Hidden Markov Models in Computational Biology

Applications to Protein Modeling

Anders Krogh[1]†, Michael Brown[1], I. Saira Mian[2]
Kimmen Sjölander[1] and David Haussler[1]‡

[1]Computer and Information Sciences
[2]Sinsheimer Laboratories
University of California, Santa Cruz,
CA 95064, U.S.A.

Hidden Markov Models (HMMs) are applied to the problems of statistical modeling, database searching and multiple sequence alignment of protein families and protein domains. These methods are demonstrated on the globin family, the protein kinase catalytic domain, and the EF-hand calcium binding motif. In each case the parameters of an HMM are estimated from a training set of unaligned sequences. After the HMM is built, it is used to obtain a multiple alignment of all the training sequences. It is also used to search the SWISS-PROT 22 database for other sequences that are members of the given protein family, or contain the given domain. The HMM produces multiple alignments of good quality that agree closely with the alignments produced by programs that incorporate three-dimensional structural information. When employed in discrimination tests (by examining how closely the sequences in a database fit the globin, kinase and EF-hand HMMs), the HMM is able to distinguish members of these families from non-members with a high degree of accuracy. Both the HMM and PROFILESEARCH (a technique used to search for relationships between a protein sequence and multiply aligned sequences) perform better in these tests than PROSITE (a dictionary of sites and patterns in proteins). The HMM appears to have a slight advantage over PROFILESEARCH in terms of lower rates of false negatives and false positives, even though the HMM is trained using only unaligned sequences, whereas PROFILESEARCH requires aligned training sequences. Our results suggest the presence of an EF-hand calcium binding motif in a highly conserved and evolutionary preserved putative intracellular region of 155 residues in the α-1 subunit of L-type calcium channels which play an important role in excitation-contraction coupling. This region has been suggested to contain the functional domains that are typical or essential for all L-type calcium channels regardless of whether they couple to ryanodine receptors, conduct ions or both.

Keywords: hidden Markov models; multiple sequence alignments; globin; kinase; EF-hand

1. Introduction

The rate of generation of sequence data in recent years provides abundant opportunities for the development of new approaches to problems in computational biology. In this paper, we apply hidden Markov models (HMMs§) to the problems of statistical modeling, database searching, and multiple alignment of protein families and protein domains. To demonstrate the method, we examine three protein families. Each family consists of a set of proteins that have the same overall three-dimensional structure but widely divergent sequences. Features of the sequences that are determinants of folding, structure and function should be present as conserved elements in the family of sequences. We consider the globins, whole proteins ranging in length from 130 to 170 residues (with few exceptions) and two domains, the protein kinase catalytic domain (250 to 300 residues) and the EF-hand calcium-binding motif (29 residues). The same

† Present address: Electronics Institute, Build 349, Technical University of Denmark, 2800 Lyngby, Denmark.

‡ Author to whom all correspondence should be addressed.

§ Abbreviations used: HMM, hidden Markov models; EM, Expectation-Maximization; ML, maximum likelihood; MAP, maximum *a posteriori*; NLL-score, negative log likelihood score.

0022-2836/94/051501–31 $08.00/0

approach can be used to model families of nucleic acid sequences as well (Krogh *et al.*, 1993*b*).

A hidden Markov model (Rabiner, 1989) describes a series of observations by a "hidden" stochastic process, a Markov process. In speech recognition, where HMMs have been used extensively, the observations are sounds forming a word, and a model is one that by its hidden random process generates these sounds with high probability. Every possible sound sequence can be generated by the model with some probability. Thus, the model defines a probability distribution over possible sound sequences. A good word model would assign high probability to all sound sequences that are likely utterances of the word it models, and low probability to any other sequence. In this paper we propose an HMM similar to the ones used in speech recognition to model protein families such as globins and kinases. In speech recognition, the "alphabet" from which words are constructed could be the set of phonemes valid for a particular language: in protein modeling, the alphabet we use is the 20 amino acids from which protein molecules are constructed. Where the observations in speech recognition are words, or strings of phonemes, in protein modeling the observations are strings of amino acids forming the primary sequence of a protein. A model for a set of proteins is one that assigns high probability to the sequences in that particular set.

The HMM we build identifies a set of positions that describe the (more or less) conserved first-order structure in the sequences from a given family of proteins. In biological terms, this corresponds to identifying the core elements of homologous molecules. The model provides additional information, such as the probability of initiating an insertion at any position in the model and the probability of extending it. The structure of the model is similar to that of a profile (Waterman & Perlwitz, 1986: Barton & Sternberg, 1990: Gribskov *et al.*, 1990: Bowie *et al.*, 1991: Lüthy *et al.*, 1991), but slightly more general. Once we have built the model from unaligned sequences, we can generate a multiple alignment of the sequences using a dynamic programming method. By employing it for database searching, the model can be used to discriminate sequences that belong to a given family from non-members. Finally, we can study the model we have found directly, and see what it reveals about the common structure underlying the various sequences in the family.

Our method of multiple alignment differs quite markedly from conventional techniques, which are usually based on pairwise alignments generated by dynamic programming schemes (Waterman, 1989: Feng & Doolittle, 1987; Barton, 1990: Subbiah & Harrison, 1989). The alignments produced by these methods often depend strongly on the particular values of the parameters required by the method, in particular the gap penalties (Vingron & Argos, 1991). Furthermore, a given set of sequences is likely to possess both fairly conserved regions and highly variable regions, yet conventional global methods assign identical penalties for all regions of the sequences. Substitutions, insertions, or deletions in a region of high conservation should ideally be penalized more than in a variable region, and some kinds of substitutions should be penalized differently in one position compared to another. That is one of the motivations for the present work. The statistical model we propose corresponds to multiple alignment with variable, position-dependent gap penalties. Furthermore, these penalties are in large part learned from the data itself. Essentially, we build a statistical model during the process of multiple alignment, rather than leaving this as a separate task to be done after the alignment is completed. We believe the model should guide the alignment as much as the alignment determines the model.

We are not the first group to employ hidden Markov models in computational biology. Lander & Green (1987) used hidden Markov models in the construction of genetic linkage maps. Other work employed HMMs to distinguish coding from non-coding regions in DNA (Churchill, 1989). Later, simple HMMs were used in conjunction with the EM algorithm to model certain protein-binding sites in DNA (Lawrence & Reilly, 1990; Cardon & Stormo, 1992) and, more recently, to model the N-caps and C-caps of alpha helices in proteins (D. Morris, unpublished results). These applications of HMMs and the EM (Expectation-Maximization) algorithm, including our own, presage a more widespread use of this technique in computational biology. During the time that we have been developing this approach, several related efforts have come to our attention. One is that of White, Stultz and Smith (White *et al.*, 1991: Stultz *et al.*, 1993), who use HMMs to model protein superfamilies. This work is more ambitious than our own, since superfamilies are harder to characterize than families. It is not yet clear how successful their work has been since no results are reported for sequences not in the training set. If there are weaknesses in their method, it is possible that these are due to the use of handcrafted models and reliance on prealigned data for parameter estimation. In contrast, our models have a simple regular structure, and we are able to estimate all the parameters of these models, including the size of the model directly from unaligned training sequences. Interestingly enough, they independently propose an alternate HMM state structure similar to ours† in section 6.3 of their paper (White *et al.*, 1991), where they discuss the relationship of their work to Bowie and co-workers (Bowie *et al.*, 1991), but they do not pursue this further. It is possible that the type of models we use may work better for characterizing superfamilies than those investigated by White *et al.* However, it is more likely that they are too simple, and that richer and more varied state

† Instead of using delete states, they have direct transitions between each pair of match states m_i and m_j with $i < j$.

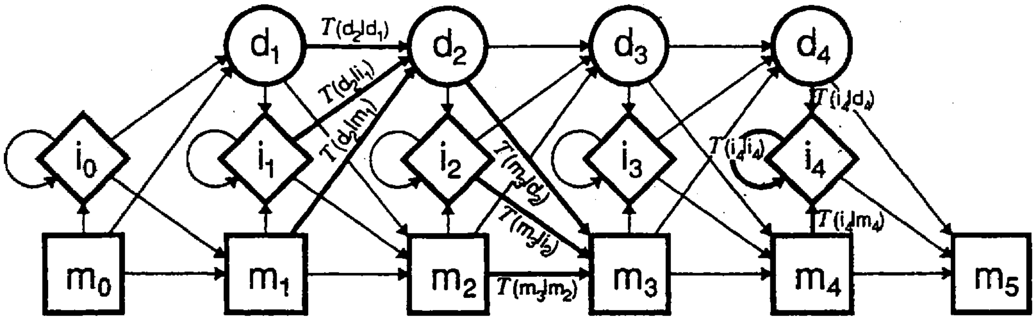


Figure 1. The model.

structure along the lines they propose is required for this problem. We recently found that Asai *et al.* (1993) have applied HMMs to the problem of predicting the secondary structure of proteins.

in terms of the probability it assigns to each protein sequence, we find that it is easier to first think of an HMM as a structure that generates protein sequences by a random process. This structure and corresponding

[redacted]

according to the probabilities $\mathcal{T}(\mathrm{m}_1|\mathrm{m}_0)$, $\mathcal{T}(\mathrm{d}_1|\mathrm{m}_0)$, and $\mathcal{T}(\mathrm{i}_0|\mathrm{m}_0)$. If m_1 is chosen, generate the first amino acid x_1 from the probability distribution $\mathcal{P}(x|\mathrm{m}_1)$, and choose a transition to the next state according to probabilities $\mathcal{T}(\cdot|\mathrm{m}_1)$, where $\cdot$ indicates any possible next state. If this

principle find the model that best describes a given set of sequences.

Given a set of training sequences $s(1), \ldots, s(n)$, one can see how well a model fits them by calculating the probability that it generates them. This probability is simply a probability estimate $\hat{\mathscr{T}}(r|q)$ is obtained by counting the number of times a transition is made from state q to r, for all paths of all training sequences, weighted by the probability of the path. The estimate $\hat{\mathscr{P}}(x|q)$ is made in a similar manner, by counting the number of times the amino acid x is aligned to the state q.

(3) In the next step of ML estimation, a new current model is created by simply replacing $\mathscr{T}(r|q)$ by $\hat{\mathscr{T}}(r|q)$ and $\mathscr{P}(x|q)$ by $\hat{\mathscr{P}}(x|q)$ for each x, q and r. In MAP EM

mizing the probability of the path and then taking the negative logarithm, it is convenient (and equivalent) to simply minimize the negative logarithm of the probability over all paths. This minimum we will call the distance from the sequence to the model,

$$\text{dist}(s, \text{model}) = \min_{paths} \{-\log \text{Prob}(s, \text{path}|\text{model})\}$$

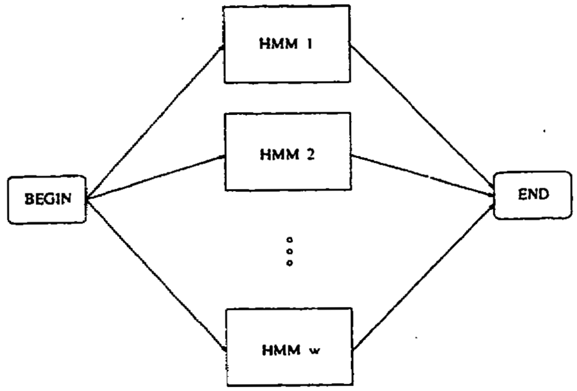


Figure 2. HMM architecture for discovering subfamilies.

components of the (composite) HMM. Presently, the number w of clusters and the initial lengths of the models for these clusters are determined empirically. We then add a new begin state with w outgoing transitions, one to each of the begin states of the component HMMs (see Fig. 2).

This new begin state is analogous to the other begin states in that it generates no amino acid. We then train this composite model with the EM algorithm as described in section (b), above. The EM re-estimation of a component model is the same as the re-estimation of a single model, except that the weight that a sequence has in the re-estimation of a component is proportional to the probability of the sequence given that component model. Thus, sequences that have better NLL-scores for a particular HMM component have greater influence in re-estimating the parameters of that component, and this causes the parameters of that component to change in such a

known subfamilies of the sequences. Experiments with the clustering of globin sequences are described in Results section (a).

(e) *Modeling protein domains with an HMM*

There are many cases when one does not want to build a statistical model of a family of whole proteins like globins, but instead to build a model of a structural motif or domain that occurs as a subsequence in many different kinds of proteins, such as the EF-hand motif (Nakayama *et al.*, 1992) or the kinase catalytic domain (Hanks & Quinn, 1991). Here we expect our model to only match a relatively small subsequence of any given protein, with many other unmatched amino acids appearing before and after this subsequence. One approach to this problem is to alter the dynamic programming method used to align a sequence to a model so that it tries all possible ways of aligning each subsequence of the sequence to a model (Waterman, 1989). We use a simpler (but almost equivalent) method in which only the HMM model is altered, so that the same standard procedures (forward-backward and Viterbi) which we use for models of whole proteins can be used without modification for models of domains.

Consider a training set of many unaligned sequences consisting not of complete proteins, but of a specific domain. Our first step is to train an HMM for these sequences exactly as described earlier. As shown in Figure 1, this HMM will have initial and final "dummy" match states m_0 and m_{N-1} (where $N+1=5$ in Fig. 1) that do not match any amino acid. To alter the HMM to represent a protein domain, we create 2 new insert states i_B and i_E, adding i_B to the model before the state m_0 and i_E at the end of the model after m_{N-1} (see Fig. 3).

We then add a new dummy BEGIN state before i_B and a new dummy END state after i_E. Eight new transitions are also added to the model. The first 4 are from BEGIN to i_B, from m_{N-1} to i_E, and the self-loops from i_B to itself comparing the distances or NLL-scores of 2 sequences with respect to the model. In addition to this penalty, all

Most proteins tend to lie on a fairly straight line (towards the top of the plot) indicating that the

(g) *Initial model, local minima, and choice of model length*

As mentioned in section (b), above, when estimating the model from the training sequences, the EM algorithm does not guarantee convergence to the best model. It is basically a steepest-descent-type algorithm that climbs the nearest peak (local maximum) of the likelihood function (or the posterior probability in MAP estimation). Since finding the globally optimal model seems to be a difficult optimization problem in general (Abe & Warmuth, 1990), we have experimented with various heuristic methods to improve the performance of the method.

Probably the best method is to give the model a hint if something is already known about the sequences, which is often the case. A good starting point makes it much more likely that the nearest peak is at least close to optimal. This is done by setting the probabilities in the initial model to values reflecting that knowledge. If, for instance, an alignment of some of the sequences is available, it is straightforward to translate that into a model by simply calculating the relative frequency of the amino acids and the transition frequencies in each position, as in the profile method (Gribskov *et al.*, 1990).

It is of course even more interesting if the model can be found from a *tabula rasa*, i.e. using no knowledge about the sequences. For that we have used an initial model where all equivalent probabilities are the same, i.e. $\mathscr{T}(m_{k+1}|m_k)$ is independent of the position k in the model, and similarly for all other transition probabilities, and $\mathscr{P}(x|m_k)$ is also independent of k. To avoid the smaller local maxima, noise is added to the model during the iteration before each re-estimation. Initially quite a lot of noise is added, but over 10 iterations the noise is decreased linearly to zero. Since noise is added directly to the model, it is not like the usual implementation of simulated annealing, but the principle is the same. The "annealing schedule" is presently rather arbitrary, but it does seem to give reasonable results† if it is applied several times, and the best of the models found is used as the final model.

It is important that the best model be selected, since suboptimal models do produce inferior alignments in general. However, when studying alignments from suboptimal globin models, we noted that they tend to align some regions well, occasionally getting better alignments in those regions than the best overall model found, while in other regions they are completely incorrect. This leaves open the intriguing possibility of combining the best solutions found for different regions into a new overall best model. We have not yet explored this possibility.

The length of the model is also a crucial parameter that needs to be chosen *a priori*. However, we have developed a simple heuristic that selects a good model length, and even helps in the problem of local maxima. The heuristic is this: after learning, if more than a fraction‡ γ_{del} of the paths of the sequences choose d_k, the delete state at position k, that position is removed from the model. Similarly, if more than a fraction γ_{ins} make insertions at position k (in state i_k), a number of new positions equal to the average number of insertions made at that position are inserted into the model after position k. After these changes in the model, it is retrained, and this cycle is repeated until no more changes are needed. We call this "model surgery".

(h) *Over-fitting and MAP estimation*

A model with too many free parameters cannot be estimated well from a relatively small data set of training sequences. If we try to estimate such a model, we run into the problem of overfitting, in which the model fits the training sequences very well, but gives a poor fit to related (test) sequences that were not included in the training set. We say that the model does not "generalize" well to test sequences. This phenomenon has been well documented in statistics and machine learning (see e.g. Geman *et al.*, 1992; Berger, 1985). One way to deal with this problem is to control the effective number of free parameters in the model by using prior information. This can be accomplished with MAP estimation. Parameters that we assume (*via* our prior distribution on models) can be well-estimated *a priori* in effect become less adaptive, because it takes a lot of data to override our prior beliefs about them, whereas those about which we have only weak prior knowledge are estimated in almost the same manner as in maximum likelihood estimation. In this way, the model can have a very large number of parameters, but a much smaller number of "effectively free" parameters. To make MAP estimation practical, we use Dirichlet distributions as priors. The details of the method are described elsewhere (Krogh *et al.*, 1993*a*; Brown *et al.*, 1993).

3. Results

(a) *Globin experiments*

The modeling was first tested on the globins, a large family of heme-containing proteins involved in the storage and transport of oxygen that have different oligomeric states and overall architecture (for a review see Dickerson & Geis (1983)). Hemoglobins are tetramers composed of two α chains and two other subunits (usually β, γ, δ or θ). Myoglobin is a single chain, some insect globins are present as dimers and some intracellular invertebrate globins occur in large complexes of many subunits.

Globin sequences were extracted from the SWISS-PROT database (release 19) by searching for the keyword "globin". Eliminating the false positives, resulted in 625 genuine globin sequences of average length 145 amino acids. We left three non-globins in the sample for illustrational purposes giving a total of 628 sequences. The sample of globins in the database is not the random sample a statistician would prefer, but is perhaps one of the best and largest collections of protein sequences from a homologous family. Searching for the words "alpha", "beta", "gamma", "delta", "theta", and "myoglobin" in the data file yielded 224 alpha, 199 beta, 16 gamma, 8 delta and 5 theta chains and 79 myoglobins, which adds up to 531 sequences. These should naturally be considered minimum numbers, but they give a good picture of how skewed the sample is.

To test our method, we trained an HMM using the method described in Methods sections (b) and

† An alternate method that also appears to give good results has been developed by Baldi *et al.* (Baldi *et al.*, 1993; Baldi & Chauvin, 1993). This method uses stochastic gradient descent in place of the EM method, which may help in avoiding local minima.

‡ Currently we choose γ_{del} and γ_{ins} each to be 1/2.

```
Helix                  AAAAAAAAAAAAAAAA  BBBBBBBBBBBBBBBBCCCCCCCCCCCC   DDDDDDDEE
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAG-KDLESIKGTA
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSA
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-FLK-GTSEVPQNNP
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG----AS---DP

Helix      EEEEEEEEEEEEEEEEEEE            FFFFFFFFFFFF  FFGGGGGGGGGGGGGGGGGGGG
HBA_HUMAN  QVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL--RVDPVNFKLLSHCLLVTLAAHLPAE
HBB_HUMAN  KVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL--HVDPENFRLLGNVLVCVLAHHFGKE
MYG_PHYCA  DLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH--KIPIKYLEFISEAIIHVLHSRHPGD
GLB3_CHITP PFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG---VTHDQLNNFRAGFVSYMKAHT--D
GLB5_PETMA DVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQYFKVLAAVIADTVAAG----
LGB2_LUPLU ELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG---VADAHFPVVKEAILKTIKEVVGAK
GLB1_GLYDI GVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGNKHIKAQYFEPLGASLLSAMEHRIGGK

Helix       HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  FTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  FTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  FGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP FA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -----DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU WSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI MNAAAKDAWAAAYADISGALISGLQS-----
```

Figure 4. Seven representative globin sequences of known structure and their alignment taken from Bashford *et al.* (1987). The letters A to H in Helix denote the 8 different α-helices. Some regions, especially CD, D and FG, are not well defined. The sequences and their SWISS-PROT identifiers are Human α (HBA_HUMAN), human β (HBB_HUMAN), sperm whale myoglobin (MYG_PHYCA), larval *Chironomous thummi* globin (GLB3_CHITP), sea lamprey globin (GLB5_PETMA), *Lupinus luteus* leghemoglobin (LGB2_LUPLU), and bloodworm globin (GLB1_GLYDI). (In SWISS-PROT 19 a $ is used instead of an "_" in the identifiers.)

(g). We used a homogeneous initial model that contained no knowledge about the globin family. Its probability parameters were derived from the prior, and were the same for all equivalent transitions (i.e. 9 different transition probabilities). All amino acid probabilities (the $\mathscr{P}$ distributions) were set equal to the distribution of the amino acids given by Krogh *et al.* (1993*a*). In the insert states we used a probability of 1/20 for all amino acids. The only model parameters set by hand are the initial transition probabilities and corresponding regularization parameters (see Krogh *et al.*, 1993*a*). From our experience the method does not seem to be very sensitive

NLL-score for the model, which was the average of the NLL-scores for the training sequences, as defined in Methods section (b). The final NLL-scores varied considerably for these runs but the best was 210·7.

We then took this model, produced ten new models by adding noise, and optimized these. These models all generated approximately the same NLL-score and we picked the model with the best NLL-score, 210·3, having a length of 147. We validated this model† in two ways: from the alignments it produced, and by its ability to discriminate between globins and non-globins. The results are was achieved by aligning these seven sequences and then aligning the rest of the 226 studied to the closest of these seven. In contrast, generating multiple alignments with HMMs requires no prior knowledge of underlying structure. Using the globin HMM, we produced a multiple alignment of all the 625 globin sequences by the Viterbi algorithm as described in Methods section (c). Figure 5 shows this alignment for the seven sequences from Bashford *et al.* (1987).

The alignment found in this experiment agrees extremely well with the structurally derived alignment of Bashford *et al.* Our alignment differs in the region between the C and E helices. However, this is a highly variable area since only some globins possess a D helix. The difference in the F/G-helices

between secondary structure elements. The last two insertions appear in the F/G region.

(ii) *Database search: discriminating globins from non-globins*

The globin HMM model we found was also tested on all the 25,044 proteins in the SWISS-PROT database release 22·0 of length less than 5000 amino acids (which is all but 2). A NLL-score and a Z-score were computed for each of these sequences as described in Methods section (f). These are plotted in Figures 6 and 7 as a scatter plot and a histogram, respectively. For the histogram (but not the scatter plot), the data were filtered as follows:

All sequences with a Z-score $>3{\cdot}5$ and either

[redacted]

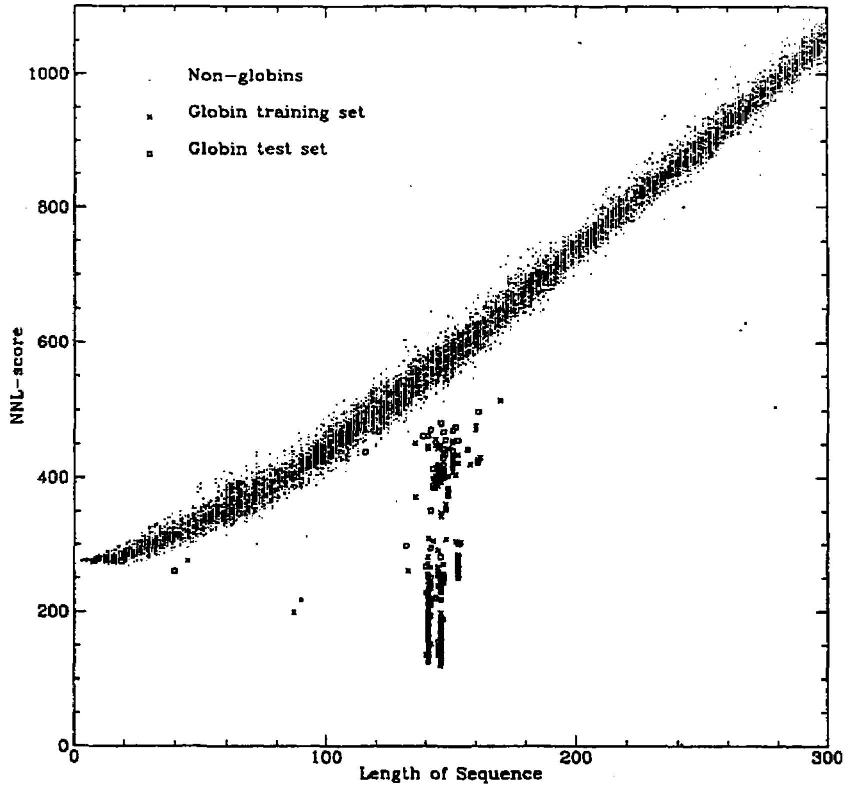
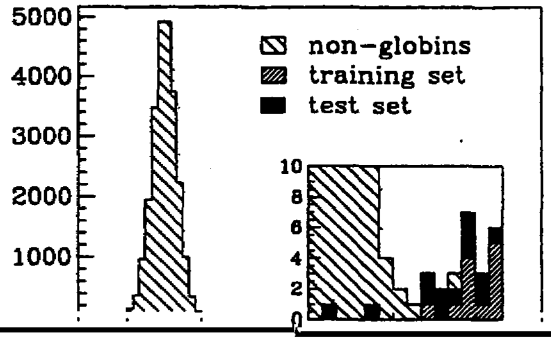


Figure 6. Plot of NLL-score *versus* sequence length for globins and non-globins. All sequences of length less than 300 from the SWISS-PROT 22 database are shown, including partial sequences and 3 false globins from the globin file, and sequences from the database containing many Xs.



so sequences with many Xs spuriously match the model very well.

Since we searched a newer release of SWISS-PROT (release 22) than the one from which the globin training set was extracted (release 19), eight new globins were found and incorporated into the test set.

Five globin fragments of length 19 to 45 were removed from the data.

Three non-globin sequences in the globin file that were identified as outliers in Figure 6 were removed. One of these non-globins was left as part of the

(GLB_PARCA and GLB_TETPY) are protozoan, whereas the other globins are metazoan. The primary sequences of these globins are similar and have little similarity with other eukaryotic globins. Note also that both of these sequences are in the test set.

(iii) *Discovering subfamilies of globins*

We also performed an experiment to automatically discover subfamilies of globins using the method described in Methods section (d). An HMM with ten component HMMs was used. The initial lengths of the components were chosen randomly between 120 and 170, but were adjusted by model surgery during training. We trained this HMM on all 628 globins and then calculated the NLL-score for each sequence for each of the ten component HMMs. A sequence was classified as belonging to the cluster represented by the component HMM that gave the lowest NLL-score, i.e. the one giving the highest probability to that sequence.† Three of these clusters were empty and the remaining seven non-empty ones represented chains from known globin subfamilies:

Class 1. 233 sequences: principally all α, a few ζ (an α-type chain of mammalian embryonic hemoglobin), π/π' (the counterpart of the α chain in major early embryonic hemoglobin P), and θ-1 chains (early erythrocyte α-like).

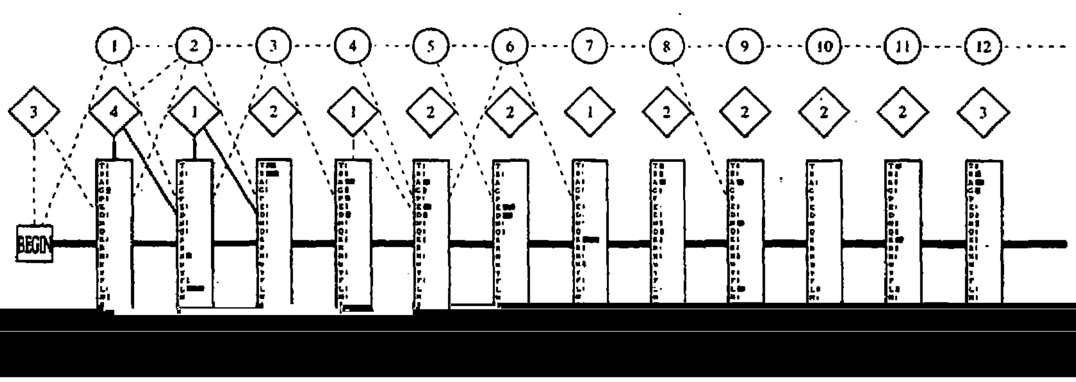
indicates what fraction of the 400 training sequences made that transition or used that particular amino acid. A broken line indicates that less than 5% of the sequences used that transition. (The continued delete is mostly due to fragments that have to make many deletions.) The histogram in a match state shows the distribution of amino acids that were matched to that state. The number in an insert shows the average length of an insertion beginning at that position.

For the amino acids the ordering proposed by Taylor (1986) is used. Starting from the top, the amino acids are medium-sized and non-polar, small and medium polar (around G and P), medium sized and polar (around K), large medium-polar (around F and Y), and finally below they are medium-large and non-polar. There does seem to be some tendency for the distributions to peak around neighboring amino acids when using this ordering, as one would expect. When one looks at the whole model, regions that are highly conserved are also readily distinguished from the more variable regions, both as a function of the probability that a position is skipped, and the entropy of the distribution of amino acids at that position.

(b) *Kinase experiments*

Protein kinases are defined as enzymes that

[redacted]


1
2
3
4
5
6
7
8
9
10
11
12
3
4
1
2
1
2
2
1
2
2
2
2
3
BEGIN

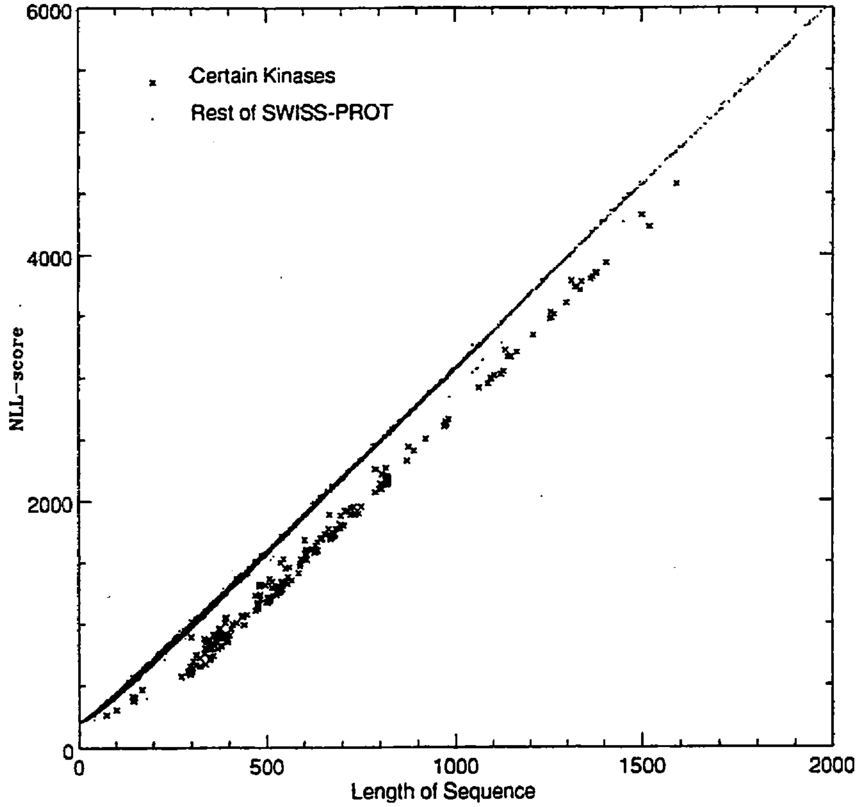


Figure 9. Scatter plot of NLL-score versus length for sequences in SWISS-PROT using the Kinase HMM.

The general issue of estimating the number of false negatives and false positives when distinguishing sequences belonging to a given family from non-members is a complex one. In the case of the globins. it is "relatively" straightforward since it is possible to identify all the globins in the database by performing a keyword or title string search. The situation for the kinase domain or the EF-hand motif (see section (c) below) is less obvious and thus more problematic. For instance, while a given protein may possess the sequence characteristics for this motif or domain. functionally, the region may not bind calcium or possess kinase activity. We have attempted to address this complicated matter as best we can as described below. However, we stress that we do not feel able to give a definitive answer as to the number of true false negatives and true false positives in our kinase or EF-hand database discrimination tests.

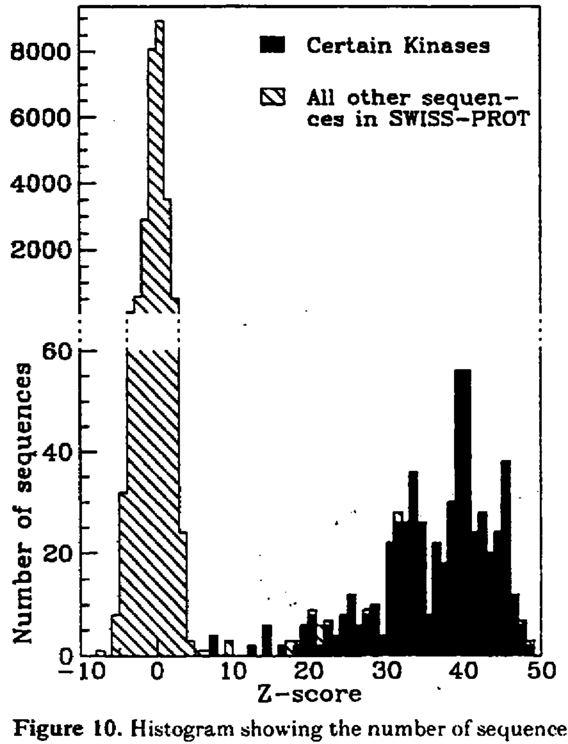


Figure 10. Histogram showing the number of sequences with a certain Z-score relative to the kinase model.

A list of potential protein kinases was created from the union of sequences designated as being kinases from four independent sources: our HMM, PROSITE (a dictionary of sites and patterns in proteins (Bairoch. 1992)), PROFILESEARCH (a technique used to search for relationships between a protein sequence and multiply aligned sequences (Gribskov *et al.*, 1990)) and a keyword search.

Two regions of the catalytic domain of eukaryotic protein kinases have been used to build PROSITE signature patterns. The first pattern corresponds to an area believed to be involved in ATP binding (PROSITE entry PROTEIN_KINASE_ATP, sequence motif [LIV]G.G.[FYM][SG].V). There are two signature patterns for the second region important for catalytic activity: one specific for serine/

```
               1        11        21        31        41        51                  71        81        91        101       111       121       131
  Subdomain    ..........<-I--------------------------->..........<-II-------------------------------->.<-III---------------------------->...<-IV---
  PROSITE      ................AAAAAAAA.A...............................................................................................
  X-ray        ..........BBBB.BB.........BBBBB..BB....................B......BBBBBBB.BAAA..AA.A..........A..A.AAA........A.A.AA.....AAA..AA.A............
  X-ray        ..........1111.11.........22222..22....................3......3333333.3BBB..BB.B..........B..C.CCC........C.C.CC.....CCC..CC.C............
 1 CAPK-ALPHA  ..........FEKI.KTLGTGSFGR.VMLVK..HK.....E..........TG.N......HYAMKIL.DKqk..VV.K..........LK.Q.IEH........T.L.NE.....KRI..LQ.AV....N...F...
 2 WEE1+       ..........FRNV.TLLGSGEFSE.VFQVE..DP.....V..........EK.TI.....KYAVKKL.KVk...FS.G..........PK.E.RNR........L.L.QE.....VSI..DR.AL....Kg..H...
 3 TIK         ..........FEDI.EEIGLGGFGQ.VFKAK..HR.....I..........DG.K......RYAIKRV.KY....--.N..........TE.K.AE-........-.-.HE.....VQA..LA.EL....N...H...
 4 SPK1        df........SIID.EVVGQGAFAT.VKKAI..EK.....T..........TG.K......TFAVKII.SKrk..VI.G..........NN.DgVT-........-.-.KE.....LEV..LQ.KL....N...K...
 5 RSK1-N      ..........FELL.KVLGQGSFGK.VFLVR..KVtrp..D..........SG.H......LYAMKVL.KKa...TL.K..........VR.D.RVR........T.K.ME.....RDI..LA.DV....N...K...
 6 PYT         ..........YSIL.KQIGSGGSSK.VFQVL..NE.....K..........K-.Q......IYAIKYV.NLe...EA.D..........NQ.T.LDS........Y.R.NE.....IAY..LN.KL....Qqh.S...
 7 PKC-ALPHA   ..........FNFL.MVLGKGSFGK.VMLAD..RK.....G..........TE.E......LYAIKIL.KKdv..VI.Q..........DD.D.VEC........T.M.VE.....KRV..LA.LL....Dk..P...
 8 PDGFR-B     ..........LVLG.RTLGSGAFGQ.VVEAT..AH.....G..........LS.Hsqatm.KVAVKML.KS....TA.R..........SS.E.KQA........L.M.SE.....LKI..MS.HL....Gp..H...
 9 PBS2        ..........LEFL.DELGHGNYGN.VSKVL..HK.....P..........TN.V......IMATKEV.RL....EL.D..........EA.K.FRQ........I.L.ME.....LEV..LH.KC....N...S...
10 KIN1        ..........FQQV.KPINESDFSF.VYHVS..SInp...P..........TE.T......VYVVKML.KKn...AA.R..........FT.G.KER........H.L.QE.....VSI..LQ.AL....Qa..C...
11 NCK1        ..........VREY.RRIGRGAFGT.VVQAY..LT.....Q..........DK.Knwlg..PFAIKKV.PA....--.N..........TE.Y.RS-........-.-.RE.....LQI..LR.IA....D...H...
12 INS.R       ..........ITLL.RELGQGSFGM.VYEGN..AR.....D..........II.Kgeaet.RVAVKTV.NE....SA.S..........LR.E.RIE........F.L.NE.....ASV..MK.GF....T...C...
13 HSVK        ..........FTIN.GALTPGSEGC.VFDSS..HP.....D..........YP.Q......RVIVKAG.WY....-T.S..........TS.-.---........-.-.HE.....ARL..LR.RL....D...H...
14 ERK1        ..........YTQL.QYIGEGAYGM.VSSAY..DH.....V..........RK.T......RVAIKKI.SP....FE.H..........QT.Y.CQR........T.L.RE.....IQI..LL.GF....R...H...
15 EGFR        ..........FKKI.KVLGSGAFGT.VYKGL..WI.....P..........EG.Ekvki..PVAIKEL.RE....AT.S..........PK.A.NKE........I.L.DE.....AYV..MA.SV....D...N...
16 ECK         ..........VTRQ.KVIGAGEFGE.VYKGM..LKt....S..........SG.Kkev...PVAIKTL.KA....GY.T..........EK.Q.RVD........F.L.GE.....AGI..MG.QF....S...H...
17 DPYK1       ae........LEFG.QTIGKGFFGE.VKRGY..WR.....E..........T-.-......DVAIKII.YRdq..FK.T..........KS.S.LVM........F.Q.NE.....VGI..LS.KL....R...H...
18 CLK         ar........YEIV.DTLGEGAFGK.VVECI..DH.....K..........VGgR......RVAVKIV.RN....--.V..........DR.Y.CEA........A.Q.SE.....IQV..LE.HL....Nct.D...
19 CDC2HS      ..........YTKI.EKIGEGTYGV.VYKGR..HR.....T..........TG.Q......VVAMKKI.RLe...SE.E..........EG.V.PST........A.I.RE.....ISL..LK.EL....K...H...
20 CAMII-ALPHA ..........YQLF.EELGKGAFSV.VRRCV..KV.....L..........AG.Q......EYAAKII.NTk...KL.S..........AR.D.HQK........L.E.RE.....ARI..CR.LL....K...H...
21 C-SRC       ..........LRLE.VKLGQGCFGE.VWMGT..WN.....G..........T-.T......RVAIKTL.KP....--.G..........TM.S.PEA........F.L.QE.....AQV..MK.KL....R...H...
22 C-RAF       ..........VMLS.TRIGSGSFGT.VYKGK..WH.....G..........D-.-......-VAVKIL.KVv...DP.T..........PE.Q.FQA........F.R.NE.....VAV..LR.KT....R...H...
23 KLSR_HUMAN  mgcgcss244LKLV.ERLGAGQFGE.VWMGY..YN.....G..........H-.T......KVAVKSL.KQ....--.G..........SM.S.PDA........F.L.AE.....ANL..MK.QL....Q...H...
24 KLSR_MOUSE  mgcvcss244LKLV.ERLGAGQFGE.VWMGY..YN.....G..........H-.T......KVAVKSL.KQ....--.G..........SM.S.PVP........F.L.AE.....ANL..MK.QL....Q...H...
25 ARRB_HUMAN  madleav190FSVH.RIIGRGGFGE.VYGCR..KR.....D..........TG.K......MYAMKCL.DKk...RI.K..........MK.Q.GET........LaL.NE.....RIM..LS.LVstg.D...C...
26 ARRB_BOVIN  madleav190FSVH.RIIGRGGFGE.VYGCR..KA.....D..........TG.K......MYAMKCL.DKk...RI.K..........MK.Q.GET........LaL.NE.....RIM..LS.LVstg.D...C...
27 BYR1_SCHPO  mfkrrrnp65LEVV.RHLGEGNGGA.VSLVK..HR.....-..........-N.I......FMARKTV.YV....GS.D..........SK.L.QKQ........I.L.RE.....LGV..LH.HC....R...S...
28 CYGR_ARBPU  mattr1157QQIF.ATIG------.-----T..YR.....G..........T-.-......ICAIHAVhKK....--.H..........ID.L.TRA........V.R.TE.....LKL..MR.DK....R...H...
29 ANPA_RAT    mpgsrrv536GSLL.TT--ECQF-Q.VFAKT..AY.....Y..........KG.N......LVAVKRVnKK....--.R..........IE.L.TRK........V.L.FE.....LKH..MR.DV....Q...N...
30 ANPA_HUMAN  mpgprrp471LEVL.ALVGSLS---.-LLGI..LI.....Vsffiyrkm70KG.N......LVAVKRVnKK....--.R..........IE.L.TRK........V.L.FE.....LKH..MR.DV....Q...N...
31 ANPB_HUMAN  malpsl1512SRLT.LSLRGSSYGS.LMTAN..GKyqifaN..........TG.Hfkgn..VVAIKHV.NK....-K.R..........IE.L.TRQ........V.L.FE.....LKH..MR.DV....Q...F...
32 ANPA_MOUSE  mprsrrv536GSLL.TT--ECQF-Q.VFAKT..AY.....Y..........KG.N......LVAVKRVnKK....--.R..........IE.L.TRK........V.L.FE.....LKH..MR.DV....Q...N...
33 ANPB_RAT    malpsl1512SRLT.LSLRGSSYGS.LMTAN..GKyqifaN..........TG.Hfkgn..VVAIKHV.NK....-K.R..........IE.L.TRQ........V.L.FE.....LKH..MR.DV....Q...F...
34 CYGS_STRPU  maharb1582QQIF.ATIG------.-----T..YR.....G..........T-.-......VCALHAVhKK....--.H..........ID.L.TRA........V.R.TE.....LKI..MR.DK....R...H...
35 VPSF_YEAST  gaqlslvv345RFL.RT--------.-CRAL..DP.....-..........HG.E......IV-IKVFiKP....--.K..........DQyS.LRP........F.L.QR.....IRA..QSfKL....Gq..L...
36 HSER_RAT    mtsllg1503----.----------R.VRQCR..YD.....-..........-R.K......RVILKDL.KHcd..GNfS..........EK.Q.RI-........-.-.-E.....LN-..-R.LL....Q...S...
37 HSER_HUMAN  mktllld445LLLV.ALL-------.-RLRK..YR.....Kdyelrqkk49DR.K......RVILKDL.KHnd..GNfT..........EK.Q.KI-........-.-.-E.....LNK..LL.QI....D...Y...
38 KR2_VZVD    mdaddtp138----.---GRGTYGR.V----..NI.....Y..........PS.S......KIAVKTN.DSr...VF.N..........RE.L.INA........I.LaSE.....GSIraGE.RL....G...I...
39 KR2_HSV13   mdesrrq157----.---GSGGYGD.VQLIR..EH.....K..........L-.-......--AVKTI.REkev.FA.V..........EL.I.ATL........L.V.GE.....C-V..LR.AG....R...T...
40 KR1_HSV11   m.........----.----------E.VYAWEtaHF.....L..........D-.-......--AAPKL.IEwe..VS.G..........TR.E.NR-........-.-.--.....---..--.--....-...-...
41 KR2_EBV     agwrsav100----.-LLGRGSYGA.VY---..AH.....A..........DR.A......TV---RL.YD....SV.T..........EL.Y.NE-........-.-.--1mvedmIQI..GK.ATaedgQ...D...
42 KRB2_VACCV  mesfkyct14WLIG.NTLYSGN-SI.LYKVR..KN.....F..........TS.Sfy....NYVRKI-.--....--.D..........HK.S.HKP........L.L.SE.....IRF..YI.SVI...D...Plti
43 KRB2_VACCC  mesfkyct14WLIG.NTLYSGN-SI.LYKVR..KN.....F..........TS.Sfy....NYVRKI-.--....--.D..........HK.S.HKP........L.L.SE.....IRF..YI.SVI...D...Plti
44 AK3_ECOLI   mseivvsk32VRL-.---------V.VLSAS..AG.....I..........TN.L......LVALAEG.LEpcerFE.R..........LD.A.IRN........I.Q.--.....FAI..LE.RL....R...Y...
45 PSP_MOUSE   mdqlgslv18ESLL.GELGS-----.---AV..NN.....-..........--.-......---LKIL.NP....-P.S..........EA.V.PQK........L.N.LD.....VEL..LQ.QAt...Sw..-...
46 DHOM_BACSU  mkairvg...LLGL.GTVGSG---V.VKIIQ..DH.....G..........DK.Lmhqvgc.PVTIKKV.LV....--.K..........DL.E.KK-........-.-.RE.....VDL..PR.EVlt..-...-...
47 FLIG_BACSU  marrd.....----.-----------.----Q..DK.....L..........TG.Kq.....KAAILMI.SL....GL.DvsasvykhltDE.E.IER........L.T.LE.....ISG..VR.SV....D...H...
48 CALQ_RABIT  mnaadrmg12LLLL.LVLGSPQSG-.-VNGE..EG.....L..........D-.-......------F.PEyd..GV.D..........RV.I.NVN........A.K.NY.....KNV..FK.KTe...-...-...
49 MYIR_PODAN  my........YSII.ISL-----IE.VVLVL..VP.....A..........L-.-......-------.-L....--.G..........IA.Y.V--........T.I.AE.....RKT..MA.SR....QrrlG...
50 RUVA_ECOLI  migrlrgi15LVLI.EVGGVG----.-YEVH..MP.....M..........T-.-......-----CF.YE....LP.E..........AG.-.---........-.-.QE.....ATV..FT.HF....-...-...
51 DISA_HSV6U  mdngvet194RTRA.RVLGVGAYGK.VF---..DL.....D..........K-.-......-VAIKTA.NE....--.D..........ES.V.ISA........F.I.A-.....---..--.--....-...-...
52 KRF1_VACCC  mgvandss94----.----TGGYGI.VFKID..NY.....Vskfvfeat79YS.G......KVFLKAF.NEr...KD.S..........IK.F.VK-........-.-.--.....--L..LS.KF....Y...P...
53 UL97_HCMVA  mssalrs335----.-RLGQGSFGE.VW---..--.....P..........LD.R......YAVVKVA.RK....--.H..........SE.T.VLT........V.W.N-.....---..--.--....-...-...
54 KKA6_ACIBA  melpniiq13SVLEpNKIGQ-SPSD.VYS-F..NR.....N..........N-.E......TFFLKRS.STly..TE.T..........TY.S.VS-........-.-.RE.....AKN..LS.WL....Se..K...
55 KKA8_ECOLI  mndidree27WARD.KV---GQSGCaVYRLH..SA.....S..........GG.S......DLFLKHG.KDafadDV.T..........DE.K.VR-........-.-.--.....---..--.-L....Kw..-...
56 KGPB_BOVIN  ..........----.----------.--HGT..LR.....-..........-D.L......QYA---L.QEk...IE.E..........LR.Q.RDA........L.I.DE.....LEL..--.--....-...-...
57 EGFR_CHICK  mgvrspla16VLVL.LLLG------.VALCS..AV.....E..........E-.-......----KKV.CQ....GT.N..........NR.-.---........-.-.--.....---..LT.QL....G...Hved
58 KKA1_ECOLI  mshiqret32----.----VGQSG-.---ATi.YR.....L..........YG.Rpdape.LF----L.KH....--.G..........KG.S.VA-........-.-.--.....---..--.--....-...-...
59 ROTK_DROME  mpsarlf393NRLS.KELGK-EFGK.EF-GK..EF.....Cpapsmslq72PG.Gpp....PFLKKKK.PR....--.P..........KG.Q.NSAprggpprswT.N.TE.....LT-..-E.AL....Q...N...
60 XPCG_HUMAN  maglgpg121----.-------YGL.VHQGK..KCsccemN..........VN.K......RC----V.RS....VP.S..........LC.G.VDH........T.E.NR.....GRL..DL.EI....R...A...
```

A (cont)

Fig. 11.

threonine kinases (PROTEIN KINASE ST OKBOG) bovine cGMP-dependent protein kinase

```
                141      151      161      171      181      191      201      211      221      231      241      251      261      271
   Subdomain    ------------------------------------------><-V-------------------------------------------------------->..........<-VIa-------------
   PROSITE      .......................................................................................................................................
   X-ray        .........................BB....B....BBB.....B..........BB..BBB..................AAA......AA...AA...A..........................AA...A.....AAA.
   X-ray        .........................44....4....444.....5..........SS..SSS..................DDD......DD...DD...D..........................EE...E.....EEE.
 1 CAPK-ALPHA   .......P......FL...V.....KLE....F....SFKD..NSN..........LY..MVM..EYVPGG.......E.MFS......HL...RR...I.....G..-..R..........FS.EP...H.....ARF.
 2 WEE1*        .......D......MI...V.....ELH....D....SVEH..GGF..........LY..MQV..ELCENG.......S.LDR......FL...EE...Q.....GqlS..R..........LD.EF...R.....VWK.
 3 TIR          .......Y......MI...V.....QYMscweG....VDYD..PENsmadtery12LF..IQM..EFCDKG.......T.LEQ......WH...RNr..H.....Q..S..K..........VD.RA...L.....ILD.
 4 SPK1         .......P......RI...V.....RLK....G....FYED..TES..........YY..MVM..EFVSGG.......D.LMD......FV...AA...H.....G..-..A..........VG.ED...A.....GRE.
 5 ASK1-K       .......P......FV...V.....KLR....Y....AFQT..EGK..........LY..LIL..DFLRGG.......D.LFT......RL...SK...E.....V..-..K..........FT.EE...D.....VKF.
 6 PYT          .......D......KI...I.....RLY....D....YEIT..DQY..........IY..MVM..ECGN-I.......D.LNS......VL...KK...K.....K..-..S..........ID.PW...E.....RKS.
 7 PKC-ALPHA    .......P......FL...T.....QLH....S....CFQT..VDR..........LY..FVM..EYVNGG.......D.LMY......HI...QQ...V.....G..-..K..........FK.EP...Q.....AVF.
 8 PDGFR-B      .......L......NV...V.....NLL....G....ACTK..GGP..........IY..IIT..EYCRYG.......D.LVD......YL...HR...N.....K..H..Tflqhhsdk98LS.YN...D.....LVG.
 9 PBS2         -......P......YI...V.....DFY....G....AFFI..EGA..........VY..MCM..EYMDGG.......S.LDK......IY...DE...Sei..G..G..-..........ID.EP...Q.....LAF.
10 RIK1         .......P......FV...V.....NLV....N....VWSY..NDN..........IF..LQL..DYCENG.......D.LSL......FL...SE...L.....GIIQ..V..........ND.PF...R.....VWK.
11 NCK1         .......P......RI...V.....KLQ....Y....FFTH..LSPqdak......VYqhLAM..ECLP-E.......T.LQI......EI...NRyvtN.....K..L..E..........RP.LK...H.....IRL.
12 INS.R        .......N......NV...V.....RLL....G....VVSK..GQP..........TL..VVM..ELMAHG.......D.LKS......YL...RS...L.....R..P..Easnnpgrpp.PT.LQ...E.....MIQ.
13 HSVK         .......P......AI...L.....PLL....D....LHVV..SGV..........TC..LVL..PKYQA-.......D.LYT......YL...SR...R.....L..N..P..........LG.RP...Q.....IAA.
14 ERK1         .......E......NV...I.....GIR....D....ILRA..PTLeamrd.....VY..IVQ..DLME-T.......D.LYK......LL...KS...Q.....Q..-..-..........LS.ND...H.....ICY.
15 EGFR         .......P......HV...C.....RLL....G....ICLT..S-T..........VQ..LIT..QLMPFG.......C.LLD......YV...RE...H.....K..D..N..........IG.SQ...Y.....LLN.
16 ECK          .......N......NI...I.....RLE....G....VISK..YKP..........MM..IIT..EYMENG.......A.LDK......FL...RE...K.....D..G..E..........FS.VL...Q.....LVG.
17 DPYK1        .......P......EV...V.....QFL....G....ACTAcrEDH..........HC..IYT..EYMGGG.......S.LRQ......FL...TQ...H.....F..-..N..........LL.EDmphI.....NLK.
```

```
                 281       291       301       311       321       331       341       351       361       371       381       391       401       411
   Subdomain     ---------------------------------------------->.<-VIb--------------------->..........<-VII------------------------------------------>.<-V
   PROSITE       ...............................................B...BBBBBBBB.B.BB..B.............................................................................
   X-ray         AAA..........A.......AAA..AAAAA.A..........A...BB...BB.......B.BB..B.B.....B..........B.......BBB.........B....BBB..........B...............
   X-ray         EEE..........E.......EEE..EEEEE.E..........E...66...66.......7.77..7.7.....8..........8.......888.........9....999..........9...............
 1 CAPK-ALPHA    YAA..........Q.......IVL..TFEYL.H..........S.L.DL...IYRDLKPE.N.LL..I.D...QQG..........Y.......IQVT.D..FGF.A....KRV..........KG...-.....--R.T..
 2 WEE1+         ILV..........E.......VAL..GLQFI.H..........N.N.NY...VHLDLKPA.N.VM..I.T...FEG..........T.......LKIG.D..FGM.A....S--..........VW...P...VPR.G..
 3 TIK           LYE..........Q.......IVT..GVEYI.H..........S.R.GL...IHRDLKPG.N.IF..L.V...DEN..........H.......IKIG.D..FGL.A....TAL..........EN...D...GAS.R..
 4 SPK1          ISR..........Q.......ILT..AIRVI.H..........S.N.GI...SHRDLKPD.N.IL..I.E...QDDpv........L.......VKIT.D..FGL.A....K-V..........QG...N...GSF.M..
 5 RSK1-N        YLA..........E.......LAL..GLDHL.H..........S.L.GI...IYRDLKPE.N.IL..L.D...EEG..........H.......IKLT.D..FGL.S....KEA..........ID...H...EKK.A..
 6 PYT           YWK..........N.......NLE..AVNTI.H..........Q.H.GI...VHSDLKPA.N.FL..I.V...-DG..........M.......LKLI.D..FGI.A....NQM..........QP...Dtt.SVV.K..
 7 PKC-ALPHA     YAA..........E.......ISI..GLFFL.H..........K.G.GI...IYRDLKLD.N.VM..L.D...SEG..........H.......IKIA.D..FGM.C....KEH..........MM...D...GVT.T..
 8 PDGFR-B       FSY..........Q.......VAN..GMEFL.A..........S.K.NC...VHRDLAAR.N.VL..I.C...EGK..........L.......VKIC.D..FGL.A....RDI..........MR...Ds..NYI.S..
 9 PBS2          IAY..........A.......VIN..GLKELKE..........Q.H.NI...IHRDVKPT.N.IL..C.Sa..NQG..........T.......VKLC.D..FGV.S....GNL..........VA...-...-SL.A..
10 KIK1          ALF..........Q.......LTQ..ALNFI.H..........L.L.EF...VHLDVKPS.N.VL..I.T...RDG..........N.......LKLG.D..FGL.A....TSL..........PV...-...SSN.V..
11 NCK1          YTY..........Q.......IAR..GMLYL.H..........G.L.GV...CHRDIKPS.N.VL..V.Dp..ETG..........V.......LKIC.D..FGS.A....KKL..........EN...N...QP-.S..
12 [illegible]
```

```
                421       431       441       451       461       471       481       491       501       511       521       531       541       551
Subdomain       III----------------------------><-II-----------------------------------------------------------------------------------------------> ...<-X------------------
PROSITE         ..........................................................................................................................................................
X-ray           ...................................................................A.AAAAAAAAAAAAAA............................................A....A.......
X-ray           ...................................................................F.FFFFFFFFFFFFFF............................................G....G.......
 1 CAPK-ALPHA   ........MTLCGT.P..EY.LAPE..IIL.........SK..........G-YNK...A.VDWWALGVLIYEMAA.G......YP...P..........F.......F....A..-DQP......I....Q.......
 2 WEE1*        ........MEREGD.C..EY.IAPE..VLA.........MR..........L-YDK...P.ADIFSLGITVFEAAAnI......VL...Pdsgqswqh17-........P....R..LSSI......D....M.......
 3 TIK          ........TRATGT.L..QY.MSPE..QLF.........LK..........N-YGK...E.VDIFALGLILAELL-.-......NT...G..........F.......T....E..-SEK......I....K.......
 4 SPK1         ........KTFCGT.L..AY.VAPE..VIR.........GKdtsvdpe12NEYSS...L.VDMWSMGCLVYVILT.G......HL...P..........F.......S....G..-STQ......D....Q.......
 5 RSK1-N       ........YSFCGT.V..EY.MAPE..VVN.........RQ..........G-NTN...S.ADWWSYGVLM-----.-......--...-..........-.......-....G..KDRK......E....T.......
 6 PYT          ........DSQVGT.V..NY.MPPE..AIKdmsssrengKSK..........SKISP...K.SDVWSLGCILYYMTY.G......KT...P..........F.......Q....Q..IINQi.....S....K.......
 7 PKC-ALPHA    ........RTFCGT.P..DY.IAPE..IIA.........YQ..........P-YGK...S.VDWWAYGVLLYEMLA.G......QP...P..........F.......D....G..-EDE......D....E.......
 8 PDGFR-B      ........KGSTFL.P..LK.WMAP..ESI.........FN..........SLYTT...L.SDVWSFGILLWEIFTlG......GT...P..........Y.......P....E..LPMN......E....Q.......
 9 PBS2         ........KTNIGC.Q..SY.MAPE..RIK.........SLnpdr......ATYTV.<.Q.SDIWSLGLSILEMAL.G......RY...P..........Y.......P....P..-ETY......DnifsQ.......
10 RIK1         ........D-LEGD.R..VY.IAPE..ILA.........SH..........N-YGK...P.ADVYSLGLSMIEAATnV......VL...Pesgvewqr16L.......P....N..LKDL......L....L.......
11 MCK1         ........ISYICS.R..FY.RAPE..LII.........GC..........TQYTT...Q.IDIWGLGCVMGEMLI.G......KA...I..........F.......Q....G..QEPL......L....Q.......
12 INS.R        ........GGKGLL.P..VR.WMAP..ESL.........KD..........GVFTT...S.SDMWSFGVVLWEITSlA......EQ...P..........Y.......Q....G..LSNE......Q....V.......
13 HSVK         ........YGIAGT.I..DT.NAPE..VLA.........GD..........P-YTT...T.VDIWSAGLVIFETAVnM......AS...L..........FsaprgphR....G..-PCD......S....Q.......
14 ERK1         ........TEYVAT.R..WY.RAPE..IML.........NS..........KGYTK...S.IDIWSVGCILAEMLS.N......RP...I..........F.......P....G..KHYL......D....Q.......
15 EGFR         ........E-GGKV.P..IK.WMAL..ESI.........LH..........RIYTH...Q.SDVWSYGVTVWELMTfG......SK...P..........Y.......D....G..IPAS......E....I.......
16 ECK          ........TSGGKI.P..IR.WTAP..EAI.........SY..........RKFTS...A.SDVWSFGIVMWEVMTyG......ER...P..........Y.......W....E..LSNH......E....V.......
17 DPYK1        ........TQSVGC.I..PY.MAPE..VFK.........GD..........S-NSE...K.SDVYSYGMVLFELLT.S......DE...P..........Q.......Q....D..NKPK......K....N.......
18 CLK          ........STLVST.R..HY.RAPE..VIL.........AL..........G-WSQ...P.CDVWSIGCILIEYYL.G......FT...V..........F.......S....T..HDSR......E....H.......
19 CDC2HS       ........TNEVVT.L..WY.RSPE..VLL.........GS..........ARYST...P.VDIWSIGTIFAELAT.K......KP...L..........F.......H....G..DSEI......D....Q.......
20 CAMII-ALPHA  ........FGFAGT.P..GY.LSPE..VLR.........KD..........P-YGK...P.VDLWACGVILYILLV.G......YP...P..........F.......W....D..-EDQ......H....R.......
21 C-SRC        ........RQGAKF.P..IK.WTAP..EAA.........LY..........GRFTI...K.SDVWSFGILLTELTTkG......RV...P..........Y.......P....G..MVNR......E....V.......
22 C-RAF        ........EQPTGS.V..LW.MAPE..VIR.........MQdan.......P-FSF...Q.SDVYSYGIVLYELMT.G......EL...P..........Y.......S....N..INNR......D....Q.......
23 KLSK_HUMAN   ........REGAKF.P..IK.WTAP..EAI.........NY..........GTFTI...K.SDVWSFGILLTEIVThG......RI...P..........Y.......P....G..MTNP......E....V.......
24 KLSK_MOUSE   ........REGAKF.P..IK.WTAP..EAI.........NY..........GTFTI...K.SDVWSFGILLTEIVThG......RI...P..........Y.......P....G..MTNP......E....V.......
25 ARBK_HUMAN   ........HASVGT.H..GY.MAPE..VLQ.........KG..........VAYDS...S.ADWFSLGCMLFKLLR.G......HS...P..........F.......R....Q..HKTK......Dkh..E.......
26 ARBK_BOVIN   ........HASVGT.H..GY.MAPE..VLQ.........KG..........VAYDS...S.ADWFSLGCMLFKLLR.G......HS...P..........F.......R....Q..HKTK......Dkh..E.......
27 BYR1_SCHPO   ........QTFVGT.S..TY.MSPE..RIR.........GG..........K-YTV...K.SDIWSLGISIIELAT.Q......EL...Pws........F.......S....N..IDDSigil..D....L.......
28 CYGR_ARBPU   ........GEHAKL.A..AKIWTAP..EHL.........REgkmhp.....G-GTP...K.GDIYSFSIILTEMYS.R......QE...P..........F.......Hen..D..LELA......D....I.......
29 ANPA_RAT     ........TLFAKK.L..--.WTAP..ELLrmaspp....AR..........G--SQ...A.GDVYSFGIILQEIALrS......GV...F. ........Yveg....L....D..LSPK......E....I.......
30 ANPA_HUMAN   ........Y-AKKL.-..--.WTAP..ELL..........RNaappv.....R-GSQ...A.GDVYSFGIILQEIAL.R......SG...V..........F.......HveglD..LSPK......E....I.......
31 ANPB_HUMAN   ........Y--AKK.L..WT.-APE..LLS..........GN..........PLPTTgmqK.ADVYSFGIILQEIAL.R......SG...P..........F.......YleglD..LSPK......E....I.......
32 ANPA_MOUSE   ........TLFAKK.L..--.WTAP..ELLrmaspp....AR..........G--SQ...A.GDVYSFGIILQEIALrS......GV...F..........Yveg....L....D..LSPK......E....I.......
33 ANPB_RAT     ........Y--AKK.L..WT.-APE..LLS..........GN..........PLPTTgmqK.ADVYSFAIILQEIAL.R......SG...P..........F.......YleglD..LSPK......E....I.......
34 CYGS_STRPU   ........GDHAKLaR..QL.WTSP..EHL..........RQegempta...G--SP...Q.GDIYSFAIILTELYS.R......QE...P..........F.......H....EneKDLA......D....I.......
35 VPSF_YEAST   fltytd..TSKRRT.-..CY.LAPE..RFN.........SKlyqdgkmns.GRLTK...E.MDIFSLGCVIAEIFAeG......RP...I..........F.......-....-..--NL......S....Q.......
36 NSER_RAT     ........KDL---.-..--.WTAP..ENL..........RQ..........ATISQ...K.GELYSFSIIAQEIIL.R......KE...T..........F.......Y....T..LSCR......D....Qneh....
37 NSER_HUMAN   ........KDL---.-..--.WTAP..ENL..........RQ..........ANISQ...K.GDVYSYGIIAQEIIL.R......KE...T..........F.......Y....T..LSCRdrn...E....K.......
38 KR2_VZVD     ........FRLVLS.H..GT.MQPP..EIL.........LDyingtglt15QRVGL...A.IDLYALGQALLEVIL1G......RL...Pgqlpisvhl4Y.......Y....G..HKLS......P....D.......
39 KR2_HSV11    qtclqe20NTLVG-.N..GY.MQPP..ELLvkylnmer15LK...........NDVGL...A.VDLYALGQTLLELVVsV......YV...Apslgvpvtr.F.......P....G..----......-....-.......
40 KR1_HSV11    ........WNPIGT.E..AY.ASPE..RSR.........DRvpdrpdsa12GTNGA...G.I--------------.-......RE...P..........Kli.....K....G..DGYR......A....H.......
41 KR2_EBV      ........KSSKGR.Q..LY.R--L..YCQ.........RE..........P-FSI...A.KDTY-----------.-......KP...Lcllekcyi24-........-....G..ADTA......L....R.......
42 KRB2_VACCV   ynedmil9NHLGAT.V..SR.RGDL..ENL.........GY..........C----...-.---------KIEWFG.G......KL...P..........W.......-....-..-KNE......S....S.......
43 KRB2_VACCC   ynedmil9NHLGAT.V..SR.RGDL..ENL.........GY..........C----...-.---------KIEWFG.G......KL...P..........W.......-....-..-KNE......S....S.......
44 AK3_ECOLI    ........DPRVVS.A..AK.RIDE..IAF.........AE..........A----...-.----------AEKATfGakvlhpAT...L..........L.......P....A..-VKS......D....I.......
45 PSP_MOUSE    ........--GIDL.T..VP.LAGE..ASL.........VL..........PFIGK...T.VDI-SVSLDLINSLS.I......KT...Maqcglpev14-........-....-..-SNT......D....K.......
46 DHOM_BACSU   ........DYEGLD.A..AR.RRA-..ILA.........RL..........G-FSN...K.VDLE-----------.-......--...-..........-dvkvkgiS....Q..ITDE......D....I.......
47 FLIG_BACSU   iqqehpq.T-----.-..--.-MAL..ILSy........LD..........PVQ--...-.-----AGQILSELN-.-......--...-..........-.......P....E..-VQA......E....V.......
48 CALQ_RABIT   ........KED---.-..--.---E..VIE.........YD..........GEFSAd..-.--------TLVEFL-.-......--...-..........-.......-....-..LDVL......E....D.......
49 RVIE_PODAE   ........GSLRST.A..QL.ISTE..LVL.........SS..........A----...-.-------ILLVIRLT.G......SL...Mlavniesq14F.......P....L..-LPV......F....I.......
50 RUVA_ECOLI   ........GALVRL.PgiGK.KTAE..RLIvemkdrfk11GD..........L-FTP...A.ADL--------VLTSpA......SP...A..........T.......D....D..AEQE......A....V.......
51 DISR_HSV6H   ........YRDACC.-..-K.VLAFhvYLI.........GI..........L----...-.--------------.-......--...-..........F.......Y....R..-DVV......F....I.......

52 KRF1_VACCC   ........S-----.-..--.----..--A.........LN..........D-FDF...S.--------------.-......--...-..........-.......-....-..-QVA......G....I.......
53 UL97_HCMVA   ........FPVAGL.R..RY.CMSE..LSA.........LG..........NVLGF...-.------CLN-----.-......--...-..........-.......-....R..LLDR......R....G.......
54 KKA6_ACIBA   ........DD----.-..--.----..--I.........DQ..........DDFDT...E.--------------.-......--...L..........W.......G....D..HKTY......LalwnE.......
55 KKA8_ECOLI   rrlhei18TTNAGL.P..ER.GSIE..AGVvdvddfdke.RE..........G-WTA...Eq---------VvEARH.R. ....LL...P..........L.......A....P..DPVV......T....H.......
56 KGPB_BOVIN   ........KQSAST.-..--.----..-LQ.........GE..........P-RTK...R.--------------.-......--...-..........-.......Q....A..ISAE......P....T.......
57 FGFR_CHICK   ........DGLRFT.P..MK.WLSE..ILN.........GG..........VKIGs...-.--------------.-......--...-..........-.......-....-..-NRRklwmIT....V
```

```
                  561       571       581       591       601       611       621       631       641       651       661       671
   Subdomain      ------------------------------------><-XI------------------------------------------------------------->..........
   PROSITE        ..........................................................................................................
   X-ray          A.AAAAA...A...........................AA........AA......A..AAAAA........................AAAAAA.................
   X-ray          G.GGGGG...G...........................HH........HH......H..HHHHH........................IIIII1.>...............
 1 CAPK-ALPHA     I.YERIV...S...GK.V.RF..........P....SH.F.SSD........LK......D..LLRNL.LQVD.LTKRF.Gnlkng.....VNDIKNHK....WF-..........
 2 WEE1+          G.SSLTSs..S...RE.T.PA..........R....SI.I.CQG........GL......Dr.VVEWN.LSPE.PRNRP.T..........IDQILATD....EV-cw........
 3 TIK            F.FESLR...K...GD.F.SN..........-....DI.F.DNK........EK......S..LLKRL.LSEK.PKDRP.E..........TSEILKT-....-L-aevrsise18
 4 SPK1           L.YKQIGr..GsyhEG.P.LK..........D....FR.I.SEE........AR......D..FIDSL.LQVD.PNNRS.T..........AAKALNHP....WI-..........
 5 RSK1-N         N.TLILKa..R...LG.N.--..........P....QF.L.STE........AQ......S..LLRAL.FKRN.PANRL.Gagpdg.....AEEIKRH-....-I-fysidvn20
 6 PYT            L.NAIIDp..N...HE.I.EF..........P....DI.P.EKQ........LQ......D..VVKCC.LKRD.PKQRI.S..........IPELLAHP....YV-..........
 7 PKC-ALPHA      L.FQSIN...E...HN.V.SY..........P....KS.L.SKE........AV......S..ICKGL.MTKH.PAKRL.Gcgpeg.....ERDVREHA....FF-..........
 8 PDGFR-B        F.YNAIKr..G...YR.N.AQ..........P....AH.A.SDE........IY......E..IMQKC.WEEK.FEIRP.P..........FSQLVLL-....-L-..........
 9 PBS2           L.SAIVD...G...PP.P.RL..........P....SDKF.SSD........AQ......D..FVSLC.LQKI.PERRP.T..........YAALTEHP....WL-..........
10 NIK1           S.NEKVQ...I...NR.V.R-..........-....--.C.AES........LQ......C..LLQRN.THPY.VDCRP.T..........TQDLLAHP....EN-if.........
11 NCK1           L.REIaK...L...LG.P.PDkrfiftsn37-....--.-.-PD........GI......D..LLNKI.LVYE.PQQRL.S..........PKKILAHQ....FF-nelrnddc11
12 INS.R          L.KFVMD...G...GY.L.DQ..........P....DN.C.PER........VT......D..LMRMC.WQFN.PNMRP.T..........FLEIVNL-....-L-..........
13 MSVK           I.TRIIRqaqV...HV.D.EFaphpesri35-....--.-.DID........VE......Y..LVCKA.LTFD.GALRP.S..........AAELLCLP....LF-..........
14 ERK1           L.NNILG...I...LG.S.PSqedlnciii31-...PR.S.DSK........AL......D..LLDKN.LTFN.PNKRI.T..........VEEALAHP....YL-eqyydptd49
15 EGFR           S.SILEK...G...ER.L.PQ..........P....PI.C.TID........VY......M..IMVKC.WMID.ADSRP.K..........FRELIIE-....-F-..........
16 ECK            K.KAIND...G...FR.L.PT..........P....MD.C.PSA........IY......Q..LMMQC.WQQE.RARRP.K..........FADIVSI-....-L-..........
17 DPYK1          A.MLAATe..S...YR.P.PI..........P....LT.T.SSK........WK......E..ILTQC.WDSN.PDSRP.T..........FKQIIVH-....-L-kemedqgv..
18 CLK            L.ANNEK...I...LG.P.LPkhmiqktr40-....--.-.-EL........LF......D..LIGRM.LEYD.PAKRI.T..........LKEALKHP....FF-yplkkht...
19 CDC2HS         L.FKIFR...A...LG.T.PNnevwpeve26-....KN.L.QEN........GL......D..LLSKM.LIYD.PAKRI.S..........GKMALNHP....YF-..........
20 CAMII-ALPHA    L.YQQIK...A...GA.Y.DF..........PspewDT.V.TPE........AK......D..LINKM.LTIN.PSKRI.T..........AAEALKHP....WI-..........
21 C-SRC          L.DQVER...G...YR.M.PC..........P....PE.C.PES........LH......D..LMCQC.WRKE.PEERP.T..........FEYLQAF-....-L-..........
22 C-RAF          IiFMVGR...G...YA.S.PDlskly.....-....KN.C.PKA........MK......R..LVADC.VKKV.KEERP.L..........FPQILSS-....-I-..........
23 KLSK_HUMAN     I.QMLER...G...YR.M.VR..........P....DN.C.PEE........LY......Q..LMRLC.WKER.PEDRP.T..........FDYLRSV-....-L-edfftate15
24 KLSK_MOUSE     I.QMLER...G...YR.M.VR..........P....DN.C.PEE........LY......H..LMMLC.WKER.PEDRP.T..........FDYLRSV-....-L-ddfftate15
25 ARKB_HUMAN     I.DRMTL...T...MA.V.EL..........P....DS.F.SPE........LR......S..LLEGL.LQRD.VNRRL.Gclgrg.....AQEVKESP....FF-rslduqm236
26 ARKB_BOVIN     I.DRMTL...T...MA.V.EL..........P....DS.F.SPE........LR......S..LLEGL.LQRD.VNRRL.Gclgrg.....AQEVKESP....FF-rslduqm236
27 BYR1_SCHPO     L.HCIVQ...E...EP.P.RL..........P....SS.F.PED........LR......L..FVDAC.LHKD.PTLRA.S..........PQQLCAMP....YF-qqalminv20
28 CYGR_ARBPU     I.ARVSR...G...EV.P.PYrpvlnavn..-....EA.A.PDC........VL......T..AIRAC.WVED.PNERP.N..........IIEVRTM-....-L-aplqkgl150
29 ANPA_RAT       I.ERVTR...G...EQ.P.PFrpsmdlqshi-....--.-.-EE........LG......Q..LMQRC.WAED.PQERP.P..........FQQIRLA-....-L-rkfnken260
30 ANPA_HUMAN     I.ERVTR...G...EQ.P.PFrpslalqshl-....--.-.-EE........LG......L..LMQRC.WAED.PQERP.P..........FQQIRLT-....-L-rkfnren260
31 ANPB_HUMAN     V.QKVRN...G...QR.P.YFrpsidrtq..-....--.L.NEE........LV......L..LMERC.WAQD.PAERP.D..........FGQIKGF-....-I-rrfnkeg261
32 ANPA_MOUSE     I.ERVTR...G...EQ.P.PFrpsmdlqshl-....--.-.-EE........LG......Q..LMQRC.WAED.PQERP.P..........FQQIRLA-....-L-rkfnken260
33 ANPB_RAT       V.QKVRN...G...QR.P.YFrpsidrtq..-....--.L.NEE........LV......L..LMERC.WAQD.PTERP.D..........FGQIKGF-....-I-rrfnkeg261
34 CYGS_STRPU     I.GRVRS...G...EV.P.PYrpilnavn..-....AA.A.PDC........VL......S..AIRAC.WPED.PADRP.N..........INAVRTK-....-L-aplqkgl286
35 VPSF_YEAST     L.FKYKSn..S...YD.V.NKeflmeemn..-....--.-.STD........LR......N..LVLDK.IQLD.PSKRL.S..........CDELLNK-yrgiFF-pdyfytl155
36 HSER_RAT       I.FRVEN...S...YG.T.RPfrpdlflei15-...--.-.--E........VY......L..LVKSC.WEED.PEKRP.D..........FKKIEST-....-L-akifglf328
37 HSER_HUMAN     I.FRVEN...S...NG.N.RPfrpdlflei15-...--.-.--E........VY......L..LVKNC.WEED.PEKRP.D..........FKKIETT-....-L-akifglf328
38 KR2_VZVD       L.ALDTL...A...YRcV.LA..........PyilpSD.I.PGD........LNympfih-..-----.-ACE.LNTRI.S..........RNSLRRI-....-F-qchavryg58
39 KR2_HSV11      -.YQYFN...N...QL.S.PDfalallay66-....VA.L.PPE........LKpll...V..LVSRL.CHTN.PC---.-..........---ARNA-....-L-a.........
40 KR1_HSV11      V.LKVIR...A...RG.T.LDlrggartell-....--.-.-DE........LI......G..LVARC.LERD.PANRP.S..........LETLVDE-....-F-ski.......
41 KR2_EBV        L.DLQSL...G...YS.L.LYgimblads38-....--.-.---........LL......E..VLSQN.WNLN.LDKGL.Tscgeapcv49VAELLADD....FF-gpdgrrg...
42 KRB2_VACCV     I.KVIKQ...R...RE.Y.RKfiatffed15-....--.-.PLE........LV......R..YIELV.YTLD.YSQTP.N..........YDRLRKL-....-F-iqd.......
43 KRB2_VACCC     I.KVIKQ...R...RE.Y.RKfiatffed15-....--.-.PLE........LV......R..YIELV.YTLD.YSQTP.N..........YDRLRRL-....-F-iqd.......
44 AK3_ECOLI      P.VFVGS...S...KD.P.RAggtlvcnk79-....TL.L.TQS........LL......N..ELSAL.CRVE.VEE--.-..........GLALVAL-....-I-gndlshac57
45 PSP_MOUSE      I.SISLL...G...RR.L.Plinsildgvs.-....TL.L.TST........LS......T..VLQNF.LC--.----P.L..........LQYVLST-....-L-npsvlqgl21
46 DNOH_BACSU     S.FSKRL...G...YT.N.KLigiaqrdg56-....--.-.-PT........AT......S..VVSDL.VAVN.KNNRL.GvtgaefvgI7PSDIYAQQ....FL-ribvkdev82
47 FLIG_BACSU     A.RRIAV...N...DR.T.--..........-....--.-.SPE........II......N..EVERI.LEQK.LSSAF.Tqdytqtgg..IEAVVEV-....-L-agvdrgt130
48 CALQ_RABIT     P.VELIE...G...EReL.QA..........F....EN.I.EDE........IK......-..LIGYF.RKKD.SENYR.A..........FKEAAEE-....-F-hpyipff197
49 NU1K_PODAN     I.FFIGS...V...AE.TnRA..........P....FD.L.AEA........ES......E..LVSGF.NTEH.AA---.-..........---VVFV-....-F-ffiaeyg137
50 RUVA_ECOLI     A.RLVAL...G...YK.-.--..........-....--.-.PQE........AS......R..MVSKI.ARPD.AS---.S..........ETLIREA-....-L-raal.....
51 UI5R_HSV6U     Y.EKLY-...D...FL.D.ER..........G....EF.G.SRD........LF......EatFLNN-.--SK.LTRRQ.P..........IREGLAS-....-L-qsseygek35
52 KRF1_VACCC     I.NKKIKn..N...FR.V.KHnwyy......-....--.-.--D........FH......F..FVHTL.LKTY.PEIEK.Die........FSTALEE-....-F-imctktdc41
53 UL97_HCMVA     L.DEVRN...G...TE.A.LL..........F....RN.A.GAA........CR......A..LENGK.LTH-.-----.-..........CSDACLL-....-I-laaqmsyg97
54 KRA6_ACIBA     L.TETRV...E...ER.L.VFahgditds25-....AG.L.ADE........FV......DisFVERC.LRED.ASE--eT..........ARIFLKH-....-L-kndrpdkr18
55 KRA8_ECOLI     G.DFSLD...N...LL.I.VEgkvvgcid20-....EE.F.EPS........LQe.....R..LVAQY.GIAD.PDRRK.Lq.........FHLLLDE-....-L-f.........
56 KGPB_BOVIN     A.FDIQD...L...SN.V.TL..........P....FY.PkSPQ........SK......D..LIKEAiLDND.FMKRL.E..........LSQIQEI-....-V-dcmypve162
57 EGFR_CHICK     LwNDIIDt..S...RR.P.LTvldfasn193-....SI.L.PVA........FLg.....D..AFTKT.LPLD.PKK-L.Dvfrt......VKEISGF-....-L-liqavpd291
58 KRA1_ECOLI     L.AQAQSrm.R...NG.L.VD..........A....SD.F.DDErngwpveqVV......K..ENNKL.LPFS.PDS--.-..........---VVTNG....DF-eldnlifd72
59 KDTK_DROME     N.GQLDL...D...LG.L.PLgppggprs39-....PG.L.PLS........NL......N..LL---.----pPAERN.N..........AAAANNH-....-L-gvrvnrta37
60 KPCG_HUMAN     F.VFNLK...P...GD.V.E-..........-....RR.L.SVE........V-......-..-----.WDWD.RTSRN.Dfmgamsfg..VSELLKAP....VD-gvylllnq46
```

A (cont)

Fig. 11.

negatives (41 to 43, 51 to 53) of which the first three fall immediately below our kinase cutoff. For PROFILESEARCH, there are 12 false negatives (23 to 26, 35, 38 to 41, 51 to 53) but it should be recalled that eight of these (those indicated by $ in Fig. 11B) do not appear in the results obtained from searching SWISS-PROT 25 provided to us by M. Gribskov (personal communication). We suspect that at least four (23 to 26) would be correctly classified as kinases by PROFILESEARCH leaving an estimate of three to eight false negatives. In the case of PROSITE, using our assumption of a kinase to be a true positive (T) sequence for any one of the three patterns, there are three false negatives (39, 42 to 43). However, the actual performance of the PROSITE patterns themselves is much worse; scans of SWISS-PROT 22 with each of the patterns PROTEIN_KINASE_ATP, PROTEIN_KINASE_ST and PROTEIN_KINASE_TYR individually yield 40, 2 and 3 false negatives, respectively.

The difficulty in quantifying the precise number of false positives and false negatives produced by the database discrimination tests may be illustrated by employing an alternative mechanism for assessing the number of false negatives. If simply

ID	Length	NLL-score	Z-score	HMM	PROFILE-SEARCH	Keyword	PROSITE		
							A	B1	B2
23 KLSK_HUMAN	509	1188.032	48.056	+	-$	+	T	-	T
24 KLSK_MOUSE	509	1193.879	47.376	+	-$	+	T	-	T
25 ARKB_HUMAN	689	1826.919	31.781	+	-$	+	*	*	-
26 ARKB_BOVIN	689	1827.514	31.720	+	-$	+	*	*	-
27 BYR1_SCHPO	340	808.153	27.540	+	+	-	N	T	-
28 CYGR_ARBPU	986	2839.392	22.121	+	+	-	-%	-	-
29 ANPA_RAT	1057	3062.107	21.418	+	+	-	-%	-	-
30 ANPA_HUMAN	1061	3072.615	21.390	+	+	-	-%	-	-
31 NPB_HUMAN	1047	3033.232	21.220	+	+	-	-%	-	-
32 ANPA_MOUSE	1057	3065.181	21.042	+	+	-	-%	-	-
33 ANPB_RAT	1047	3038.053	20.633	+	+	-	-%	-	-
34 CYGS_STRPU	1125	3277.621	18.745	+	+	-	-%	-	-
35 VPSF_YEAST	1454	4263.173	17.896	+	-	+	N	T	-
36 HSER_RAT	1075	3143.529	17.681	+	-	-	-%	-	-
37 HSER_HUMAN	1073	3139.039	17.552	+	-	-	-%	-	-
38 KR2_VZVD	510	1521.597	9.615	+	-	+	N	T	-
39 KR2_HSV11	518	1548.949	9.042	+	-	+	N	-	-
40 KR1_HSV11	230	710.448	6.773	+	-$	+	N	T	-
41 KR2_EBV	455	1393.761	4.935	-	-	+	T	-	T
42 KRB2_VACCV	283	880.650	4.848	-	+	+	N	N	-
43 KRB2_VACCC	283	880.753	4.838	-	+	+	N	N	-
44 AK3_ECOLI	449	1385.412	3.900	-	-	-	-	-	-
45 PSP_MOUSE	235	754.545	3.804	-	-	-	-	-	-
46 DHOM_BACSU	433	1340.413	3.706	-	-	-	-	-	-
47 FLIG_BACSU	338	1055.096	3.699	-	-	-	-	-	-
48 CALQ_RABIT	395	1229.120	3.487	-	-	-	-	-	-
49 NU1M_PODAN	368	1149.759	3.415	-	-	-	-	-	-
50 RUVA_ECOLI	203	667.519	3.413	-	-	-	-	-	-
51 U15R_HSV6U	562	1728.770	3.171	-	-$	+	T	-	T
52 KRF1_VACCC	439	1366.011	2.900	-	-$	+	N	T	-
53 UL97_HCMVA	707	2165.296	2.854	-	-$	+	N	-	T
54 KKA6_ACIBA	259	838.469	2.370	-	-	-	-	-	T
55 KKA8_ECOLI	271	885.548	1.182	-	-	-	-	-	T
56 KGPB_BOVIN	293	953.735	0.684	-	-	+	P	P	-
57 EGFR_CHICK	703	2179.703	0.065	-	-	+	P	-	P
58 KKA1_ECOLI	271	902.461	-0.467	-	-	-	-	T	-
59 KDTK_DROME	753	2334.760	-0.523	-	-	+	N	-	N
60 KPCG_HUMAN	318	1051.016	-1.486	-	-	+	P	P	-

B

Figure 11. A, Multiple sequence alignment generated by our kinase HMM of some of the sequences used to train the HMM (1 to 22) and test sequences from the SWISS-PROT 22 database (23 to 60) (see Results section (b)). Numerals appearing in the alignments indicate the number of amino acids to be inserted at that point, otherwise the notation follows the convention of Fig. 5. In Subdomain, the Roman numerals and * refer to the subdomains and residues conserved across 75 serine/threonine kinases given by Hanks & Quinn (1991). A and B in PROSITE refer to the ATP binding and catalytic regions, respectively, used to create 2 different signature patterns for kinases. X-ray identifies the location of the α-helices AA-AI and β-strands B1-B9 (read vertically) derived from the 2·7 Å crystal structure of the catalytic subunit of cAMP-dependent protein kinase (sequence 1) (Knighton *et al.*, 1991). Sequences 1 to 22 are representative kinases taken from the March 1992 Protein Kinase Catalytic Domain Database (Hanks & Quinn, 1991). These are: CAPK-ALPHA, cAMP-dependent protein kinase catalytic subunit, α-form; WEE1+, reduced size at division mutant wild-type allele gene product; TIK, mouse serine/threonine kinase; SPK1, *S. cerevisiae* kinase cloned with anti-p-Tyr antibodies; RSK1-N, amino domain of type 1 ribosomal protein S6 kinase; PYT, putative serine/threonine kinase cloned with anti-p-Tyr antibodies; PKC-ALPHA, protein kinase C, α-form; PDGFR-B, platelet-derived growth factor receptor B type; PBS2, polymix in B antibiotic resistance gene product; MIK1, *S. pombe mik1* acts redundantly with *wee1+*; MCK1, *S. cerevisiae* protein kinase; INS.R, insulin receptor; HSVK, Herpes simplex virus-US3 gene product; ERK1, rat insulin-stimulated protein kinase; EGFR, epidermal growth factor receptor (cellular homolog of *v-erbB*); ECK, receptor-like tyrosine kinase detected in epithelial cells; DPYK1, developmentally regulated tyrosine kinase in *D. discoideum*; CLK, mouse serine/threonine/tyrosine kinase; CDC2HS, human functional homolog of yeast cdc2+/CDC28; CAMII-ALPHA, calcium/calmodulin-dependent protein kinase II, α-subunit; C-SRC, cellular homolog of *v-src*; and C-RAF, cellular homolog of *v-raf/mil*. Sequences 2 to 4, 6, 10, 11, 14, 17 and 18 are the candidate dual-specificity protein kinases as defined by Lindberg *et al.* (1992). Sequences 23 to 40 are the SWISS-PROT 22 sequences designated as kinases by our HMM (*Z*-score >6·0) but not by all 3 other methods, PROSITE, PROFILESEARCH and the keyword search. Sequences 41 to 50 are the top 10 sequences below our cutoff of 6·0 and 41 to 43 and 51 to 60 are sequences that were not classified as kinases by the HMM but were so by one or more (but not all) of the 3 other methods. Note that sequences identified as kinases by all 4 methods are not shown. All sequences that are less than 200 residues in length

the number of sequences denoted as kinases only by all three other methods is evaluated, the number of false negatives for each of the techniques differ from the more detailed analysis: two for the HMM (42 to 43), seven for PROFILESEARCH (23 to 26, 35, 38, 40) and none for PROSITE (ignoring known false negatives as above). This general problem is further highlighted by the guanylyl cyclases (indicated by % in Fig. 11B). If the definition of a kinase is based upon function and not possession of particular sequence patterns, then the guanylyl cyclases are the only false positives for both the HMM and PROFILESEARCH. The PROSITE patterns PROTEIN_KINASE_ATP, PROTEIN_KINASE_ST and PROTEIN_KINASE_TYR produce eight, none and two false positives, respectively, giving some indication of the actual PROSITE performance.

Overall, both the HMM and PROFILESEARCH appear to perform generally better than PROSITE in the discrimination tests, with the HMM possibly having a slight advantage over PROFILESEARCH.

The HMM database search did not suggest any new putative kinases in SWISS-PROT 22. However, a comparative examination of the HMM produced multiple sequence alignment and the crystal structure of the catalytic subunit of cAMP-dependent protein kinase (Knighton *et al.*, 1991) (sequence 1), a template for the protein kinase family, yields insights into the conserved regions and their functions in kinases of unknown structure. Figure 11A displays the location of secondary structure elements obtained from this crystal structure. An invariant Asp in subdomain VIb (Asp166 in Knighton *et al.*, 1991) that is proposed to be the catalytic base is known to diverge in guanylyl cyclases (28 to 34, 36 to 37) even though the immediate region is highly conserved (Garbers, 1992). Our results indicate that other invariant residues appear to be replaced as well. In the sea urchin spermatozoan cell-surface receptor for the chemotactic peptide "resact" (sequences 28 and 34), a Lys in subdomain II (Lys72) that forms part of the ATP α- and β-phosphate binding site is changed to His. The heat-stable entertoxin receptor of rat (36) replaces an Asp in subdomain IX (Asp200) that contributes directly to stabilization of the catalytic loop by Glu. Yeast VPS15 (sequence 35), a probable serine/threonine kinase that is autophosphorylated, lacks many of the residues in subdomain I. In addition, a conserved ion-pair that stabilizes ATP (Glu91-Lys72) would be disrupted in VPS15 because the Glu in subdomain III is altered to Arg resulting in the apposition of two positively charged residues. In the putative B12 kinases of two strains of vaccinia virus (42 to 43), the proposed Asp catalytic base is replaced by Lys (cf. guanylyl cyclases). This is accompanied by a further change in the "general" sequence of the catalytic loop: the normally positively charged residue at $n + 2$ has been altered to Glu. In general, all the sequences below our cutoff and the last one above it (40 to 60) appear to lack α-helix F (see X-ray in Fig. 11A). The functional and or structural consequences of these modifications on any kinase activity are not clear.

(c) *EF-hand experiments*

For these experiments we used the June 1992 database of EF-hand sequences maintained by Kretsinger and co-workers (Nakayama *et al.*, 1992). Sequences in this database are proteins containing one or more copies of the EF-hand motif, a 29 residue structure present in cytosolic calcium-modulated proteins (Nakayama *et al.*, 1992; Persechini *et al.*, 1989; Moncrief *et al.*, 1990). These proteins bind the second messenger calcium and in their active form function as enzymes or regulate other enzymes and structural proteins. The motif consists of an α-helix, a loop binding a Ca^{2+} followed by a second helix. Although a number of proteins possess the EF-hand motif, some of these regions have lost their calcium-binding property.

For our training set, we extracted the EF-hand structures from each of the 242 sequences in the

have been removed. B, Details on sequences 23 to 60 shown in the alignment (arranged in order of decreasing *Z*-score). NLL-score and *Z*-score are measures of how well the kinase HMM fits these SWISS-PROT 22 test sequence that were not present in the training set (see Results section (b) for more details). In HMM, PROFILESEARCH and Keyword, + denotes sequences that are classified as containing a kinase domain and − those that do not. For PROFILESEARCH, -$ identifies sequences that do not appear in the results obtained from searching SWISS-PROT 25 (not 22 as in HMM, Keyword and PROSITE) provided to us by M. Gribskov (personal communication). Two PROSITE signature patterns for eucaryotic protein kinases have been derived and these are labeled A and B in the alignment. A is the region believed to be involved in ATP binding (PROSITE entry PROTEIN_KINASE_ATP) while B1 and B2 indicate the area important for catalytic activity in serine/threonine kinases (PROTEIN_KINASE_ST) and tyrosine kinases (PROTEIN_KINASE_TYR), respectively. In all instances, T signifies a true positive; N a false negative (a sequence which belongs to the set under consideration but which is not picked up by the pattern); P a "potential" hit (a sequence that belongs to the set but which is not picked up because the region that contains the pattern is not yet available in the data bank, i.e. a partial sequence); and ? an unknown (a sequence which possibly could belong to the set). * Indicates SWISS-PROT files which contain a cross reference to the specified PROSITE pattern, but these PROSITE entries do not contain a corresponding pointer to the SWISS-PROT file. − Signifies sequences that do not satisfy the kinase patterns and % denotes particulate forms of guanylyl cyclase receptors which contain an intracellular protein kinase-like domain but which have not been shown to possess kinase activity to date (reviewed by Garbers, 1992).

database. obtaining 885 EF-hand motifs having an average length of 29. For our first experiment we trained five HMMs on all 885 EF-hand motifs using

```
                1         11        21        31        41        51        61        71
   Structure    ..........H..H.HHHHH.H.H..........H....LL...LLL.LLLL....H.HH.HHHHHHH..........
   PROSITE      ..................................*....**...***.****....*.**..................
   Ca-binding   ..................................X.....Y....Z..y.x.......z...................
 1 CAMHS        ..........E..F.KEAFS.L.F..........D....KD...GDG.TITT....K.EL.GTVMRSL-.........
 2 aACTGG       ..........E..F.RASFN.H.F..........D....RR...RTG.NRDC....E.DF.RACLISR-.........
 3 VISININ      ..........E..L.SRWYE.G.F..........Dr...QC...SDG.RIRC....D.EF.ERIYGNF-.........
 4 TPP24CF      ..........G..L.ARFFR.R.L..........D....RD...RSR.SLDS....K.EL.QRGLAEL-.........
 5 TPHUCS       ..........E..F.KAAFD.R.F..........D....AD...GGG.DISV....K.EL.GTVMRML-.........
 6 TPAFI        ..........A..L.QKAFD.S.F..........D....TD...SKG.FITP....E.TV.GIILRMH-.........
 7 TCBP25       ..........V..A.RRIFE.M.Y..........D....KG...RKG.RIEH....T.OC.VPRITEA-.........
 8 SPEC2A       ..........L..F.KSSFK.S.E..........D....TD...GDG.KITS....E.EL.RAAFRSI-.........
 9 SCBPBLI      ..........K..I.KFTFD.F.FI.........D....YH...KDG.SIQW....E.DF.EENIKRY-.........
10 QUIDLN       ..........E..I.KDAFD.M.F..........D....ID...GDG.QITS....R.EL.RSVMKSL-.........
11 HOHSCR       ..........E..F.KEAFT.I.M..........D....QN...RDG.FIDK....N.DL.RDTFAAL-.........
12 HOHSAI       ..........E..F.KEAFL.L.F..........D....ST...GDS.KIIL....S.QV.GDVLRAL-.........
13 LPS1A        *.........A..L.KQEFR.DnY..........D....TN...KDG.TVSC....A.EL.VKLMMWT-.........
14 LAVI         ..........A..L.VADFR.K.I..........D....TN...SNG.TLSR....K.EF.REHFVRL-.........
```

[redacted]

```
18 CMSE         ..........R..L.RKRFD.R.W..........D....FD...GNG.ALER....A.DF.EKEAQHI-.........
19 CDPK         ..........G..L.KELFK.M.I..........D....TD...NSG.TITF....D.EL.KDGLKRV-.........
20 CDC31        ..........E..I.VEAFS.L.F..........D....NN...NDG.FLDY....H.EL.KVAMKAL-.........
21 CALPLMS      ..........T..C.RSMVA.V.M..........D....SD...TTG.KLGF....E.EF.KYLWNNI-.........
22 CALCIB       ..........R..L.GKRFK.K.L..........D....LD...NSG.SLSV....E.EF.NS-LPEL-.........
23 CALBNGG      ..........Q..F.FEIWN.H.Y..........D....SD...GNG.YMDG....K.EL.QNFIQEL-.........
24 CALICE       ..........E..F.REAFR.M.F..........D....KD...GNG.TIST....K.EL.GIAMRSL-.........
25 BCMS         ..........A..L.IDVFH.Q.Y..........Sg...KE...GDKhKLKK....S.EL.KELINNE-.........
26 AEQAVI       ..........R..H.KHMFN.F.L..........D....VN...HNG.KISL....D.EM.VYKASDI-.........
27 1FO          ..........R..R.IELFK.K.F..........D....KN...ETG.KLCY....D.EV.NSGCLEV-.........
28 CALM_ASPNI   adalteeqvsE..Y.KEAFS.L.F..........D....KD...GDG.QITT....K.EL.GTVMRSL-gqnpsesl09
29 MLE1_HUMAN   apkkdvkl29D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
30 MLE1_RABIT   apkkdvkl27D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkees35
31 MLEV_HUMAN   apkkpepl30D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gerltede35
32 MLEC_CHICK   ppkkpepl29D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLRTL-gerlteee35
33 MLEV_RAT     apkkpepl35D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gerltede35
34 MLE1_CHICK   pkkdvkkl26D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmteee35
35 MLE1_RAT     apkkdvkl24D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
36 MLE1_MOUSE   apkkdvkl23D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
37 MLEF_HUMAN   apkkpepl32D..F.VEGLR.V.F..........D....KE...SNG.TVMG....A.EL.RHVLATL-gekmteae35
38 MLEF_RAT     ppkkpepl28D..F.VEGLR.V.F..........D....KE...SNG.TVMG....A.EL.RHVLATL-gekmseae35
39 MLEF_MOUSE   ppkkpepl28D..F.VEGLR.V.F..........D....KE...SNG.TVMG....A.EL.RHVLATL-gekmseae35
40 MLEX_CHICK   mplkkpdl2lD..F.VEGLR.V.F..........D....KE...GNG.LVMG....A.EL.RHVLVTL-gekmteee35
41 MLE3_HUMAN   afaadqia85D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
42 MLEY_HUMAN   mppkkdvl44D..Y.LEGFR.V.F..........D....KE...GNG.KVMG....A.EL.RHVLTTL-gekmteee35
43 MLE3_RABIT   afaadqia85D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
44 MLE3_RAT     afaadqia85D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
45 MLE3_MOUSE   afaadqia85D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmkeee35
46 MLE3_CHICK   afapdqid85D..F.VEGLR.V.F..........D....KE...GNG.TVMG....A.EL.RHVLATL-gekmteee35
47 AACT_HUMAN   mdhydsq749E..F.RASFN.H.F..........D....RD...NSG.TLGP....E.EF.KACLISL-gydigndll4
```

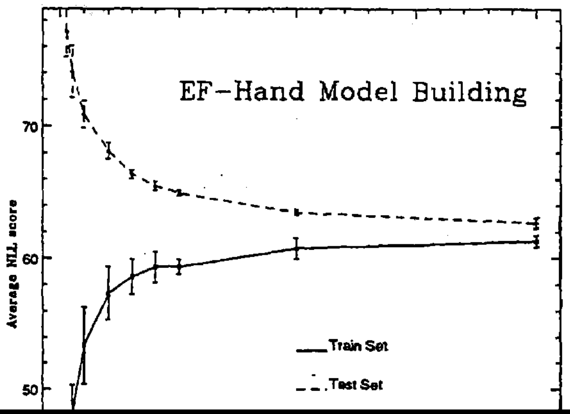
```
79 CAP2_HUMAN   magiaakS75T..C.KIKVD.K.L..........D....SD...GSG.KLGL....K.EF.YILWTKl-qkyqkiyr96
80 ADCL_PIG     mskergll57I..L.QENHK.E.I..........D....YD...GSG.SYSL....A.EV.LRAGATT-vpllvl1548
81 SCPA_PENSP   aysvdar103F..I.ANQFK.A.I..........D....VN...GDG.KYGL....D.EYrLDCITRS-afarvhei59
82 SCPB_PENSP   aysvdarv59L..V.HEIAE.L.A..........D....FN...KDG.EYTV....D.EF.KQAVQKH-ckgkafa104
83 IPYR_ARATH   mseikde169E..I.RAFFE.D.Y..........K....KN...EHK.KVDV....E.AF.LPAQAAI-daikdwmd65
84 SCP1_BRALA   glndfqk105R..I.PFLFK.G.M..........D....VS...GDG.IVDL....E.EF.QNYCKNF-qlqcadvp51
85 SCP2_BRALA   glndfqk105K..I.PFLFK.G.M..........D....VS...GDG.IVDL....E.EF.QNYCKNF-qlqcadvp51
86 PIP3_RAT     mdsgrdf143V..I.HSCLR.K.A..........D....KN...KDN.KKMF....K.EL.KDFLKEL-niqvddg504
87 AACT_CHICK   mdhhydp706E..F.ARINS.I.V..........D....PN...RHG.VVTF....Q.AF.IDFMSRE-tadtdtad73
88 CAB_MOUSE    marpleea53A..F.QKVHS.N.L..........D....SN...RDN.EVDF....Q.EY.CVFLSCI-ammcaeff19
```

ID	Length	NLL-score	Z-score	HMM	PROFILESEARCH		Keyword	Prosite
					Gribskov	HMM		
28 CALM_ASPNI	148	398.961	12.975	+	-	-	+	T
29 MLE1_HUMAN	193	542.924	11.662	+	+	+	-	%
30 MLE1_RABIT	191	537.011	11.661	+	+	+	-	%
31 MLEV_HUMAN	194	546.027	11.631	+	+	+	-	%
32 MLEC_CHICK	193	543.095	11.605	+	+	+	-	%
33 MLEV_RAT	199	561.007	11.561	+	+	+	-	%
34 MLE1_CHICK	190	534.042	11.516	+	+	+	-	%
35 MLE1_RAT	188	528.051	11.262	+	+	+	-	%
36 MLE1_MOUSE	187	525.056	11.224	+	+	+	-	%
37 MLEF_HUMAN	196	554.316	11.005	+	+	+	-	%
38 MLEF_RAT	192	542.332	10.892	+	+	+	-	%
39 MLEF_MOUSE	192	542.332	10.892	+	+	+	-	%
40 MLEX_CHICK	185	521.797	10.342	+	+	+	-	%
41 MLE3_HUMAN	149	411.100	10.201	+	+	+	-	%
42 MLEY_HUMAN	208	588.847	10.194	+	+	+	-	%
43 MLE3_RABIT	149	411.179	10.177	+	+	+	-	%
44 MLE3_RAT	149	411.207	10.169	+	+	+	-	%
45 MLE3_MOUSE	149	411.208	10.169	+	+	+	-	%
46 MLE3_CHICK	149	411.206	10.169	+	+	+	-	%
47 AACT_HUMAN	892	2642.237	9.957	+	-	+	+	T
48 MLE_HALRO	151	418.497	9.918	+	+	+	-	%
49 MLES_HUMAN	151	418.627	9.879	+	+	+	-	%
50 MLEN_HUMAN	151	418.627	9.879	+	+	+	-	%
51 MLEN_CHICK	150	415.631	9.798	+	+	+	-	%
52 MLEM_CHICK	150	415.631	9.798	+	-	-	-	%
53 MLEG_HUMAN	94	248.725	9.735	+	+	+	-	%
54 MLE_PATYE	156	433.703	9.629	+	+	+	-	%
55 MLE_AEQIR	156	433.703	9.629	+	+	+	-	%
56 AACT_DROME	895	2653.286	9.130	+	-	+	+	T
57 RECO_CHICK	192	548.396	8.848	+	-	-	+	T
58 MLE_DICDI	166	465.170	8.834	+	+	+	-	T
59 SPCA_DROME	2415	7205.568	8.787	+	-	+	+	T
60 MLR_DICDI	161	451.967	8.678	+	+	+	+	-
61 MLE_TODPA	159	446.406	8.616	+	+	+	-	%
62 SPCN_CHICK	2477	7392.895	8.157	+	-	+	-	T
63 CL1L_MOUSE	96	263.095	7.516	+	+	+	-	%
64 AACS_CHICK	897	2663.548	7.446	+	-	-	+	-
65 CL1L_RAT	94	257.103	7.423	+	+	+	-	%
66 LAV1_PHYPO	355	1039.236	7.298	+	-	+	-	T

[redacted]

ID	Length	NLL-score	Z-score	HMM	PROFILESEARCH		Keyword	Prosite
					Gribskov	HMM		
81 SCPA_PENSP	192	556.636	6.071	+	-	+	+	T
82 SCPB_PENSP	192	557.071	5.924	+	-	+	+	T
83 IPYR_ARATH	263	769.241	5.909	+	-	-	-	-
84 SCP1_BRALA	185	535.787	5.827	+	-	+	+	T
85 SCP2_BRALA	185	535.816	5.818	+	-	+	+	T
86 PIP3_RAT	756	2244.255	5.713	+	-	-	-	?
87 AACT_CHICK	888	2641.411	5.684	+	-	-	+	N
88 CAB_MOUSE	101	284.695	5.589	+	-	-	+	-
89 TEGU_SCHMA	190	552.242	5.469	+	-	+	-	?
90 CAB_RAT	101	285.488	5.369	+	-	-	-	-
91 G19P_HUMAN	527	1560.198	5.330	+	-	-	-	T
92 TCH2_ARATH	45	116.235	5.321	+	-	-	+	T
93 KDGL_HUMAN	735	2182.343	5.301	+	-	-	+	T
94 PIP3_BOVIN	695	2063.206	5.034	+	-	-	-	?
95 CALM_LYTPI	30	67.341	4.942	+	-	-	+	P
96 CAP1_HUMAN	714	2120.342	4.924	+	-	+	+	T
97 CIC1_CYPCA	1852	5530.321	4.714	-	+	-	-	-
98 GUNF_CLOTM	739	2196.618	4.602	-	-	-	-	?
99 CIC1_RABIT	1873	5593.640	4.550	-	+	-	-	-
100 V57A_BPT4	80	224.359	4.470	-	-	-	-	-
101 CALG_CHICK	65	178.908	4.438	-	+	+	+	T
102 NIFH_NOSCO	86	243.556	4.347	-	-	-	-	-
103 ARFL_DROME	180	524.609	4.300	-	-	-	-	-
104 AROA_KLEPN	427	1264.280	4.296	-	-	-	-	-
105 REL1_HUMAN	185	540.676	4.249	-	-	-	-	-
106 H11_BOVIN	104	298.227	4.240	-	-	-	-	-
107 YCSX_CHLPY	110	316.022	4.210	-	-	-	-	-
108 DP3X_ECOLI	643	1910.667	4.186	-	-	-	-	-
109 AROA_SALTY	427	1264.760	4.130	-	-	-	-	-
110 ANX1_CAVCU	346	1022.514	4.043	-	-	-	-	-
111 CICC_RAT	2169	6481.468	4.011	-	+	-	-	-
112 CICC_RABIT	2171	6487.460	4.010	-	+	-	-	-
113 LACA_LACLA	141	407.967	3.986	-	-	-	-	-
114 AROA_BORPE	442	1310.475	3.985	-	-	-	-	-
115 AROA_SALTI	427	1265.295	3.945	-	-	-	-	-
116 AROA_SALGL	427	1265.295	3.945	-	-	-	-	-
117 CAP1_CHICK	704	2093.590	3.888	-	-	-	+	T

There is considerable overlap between this training set and the EF-hand motifs found in SWISS-PROT 22, so in order to provide some clearer cross validation of our results we also did another series of experiments. In these experiments, models were estimated using training sets consisting of different numbers of randomly chosen EF-hand sequences from the database of 885 EF-hand sequences. For training sets consisting of 5, 10, and 20 random EF-hand sequences, 15 models were estimated, each using a different randomly chosen training set. For training sets consisting of 40, 80, 100, 200, and 400 random EF-hand sequences, five models were estimated. In all, 70 models were estimated. A model's performance after training was gauged on how well



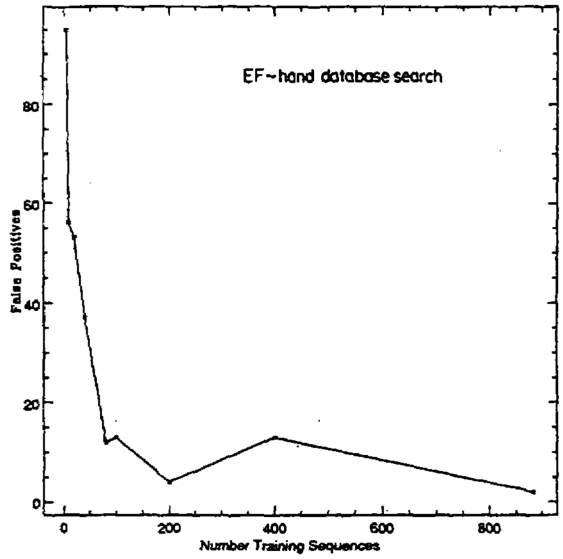


Figure 15. EF-hand database search false positives for models trained with 5, 10, 20, 40, 80, 100, 200, 400 and 885 sequences.

appears substantially when the training set size reaches about 100 sequences.

4. Discussion

A new method to model protein families using hidden Markov models has been introduced. The method is capable of tapping into the tremendous

[redacted]

might even be lowered further. However, there will be a limit on how small the number of available sequences can be if one hopes to obtain a reasonable model starting from a *tabula rasa*.

We believe that the answer to the problem of small training sets is to add more prior knowledge into the training process. One way to do this is by starting with a better initial model. We have performed several experiments in which we have started with a model obtained from a small set of aligned sequences, and then trained the model further using a larger set of unaligned sequences. These will be reported in a future paper. We find that this technique can often give better results. This also suggests that one application of HMMs may be in maintaining multiple alignments as the number of sequences in the alignment grows. Each time new sequences are added to a dataset of homologous sequences, we can begin with the HMM based on the alignment of the previous set of sequences, train it with the larger dataset that includes the new sequences, and then create a new multiple alignment for the larger dataset from this HMM. Not only will the new sequences be included in the new alignment, but the alignment of the old sequences may be improved by utilizing the statistical information present in the larger dataset.†

Another way to add more prior knowledge into the training process is to use a more sophisticated Bayesian prior. We are currently exploring the use of a prior on the probability distribution over the amino acids in a match state of the model consisting of a mixture of Dirichlet priors (Brown *et al.*, 1993). Using such a prior is like "soft-tying" the distribuprobabilities between amino acids into account. It also remains to be seen whether or not incorporating any of these extensions into the HMM approach will yield even better results.

We also believe that some of the errors made by our HMM models are due to the fact that these models are suboptimal, in the sense that their NLL-scores are not as low as they could be. This is because the EM procedure is not guaranteed to find the globally optimal model for a given training set. In other experiments, reported by Haussler *et al.* (1993), we trained an HMM for globins beginning with a model derived from the Bashford *et al.* (1987) alignment, and obtained a slightly lower NLL-score than any model from our experiments using EM on unaligned training sequences (208 compared to 210·3). Hence, we know that EM is not locating the globally optimal model in this case. An important open problem is to find a reliable way to prevent EM from getting stuck and returning a suboptimal solution.

Another issue is the adequacy of the hidden Markov model itself as a statistical model of the sequence variation within a protein family. Clearly an HMM provides at best a "first order" model of sequence variation. There are many kinds of interactions in proteins that are not easily modeled by HMMs, for example, pairwise correlations between amino acid distributions in positions that are widely separated in the primary sequence, but close in the three-dimensional structure (see e.g. Klinger & Brutlag (1993)). It would be very valuable to have more general models that incorporate such interactions while still remaining computationally tractable. We are currently exploring the potential of one model class of this type to capture the base-

of the PROSITE-indexed domains in a single long protein, using the Viterbi algorithm. The remaining portions of the sequence could be marked as "unknown". While this would not constitute a complete parse of the sequence, it would be very useful in providing some automatic annotation of new sequences, which is of critical importance as the rate of growth of the protein databases continues to accelerate. A related approach to protein annotation is given by Stultz *et al.* (1993), and a related HMM-based DNA parser for *E. coli* is described by Krogh *et al.* (1993*b*).

A comparative examination of the HMM produced kinase multiple sequence alignment and the crystal structure of the catalytic subunit of cAMP-dependent protein kinase (Knighton *et al.*, 1991) indicates a number of conserved residues in kinases of unknown structure that may be suitable for further experimental study (see Results section (b)). Results from our database discrimination tests suggest the presence of an EF-hand calcium-binding motif in a highly conserved and evolutionary preserved putative intracellular region of 155 residues in the α-1 subunit of L-type calcium channels which play an important role in excitation-contraction coupling (see Results section (c)). This region has been suggested to contain the functional domains that are typical or essential for all L-type calcium channels regardless of whether they couple to ryanodine receptors, conduct ions or both. Our EF-hand HMM indicates the following proteins may also possess this motif: chicken myosin light chain alkali (smooth muscle), bovine calpactain I light chain, *Arabidopsis thaliana* inorganic pyrophosphatase, rat placental calcium-binding protein and rat and bovine 1-phosphatidylinositol-4,5-bisphosphate

References

Abe, N. & Warmuth, M. (1990). On the computational complexity of approximating distributions by probabilistic automata. In *Proceedings of the 3rd Workshop on Computational Learning Theory*, pp. 52–66. Morgan Kaufmann, Rochester, NY.

Allison, L., Wallace, C. S. & Yee, C. N. (1992). Finite-state models in the alignment of macromolecules. *J. Mol. Evol.* 35, 77–89.

Asai, K., Hayamizu, S. & Onizuka, K. (1993). HMM with protein structure grammar. In *Proceedings of the Hawaii International Conference on System Sciences*, pp. 783–791. IEEE Computer Society Press, Los Alamitos, CA.

Bairoch, A. (1992). Prosite: a dictionary of sites and patterns in proteins. *Nucl. Acids Res.* 20, 2013–2018.

Baldi, P. & Chauvin, Y. (1993). A smooth learning algorithm for hidden Markov models. *Neural Computation*, in the press.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1993). Hidden Markov models in molecular biology: new algorithms and applications. In *Advances in Neural Information Processing Systems 5* (Hanson, Cowan & Giles, eds), pp. 747–754, Morgan

their endocrine, paracrine and autocrine ligands. *Cell*, 71, 1–4.

Geman, S., Bienenstock, E. & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1–58.

Grabner, M., Friedrich, K., Knaus, H.-G., Striessnig, J., Scheffauer, F., Staudinger, R., Koch, W. J., Schwartz, A. & Glossmann, H. (1991). Calcium channels from *Cyprinus carpio* skeletal muscle. *Proc. Nat. Acad. Sci., U.S.A.* 88, 727–731.

Gribskov, M., Lüthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* 183, 146–159.

Hanks, S. K. & Quinn, A. M. (1991). Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* 200, 38–62.

Hanks, S. K., Quinn, A. M. & Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domain. *Science*, 241, 42–52.

Haussler, D. & Krogh, A. (1992). Protein alignment and clustering. Presented at the conference Neural Networks for Computing.

Haussler, D., Krogh, A., Mian, I. S. & Sjölander, K.

Dual-specificity protein kinases: will any hydroxyl do? *Trends Biochem. Sci.* 17. 114–119.

Lüthy, R., McLachlan, A. D. & Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring table in searching protein sequence databases for structural similarities. *Proteins: Struct. Funct. Genet.* 10, 229–239.

Moncrief, N. D., Kretsinger, R. H. & Goodman, M. (1990). Evolution of EF-hand calcium-modulated proteins. I. Relationships based on amino acid sequences. *J. Mol. Evol.* 30, 522–562.

Nakayama, S., Moncrief, N. D. & Kretsinger, R. H. (1992). Evolution of EF-hand calcium-modulated proteins. II. Domains of several subfamilies have diverse evolutionary histories. *J. Mol. Evol.* 34, 416–448.

Nowlan, S. (1990). Maximum likelihood competitive learning. In *Advances in Neural Information Processing Systems* (Touretsky, D., ed). vol. 2, pp. 574–582. Morgan Kaufmann. San Mateo. CA.

Nowlan, S. J. & Hinton, G. E. (1992). Soft weight-sharing. In *Advances in Neural Information Processing Systems 4* (Moody. Hanson & Lippmann. eds). Morgan Kauffmann Publishers. San Mateo, CA.

Persechini, A., Moncrief, N. D. & Kretsinger, R. H. (1989). The EF-hand family of calcium-modulated proteins. *Trends Neurosci.* 12 (11), 462–467.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77 (2). 257–286.

Sakakibara. Y.. Brown, M., Underwood, R.. Mian. I. S. & Haussler. D. (1993). Stochastic context-free grammars for modeling RNA. Technical Report UCSC-CRL-93-16 University of California at Santa Cruz, Computer Science Dept., Santa Cruz, CA 95064.

Sibbald, P. & Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* 216, 813–818.

Stultz, C. M., White, J. V. & Smith, T. F. (1993). Structural analysis based on state-space modeling. *Protein Sci.* 2, 305–315.

Subbiah. S. & Harrison, S. C. (1989). A method for multiple sequence alignment with gaps. *J. Mol. Biol.* 209, 539–548.

Tanaka, H., Ishikawa, M., Asai, K. & Konagaya, A. (1993). Hidden Markov models and iterative aligners. In *First International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park.

Taylor, W. R. (1986). The classification of amino acid conservation. *J. Theoret. Biol.* 119. 205–218.

Vingron. M. & Argos. P. (1991). Motif recognition and alignment for many sequences by comparison of dot-matrices. *J. Mol. Biol.* 218, 33–43.

Waterman, M. S. (1989). Sequence alignments. In *Mathematical Methods for DNA Sequences* (Waterman, M. S., ed.). CRC Press, Boca Raton, FL.

Waterman. M. S. & Perlwitz, M. D. (1986). Line geometries for sequence comparisons. *Bull. Math. Biol.* 46. 567–577.

White, J. V., Stultz, C. M. & Smith, T. F. (1991). Protein classification by nonlinear optimal filtering of amino-acid sequences. Unpublished manuscript.

Edited by F. Cohen

(Received 4 January 1993; accepted 23 September 1993)