. Rev. Biochem.* **61**, 355 (1992); W. R. Schafer and J. Rine, *Annu. Rev. Genet.* **30**, 209 (1992).
37. S. F. Altschul, R. J. Carroll, D. J. Lipman, *J. Mol. Biol.* **207**, 647 (1989); M. Vingron and P. R. Sibbald, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 8777 (1993).
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 shuffled 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. Titterton, *J. R. Stat. Soc. B* **51**, 459 (1989)].
39. M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, in *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, Ed. (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3, pp. 345–352; R. M. Schwartz and M. O. Dayhoff, in *ibid.*, pp. 353–358; S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991); S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915 (1992); D. F. Feng, M. S. Johnson, R. F. Doolittle, *J. Mol. Evol.* **21**, 112 (1985).
40. M. Gribskov, A. D. McLachlan, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4355 (1987); R. F. Smith and T. Smith, *ibid.* **87**, 118 (1990).
41. M. Brown *et al.*, in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, L. Hunter, D. Searls, J. Shavlik, Eds. (AAAI Press, Menlo Park, CA, 1993), pp. 47–55.
42. A. P. Dempster, N. M. Laird, D. B. Rubin, *J. R. Stat. Soc. B* **39**, 1 (1977).
43. D. Haussler, A. Krogh, S. Mian, K. Sjolander, *CIS Tech. Rep. No. UCSC-CRL-92-23* (University of California, Santa Cruz, 1992).
44. For a representative data set, such as that from the HTH problem studied above, $T \approx 50$, $N \approx 30$, $L \approx 250$, and $W \approx 20$, yielding $\sim 8,000,000$ iterations of the innermost loop. The number of possible solutions for such a problem is $>10^{65}$. Executed on an Indigo R4000 workstation (Silicon Graphics Corp.), the HTH example (Figs. 1 to 3) required 2 CPU seconds on average, while the examples of Figs. 4 and 5 required 3 and 31 CPU seconds, respectively.
45. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
46. J. C. Wootton and S. Federhen, *Comput. Chem.* **17**, 149 (1993).
47. J.S.L. is supported in part by NIMH grant MH-31-154.

14 May 1993; accepted 3 September 1993

LINKED CITATIONS

- Page 1 of 5 -



You have printed the following article:

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

Science, New Series, Vol. 262, No. 5131. (Oct. 8, 1993), pp. 208-214.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819931008%293%3A262%3A5131%3C208%3ADSSSAG%3E2.0.CO%3B2-0>

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

References and Notes

¹ **Trees, Stars, and Multiple Biological Sequence Alignment**

Stephen F. Altschul; David J. Lipman

SIAM Journal on Applied Mathematics, Vol. 49, No. 1. (Feb., 1989), pp. 197-209.

Stable URL:

<http://links.jstor.org/sici?sici=0036-1399%28198902%2949%3A1%3C197%3ATSAMBS%3E2.0.CO%3B2-0>

¹ **The Multiple Sequence Alignment Problem in Biology**

Humberto Carrillo; David Lipman

SIAM Journal on Applied Mathematics, Vol. 48, No. 5. (Oct., 1988), pp. 1073-1082.

Stable URL:

<http://links.jstor.org/sici?sici=0036-1399%28198810%2948%3A5%3C1073%3ATMSAPI%3E2.0.CO%3B2-0>

¹ **A Tool for Multiple Sequence Alignment**

David J. Lipman; Stephen F. Altschul; John D. Kececioglu

Proceedings of the National Academy of Sciences of the United States of America, Vol. 86, No. 12. (Jun. 15, 1989), pp. 4412-4415.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819890615%2986%3A12%3C4412%3AATFMSA%3E2.0.CO%3B2-J>

NOTE: The reference numbering from the original has been maintained in this citation list.

LINKED CITATIONS

- Page 2 of 5 -



¹ **Simultaneous Comparison of Three Protein Sequences**

M. Murata; J. S. Richardson; Joel L. Sussman

Proceedings of the National Academy of Sciences of the United States of America, Vol. 82, No. 10. (May 15, 1985), pp. 3073-3077.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819850515%2982%3A10%3C3073%3ASCOTPS%3E2.0.CO%3B2-L>

¹ **Minimal Mutation Trees of Sequences**

David Sankoff

SIAM Journal on Applied Mathematics, Vol. 28, No. 1. (Jan., 1975), pp. 35-42.

Stable URL:

<http://links.jstor.org/sici?sici=0036-1399%28197501%2928%3A1%3C35%3AMMTOS%3E2.0.CO%3B2-2>

³ **Finding Sequence Motifs in Groups of Functionally Related Proteins**

Hamilton O. Smith; Thomas M. Annau; Srinivasan Chandrasegaran

Proceedings of the National Academy of Sciences of the United States of America, Vol. 87, No. 2. (Jan., 1990), pp. 826-830.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%28199001%2987%3A2%3C826%3AFSMIGO%3E2.0.CO%3B2-J>

⁵ **Identifying Protein-Binding Sites from Unaligned DNA Fragments**

Gary D. Stormo; George W. Hartzell

Proceedings of the National Academy of Sciences of the United States of America, Vol. 86, No. 4. (Feb. 15, 1989), pp. 1183-1187.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819890215%2986%3A4%3C1183%3AIPSFUD%3E2.0.CO%3B2-S>

¹² **An ATPase Domain Common to Prokaryotic Cell Cycle Proteins, Sugar Kinases, Actin, and hsp70 Heat Shock Proteins**

Peer Bork; Chris Sander; Alfonso Valencia

Proceedings of the National Academy of Sciences of the United States of America, Vol. 89, No. 16. (Aug. 15, 1992), pp. 7290-7294.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819920815%2989%3A16%3C7290%3AAADCTP%3E2.0.CO%3B2-W>

NOTE: The reference numbering from the original has been maintained in this citation list.

LINKED CITATIONS

- Page 3 of 5 -



¹⁵ **Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models**

Leo A. Goodman

Biometrika, Vol. 61, No. 2. (Aug., 1974), pp. 215-231.

Stable URL:

<http://links.jstor.org/sici?sici=0006-3444%28197408%2961%3A2%3C215%3AELSAUB%3E2.0.CO%3B2-F>

¹⁷ **Optimization by Simulated Annealing**

S. Kirkpatrick; C. D. Gelatt; M. P. Vecchi

Science, New Series, Vol. 220, No. 4598. (May 13, 1983), pp. 671-680.

Stable URL:

<http://links.jstor.org/sici?sici=0036-8075%2819830513%293%3A220%3A4598%3C671%3AOSBA%3E2.0.CO%3B2-8>

¹⁸ **The Calculation of Posterior Distributions by Data Augmentation**

Martin A. Tanner; Wing Hung Wong

Journal of the American Statistical Association, Vol. 82, No. 398. (Jun., 1987), pp. 528-540.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28198706%2982%3A398%3C528%3ATCOPDB%3E2.0.CO%3B2-M>

²⁰ **The Calculation of Posterior Distributions by Data Augmentation**

Martin A. Tanner; Wing Hung Wong

Journal of the American Statistical Association, Vol. 82, No. 398. (Jun., 1987), pp. 528-540.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28198706%2982%3A398%3C528%3ATCOPDB%3E2.0.CO%3B2-M>

²⁷ **The Amino-Terminal Domain of LexA Repressor is α -helical but Differs from Canonical Helix-Turn-Helix Proteins: A Two-Dimensional ¹H NMR Study**

R. M. J. N. Lamerichs; A. Padilla; R. Boelens; R. Kaptein; G. Otteleben; H. Rüterjans; M. Granger-Schnarr; P. Oertel; M. Schnarr

Proceedings of the National Academy of Sciences of the United States of America, Vol. 86, No. 18. (Sep. 15, 1989), pp. 6863-6867.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819890915%2986%3A18%3C6863%3ATADOLR%3E2.0.CO%3B2-U>

LINKED CITATIONS

- Page 4 of 5 -



²⁷ **The Molecular Structure of Wild-Type and a Mutant Fis Protein: Relationship Between Mutational Changes and Recombinational Enhancer Function or DNA Binding**

Hanna S. Yuan; Steven E. Finkel; Jin-An Feng; Maria Kaczor-Grzeskowiak; Reid C. Johnson; Richard E. Dickerson

Proceedings of the National Academy of Sciences of the United States of America, Vol. 88, No. 21. (Nov. 1, 1991), pp. 9558-9562.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819911101%2988%3A21%3C9558%3ATMSOWA%3E2.0.CO%3B2-E>

³⁰ **Human Brain Prostaglandin D Synthase has been Evolutionarily Differentiated from Lipophilic-Ligand Carrier Proteins**

Akishisa Nagata; Yasuhiko Suzuki; Makoto Igarashi; Naomi Eguchi; Hiroyuki Toh; Yoshihiro Urade; Osamu Hayaishi

Proceedings of the National Academy of Sciences of the United States of America, Vol. 88, No. 9. (May 1, 1991), pp. 4020-4024.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819910501%2988%3A9%3C4020%3AHBPDSH%3E2.0.CO%3B2-Y>

³⁷ **Weighting in Sequence Space: A Comparison of Method in Terms of Generalized Sequences**

Martin Vingron; Peter R. Sibbald

Proceedings of the National Academy of Sciences of the United States of America, Vol. 90, No. 19. (Oct. 1, 1993), pp. 8777-8781.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819931001%2990%3A19%3C8777%3AWISSAC%3E2.0.CO%3B2-2>

³⁹ **Amino Acid Substitution Matrices from Protein Blocks**

Steven Henikoff; Jorja G. Henikoff

Proceedings of the National Academy of Sciences of the United States of America, Vol. 89, No. 22. (Nov. 15, 1992), pp. 10915-10919.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819921115%2989%3A22%3C10915%3AAASMFP%3E2.0.CO%3B2-E>

⁴⁰ **Profile Analysis: Detection of Distantly Related Proteins**

Michael Gribskov; Andrew D. McLachlan; David Eisenberg

Proceedings of the National Academy of Sciences of the United States of America, Vol. 84, No. 13. (Jul. 1, 1987), pp. 4355-4358.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%2819870701%2984%3A13%3C4355%3APADODR%3E2.0.CO%3B2-M>

NOTE: The reference numbering from the original has been maintained in this citation list.

LINKED CITATIONS

- Page 5 of 5 -



⁴⁰ **Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences**

Randall F. Smith; Temple F. Smith

Proceedings of the National Academy of Sciences of the United States of America, Vol. 87, No. 1.
(Jan., 1990), pp. 118-122.

Stable URL:

<http://links.jstor.org/sici?sici=0027-8424%28199001%2987%3A1%3C118%3AAGOPSP%3E2.0.CO%3B2-V>