Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 1059–1063, February 1994
Biochemistry

Hidden Markov models of biological primary sequence information

(multiple sequence alignments/protein modeling/adaptive algorithms/sequence classification)

PIERRE BALDI*†, YVES CHAUVIN‡§, TIM HUNKAPILLER*¶, AND MARCELLA A. MCCLURE||**

*Division of Biology, California Institute of Technology, Pasadena, CA 91125; ‡NetID, Inc., San Francisco, CA 94107; ¶Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195; ||Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717; †Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109; and §Department of Psychology, Stanford University, Stanford, CA 94025

Communicated by Leroy Hood, October 12, 1993 (received for review January 14, 1993)

ABSTRACT **Hidden Markov model (HMM) techniques are used to model families of biological sequences. A smooth and convergent algorithm is introduced to iteratively adapt the transition and emission parameters of the models from the examples in a given family. The HMM approach is applied to three protein families: globins, immunoglobulins, and kinases. In all cases, the models derived capture the important statistical characteristics of the family and can be used for a number of tasks, including multiple alignments, motif detection, and classification. For K sequences of average length N, this approach yields an effective multiple-alignment algorithm which requires $O(KN^2)$ operations, linear in the number of sequences.**

Comparative analysis of primary sequence information is a major tool in the elucidation of the molecular mechanisms of replication and evolution of organisms and the structure and function of proteins. For the simple case of pairwise sequence comparison, good algorithms exist (see refs. 1 and 2 for recent reviews) that can align two sequences of length N in roughly $O(N^2)$ steps. Most of these algorithms are based on dynamic programming (3), with location-independent substitution and gap penalties. Unfortunately, when dynamic programming is applied to a family of K sequences its behavior scales like $O(N^K)$, exponentially in the number of sequences (4).

A number of algorithms have been devised to try to tackle the multiple alignment problem (see refs. 5–7 for some of the most recent ones). Most protein sequence relationships exhibiting >50% identical residues can be aligned by several of these algorithms. Many of the most interesting protein families, however, exhibit conservation far below 50% identity. To date, alignment methods have not been developed that can correctly identify all the motifs that define each protein family (2).

Here, we apply a different approach, based on hidden Markov models (HMMs), to the problem of modeling and aligning a family by using primary structure information only. Initial results were presented (8). Markov models and the related expectation–maximization (EM) (9) algorithm in statistics have already been applied to biocomputational problems (10–13). Krogh *et al.* (14) were the first to demonstrate the power of a similar method on the globin family. Rather than starting from pairwise alignments, the approach seeks to take advantage of the massive amount of information typically present in a family with a flexible use of position-dependent parameters. A new algorithm is introduced for the iterative adjustments of the parameters of the models. The algorithm is used here to model three protein families: globins, immunoglobulins, and kinases.††

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

HMMs and Learning

A first-order discrete HMM (15) is completely defined by a set of states S, an alphabet of m symbols, a probability transition matrix $\mathbf{T} = (t_{ij})$, and a probability emission matrix $\mathbf{E} = (e_{i\alpha})$. When the system is in state i, it has a probability t_{ij} of moving to state j and a probability $e_{i\alpha}$ of emitting symbol α. Only the output string is observed, one of the goals being the reconstruction of the underlying hidden transitions.

As in the application of HMMs to speech recognition, a family of biological sequences can be seen as a set of different utterances of the same word generated by a common underlying HMM with a left–right architecture, with $m = 4$ for DNA or RNA and $m = 20$ for proteins. Common knowledge about evolutionary mechanisms suggests to introduce three classes of states (in addition to the start and end states): the main states, the delete states, and the insert states with S = start, $m_1, \ldots, m_N$, $i_1, \ldots, i_{N+1}$, $d_1, \ldots, d_{N+1}$, end (Fig. 1). N is the length of the model. The main and insert states always emit a letter of the alphabet, whereas the delete states are mute. The linear sequence of main state transitions is the backbone of the model. Self loops on the insert states allow for multiple insertions. Architectural variations are possible and may be tailored to particular problems when additional information is available.

Given a set of training sequences, the parameters of a model can be iteratively modified to optimize the fit of the model to the data according to some measure, usually the product of the likelihoods of the sequences. Different algorithms are available for HMM training, including the classical Baum–Welch algorithm (15). Here, we introduce a smooth algorithm which is particularly simple and can be used on-line—i.e., after the presentation of each example. The mathematical properties of this algorithm and its relation to other approaches have been studied (16). We first reparametrize the model by using a new set of variables w_{ij} and $v_{i\alpha}$ in the form $t_{ij} = e^{w_{ij}}/\Sigma_k\, e^{w_{ik}}$ and $e_{i\alpha} = e^{v_{i\alpha}}/\Sigma_\beta\, e^{v_{i\beta}}$. This reparametrization has two advantages: (*a*) modification of the w and v parameters automatically preserves the normalization constraints on probability distributions and (*b*) transition and emission probabilities can never reach the absorbing value 0. We then iteratively cycle through the training set and compute, for each sequence, the corresponding most likely path through the model. This can be done efficiently in $O(N^2)$ steps by using a dynamic programming scheme known as the Viterbi algorithm. Being in a state i along a Viterbi path, we update the parameters of the model according to $\Delta w_{ij} = \eta(T_{ij} - t_{ij})$ and $\Delta v_{i\alpha} = \eta(E_{i\alpha} - e_{i\alpha})$, where η is the learning rate. At

Abbreviation: HMM, hidden Markov model.

**Present address: Department of Biological Sciences, University of Nevada, Las Vegas, NV 89154.

††All programs, technical reports, data sets, complete alignments, and other results are available, for noncommercial purposes, from the authors upon request.

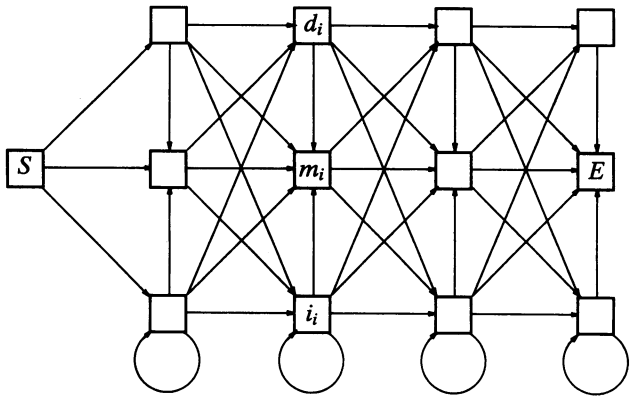

FIG. 1. HMM architecture. S and E are the start and end states. Sequence of main states m_i is the backbone. Side states d_i (resp. i_i) correspond to deletions (resp. insertions).

each step of a Viterbi path, and for any state i on the path, $T_{ij} = 1$ (resp. $E_{i\alpha} = 1$) if the $i \rightarrow j$ transition (resp. emission of α from i) is used and 0 otherwise. Parameters of a loop are updated every time the loop is traversed. These rules are repeated for each example until no significant variations occur. This algorithm approximates a gradient descent procedure on the negative log-likelihood of the sequences (16). As such, it can be expected to converge to a possibly local maximum likelihood estimator.

Once a HMM has been successfully trained, it can be used for a number of tasks. Multiple alignments result immediately from aligning the corresponding optimal paths. Discrimination of whether any given sequence belongs to the family can be based on its likelihood according to the model. Structural properties of the model revealed—for instance, by an entropy plot of the emission distribution along the backbone—are also useful.

Experiments and Results

Experiments have been performed with several protein families, including globins, immunoglobulins, kinases, G-protein-coupled receptors, aspartic proteases, and human immunodeficiency virus membrane proteins. The focus here will be on globins, immunoglobulins, and kinases. In all three experiments, the length of the model is initially set to the average length of the sequences in the family and then kept constant for the entire experiment. The probability parameters of the models are initialized uniformly prior to training. A random subset of sequences is typically used for training with a learning rate $\eta = 0.1$. The remaining sequences are used for validation. After training, all the sequences in the training and validation set are aligned to the model. For brevity, only a portion of the resulting multiple alignments is displayed, by using a subset of phylogenetically representative sequences covering both the training and validation sets.

Globins. The globins form a well-known family of heme-containing proteins that reversibly bind oxygen and are involved in its storage and transport. From crystallographic studies, all globins have similar three-dimensional structure characterized by seven α-helices, labeled A, B, C, E, F'F, G, and H (some structures also have a short D helix) (17). The globin sequences used here were extracted from the nonredundant (NR) data base composed of Protein Identification Resource 34.0, Swiss-Prot 23, and GenPept (translated Gen-

```
                                        1         2         3         4
  .............................AAAAAAAAAAAAAAAAA.BBBBBBBBBBBBBBBBBBBBBBBBBBBBBCCCCCCC..
V HAHU      ------------------VLSPA-DKTNVKAAWGKVGAHAgEYGAE-A---L--E-R--MFL----SFPTTKTYFP
T HAOR      ------------------MLTDA-EKKEVTALWGKAAGHGeEYGAE-A---L--E-R--LFQ----AFPTTKTYFS
V HADK      ------------------VLSAA-DKTNVKGVFSKIGGHAeEYGAE-T---L--E-R--MFI----AYPQTKTYFP
V HBHU      ------------------VHLTP-EEKSAVTALWGKVNVD-EVGGE-A---L--G-R--LLV----VYPWTQRFFE
T HBOR      ------------------VHLSG-GEKSAVTNLWGKVNIN-ELGGE-A---L--G-R--LLV----VYPWTQRFFE
V HBDK      ------------------VHWTA-EEKQLITGLWGKVNVA-DCGAE-A---L--A-R--LLI----VYPWTQRFFA
V MYHU      ------------------GLSDG-EWQLVLNVWGKVEADIpGHGQE-V---L--I-R--LFK----GHPETLEKFD
V MYOR      ------------------GLSDG-EWQLVLKVWGKVEGDLpGHGQE-V---L--I-R--LFK----THPETLEKFD
V IGLOB     mkffavlalcivgaias-PLTA-DEASLVQSSWKAVSHN-EV--E-I---L--A-A--VFA----AYPDIQNKFS
V GPYL      ------------------GVLTDvQVALVKSSFEEFNANIpKN---t-hr--fft-L--VLE----IAPGAKDLF-
V GPUGNI    ------------------ALTEK-QEALLKQSWEVLKQNIpAH---s-l-------R--LFAliieAAPESKYVF-
V GGZLB     ------------------ML-D----QQTINIIKATVPVLkEHGVT-Ittt-f--y-knLFA----KHPEVRPLFD
  ..........................................................*.................*.....*.
  ..........................................................................I....

                5         6         7         8                   9
  ..............DD++++++EEEEEEEEEEEEEEEEEE.................FFFFFFFFFFF...........G
HAHU      HF-DLSHGSA------QVKGHGKKVADALTNAVAHVDD------M-P--NAL---SALSDLHAH--K-L-R--VD
HAOR      HF-DLSHGSA------QIKAHGKKVADALSTAAGHFDD------M-D--SAL---SALSDLHAH--K-L-R--VD
HADK      HF-DLSHGSA-------QIKAHGKKVAAALVEAVNHVDD-------I-A--GAL---SKLSDLHAQ--K-L-R--VD
HBHU      SFGDLSTPDAVM-GNPKVKAHGKKVLGAFSDGLAHLDN------L-K--GTF---ATLSELHCD--K-L-H--VD
HBOR      AFGDLSSAGAVM-GNPKVKAHGAKVLTSFGDALKNLDD------L-K--GTF---AKLSELHCD--K-L-H--VD
HBDK      SFGNLSSPTAIL-GNPMVRAHGKKVLTSFGDAVKNLDN------I-K--NTF---AQLSELHCD--K-L-H--VD
MYHU      KFKHLKSEDEMK-ASEDLKKHGATVLTALGGILKKKGH------H-E--AEI---KPLAQSHAT--K-H-K--IP
MYOR      KFKGLKTEDEMK-ASADLKKHGGTVLTALGNILKKKGQ------H-E--AEL---KPLAQSHAT--K-H-K--IS
IGLOB     QF-AGKDLASIK-DTGAFATHATRIVSFLSEVIALSGNtsnaaaV-N--SLV---SKLGDDH--k--a-r-g-VS
GPYL      SF--LKGSSEVPqNNPDLQAHAGKVFKLTYEAAIQLEVngavas--D--ATL---KSLGSVH--v--s-k-gvVD
GPUGNI    SF--LKDSNEIPeNNPKLKAHAAVIFKTICESATELRQkghav--w-dnNTL---KRLGSIH--lk-n-k-i-TD
GGZLB     MG--RQESLE------QPKALAMTVLAA----AQNIEN------L-P--AILpavKKIAVKH--c--q-a-g-VA
  .............................*.....................................*...*.............
  .............................II.....................................III.............

          0         1         2         3         4
  ..........GGGGGGGGGGGGGGGGG.....HHHHHHHHHHHHHHHHHHHHHHHHHHHHH.....................
HAHU      PVNFKLLSHCLLVTLAAHLP-AEFTPAVHASLDKFLAS-VSTV-------LTSK-----Yr-------
HAOR      PVNFKLLAHCILVVLARHCP-GEFTPSAHAAMDKFLSK-VATV-------LTSK-----Yr-------
HADK      PVNFKFLGHCFLVVVAIHHP-AALTPEVHASLDKFMCA-VGAV-------LTAK-----Yr-------
HBHU      PENFRLLGNVLVCVLAHHFG-KEFTPPVQAAYQKVVAG-VANA-------LAHK-----Yh-------
HBOR      PENFNRLGNVLIVVLARHFS-KDFSPEVQAAWQKLVSG-VAHA-------LGHK-----Yh-------
HBDK      PENFRLLGDILIIVLAAHFT-KDFTPECQAAWQKLVRV-VAHA-------LARK-----Yh-------
MYHU      VKYLEFISECIIQVLQSKHP-GDFGADAQGAMNKALEL-FRKD-------MASN-----Ykelgfqg-
MYOR      IKFLEYISEAIIHVLQSKHS-ADFGADAQAAMGKALEL-FRND-------MAAK-----Ykefgfqg-
IGLOB     AAQFGEFRTALVAYLQANVSwGD---NVAAAWNKALDNtFAIV-------VPRL-------------
GPYL      AH-FPVVKEAILKTIKEVVG-DKWSEELNTAWTIAYDE-LAII-------IKKEmkdaa---------
GPUGNI    PH-FEVMKGALLGTIKEAIK-ENWSDEMGQAWTEAYNQ-LVAT-------IKAE-----Mke------
GGZLB     AAHYPIVGQELLGAIKEVLG-DAATDDILDAWGKAYGV-IADVfiqveadLYAQ-----Ave------
  .....................................................................................
  ..................IV.................V.....................................
```

FIG. 2. Alignment of 12 phylogenetically representative globins to the HMM of length 145 trained on a random subset of 275 globin sequences. HAHU (human), HAOR (duckbill platypus), and HADK (duck) are α-globin chains and HBHU (human), HBOR (duckbill platypus), and HBDK (duck) are β-globin chains, respectively. MYHU (human) and MYOR (duckbill platypus) are myoglobins. The remaining hemoglobin sequences are IGLOB (insect, *Chironomus thummi*), GPYL (legume, yellow lupine), GPUGNI (non-legume, swamp oak), and GGZLB (bacteria, *Vitreoscilla* sp.). T (resp. V) in the first column refers to sequences in the training (resp. validation) set. Approximate positions of the α-helices A–H are indicated on the top of the alignment (plus signs indicate a region of overlap between the D and E helices). Numbers above the alignment indicate every 10 main states. Letters emitted from insert states are lowercase. A dash signals a transition through a delete state or a position where other sequences in the family are using an insert state. Highly conserved residues are marked with stars. Roman numerals indicate the five conserved or semiconserved motifs.

Bank 73). Partial sequences were removed. The set contains 483 sequences, with minimum length 116, average length 145, and maximum length 170.

A random subset of 275 sequences was used to train a model of length 145. Alignment of 12 phylogenetically representative sequences to the model is given in Fig. 2. The percentage of identical residues in this set ranges from 10% to 70%. Correct identification of five motifs that are conserved or semiconserved throughout this phylogenetic distribution is used as an indicator of a good alignment. Motif I is essentially helical region C. Motifs II and III, in helical regions E and F, respectively, are within the heme-binding region. Motifs IV and V are in helical regions G and H, respectively (Fig. 2).

To test the classification abilities of the globin model, we generated 300 random sequences of length 100, 120, 140, 160, 180, and 200 (50 random sequences at each length) with the same amino acid composition as the average computed over the entire data base. The negative logarithms of the likelihoods of the Viterbi paths associated with these random sequences are essentially a linear function of the length (P.B. and Y.C., unpublished work) (Fig. 3). A regression line was computed together with the histograms of the residuals of the globins and of the random sequences.

Immunoglobulins. Immunoglobulins or antibodies are proteins produced by B cells that bind with specificity to foreign antigens in order to neutralize them or target their destruction by other effector cells. The various classes of immunoglobulins are defined by pairs of light and heavy chains that are held together principally by disulfide bonds. Each light or heavy chain contains one variable (V) region and one (light) or several (heavy) constant (C) regions. The V regions provide the specificity of the antigen recognition. Our data base consists of human and mouse immunoglobulin heavy-chain V-region sequences from the Protein Identification Resource data base. It corresponds to 224 sequences, with minimum length 90, average length 117, and maximum length 254. The variation in length results from including any sequence with a V region: those with or without signal or leader sequences, germ-line sequences not rearranged to any joining (J) segment, and some that contain the C region as well.

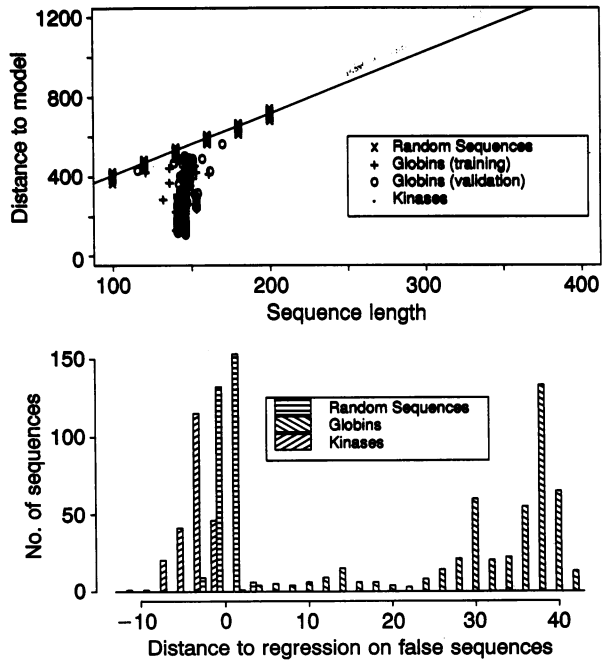

FIG. 3. (*Upper*) Regression line for the negative log-likelihood of the Viterbi paths associated with 300 random sequences of similar amino acid composition to globins. (*Lower*) Histogram of residuals measured in approximate standard deviations of random sequences. Most globins are separated by 5 standard deviations or more. Modes in globin distribution correspond to different subclasses.

A model of length 117 was trained by using a random subset of 150 sequences. A multiple alignment of a representative subset of 20 sequences is given in ref. 8. The algorithm detects all the main regions of highly conserved residues. The cysteine residues toward the beginning and the end of the region, responsible for the disulfide bonds, are perfectly aligned. The only exception (PH0097), which has a serine residue in its terminal portion, is a rare but recognized exception to the conservation of this position. We did not remove the headers (transport signal peptide) attached to some sequences prior to training. The model detects and accommodates these headers by treating them as initial repeated inserts.

Kinases. Eukaryotic protein kinases constitute a large family of proteins that regulate the most basic of cellular processes through phosphorylation, by transferring phosphate groups usually from an ATP or GTP, onto tyrosine, serine, or threonine residues of many different proteins. They have been termed the "transistors" of the cell (18). We have used the sequences available in the kinase data base maintained at the Salk Institute. Our basic set consists of 223 sequences, with minimum length 156, average length 287, and maximum length 569.

We trained a model of length 287 by using a random subset of 150 kinase sequences. Fig. 4 displays the alignment for a subset of 12 phylogenetically representative sequences. The percentage of identical residues within the kinase data sets ranges from 8% to 30%, suggesting that only those residues involved in catalysis are conserved among the most divergent sequences. All the 12 characteristic catalytic domains or subdomains described in refs. 19 and 20 are easily recognizable. We have indicated the unvaried or rarely varied residues used by the authors of refs. 19 and 20 to characterize each domain. For instance, the initial hydrophobic consensus Gly-Xaa-Gly-Xaa-Xaa-Gly, together with the Lys located 15–20 residues downstream, is part of the ATP/GTP binding site. The carboxyl terminus is characterized by the presence of an unvaried Arg residue. Conserved residues in proximity to the acceptor amino acid are found in the VIb (Asp), VII (Asp-Phe-Gly), and VIII (Ala-Pro-Glu) domains. Crystallographic studies of the cAMP-dependent protein kinase confirm that most of the conserved motifs of the protein kinase core are clustered in the regions of the protein involved in nucleotide binding and catalysis (21).

A classification test, similar to the one for the globins, was done by generating 140 random sequences of length 150, 200, 250, 300, 350, 400, and 450 (20 random sequences at each length) with the same average amino acid composition as the kinases. Negative log-likelihoods associated with optimal paths in the kinase HMM model are plotted in Fig. 5, together with the histogram of the residuals to the regression line of the random sequences. The separation achieved by the kinase model exceeds the one achieved by the globin model.

Discussion

The experiments show that the HMM approach can capture the important statistical properties of protein families. While providing global alignments for entire families, the method uses locally adjustable parameters, equivalent to variable gap penalties, in a way that is efficient in terms of both computation and motif detection. At each learning iteration, with K sequences of average length N, the key step is the application of the Viterbi procedure, which requires $O(N^2)$ operations. A small number c, 2–15, of training cycles is usually sufficient to produce stable alignments. Therefore the total number of

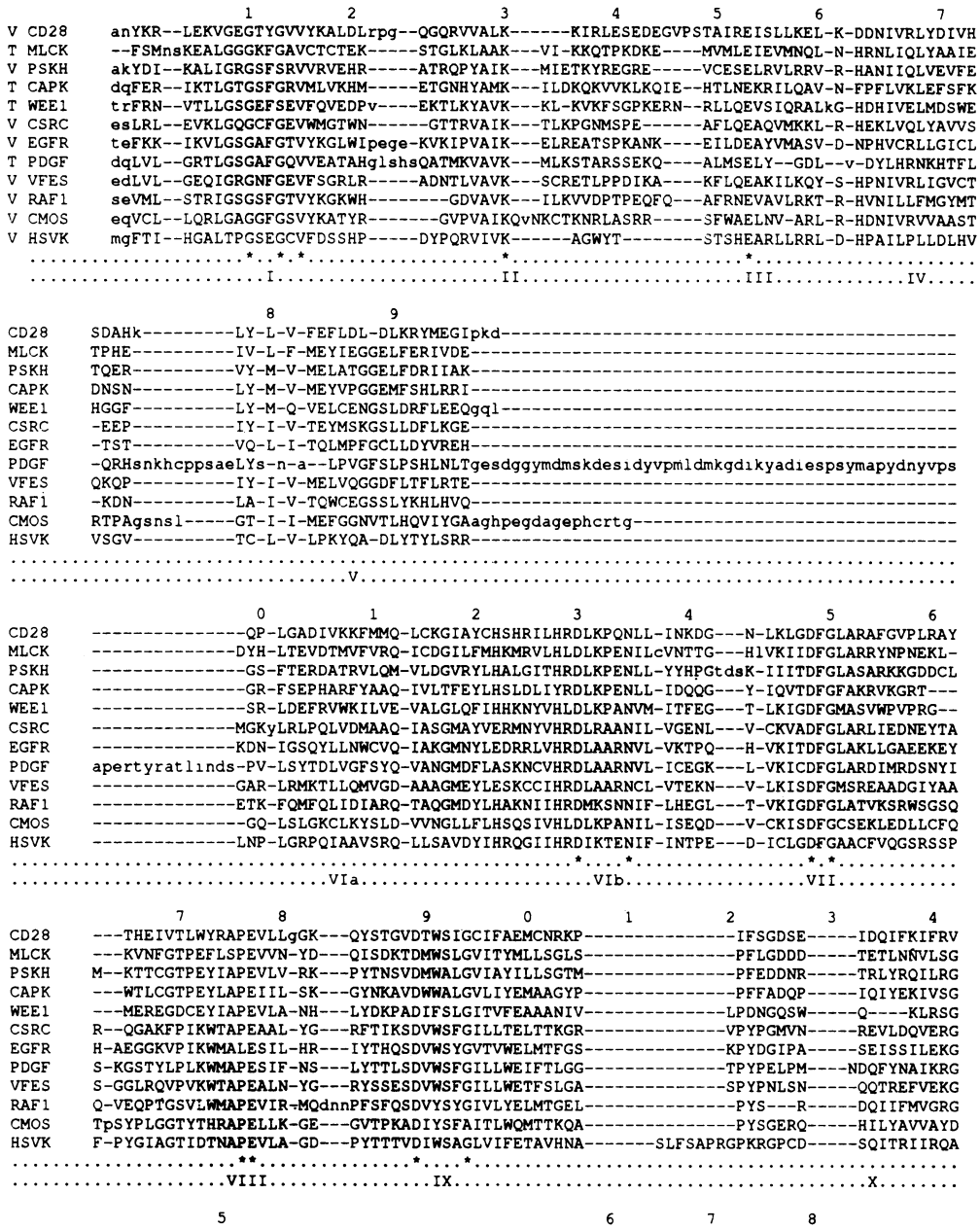

FIG. 4. Alignment of 12 phylogenetically representative kinase sequences to the HMM of length 287 trained on a random subset of 150 kinases. CD28 (CDC28 of *Saccharomyces cerevisiae*), MLCK (myosin light-chain kinase of rat skeletal muscle), PSKH (HeLa cell), CAPK (bovine cardiac muscle), and CMOS and RAF1 (human oncogenic proteins) are serine/threonine-specific kinase proteins. WEE1 is a dual-specificity kinase from *Schizosaccharomyces pombe*. CSRC (chicken oncogenic protein), EGFR (human epidermal growth factor receptor), PDGMR (mouse platelet-derived growth factor receptor), VFES (feline sarcoma virus oncogenic protein), and HSVK (herpes simplex virus kinase) are tyrosine-specific kinase proteins. Roman numerals at the bottom indicate domain designations as described (19). Characteristic invariant or quasi-invariant residues used by those authors are also marked by stars. All other designations are as in Fig. 2.

steps scales like $O(cK\,N^2) = O(K\,N^2)$. Alignment of K sequences to the model takes also $O(K\,N^2)$ steps associated with K applications of dynamic programming. Thus, a solution to the multiple-alignment problem can be derived in time which grows linearly with the number of sequences.

To be successful, the HMM approach requires a representative training set capable of constraining the parameters sufficiently. The architecture used here has $52N + 23 \approx 52N$ transition and emission parameters. The number of effective parameters is smaller but difficult to estimate. In a training

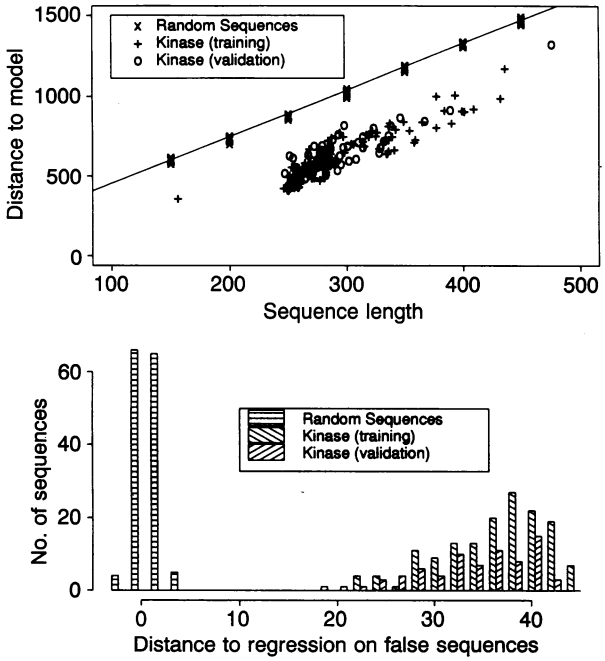

FIG. 5. Similar to Fig. 3 for the kinase model and 140 random sequences of amino acid composition similar to kinases.

sequence of length N, each letter and each transition from one letter to the next provide a constraint on the parameters. Thus, a relatively small number of training sequences, perhaps on the order of 30 or so, already provides a number of constraints comparable to the number of parameters in the model. In a recent experiment on human immunodeficiency virus membrane proteins, good alignments were obtained with 42 sequences only. In our experience, the quality of the alignments is robust with respect to variations in the initial conditions or learning parameters. The smoothness of the learning algorithm introduced seems important for the orderly learning behavior we observe. Discontinuous learning rules, such as the Baum–Welch algorithm, may be more prone to overfitting.

The HMM approach can be extended in several directions. We are applying it to other protein families and to DNA families (26) with more complex versions of the algorithm, as well as to the problem of phylogenetic reconstruction. Although random sequences have been used here for demonstration purposes, it is clear that HMMs can be used for classification experiments over entire data bases. In an experiment reported elsewhere (P.B. and Y.C., unpublished work), a model trained on 142 G-protein-coupled receptors has been used to find all remaining known G-protein-coupled receptors in the Swiss-Prot data base. Additional information can also be incorporated in the method with a Bayesian approach, to speed up learning, improve the models, or compensate for the sparseness of the training data. Prior information may be obtained from previous alignments and from scoring matrices (22). New scoring matrices could also be generated directly from the trained HMMs. Higher-order correlations as well as secondary or tertiary structure information could also be integrated in the models.

One additional useful feature of HMMs, which has been exploited in speech recognition, is that they can be organized in a modular, hierarchical fashion, to recognize speech segments of increasing length and complexity. It remains to be seen whether a similar approach is also useful for biological sequences by building HMMs for motifs, families, and superfamilies (23). It may be of interest, for instance, to train different models for protein-tyrosine kinases and protein-serine/threonine kinases, or with different subclasses of globins or immunoglobulins, and then try to merge them at a higher level. Conversely, a bank of parallel HMMs could also be trained simultaneously, via some form of competitive learning, to progressively segregate different subclasses from a family (14). In this modular context, it has been conjectured (24, 25) that the total number of superfamilies of proteins may be relatively small, perhaps on the order of a thousand. If true, it would then be possible to train one or a small number of HMMs for each superfamily. Training of a typical model takes only a few hours on a workstation, and HMMs naturally lend themselves to parallel implementations. Such a battery of models could have a wide range of applications in biology.

1. Chan, S. C., Wong, A. K. C. & Chiu, D. K. Y. (1992) *Bull. Math. Biol.* **54**, 563–598.
2. McClure, M. A., Vasi, T. K. & Fitch, W. F. (1993) *Mol. Biol. Evol.*, in press.
3. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
4. Maier, D. (1978) *J. Assoc. Comput. Mach.* **25**, 2, 322–336.
5. Vingron, M. & Argos, P. (1991) *J. Mol. Biol.* **218**, 33–43.
6. Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992) *Comput. Appl. Biosci.* **8**, 189–191.
7. Gusfield, D. (1993) *Bull. Math. Biol.* **55**, 141–154.
8. Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1993) in *Advances in Neural Information Processing Systems*, eds., Hanson, S. J., Cowan, J. D. & Lee Giles, C. (Kaufmann, San Mateo, CA), Vol. 5, pp. 747–754.
9. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *J. Royal Stat. Soc. B* **39**, 1–22.
10. Churchill, G. A. (1989) *Bull. Math. Biol.* **51**, 79–94.
11. Lawrence, C. E. & Reilly, A. A. (1990) *Proteins Struct. Funct. Genet.* **7**, 41–51.
12. Thorne, J. L., Kishino, H. & Felsenstein, J. (1991) *J. Mol. Evol.* **33**, 114–124.
13. Cardon, L. R. & Stormo, G. D. (1992) *J. Mol. Biol.* **223**, 159–170.
14. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994) *J. Mol. Biol.* **235**, 1501–1531.
15. Rabiner, L. R. (1989) *Proc. IEEE* **77**, 257–286.
16. Baldi, P. & Chauvin, Y. (1994) *Neural Comput.* **6**, 305–316.
17. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196**, 199–216.
18. Hunter, T. (1987) *Cell* **50**, 823–829.
19. Hanks, S. K. & Quinn, A. M. (1991) *Methods Enzymol.* **200**, 38–62.
20. Lindberg, R. A., Quinn, A. & Hunter, T. (1992) *Trends Biochem. Sci.* **17**, 114–119.
21. Knighton, D. R., Zheng, J., Ten Eyck, L. F., Ashford, V. A., Xuong, N. H., Taylor, S. S. & Sovadski, J. M. (1991) *Science* **253**, 407–414.
22. States, D. J., Gish, W. & Altschul, S. F. (1991) *Methods Companion Methods Enzymol.* **3**, 66–70.
23. Doolittle, R. F. (1981) *Science* **214**, 149–159.
24. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91**, 524–545.
25. Zuckerkandl, E. (1976) *J. Mol. Evol.* **7**, 167–183.
26. Beldi, P., Bzunek, S., Chauvin, Y., Engelbrecht, J. & Krogh, A. (1994) in *Advances in Neurol Information Processing Systems*, eds. Cowan, J. D., Tesouro, G. & Alspector, J. (Kaufmann, San Mateo, CA), Vol. 6.