

Hidden Markov Models of Biological Primary Sequence Information

P Baldi, Y Chauvin, T Hunkapiller, and MA McClure

PNAS 1994;91;1059-1063 doi:10.1073/pnas.91.3.1059

This information is current as of January 2007.

	This article has been cited by other articles: www.pnas.org#otherarticles
E-mail Alerts	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.
Rights & Permissions	To reproduce this article in part (figures, tables) or in entirety, see: www.pnas.org/misc/rightperm.shtml
Reprints	To order reprints, see: www.pnas.org/misc/reprints.shtml

Notes:

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 positiondependent 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 $\Delta w_{ij} = \eta (T_{ij})$ $(-t_{ii})$ 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 hemecontaining 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
		AA	AAAAAAA	AAAAAA . BE	BBBBBBBBBBB	BBBBBBB	BBBBBBBBB	BBCCCCCCC.
V HAHU		VLSPA	-DKTNVK	AWGKVGAHA	GEYGAE-A-	LE-	RMFL	-SFPTTKTYFE
T HAOR		MLTD	-EKKEVTA	LWGKAAGHG	GEYGAE-A-	LE-	RLFQ	-AFPTTKTYFS
V HADK		VLSA	-DKTNVKO	VFSKIGGHA	EYGAE-T-	LE-	RMFI	-AYPOTKTYF
V HBHU		VHLTI	-EEKSAV1	ALWGKVNVI	-EVGGE-A-	LG-	RLLV	-VYPWTORFFI
T HBOR		VHLSO	G-GEKSAVI	NLWGKVNIN	-ELGGE-A-	LG-	RLLV	-VYPWTORFFI
V HBDK		VHWT	-EEKOL IT	GLWGKVNV	-DCGAE-A-	1A-	BLI.T	-VYPWTORFF
V MYHU		GLSD	-EWOLVLA	WWGKVEAD	DGHGOE-V-	LT-	RLFK	-GHPETLEKFI
V MYOR		GLSD	-EWOLVLE	WWGKVEGDI	DGHGOE-V-	TT-	RLFK	-THPETLEKE
V TGLOB	mkffaylalci	voaias-PLT	-DEASLUC	SSWKAVSH	J-EVE-T-	12-	AVFA	-AVPDIONE
V GPVI.		GVI.TI	WOVALVKS	SFEEFNANI	DKNt-h		LVLE	-TAPCAKDLE
V CRICNT		AT.TE	-OFALLK	SWEVLKON	nAHe-l		P	ALADESEVUT.
V GFUGNI		MI -D		JTTKATUDUI	FUCUT-T+	++_f	-koleani	-KAPESKIVF
A GGUTD		PIL-D-		IIIKAIVPVI	SKENGVI-IC	*	-KILLE A	
• • • • • • •	• • • • • • • • • • • • • • •	•••••	•••••	••••••	• • • • • • • • • • •	•••	•••••	••••"••••"
• • • • • • •	••••••	••••	•••••	•••••	• • • • • • • • • • •	• • • • • • • •	•••••	
	F	e	-	•			•	
	5		/	0			9	
		TELEELEELE						· · · · · · · · · · · · · · · · · · ·
HAHU	HF-DLSHGSA	QVKGHGK	VADALIN	AVAHVDD	M-PN	ALSA	LSDLHAH	K-L-RVD
HAOR	HF-DLSHGSA	QIKAHGK	KVADALSTA	AGHPDD	M-DS	ALSA	LSDLHAH	K-L-RVD
HADK	HF-DLSHGSA	QIKAHGK	KVAAALVEA	AVNHVDD	1-AG	ALSK	LSDLHAQ	K-L-RVD
HBHU	SFGDLSTPDAVM	GNPKVKAHGK	KVLGAF SDO	SLAHLDN	L-KG	TFAI	LSELHCD	K-L-HVD
HBOR	AFGDLSSAGAVM-	-GNP KVKAHGA	KVLTSFGD/	ALKNLDD	L-KG	TFAK	LSELHCD	K-L-HVD
HBDK	SFGNLSSPTAIL-	-GNPMVRAHGK	KVLTSFGD/	AVKNLDN	I-KN	TFAC	LSELHCD	K-L-HVD
MYHU	KFKHLKSEDEMK-	-ASEDLKKEGA	TVLTALGG:	ILKKKGH	н-еа	EIKP	LAQSHAT	K-H-KIP
MYOR	KFKGLKTEDEMK-	-ASADLKKHGG	TVLTALGN:	ILKKKGQ	н-еа	ELKP	LAQSHAT	K-H-KIS
IGLOB	QF-AGKDLASIK-	-DTGAFATHAT	RIVSFLSE	VIALSGNts	naaaV-NS	LVSF	LGDDHk-	-a-r-g-VS
GPYL	SFLKGSSEVPC	NNPDLQAHAG	KVFKLTYE	AIQLEVnga	ivasDA	TLKS	LGSVHv-	-s-k-gvVD
GPUGNI	SFLKDSNEIP	ENNP KLKAHAA	VIFKTICE	SATELRQkgl	navw-dnN	TLKR	LGSIH1k	-n-k-i-TD
GGZLB	MGRQESLE	QPKALAM	TVLAA	-AQNIEN	L-PA	ILpavKF	IAVKHc-	-q-a-g-VA
		*					**	
		II.					.III	
	0 1	2	:	3	4			
	. GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	GGGGGG	. нананны	аннананы	внанныннын	BHB		••
HAHU	PVNFKLLSHCLL	/TLAAHLP-AE	FTPAVHAS	LDKFLAS-V	STV	LTSK	Yr	
HAOR	PVNFKLLAHCIL	/VLARHCP-GE	FTPSAHAAI	MDKFLSK-V	ATV	LTSK	Yr	
HADK	PVNFKFLGHCFL	/VVAIHHP-AA	LTPEVHAS	LDKFMCA-V	GAV	LTAK	Yr	
HBHU	PENFRLLGNVLV(CVLAHHFG-KE	FTPPVQAA	YQKVVAGV	ANA	LAHK	Yh	
HBOR	PENFNRLGNVLIV	/VLARHFS-KD	FSPEVQAA	QKLVSG-V	АНА	LGHK	Yh	
HBDK	PENFRLLGDILI	IVLAAHFT-KD	FTPECQAA	QKLVRV-V	АНА	LARK	Yh	
MYHU	VKYLEFISECII(VLQSKHP-GD	FGADAQGAI	MNKALEL-FI	RKD	MASN	Ykelgfq	g-
MYOR	IKFLEYISEAII	IVLQSKHS-AD	FGADAQAAI	GRALEL-FI	RND	MAAK	Ykefgfg	g-
IGLOB	AAQFGEFRTALV	YLOANVSwGD	NVAAA	WNKALDNEF	AIV	VPRL		
GPYL	AH-FPVVKEAILI	TIKEVVG-DK	NSEELNTA	TIAYDE-L	AII	IKKEmkc	laa	
GPUGNI	PH-FEVMKGALL	GTIKEAIK-EN	WSDEMGQA	TEAYNQ-L	VAT	IKAE	Mke	
GGZLB	AAHYPIVGQELLO	GAIKEVLG-DA	ATDDILDA	WGKAYGV-I	ADVfiqvead	LYAQ	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



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.

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.

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 Ksequences 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

W 6039	1 2 3 4 5 6 7
V CD28	anikk-lekveesiigvikaldigg-veevvalakiklesedesystäkeisiikei-k-DDNIVKLIDIV
V DCKU	
T CAPK	drift RAIGROST BAVAVENT
T WEE1	trFRN-VTLLGSGEFSEVFOVEDPvEKTLKYAVKKL-KVKFSGPKERNRLGEVSIOPAILC-HPHIVEIMOSM
V CSRC	esLRLEVKLGGCFGEVWMGTWNGTTRVAIKTIKPGNMSPEAFLGEAOVMKKI-R-HEKIVOLYAV/
V EGFR	teFKKIKVLGSGAFGTVYKGLWIDege-KVKIPVAIKELREATSPKANKEILDEAYVMASV-D-NPHVCRLIGIC
T PDGF	dgLVLGRTLGSGAFGQVVEATAHglsbsOATMKVAVKMLKSTARSSEKOALMSELYGDLV-DYLHRNKHTF
V VFES	edLVLGEQIGRGNFGEVFSGRLRADNTLVAVKSCRETLPPDIKAKFLOEAKILKOY-S-HPNIVRLIGVC
V RAF1	seVMLSTRIGSGSFGTVYKGKWHGDVAVKILKVVDPTPEOFOAFRNEVAVLRKT-R-HVNILLFMGYM
V CMOS	eqVCLLORLGAGGFGSVYKATYRGVPVAIKOvNKCTKNRLASRRSFWAELNV-ARL-R-HDNIVRVVAAS
V HSVK	mgFTIHGALTPGSEGCVFDSSHPDYPQRVIVKAGWYTSTSHEARLLRRL-D-HPAILPLLDLHV
	* * * * * * * * * * * * * * * * * * * *
	II
	8 9
CD28	SDAHkLY-L-V-FEFLDL-DLKRYMEGIpkd
MLCK	TPHEIV-L-F-MEYIEGGELFERIVDE
PSKH	TQERVY-M-V-MELATGGELFDRIIAK
CAPK	DNSNLY-M-V-MEYVPGGEMFSHLRRI
WEE1	HGGFLY-M-Q-VELCENGSLDRFLEEQgql
CSRC	-EEPIY-I-V-TEYMSKGSLLDFLKGE
EGFR	-TSTVQ-L-I-TQLMPFGCLLDYVREH
PDGF	-QRHsnkhcppsaeLYs-n-aLPVGFSLPSHLNLTgesdggymdmskdesidyvpmldmkgdikyadiespsymapydnyvps
VFES	QKQPIY-I-V-MELVQGGDFLTFLRTE
RAF1	-KDNLA-I-V-TQWCEGSSLYKHLHVQ
CMOS	RTPAgsnslGT-I-I-MEFGGNVTLHQVIYGAaghpegdagephcrtgGT-I-I-MEFGGNVTLHQVIYGAaghpegdagephcrtgGT-I-I-MEFGGNVTLHQVIYGAaghpegdagephcrtg
HSVK	VSGVTC-L-V-LPKYQA-DLYTYLSRR
• • • • • • •	······
CD 20	0 1 2 3 4 5 6
CD28	
MLCK	C C C C C C C C C C C C C C C C C C C
CADE	CD ECEDWARKUCH-VLDGVKILHALGIIHKDLKPENLL-IHRGCOSK-IIITDFGLASARKKGDDCL
WEEL	
CEDC	
ECED	
LGFR	
VEES	apertyratiing=PV-LSITDLVGFSiQ-VANGMDFLASKNCVHRDLAARNVL-ICEGKL-VKICDFGLARDIMRDSNYI
VFES	apertyratlinds-PV-DSitDLVGFSiQ-VANGMDFLASKNCVHKULAARNVL-ICEGKL-VKICDFGLARDINKDSNYI GAR-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGMSREAADGIYAA
VFES RAF1	apertyratlinds-PV-LSitDLVGFSiQ-VANGMDFLASKNCVHKULAARNUL-ICEGKL-VKICDFGLARDIARDSNYI GAR-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNUL-VTEKNV-LKISDFGMSREAADGIYAA ETK-FQMFQLIDIARQ-TAQGMDYLHAKNIIHRDMKSNNIF-LHEGLT-VKIGDFGLATVKSRWSGSQ
VFES RAF1 CMOS HSVK	apertyratilnds-PV-DSIDLVGSIDGSIG-VANGMUFLASKNCVHVULARKNVL-ICEUKL-VKICDFGLARDIMKDSNY1 GR-LRMKTULQMYGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGMSREAABGIYAA GR-LSLGKCLKYSLD-VVNGLLFLBSQSIVHLDKFNNIL-ISEQDV-CKISDFGCSEKLEDLLCFQ
VFES RAF1 CMOS HSVK	apertyratilnds-PV-DSIDLVGFSIQ-VANGMDFLASKNCVHRULAARNVL-ICECKL-VKICDFGLARDIMADSNY1 GRA-LRMKTLLQMVGD-AAAGMEYLESKCCIHRULAARNCL-VTEKNV-LKISDFGMSREAADGIYAA ETK-FQMFQLIDIARQ-TAQGMDYLHAKNIHRDMKSNNIF-LHECLT-VKIGDFGLATVKSRWSGSQ GQ-LSLGKCLKYSLD-VVNGLLFLHSQSIVHLDLKPANIL-ISEQDV-CKISDFGCSEKLEDLLCFQ
VFES RAF1 CMOS HSVK	apertyratilnds-PV-DSITDLVGFSIQ-VANGMDFLASKNCVHKULAARNVL-ICECKL-VKICDFGLARDIMKDSNYI GAR-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGMSREAADGIYAA BTK-FCMFQLIDIARQ-TAQGMDYLHAKNIHRDMKSNNIF-LHEGLT-VKIGDFGLATVKSRWSGSQ GQ-LSLGKCLKYSLD-VVNGLLFLHSQSIVHLDLKPANIL-ISEQDV-CKISDFGCSEKLEDLLCFQ LNP-LGRPQIAAVSRQ-LLSAVDYIHRQGIHRDIKTENIF-INTPED-ICLGDFGAACFVQGSRSSP
VFES RAF1 CMOS HSVK	apertyratlinds-PV-DSITDLVGFSIQ-VANGMDFLASKNCVHVDLAARNVL-ICECKL-VKICDFGLARDINKDSMYI GAR-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGMSREAADGIYAA BTK-FQMFQLIDIARQ-TAQGMDYLHAKNIHRDMKSNNIF-LHECLT-VKIGDFGLATVKSRWSGSQ GQ-LSLGKCLKYSLD-VVNGLLFLHSQSIVHLDLKPANIL-ISEQDV-CKISDFGCSEKLEDLLCFQ LNP-LGRPQIAAVSRQ-LLSAVDYIHRQGIIHRDIKTENIF-INTPED-ICLGDFGAACFVQGSRSSP ****
VFES RAF1 CMOS HSVK	apertyratlinds-PV-DSTDLVGFSTQ-VANGMDFLASKNCVHVDLARKNVL-ICEKKL-VKICDFGLARDIMKDSNTI GR-LENKTLLQMYGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGGSEAAGFIYAA GR-LSLGKCLKYSLD-VVNGLLFLESKCIHRDLARNNL-ISEQDV-CKISDFGCSEKLEDLCFQ
VFES RAF1 CMOS HSVK 	apertyratlinds-PV-DSTDLVGFSTQ-VANGMDFLASKNCVHVDLARKNVL-ICEKKL-VKICDFGLARDIMKDSMY1 GR-LFNKTKLLQMYCD-AAAGMEYLESKCCIHRDLAARNCL-VEKNV-VKISDFGGLARDIMKDSMY1
VFES RAF1 CMOS HSVK CD28 MLCK	apertyratlinds=PV=DsiftLvGFSiQ-VANGMDFLASKNVHRULARKNVL=ICECKL-VKICDFGLARDIMADSMY1 GAR-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGMSREAADGIYAA
CD28 MLCK PSKH	apertyratilnds-PV-DSTIDLVGS SIQ-VANGMUF LASKNOVHNULARKNVL-ICEGKL-VKICDFGLARDIMNDSMY1 CAR-LRMKTLLQMYGD-AAAGMEYLLSSKCIHRDLAARNCL-VTEKNV-LKISDFGGSEALEDILGFQ ETK-FQMFQLIDIARQ-TAQGMDYLHAKNIIHRDMKSNNIF-LHEGLT-VKIGDFGLATVKSRWSGSQ ETK-FQMFQLIDIARQ-TAQGMDYLHAKNIIHRDMKSNNIF-LHEGLT-VKIGDFGLATVKSRWSGSQ
CD28 MLCK PSKH CAPK	apertyratilnds-PV-DSTIDLVGSTQ-VANGMUFLASKNCVHRULARNVL-ICEGKL-VKICDFGLARDIMADSMY1 GR-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLARNVL-ICEGKL-VKIGDFGLATVKSRWSGSQ
CD28 MLCK PSKH CAPK WEE1	apertyratilnds-PV-DSTIDLVGS STQ-VANGMUF LASKNCVHVDLAARNUL-ICECKL-VKICDF GLARDIMKDSNY1 GAR-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLAARNCL-VTEKNV-LKISDFGGSEAADGIYAA
CD28 MLCK PSKH CAPK WEE1 CSRC	apertyratilnds-PV-DSTIDLVGFSTQ-VANGMDFLASKNCVHRULAARNVL-ICECKL-VKICDFGLARDIMADSMY1 GRA-LRMKTLLQMVGD-AAAGMEYLESKCCIHRDLAARNVL-ICECKU-VKIGDFGLATVKSRMSGSQ
CD28 MLCK PSKH CAPK WEE1 CSRC EGFR	apertyratilids-PV-DSTIDLYGSTQ-VANGMUFLASKNCVHRULAARNUL-ICEGKL-VKICDFGLARDIMKDSMTI CRA-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLAARNUL-VTEKNV-LKISDFGGSREAAAGGIYAA GQ-LSLGKCLKYSLD-VVNGLLFLHSSKCIHRDLARNIL-ISEQDV-CKISDFGCSEKLEDLCFQ LNP-LGRPQIAAVSRQ-LLSAVDYIHRQGIIHRDIKTENIF-INTPED-ICLGDFGAACFVQGSRSSP
CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF	apertyratilnds-PV-DSTIDLYGSTQ-VANGMUFLASKNCVHRULARKNVL-ICEGKL-VKICDFGLARDIMADSMY1 GR-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLARKNVL-ICEGKVKIGDFGLATVKSRWSGSQ GR-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLARKNVL-ICEGKVKIGDFGLATVKSRWSGSQ
CD28 MLCK PSK CD28 MLCK PSKH CSRC EGFR WEE1 CSRC EGFR VFES	apertyratilnds-PV-DSTIDLVGFSTQ-VANGMUFLASKNCVHVDLAARNUL-ICEDKL-VKICDFGLARDIMKDSMY1 GR-LRMKTTLQMYGD-AAAGMEYLBSKCCIHRDLAARNUL-ICEDKL-VKIGDFGLATVKSREMADGIYAA
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK VEE1 CSRC EGFR PDGF PDGF RAF1	apertyratilids-PV-DSTIDLVGFSTQ-VANGMUFLASKNCVHRULAARNUL-ICEUKL-VKICDFGLARDIMADSMTI CRA-LRMKTILQMYGD-AAAGMEYLBSKCCIHRULAARNUL-ICEUKL-VKIGDFGLATVKSRWSGSQ
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC CSRC EGFR PDGF RAF1 CMOS	apertyratilids-PV-DSTIDLVGSTQ-VANGMUFLASKNCVHRULAARNUL-ICEUKL-VKICDFGLARDIMADSMYI GR-LRMKTLLQMYGD-AAAGMEYLESKCCIHRDLAARNUL-ICEUKL-VKIGDFGLATVKSRWSGSQ GQ-LSLGKCLKYSLD-VVNGLLFLHSQSIVHLDLKPANIL-ISEQDV-CKISDFGCSEKLEDLLCFQ
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CSRC CAPK WEE1 CSRC PDGF VFES RAF1 CMOS HSVK	apertyratilnds-PV-DSTIDLVGSTQ-VANGMDFLASKNVHRULARKNVL-ICEGKL-VKICDFGLARDIMADSMY1 GR-LRMKTLLQMYGD-AAAGMEYLBSKCCIHRDLARRNCL-VTEKNV-LKISDFGGSEKLEDLCFQ
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF RAF1 CMOS HSVK	apertyratlinds-PV-DSTIDLVGSTQ-VANGMDFLASKNCVHRULAARNVL-ICEUKL-VKICDFGLARDIMKDSMYI GR-LRMKTLLQMYGD-AAAGMEYLBSKCCIHRDLAARNCL-VTEKNV-LKISDFGGSEALADGIYAA GQ-LSLGKCLKYSLD-VVNGLLFLHSQSIVHLDLKPANIL-ISEQDV-CKISDFGCSEKLEDLLCFQ
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES RAF1 CMOS HSVK	apertyratilids-PV-DSTIDLVGSTQ-VANGMUFLASKNCVHRULAARNUL-ICEUKL-VKICDFGLARDIMKDSMTI CRA-LRMKTLQMYGD-AAAGMEYLBSKCCIHRDLAARNUL-VTEKNV-LKISDFGGSREAAAGGIYAA
CD28 RAF1 CMOS HSVK CD28 MLCK CAPK WEE1 CSRC CSRC CSRC CSRC CSRC CSRC CSRC CSR	apertyratilnds-PV-DSTIDLVGSTQ-VANGMDFLASKNCVHRULAARNUL-ICEGKL-VKICDFGLARDIMKDSMTI GR-LRMKTLLQMYGD-AAAGMEYLBSKCCIHRDLAARNCL-VTEKNV-LKISDFGGSEKLEDLCFQ
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CSRC CAPK WEE1 CSRC PDGF VFES RAF1 CMOS HSVK 	apert yrat 11 hds - PV - DS TIDLVGS TQ - VANGHDF LASKNCVHRULAARNUL - ICEUK L-VKICDF GLARDIHKDSNYT
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF PDGF RAF1 CMOS HSVK CSRC CSRC CSRC EGFR PDGF RAF1 CMOS HSVK	apert yrat 11hds - PV - DS TIDLVGF STQ-VANGMDF LASKNCVHRULAARNUL - ICEUK L-VKICDF GLARDIMKDSNY1 GA-LSMKTLLQMYGD-AAAGMEY LESKCCIHRDLAARNUL - ICEUK
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES RAF1 CMOS HSVK CD28 MLCK CSRC EGFR PDGF PDGF RAF1 CMOS HSVK	apertyrat 11nds-PV-DSTIDLVGSTQ-VANGMUF LASKNOVHRULAARNUL-ICEUKL-VKICDF GLARDIMKDSNY1 GR-LEMKTLQMYGD-AAAGMEYLBSSKCIHRDLAARNUL-VTEKNV-LKISDFGGSEKLEDLLCFQ
CD28 PSCH CD28 CD28 CD28 CD28 MLCK PSCH CSRC CSRC CSRC CSRC CMOS HSVK CMOS HSVK CD28 MLCK PSCH CSRC CSC CSC CSC CSC CSC CSC CSC CSC CS	apert yrat 11hds-PV-DSTIDLVGSTQ-VANGMDFLASKNCVHRULAARNUL-ICEGKL-VKICDFGLARDIMKDSMY1 GR-LRMKTULQMYGD-AAAGMEYLESKCCIHRDLAARNUL-ICEGKVKIGDFGLATVKSRWSGSQ
CD28 MLCK PSKH CSD28 MLCK PSKH CSRC CAPK WEE1 CSRC CAPK WEE1 CSRC CMOS HSVK CD28 MLCK PSKH CAPK CAPK CAPK MLCK PSKH CAPK	apert yrat 11hds - PV - DS TIDLVGF STQ-VANGMDF LASKNOVHRULAARNOL - ICEDK L-VKICDF GLARDIMADSNY1 GR-LRMKTLLQMYGD-AAAGMEY LESKCCI HRDLAARNOL - VKISDFGGSEKLEDLCFQ
CD28 MLCK PSKH CAPK WEEL CSRC EGFR PDGF CSRC CSRC CSRC CSRC CSRC CSRC CSRC CSR	apertyratilids-PV-DSTIDLVGFSTQ-VANGMUFLASKNCVHRULAARNUL-ICEGKL-VKICDFGLARDIMKDSNTI CRA-LRMKTILQMYGD-AAAGMEYLBSKCCIHRDLAARNUL-VTEKNV-LKISDFGGSEKLEDLLCFQ
CD28 CD28 MLCK PSKH CAPK WEE1 CSRC	apertyratiinds-PV-LSTIDLOGPSTO-VANGMOPLASKMCVHKULAARNUL-ICEGKL-VKIGDFGLARUIMKDSNTI
CD28 MLCK PSKH CAPK CSRC CAPK WEE1 CSRC CMOS HSVK CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES CMOS HSVK CD28 MLCK PSKH CAPK CAPK CAPK	apertyratiinds-PV-LSTIDLOGFSUC-VANGMOPLASKMCVHKULAARNUL-ICEGKL-VKIGDFGLAKUIMKUSNTI
CD28 RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF CSRC EGFR HSVK CC28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF PDGF PDGF VFFS	apertyratiinds-rv-LstribuvGrstQ-VANGMD LASKNC-HKDLAARNCL-UTEKNV-KLIDFGLARDITMADSNI CRA-LRMKTLQMVGD-AAAGMEYLESKCCIHADLAARNCL-UTEKNV-KLISDFGNSERADGIYAA
CD28 MLCK PSKH CAPK WEEL CSRC EGFR PDGF VFES RAF1 CMOS MLCK PSKH CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK WEEL CSRC CAPK CAPK CAPK CAPK CAPK CSRC CSRC CSRC CSRC CSRC CSRC CSRC CSR	apertyratiinds-rv-Lstruckyrg-Janadmsyleskkovikulaaknvcl-UtekkVkiEDFGJARDITMKDSNII
CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES RAF1 CMOS HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES RAF1 CSRC EGFR PDGF VFES RAF1	apertyratiinds-rv-Lstribuverstu-Vangmbrists
CD28 HSVK CD28 HSVK CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF RAF1 CMOS RAF1 CMOS CSRC CSRC CSRC CSRC CSRC CSRC CSRC CS	apertyratiinds=Pv=LsindvGashadbyLeskw.VikbUAARWC=rUCBKV-UKLSDFGMSREAADGIYAA
CD28 MLCK PSKH CAPK WEE1 CSRC EGFR PDGF VFES RAF1 CMOS HSVK WEE1 CSRC CAPK WEE1 CSRC CAPK WEE1 CSRC CSRC CAPK WEE1 CSRC CSRC CAPK WEE1 CSRC CSRC CAPK WEE1 CSRC CSRC CSRC CSRC CAPK WEE1 CSRC CSRC CSRC CSRC CSRC CAPK WEE1 CSRC CSRC CSRC CAPK WEE1 CSRC CSRC CAPK CAPK CAPK CAPK CSRC CSRC CSRC CSRC CSRC CSRC CSRC CSR	apercyratings-Pv-Lstibuvgstouvakamer Laskm.VinkDLAAKWU-ruckGkrL-VICUP GLAANDIMRUSNIT

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 proteinserine/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.

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Bashford, D., Chothia, C. & Lesk, A. M. (1987) J. Mol. Biol. 196, 199-216.
- 18. Hunter, T. (1987) Cell 50, 823-829.
- Hanks, S. K. & Quinn, A. M. (1991) Methods Enzymol. 200, 38-62.
- Lindberg, R. A., Quinn, A. & Hunter, T. (1992) Trends Biochem. Sci. 17, 114–119.
- 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.
- 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.
- 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.