REFERENCES AND NOTES

1. 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.

injection of KLH conjugate in alum was given 2 weeks later. One month after the second injection, the mouse with the highest titer was injected intravenously with 50 μg of KLH-conjugate; 3 days later, the spleen was taken for the preparation of hybridomas. Spleen cells (1.0 × 10^8) were fused with 2.0 × 10^7 SP2/0 myeloma cells. Cells

fluoroacetic acid)/acetonitrile, 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.

23. 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 manu-

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 se-

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 previ-

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 descrip-

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

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

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

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).

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 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 interelement 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 additional element is not

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

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

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 $A_x/\sum_i Ai$, 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. Dumsmore, *Statistical Predic-*

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.

32, 457 (1992); M. Z. Papiz *et al.*, *Nature* 34, 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. Titterington, *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,