

**Running title:** Maximum discrimination HMMs

# Maximum Discrimination Hidden Markov Models of Sequence Consensus

Sean R. Eddy\*, Graeme Mitchison, and Richard Durbin

Laboratory of Molecular Biology

Hills Road

Cambridge CB2 2QH

England

Phone: +44-223-402010

FAX: +44-223-402008

sre@mrc-lmb.cam.ac.uk, gjm@mrc-lmb.cam.ac.uk, rd@sanger.ac.uk

December 6, 1994

## **Abstract**

We introduce a maximum discrimination method for building hidden Markov models (HMMs) of protein or nucleic acid primary sequence consensus. The method compensates for biased representation in sequence data sets, superseding the need for sequence weighting methods. Maximum discrimination HMMs are more sensitive for detecting distant sequence homologues than various other HMM methods or BLAST when tested on globin and protein kinase catalytic domain sequences.

Keywords: hidden Markov model, database searching, sequence consensus, sequence weighting

# Introduction

Only about a third of new predicted protein sequences are convincingly similar to another sequence (Bork *et al.*, 1992; Green *et al.*, 1993) using current database searching algorithms based on pairwise comparisons (Altschul *et al.*, 1990; Pearson and Lipman, 1988). A promising approach towards improving the sensitivity of database searching is to use consensus models built from multiple sequence alignments of protein sequence families (Bairoch, 1993; Baldi *et al.*, 1994; Barton, 1990; Gribskov *et al.*, 1987; Henikoff and Henikoff, 1994a; Krogh *et al.*, 1994; Taylor, 1986). In principle, a consensus model can exploit additional information about a protein structure, such as the location and identity of conserved residues and varying probabilities of insertion and deletion.

In practice, unbiased consensus models are difficult to build, because sequence collections are rarely a fair representation of the diversity of sequences consistent with a given protein structure. For various reasons, including the current interests of the research community, sequence datasets are usually biased by the overrepresentation of certain sequence subfamilies. For instance, 65% of available globin sequences are vertebrate  $\alpha$ - and  $\beta$ -globins, but other distinct subfamilies of globins are represented by only a few sequences. A consensus model naively built from a multiple sequence alignment of all known globins tends to describe the vertebrate hemoglobin sequences well, but invertebrate globins (for instance) are poorly recognized. The obvious, intuitive solution is to weight distant sequences more highly than closely similar ones – but how best to calculate the weights? Several different weighting schemes have been proposed (Altschul *et al.*, 1989; Gerstein *et al.*, 1994; Henikoff and Henikoff, 1994b; Thompson *et al.*, 1994; Sibbald and Argos, 1990). All generally tend to produce more sensitive consensus models (Thompson *et al.*, 1994; Luthy *et al.*, 1994).

If one’s goal in building the consensus model is to recognize distant sequence homologues, a practical objective definition of the best consensus model is the one which discriminates as many as possible of the known example sequences of a family from unrelated sequences. This idea of “maximum discrimination” is mathematically well-defined. It is different from the usual maximum likelihood parameter estimation criterion, which seeks to maximize the overall likelihood of the data set and will be biased by overrepresented sequences. A maximum discrimination consensus model trades off the probability of easily recognized sequences in order to distinguish distantly related sequences from the background of unrelated sequences. The more difficult it is to distinguish a known example sequence from background, the more attention is paid to that sequence when constructing the model. All the proposed weighting methods work in the right direction, but they

are to varying degrees only approximations to the true goal of a maximum discrimination consensus model.

We describe algorithms for building probabilistic maximum discrimination consensus models using hidden Markov models(HMMs). HMMs have been introduced from the speech recognition field as a powerful framework for modeling primary sequence consensus (Krogh *et al.*, 1994). Other descriptions of primary sequence consensus, such as profiles (Grib-skov *et al.*, 1987), flexible patterns (Barton, 1990), templates (Taylor, 1986), and blocks (Henikoff and Henikoff, 1994a), can be described within the HMM formalism. In contrast to these previous methods, HMMs allow a consistent probabilistic treatment of deletions and insertions, and they have the ability to iteratively refine an imperfect initial multiple alignment during the model training process (Krogh *et al.*, 1994). Another advantage, which we exploit here, is that HMMs are well-grounded in probability theory, allowing us to explore alternative model training criteria with different probabilistic interpretations. HMMs of sequence consensus have been trained previously by maximum likelihood criteria (Baldi *et al.*, 1994; Krogh *et al.*, 1994). Using HMMs, it is straightforward to write down an objective function to represent the maximum discrimination concept and to find algorithms for optimizing this objective function. One algorithm we describe here yields what can be interpreted as a set of sequence weights, arrived at by a method quite unrelated to other proposed weighting rules. We have compared the performance of maximum discrimination HMMs to other HMM methods and BLAST (Altschul *et al.*, 1990), using globin sequences as an example.

# Methods and algorithms

## Hidden Markov models

An HMM is a probabilistic model composed of a number of interconnected states, each of which emits an observable output (here, an amino acid or a nucleotide). Each state has two kinds of parameters. Symbol emission probabilities describe the probabilities of the possible outputs from the state, and state transition probabilities specify the probability of moving to a new state from the current one. An observed sequence is generated by starting at some initial state and moving probabilistically from state to state until some terminal state is reached, emitting observables from each state that is passed through. A sequence of states is a first-order Markov chain, but this state sequence is “hidden”, only the sequence of symbols that it emits being observable; hence the term “hidden Markov model” (Rabiner, 1989).

Figure 1 diagrams the structure of a hidden Markov model for modeling multiple alignments of biological sequences, as introduced by (Krogh *et al.*, 1994). One “match” state is assigned to each consensus column of the multiple alignment. “Insert” states between the match states allow for insertions relative to the consensus, and “delete” states allow for consensus positions to be skipped. The structure is repetitive, and only the number of match states needs to be specified to define all the states and how they are interconnected by state transitions. To build an HMM from a multiple sequence alignment, columns in the alignment are assigned either to match/delete or to insert states. Either the symbols in a column were produced by a match state and the gaps by a delete state, or the symbols were produced by an insert state and the gaps don’t have to be accounted for. This choice can be made heuristically based on the frequency of gaps (Krogh *et al.*, 1994), or an optimal (maximum *a posteriori*, MAP) structure can be found (S.R.E. and R.D., manuscript in preparation). This gives an assignment of the amino acids of every column onto either a match state or an insert state, so observed counts  $c_{x,k}$  are known for every amino acid  $x$  at every match or insert state  $k$ . Similarly, counts for state transitions from every state are determined directly from the alignment.

In this paper, the multiple alignment is not changed during the process of determining the optimal symbol emission and state transition probability parameters. The observed counts of symbols and state transitions are completely known from the outset. Our iterative algorithms are not to be confused with the usual training of HMMs from *unaligned* data by expectation maximization (EM) or gradient descent, such as the iterative multiple sequence

alignment procedures described for sequence HMMs (Baldi *et al.*, 1994; Krogh *et al.*, 1994), in which the counts of symbol emissions and state transitions are initially unknown. We are solely concerned with the problem of estimating probability parameters from observed counts. For simplicity, we will only describe equations for finding the symbol emission probabilities  $p_x$  for one state  $k$  given observed amino acid counts  $c_x$ . The equations for the remaining states and for the state transition probabilities are analogous.

## Maximum likelihood and MAP parameter estimation

Here, we briefly sketch the usual methods of estimating parameters on an HMM or other probabilistic models, before introducing maximum discrimination estimation. According to Bayes' theorem, the probability of a model  $M$  given some sequences  $S = (S_1, \dots, S_N)$  is:

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)} \quad (1)$$

$P(S | M)$  is referred to as the likelihood of the model, and is defined by  $P(S | M) = \prod_i P(S_i | M)$  where the  $P(S_i | M)$  are the likelihood of the model for each sequence.  $P(M)$  is the prior probability of the model.  $P(S)$  is  $P(S | M_j)P(M_j)$  integrated over all possible alternative models for the data  $M_j$ .

The alignment of the sequence to the model is known from the multiple alignment, so we calculate  $P(S_i | M)$  as the product of all the probabilities for symbol emissions and state transitions in that alignment. Hence, we use the approximation that  $P(S_i | M)$  is equal to the likelihood of a single alignment, rather than the sum over all possible alignments.

In probabilistic modeling, one is generally seeking a model that optimizes  $P(M | S)$ . However,  $P(S)$  is not trivial to calculate, as it generally requires a difficult integration. Approximations are often made in order to find good probabilistic models. Maximum likelihood (ML) estimation seeks to optimize the likelihood  $P(S | M)$  instead of the probability  $P(M | S)$ ; or, in terms of the log likelihoods of individual sequences, it seeks to optimize the sum of their log likelihoods  $\sum_i \log P(S_i | M)$ .

The maximum likelihood estimation of the symbol emission probabilities  $p_x$  sets them to the observed frequencies,  $c_x / \sum_{y=1}^{20} c_y$ . This has the undesirable property that unobserved amino acids will be assigned a probability of zero. As discussed by (Krogh *et al.*, 1994), this problem can be avoided by including the term  $P(M)$  and thus incorporating prior information about expected amino acid frequencies, setting  $p_x = (c_x + \alpha m_x) / \sum_{y=1}^{20} (c_y + \alpha m_y)$ .  $m_x$  and  $\alpha$  define Dirichlet prior distributions on the parameters;  $m_x$  are the expected mean frequencies of amino acids and  $\alpha$  is a constant that defines how much weight is attached

to the prior relative to the data. This is the maximum *a posteriori* (MAP) estimate for  $p_x$ . The terms  $\alpha m_x$  are sometimes referred to as “pseudocounts” for symbols  $x$ .

A weighted maximum likelihood or weighted MAP solution assigns weights  $w_i$  to each sequence  $i$  of  $N$  total sequences in the training data set. These weights are chosen so as to try to compensate for underrepresented subfamilies in the training data. There are a number of heuristic weighting rules for sequence data (Altschul *et al.*, 1989; Gerstein *et al.*, 1994; Henikoff and Henikoff, 1994b; Thompson *et al.*, 1994; Sibbald and Argos, 1990). The counts become  $\sum_{i=1}^N w_i \delta_{i,x}$ , where  $\delta_{i,x}$  is 1 if sequence  $i$  has a symbol  $x$  at this state, and 0 otherwise. The MAP estimate for  $p_x$  for the weighted case is:

$$p_x = \frac{\sum_{i=1}^N w_i \left( \delta_{i,x} + \frac{\alpha}{N} m_x \right)}{\sum_{y=1}^{20} \sum_{i=1}^N w_i \left( \delta_{i,y} + \frac{\alpha}{N} m_y \right)} \quad (2)$$

Note that the prior is apportioned to each individual training sequence, a detail which appears again in the maximum discrimination case. The relative weight of prior versus data remains the same as in the unweighted case.

## Maximum discrimination parameter estimation

In maximum discrimination (MD), we imagine that sequences fall into two classes: family members and unrelated proteins. We consider the case in which all the example sequences belong to the first class (the sequence family), and we will seek an HMM  $M$  that optimally distinguishes all these examples from a fixed model  $R$  of the second class (which we will refer to as the “random model”). One can envisage more complicated discriminative training, such as using real examples of unrelated sequences, but we have not explored this avenue.

The key idea of MD is that we will require every example sequence to match the model. We seek a model that maximizes the probability that *all* of the sequences match, as opposed to all other combinations of match/doesn’t match for the individual sequences. In maximum likelihood or MAP estimation, a few unlikely sequences does not greatly affect parameter estimation, because the total log likelihood summed over all the data can remain high. In maximum discrimination, even a single unrecognized sequence can have a dominant effect on the probability that all of the sequences are recognized. An MD model must pay particularly close attention to the difficult, divergent sequences in the example data in order to successfully distinguish all the examples from the competing random model.

The probability that an individual sequence  $S_i$  matches the model  $M$  as opposed to the competing random model  $R$  is calculated from Bayes' theorem:

$$P(M | S_i) = \frac{P(S_i | M)P(M)}{P(S_i | M)P(M) + P(S_i | R)P(R)} \quad (3)$$

We assume that the random model  $R$  produces sequences of the same amino acid composition of the overall SwissProt protein database. This random model is a standard assumption in sequence analysis; it underlies log-odds based scoring calculations such as the Dayhoff PAM matrices for scoring pairwise sequence alignments (Altschul, 1991). There are no state transitions in the random model.

For simplicity, we further assume that the models  $M$  and  $R$  are *a priori* equally likely, so that the prior terms  $P(M)$  and  $P(R)$  cancel. Lower prior ratios of  $P(M)/P(R)$  would force maximum discrimination to discriminate example sequences even more strongly from the random model, but we have not explored this. This assumption ignores the prior on the parameters of the model, which has the important feature in MAP estimation of preventing any probability from becoming zero. We have not seen any entirely satisfactory way of incorporating a Dirichlet prior on the parameters of the model into MD. We come back to this issue later (see below).

Let  $D$  be the probability that *all* the individual sequences are matched by the model and none by the random model, as opposed to all other  $2^N - 1$  possible ways of labeling the individual sequences as matching the model or the random model:

$$D = \prod_{i=1}^N \frac{P(S_i | M)}{P(S_i | M) + P(S_i | R)} \quad (4)$$

The goal of MD estimation is to find a model which maximizes  $D$  with respect to the model's parameters. There is apparently no analytical solution, as there was for the ML or MAP cases, but we can use an iterative re-estimation method akin to expectation maximization (Dempster *et al.*, 1977). The re-estimation equation for each iteration of the algorithm is obtained by calculating the partial derivatives of  $D$  with respect to all the parameters and using Lagrange multipliers to find a constrained optimum for  $D$ , the constraints being that all probability vectors must sum to 1. This gives equations of the following form for re-estimating symbol emission probabilities  $p'_x$  for the 20 amino acids, based on the observed occurrences of symbols  $\delta_{i,x}$ :

$$p'_x = \frac{\sum_{i=1}^N (1 - P(M | S_i)) \delta_{i,x}}{\sum_{y=1}^{20} \sum_{i=1}^N (1 - P(M | S_i)) \delta_{i,y}} \quad (5)$$

Interestingly, this means the discriminative rule implies re-estimation equations which are analogous to the solution for a weighted ML solution. The weight  $w_i$  is replaced by the probability that sequence  $i$  is not recognized by the model,  $1 - P(M | S_i)$ . Thus well-recognized sequences hardly contribute to an MD model; the model is constructed from a selected set of divergent, hard-to-recognize sequences which suffice to optimally recognize the entire set.

We now exploit this analogy to the weighted ML solution to bring Dirichlet prior terms into the MD parameter re-estimation equation in the same way they appear in the weighted MAP solution. This is *ad hoc* but it is useful for assuring non-zero parameters:

$$p'_x = \frac{\sum_{i=1}^N (1 - P(M | S_i)) \left( \delta_{i,x} + \frac{\alpha}{N} m_x \right)}{\sum_{y=1}^{20} \sum_{i=1}^N (1 - P(M | S_i)) \left( \delta_{i,y} + \frac{\alpha}{N} m_y \right)} \quad (6)$$

In our experiments,  $m_x$  are set to the average occurrence frequency of amino acids in the SwissProt database, and  $\alpha$  is set to 20 for match states and 100000 for insert states (which has the effect of fixing the insert state emission distribution to the background distribution). Our state transition priors follow those of (Krogh *et al.*, 1994).

Iterative re-estimation of the parameter values and rescoring of the example sequences using the re-estimated model converges to a local optimum of  $D$ . We found we needed to damp out oscillatory behavior in the convergence, which we did by constructing the model for the next iteration as a weighted average of the previous model and the newly reestimated model. A weighted average of 0.99 of the previous model and 0.01 of the reestimated model gives stable convergence in all the cases we have studied.

An example of the MD rule applied to a toy data set is shown in Table 1.

## Maximin algorithm

The fact that the MD rule concentrated almost exclusively on raising the score of poorly recognized sequences suggests an alternative gradient descent training algorithm which we call the “maximin” algorithm. This algorithm seems to converge somewhat slower than the MD algorithm, but we give it here because it may be more widely applicable to other, non-probabilistic models such as profiles (Gribskov *et al.*, 1987).

The maximin idea is to maximize the score of the minimum-scoring sequence. One determines the score  $P(M|S_i)$  for all the sequences with the current model (starting with a random one); finds the sequence or sequences that have the lowest scores; and bumps the probabilities of the model so as to better recognize those sequences, by including a fraction

of their symbol emission and state transition counts in a reestimated model. Reiteration of this converges towards a model which maximizes the score of the minimum-scoring sequences. Specifically, if the old model contained probabilities  $p_x$  and these probabilities are to be modified by the occurrences  $\delta_{i,x}$  of symbols  $x$  in each sequence  $i$  of  $M$  lowest-scoring sequences, we re-estimate new probabilities  $p'_x$  by:

$$p'_x = (1 - \epsilon)p_x + \epsilon \frac{\sum_{i=1}^M \left( \delta_{i,x} + \frac{\alpha}{N} m_x \right)}{\sum_{y=1}^{20} \sum_{i=1}^M \left( \delta_{i,y} + \frac{\alpha}{N} m_y \right)} \quad (7)$$

where  $\epsilon$  sets the learning rate; values of 0.01 - 0.1 seem to work well. The algorithm works best when several sequences within some arbitrary window of the minimum are included in the  $M$  sequences incorporated at the update (we use a window size of 10% of the current average score).

Another way to implement the maximin algorithm for HMMs would be to follow the incremental gradient descent training methods of (Baldi *et al.*, 1994), except that one would always choose the next training sequence from among the current worst scoring sequences instead of in regular or random order.

Maximin is performing a different optimization than the MD algorithm; it would also work with likelihoods and so does not need to assume a background random sequence model. Moreover, it could probably be made to work on models that score by non-probabilistic criteria. However, in practice, the two algorithms have given closely similar results on the globins and several other sequence data sets we have studied. We have used the MD algorithm for all the data we present here.

## Database searching and scoring

The likelihood  $P(S_i | M)$  that is obtained from a dynamic programming alignment of the HMM to each sequence  $S_i$  in a sequence database is strongly dependent on the length of the sequence. In previous work with HMMs of sequence families, significance of alignment scores has been calculated by fitting a regression line to a graph of negative log likelihoods (NLLs) of unrelated or random sequences ( $-\log P(S_i | M)$ ) versus their length and calculating a Z score from each NLL score as the number of standard deviations away from this regression line the NLL score falls.

Instead, we calculate a log odds score:

$$score = \log_2 \frac{P(S_i | M)}{P(S_i | R)} \quad (8)$$

This score is related monotonically to  $P(M | S_i)$ . It corrects for sequence length. Importantly, it is also easily calculated on the fly during an alignment. Before alignment, emission probabilities in the model are preconverted to log odds scores and state transition scores are converted to log probabilities. We use log base 2, and so report scores in bits. This scoring scheme has a relationship to the information theoretic interpretation of PAM scoring matrices, which are implicitly a log-odds alignment scoring scheme (Altschul, 1991). Scores above zero are a more likely match to the model than to the random model if the HMM and the random model are equally likely *a priori*.

## Implementation

Our implementation of hidden Markov models follows (Krogh *et al.*, 1994) in most other details. A fuller description of our implementation will be published elsewhere (S.R.E. and R.D., manuscript in preparation). The maximum discrimination rule was implemented in the program `hmmb`, which builds models from multiple sequence alignments. The heuristic weighting rule used for comparison was implemented in `hmmb` according to (Gerstein *et al.*, 1994). The source code and documentation for the full suite of programs, including multiple alignment and database searching tools, is freely available via anonymous ftp to `cele.mrc-lmb.cam.ac.uk` in `pub/sre` and on the Web at `http://logi.mrc-lmb.cam.ac.uk`. The software is written in C and is known to be portable among a variety of UNIX platforms. Contact S.R.E. (`sre@mrc-lmb.cam.ac.uk`) for further details.

## Results

We have tested maximum discrimination HMMs on the globin sequence family. Globins are familiar for their role as oxygen-carrying molecules such as the vertebrate hemoglobins and myoglobins. The globin family is particularly attractive for this work because it is one of the few large protein families in which members have been identified primarily by biochemical and structural criteria rather than by sequence similarity, and which contains several members which are extremely divergent in sequence. Distantly related globin sequences include various prokaryotic and lower eukaryotic globins (Vasudevan *et al.*, 1991; Manning *et al.*, 1990; Potts *et al.*, 1992; Wakabayashi *et al.*, 1986; Zhu and Riggs, 1992), extracellular nematode globins of unknown function (de Baere *et al.*, 1992), and bacterial and unicellular eukaryotic two-domain proteins which may use a globin domain as an oxygen sensor coupled to an effector domain (Gilles-Gonzalez *et al.*, 1991; Iwaasa *et al.*, 1992). The known examples of the globin sequence family are heavily biased towards vertebrate alpha and beta hemoglobin sequences (Figure 2a).

A data set of 629 globin sequences was assembled from SwissProt 25, by searching for GLOBIN and other keywords and excluding sequence fragments. 400 globin sequences were randomly chosen from this data set and the remaining 229 sequences were held out as independent test data. 29 more globin sequences from SwissProt 27 and PIR 39 that were found subsequently from more keyword searches and from the literature were added to the test set, making a total test set of 258 sequences. The data set is biased towards vertebrate hemoglobins; it contains 239  $\alpha$ -hemoglobins and 201  $\beta$ -hemoglobins.

The 400 training sequences were aligned automatically using maximum-likelihood HMM training (Krogh *et al.*, 1994), beginning with an initial model built from a published structural alignment of seven divergent globins (Bashford *et al.*, 1987). The final automatic alignment was spot-checked and found to be essentially identical to the structural alignment and to smaller manual alignments. This alignment of 400 globins was used to construct three different HMMs; a standard maximum likelihood HMM (GLOB-ML), a weighted maximum likelihood HMM (GLOB-W) using the weighting rule of Gerstein *et al.* (Gerstein *et al.*, 1994), and a maximum discrimination HMM (GLOB-MD). The weighting rule used for GLOB-W is a tree-based rule which gives “intuitively correct” weights, increasing the representation of distant sequences (Gerstein *et al.*, 1994). It gives weights similar to other commonly used sequence weighting methods, such as Voronoi weights (Sibbald and Argos, 1990). All three of these models were then used to search SwissProt release 27.

Because the maximum discrimination rule implies a set of “weights” (see Methods),

one can get an intuitive picture of how the model is achieving its goal of maximum discrimination (Figure 2bc) using a two-dimensional representation of globin sequence space (Schiffman *et al.*, 1981). The “weights” assigned by the MD rule are extreme, relative to all the weighting rules; most of the  $\alpha$ - and  $\beta$ -globins are assigned vanishingly small weights, for instance. High weights are assigned to a particular set of divergent outlying sequences by the MD algorithm. The model is built primarily from these outliers.

Figure 3 shows a histogram of scores from the database searches. Compared to the ML model, both the weighted and the maximum discrimination models more cleanly separate the known globins from other sequences. The separation achieved by the MD model is significantly cleaner than that achieved by the weighting rule.

Table 2 shows how well each model recognizes eleven particularly divergent test globin sequences. These sequences were chosen because none of them has better than 30% sequence identity to any sequence in the training data set. BLASTP (Altschul *et al.*, 1990), using the training data as the database, detects five of the eleven sequences, and two of the remaining sequences produce low-scoring maximal segment pairs (MSPs) consistent with a single gapped alignment that can be weakly detected by a BLAST post-processor, MSPcrunch (Sonnhammer and Durbin, 1994), which was written to boost the sensitivity of BLAST; the remaining four sequences produced no MSPs and would not be detected by a routine BLASTP search. The maximum likelihood model detects seven of these sequences above the first non-globin, the weighted model detects nine, and the maximum discrimination model detects ten of the eleven, missing only GLBN\_NOSCO, a cyanobacterial globin. Interestingly, though BLASTP misses several real globins, it does not miss GLBN\_NOSCO. This is because there is an isolated divergent *Tetrahymena* sequence GLB\_TETPY in the training data set which has significant similarity (27.9% overall pairwise identity) to GLBN\_NOSCO. GLB\_TETPY received an MD weight of only 0.04. After inspecting BLAST, Smith/Waterman, and HMM alignments of these two sequences, we believe that GLB\_TETPY shares sufficient similarity with the overall globin consensus that it received a low MD weight, and that its pairwise similarity to GLBN\_NOSCO is largely in regions separate from the common globin consensus. To some extent, pairwise methods such as BLAST and consensus methods such as HMMs are complementary.

We were concerned that the power of maximum discrimination might also be a weakness. If a training sequence is accidentally included which does not belong to the family in question, maximum likelihood will tend to ignore it but maximum discrimination will make every effort to modify the model and accommodate it. To test how sensitive MD estimation is to bad data, we added 20 completely random sequences of length 145 to the training

set of 400 globins and aligned the “poisoned” training set with the original ML HMM. We used ML, weighted, and MD models built from this poisoned alignment as well as the unpoisoned alignment to search SwissProt 28. The results of these experiments are summarized in Table 3. MD may be slightly more sensitive than the heuristic sequence weighting procedure to incorrect sequences, but, perhaps surprisingly, MD is fairly insensitive to the presence of false positive random sequences. The poisoned MD model still outperformed the unpoisoned maximum likelihood model. If we had instead added real sequences from another family instead of random sequences, though, we expect that an MD model would strongly recognize homologues of the added sequences in addition to the globins. HMMs, particularly MD HMMs, are quite capable of recognizing multiple sequence families if that is what they are given as training data.

We also evaluated the method on the protein kinase sequence family. Unfortunately, the protein kinases are not as attractive a family as the globins for comparisons of database search sensitivity. Many (perhaps most) kinases have been defined by sequence similarity rather than by biochemistry and structure. The difficulties in defining true kinases for the purposes of evaluating database search performance have been discussed extensively by (Krogh *et al.*, 1994). We aligned the 261 protein kinase catalytic domains in the April 1993 Hanks and Quinn database (Hanks and Quinn, 1991) using a simulated annealing procedure in conjunction with maximum likelihood HMM training (S.R.E. and R.D., manuscript in preparation). ML, weighted, and MD HMMs were built from this alignment and used to search SwissProt 28 using a Smith/Waterman HMM alignment algorithm that allows partial alignments, so that subsequences comprising the catalytic domain could be recognized (S.R.E. and R.D., manuscript in preparation). A list of 447 true kinases in SwissProt 28 was assembled from PROSITE 12 lists for the protein kinase ATP binding site motif and the active site motifs for the serine/threonine and the tyrosine kinases. There are also some sequences for which it is not clear whether they should be considered as kinases or not. We made a list of 84 such sequences which were excluded from the performance comparison statistics. This list included 45 kinase fragments, 28 proteins which match the guanylate cyclase PROSITE signature (guanylate cyclases contain an intracellular protein kinase-like domain which has not been demonstrated to have kinase activity) (Krogh *et al.*, 1994), 9 aminoglycoside (antibiotic) kinases which match PROSITE kinase patterns but seem to have little if any other similarity to protein kinase catalytic domains, and 2 *C. elegans* hypothetical genes (discussed further below). The results of these three database searches are summarized in Table 4. As in the globins, the weighted ML model (17 missed kinases below the highest non-kinase score) slightly outperforms the ML model (19 missed kinases),

and MD outperforms the weighted ML model (13 missed kinases). However, the sequences which are being recognized only by the MD model are hypothetical kinases, annotated based on PROSITE matches and pairwise similarity to other (probably also hypothetical!) kinases, so these results are less firm than the globin results.

Two interesting sequences were recognized by the MD kinase model, and not by the ML or weighted models, with fairly good separation from the noise. These sequences are YMX8\_CAEEL and YOO2\_CAEEL, hypothetical gene predictions from the *Caenorhabditis elegans* genome sequencing project (Wilson *et al.*, 1993). Neither had been annotated as a putative kinase by the BLAST analysis done by the nematode sequencing groups. Interestingly, though, both predicted genes are immediately adjacent to other predicted nematode kinase genes. YOO2\_CAEEL, in particular, is in an obviously rearranged and duplicated stretch of the genome in the cosmid ZK507 that includes the adjacent predicted kinase gene. A high degree of gene duplication has been observed in the *C. elegans* genome sequence thus far; when a duplication occurs, it is often adjacent to the original copy (S.R.E., R.D., unpublished). It seems likely that both YMX8\_CAEEL and YOO2\_CAEEL arose by duplication of adjacent kinase genes. We cannot say if any of the kinase genes in question are functional, and further analysis will be necessary to look at the genomic rearrangements that may have generated them. We added these two sequences to the “to be ignored” list of possible kinases for the preceding performance comparisons. They give an example of an interesting biological result that has been obtained with maximum discrimination HMMs but was not obvious with several other methods.

## Discussion

This work introduces a maximum discrimination optimization algorithm for training hidden Markov models of biological sequence consensus. The principal advantage of this method is that it compensates for the biased representation that is common in biological sequence data. Various sequence weighting schemes have previously been proposed to deal with this problem by giving more weight to poorly represented sequences, but all these methods rely on indirect and usually heuristic schemes (Altschul *et al.*, 1989; Gerstein *et al.*, 1994; Thompson *et al.*, 1994; Sibbald and Argos, 1990). The maximum discrimination criterion, in contrast, directly optimizes for the desired quality: that as many as possible of the known sequences should be correctly distinguished from the background. One effect is that all the example sequences end up being assigned a narrower range of probability scores. Every example sequence is considered to be an almost equally likely representative of the family.

We argue that, for database searching, maximum discrimination is the best optimization criterion for a consensus model. The “weights” that are assigned by the MD algorithm are therefore arguably the best weighting scheme. We view other proposed weighting schemes in this context as approximations to MD. However, weighting rules are useful for other purposes, such as in collecting various sorts of statistics from sequence families. We do not necessarily recommend the MD rule in these other contexts. Weighting rules alter the observed sequence data by making reasonable assumptions about the way the data are biased rather than directly optimizing for an operational criterion such as maximum discrimination, and hence are probably more appropriate for many other problems.

Hidden Markov models for modeling biological sequences were adapted from the speech recognition field (Krogh *et al.*, 1994). In speech recognition, as in sequence analysis, maximum likelihood models are typically used (Rabiner, 1989). Alternative training criteria similar in some ways to our maximum discrimination criterion have been explored in speech recognition applications (Rabiner, 1989; Renals and Morgan, 1992). The MD objective function is probably most similar to the logistic transfer functions used in multi-layered perceptrons trained to discriminate classes in data, given labeled training data (Hertz *et al.*, 1991). We expect that the MD algorithm is applicable to many other statistical modeling methods in which biased representation in the training data is a problem.

All one-dimensional consensus models, such as profiles, templates, flexible patterns, and even simple patterns (regular expressions) can be re-expressed or approximated using the hidden Markov model formalism. The maximum discrimination criterion could therefore

be applied to many other, if not all, consensus models. We expect that the alternative “maximin” training algorithm may be applicable to non-probabilistic models.

There are interesting parallels with sequence “templates”, which have been explored as an approach to the problem of detecting distant structural homologues (Taylor, 1986). In a template, each amino acid that has been observed in an aligned column is assigned an equal score, along with other amino acids that fit in the same structural class. For instance, if a column contained W,W,W,W,F, probability scores might be set to be  $1/3$  W,  $1/3$  F, and  $1/3$  Y for the other large hydrophobic residue. The idea is to enumerate the possible amino acids for a given structural position in the protein and treat them equally, regardless of the observed frequencies (which may be biased by representation or by evolution) in order to permit detection of distant structural homologues. However, one or two exceptional amino acids dilute the information in the column’s scores, and cause the rule to break down; the larger the data set, the more spurious amino acids in the aligned columns, and the less well templates work. Impromptu criteria are used to alleviate this problem. In some ways, such as the tendency towards assigning equal probabilities to all the example sequences, maximum discrimination HMMs are like templates and similarly tend to be more sensitive to distant structural homologues than other sorts of models. On the other hand, by trying to equalize the scores of whole sequences rather than symbols in individual columns, maximum discrimination HMMs handle exceptional amino acids smoothly, minimizing (in a strict sense) the amount of information lost.

There is another kind of consensus model which does not suffer a serious biased representation problem. PROSITE motif patterns (Bairoch, 1993) are hand-optimized consensus patterns (regular expressions) which describe a small motif shared amongst all the known examples of a protein family. The pattern is deliberately designed to discriminate every known member of the family from unrelated sequences. Probably for this reason, PROSITE patterns occasionally outperform the theoretically more powerful statistical modeling techniques (unpublished observations), illustrating the utility of discriminative optimization.

Although the size of sequence databases continues to grow exponentially, many new protein sequences seem to be falling into a relatively small number (perhaps on the order of a thousand) of sequence families (Green *et al.*, 1993; Chothia, 1992). Matching new sequences against a thousand consensus models instead of 10,000–100,000 individual protein sequences would be advantageous in terms of both computational effort and discriminatory power. Hidden Markov models provide a means of accomplishing this goal with more sensitivity and specificity than other consensus methods (Baldi *et al.*, 1994; Krogh *et al.*, 1994). The maximum discrimination rule we describe here is a further improvement on the

sensitivity of HMM-based consensus modeling methods, and MD HMMs seem to be one of the most sensitive means yet for detecting distant homologues by sequence information alone.

## Acknowledgments

S.R.E. gratefully acknowledges financial support from a Human Frontier Science Program long-term postdoctoral fellowship. We thank the members of the Cambridge computational molecular biology discussion group for feedback, and especially Cyrus Chothia for his expertise and advice regarding the globin sequence family.

## References

- Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.
- Altschul, S. F., Carroll, R. J., and Lipman, D. J. 1989. Weights for data related by a tree. *J. Mol. Biol.* 207, 647–653.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Bairoch, A. 1993. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids Res.* 21, 3097–3103.
- Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. 1994. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA*, 91, 1059–1063.
- Barton, G. J. 1990. Protein multiple sequence alignment and flexible pattern matching. *Meth. Enzymol.* 183, 403–427.
- Bashford, D., Chothia, C., and Lesk, A. M. 1987. Determinants of a protein fold: Unique features of the globin amino acid sequences. *J. Mol. Biol.* 196, 199–216.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. 1992. Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Science*, 1, 1677–1690.
- Chothia, C. 1992. One thousand families for the molecular biologist. *Nature*, 357, 543–544.
- de Baere, I., Liu, L., Moens, L., Beeumen, J. V., Gielens, C., Richelle, J., Trotman, C., Finch, J., Gerstein, M., and Perutz, M. 1992. Polar zipper sequence in the high-affinity hemoglobin of *Ascaris suum*: Amino acid sequence and structural interpretation. *Proc. Natl. Acad. Sci. USA*, 89, 4638–4642.
- Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1–38.
- Gerstein, M., Sonnhammer, E., and Chothia, C. 1994. Volume changes in protein evolution. *J. Mol. Biol.* 235, 1067–1078.

- Gilles-Gonzalez, M. A., Ditta, G. S., and Helinski, D. R. 1991. A haemoprotein with kinase activity encoded by the oxygen sensor of *Rhizobium meliloti*. *Nature*, 350, 170–172.
- Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., and Claverie, J.-M. 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science*, 259, 1711–1716.
- Gribskov, M., McLachlan, A. D., and Eisenberg, D. 1987. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, 84, 4355–4358.
- Hanks, S. K. and Quinn, A. M. 1991. Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members. *Meth. Enzymol.* 200, 38–62.
- Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915–10919.
- Henikoff, S. and Henikoff, J. G. 1994a. Protein family classification based on searching a database of blocks. *Genomics*, 19, 97–107.
- Henikoff, S. and Henikoff, J. G. 1994b. Position-based sequence weights. *J. Mol. Biol.* 243, 574–578.
- Hertz, J., Krogh, A., and Palmer, R. G. 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Reading, Massachusetts.
- Iwaasa, H., Takagi, T., and Shikama, K. 1992. Amino acid sequence of yeast hemoglobin. *J. Mol. Biol.* 227, 948–954.
- Karlin, S. and Altschul, S. F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87, 2264–2268.
- Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235, 1501–1531.
- Luthy, R., Xenarios, I., and Bucher, P. 1994. Improving the sensitivity of the sequence profile method. *Protein Science*, 3, 139–146.

- Manning, A., Trotman, C., and Tate, W. 1990. Evolution of a polymeric globin in the brine shrimp *Artemia*. *Nature*, 348, 653–656.
- Pearson, W. and Lipman, D. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85, 2444–2448.
- Potts, M., Angeloni, S. V., Ebel, R. E., and Bassam, D. 1992. Myoglobin in a cyanobacterium. *Science*, 256, 1690–1692.
- Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77, 257–286.
- Renals, S. and Morgan, N. 1992. Connectionist probability estimation in HMM speech recognition. Technical Report TR-92-081 International Computer Science Institute.
- Schiffman, F., Reynolds, M., and Young, F. 1981. *Introduction to Multidimensional Scaling*. Academic Press, New York.
- Sibbald, P. R. and Argos, P. 1990. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* 216, 813–818.
- Sonnhammer, E. and Durbin, R. 1994. A workbench for large scale sequence homology analysis. *Comput. Applic. Biosci.*, in press.
- Taylor, W. R. 1986. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188, 233–258.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Applic. Biosci.* 10, 19–29.
- Vasudevan, S. G., Armarego, W. L., Shaw, D. C., Lilley, P. E., Dixon, N. E., and Poole, R. K. 1991. Isolation and nucleotide sequence of the *hmp* gene that encodes a haemoglobin-like protein in *Escherichia coli* K-12. *Mol. Gen. Genet.* 226, 49–57.
- Wakabayashi, S., Matsubara, H., and Webster, D. 1986. Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature*, 322, 481–483.
- Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M., *et al.* 1993. The *C. elegans* genome project: Nucleotide sequence of over two megabases from chromosome III. Submitted to *Nature*.

Zhu, H. and Riggs, A. F. 1992. Yeast flavohemoglobin is an ancient protein related to globins and a reductase family. *Proc. Natl. Acad. Sci. USA*, 89, 5015–5019.

# Figure legends

## Figure 1

Consensus models of a toy example alignment. A consensus sequence consists of just the highly conserved positions in the alignment. More complex consensus models such as profiles or templates assign position specific scores (possibly probabilities) to most columns in the alignment. A hidden Markov model adds a probabilistic treatment of insertions and deletions by imagining an arrangement of sequence-emitting states aligned to the sequences, and state transition probabilities (arrows) between them for moving to a new state. Match states (M) emit the consensus of the alignment, similar to a profile; delete (D) states emit nothing and permit a consensus column to be skipped; insert (I) states emit one or more random symbols (here represented by X) between consensus columns.

## Figure 2

A. A two-dimensional representation of globin sequence space for 629 globin sequences in SwissProt 25, showing the tight overrepresented clusters of certain globin subfamilies. Sequences are placed on the 2D plot roughly according to how similar they are to all the other sequences, so that similar sequences are clustered and dissimilar sequences are far apart. The plot is constructed by multidimensional scaling; we took the fractional dissimilarity of every pair of globin sequences to be a distance, and then found the most consistent two-dimensional plot by minimizing the squared difference between the 2D distances and the dissimilarities. An unbiased sample of a large region of sequence space does not show such tight clusters. Open diamonds:  $\alpha$ -globins; filled diamonds:  $\beta$ -globins; triangles: other hemoglobins; left triangles: leghemoglobins; inverted triangles: myoglobins; open circles: other globins, including the distantly related prokaryotic and lower eukaryotic globins.

B,C. Comparison of the weights for the 400 training sequences produced by the maximum discrimination method (B) and the weighting rule of (Gerstein *et al.*, 1994) (C), using the same 2D coordinate positions as in Figure 2a. The weights sum to 400, so a weight of 1 is average. Grey circles: sequences with weight  $< 1$ ; open circles: weights  $> 1$  and  $< 10$ ; black circles: weights  $> 10$ . The MD algorithm assigned fairly extreme weights to a set of divergent sequences around the border of globin sequence space. The effect of the weighting rule is more gentle, smoothing out the strong clustering in the data. MD weights ranged from 32.6 (on spoonworm globin HBF1\_URECA) to zero on 243 of the 400 sequences (including most of the  $\alpha$ - and  $\beta$ -globins). The Gerstein weights ranged from 4.8

(on *Tetrahymena* globin GLB\_TETPY) to 0.08 (on the  $\alpha$ -globin HBA\_MACSI).

### **Figure 3**

Histograms of log odds scores obtained in HMM searches of the SwissProt 27 protein sequence database; scores for the sequences in the training set, the independent test set of all other known globins, and the non-globins are shown separately. In order to emphasize the isolated low-scoring sequences, the vertical scale has been reduced; thus the peaks where many of the test and training sequences fall have been truncated, particularly for the MD plot. Note how the MD model trades off overall score in return for increased discrimination of distant sequences.

| Maximum likelihood     |          |       |                    |               |
|------------------------|----------|-------|--------------------|---------------|
| alignment              | weight   | $p_x$ | $\log P(S_i   M)$  | $P(M   S_i)$  |
| AAAAAAAAAA             | 1.0      | .40   | -9.162             | 0.991         |
| AAAAAAAAAA             | 1.0      | .40   | -9.162             | 0.991         |
| CCCCCCCCCC             | 1.0      | .40   | -9.162             | 0.991         |
| CCCCCCCCCC             | 1.0      | .40   | -9.162             | 0.991         |
| GGGGGGGGGG             | 1.0      | .20   | -16.094            | 0.097         |
|                        |          |       | $\Sigma = -52.742$ | $\Pi = 0.094$ |
| Voronoi weighting      |          |       |                    |               |
| alignment              | weight   | $p_x$ | $\log P(S_i   M)$  | $P(M   S_i)$  |
| AAAAAAAAAA             | .833     | .33   | -11.087            | 0.941         |
| AAAAAAAAAA             | .833     | .33   | -11.087            | 0.941         |
| CCCCCCCCCC             | .833     | .33   | -11.087            | 0.941         |
| CCCCCCCCCC             | .833     | .33   | -11.087            | 0.941         |
| GGGGGGGGGG             | 1.667    | .33   | -11.087            | 0.941         |
|                        |          |       | $\Sigma = -55.433$ | $\Pi = 0.739$ |
| Maximum discrimination |          |       |                    |               |
| alignment              | “weight” | $p_x$ | $\log P(S_i   M)$  | $P(M   S_i)$  |
| AAAAAAAAAA             | .851     | .34   | -10.776            | 0.956         |
| AAAAAAAAAA             | .851     | .34   | -10.776            | 0.956         |
| CCCCCCCCCC             | .851     | .34   | -10.776            | 0.956         |
| CCCCCCCCCC             | .851     | .34   | -10.776            | 0.956         |
| GGGGGGGGGG             | 1.597    | .32   | -11.413            | 0.920         |
|                        |          |       | $\Sigma = -54.517$ | $\Pi = 0.770$ |

Table 1: Models of a toy example DNA alignment using ML, weighted ML (Gerstein *et al.*, 1994), and MD parameter estimation schemes. The probabilities  $p_x$  assigned by each model to the nucleotides A, C, and G are shown, as are the values of the objective functions for both maximum likelihood (which maximizes the total log likelihood of the sequences given the model,  $\sum_{i=1}^N \log P(S_i | M)$ ) and maximum discrimination (which maximizes the probability that all the sequences are recognized by the model,  $\prod_{i=1}^N P(M | S_i)$ ). The values of these objective functions are shown below each block. The value of the MD objective function is shown in probabilities, rather than log probabilities, to emphasize the dominant effect of even just one poorly represented sequence.  $P(M | S_i)$  is calculated as described in Methods, assuming a random sequence model in which all four nucleotides are equiprobable (0.25). For simplicity, no prior is used in calculating  $p_x$  for this example. A maximum likelihood model fails to recognize the underrepresented G<sub>10</sub> sequence well (as evidenced by its low probability  $P(M | S_i)$ ). Both the weighting schemes and the maximum likelihood method produce models which recognize G<sub>10</sub>, but the MD solution scores better for both objective functions.

| sequence   | source     | best id | BLAST |    | ML      | W | MD      |    |        |    |
|------------|------------|---------|-------|----|---------|---|---------|----|--------|----|
| GLBH_TRICO | nematode   | 15.7%   | -     |    | -86.71  | * | -28.86  | *  | -7.15  | *  |
| GLBH_CAEEL | nematode   | 16.2%   | -     |    | -117.22 |   | -47.16  | *  | -10.95 | *  |
| GLB_ASCSU  | nematode   | 18.4%   | 54    | *  | -107.63 |   | -41.56  | *  | -4.93  | *  |
| HB_CANNO   | yeast      | 19.1%   | 46    | *  | -75.78  | * | -14.52  | *  | 8.74   | ** |
| GLB_PSEDC  | nematode   | 19.6%   | -     |    | -91.03  | * | -15.36  | *  | 19.54  | ** |
| BAHG_VITSP | bacterial  | 20.6%   | 102   | ** | -47.92  | * | 18.73   | ** | 43.64  | ** |
| HMP_ECOLI  | E. coli    | 21.2%   | -     |    | -67.60  | * | -8.37   | *  | 13.68  | ** |
| B45383     | yeast      | 22.1%   | 101   | ** | -72.78  | * | -6.86   | *  | 14.70  | ** |
| GLBN_NOSCO | cyanobact. | 27.9%   | 170   | ** | -170.41 |   | -107.02 |    | -41.20 |    |
| GLB3_LUMTE | earthworm  | 28.5%   | 389   | ** | -45.38  | * | 42.95   | ** | 83.88  | ** |
| S31726     | Chlamy.    | 29.9%   | 177   | ** | -123.90 |   | -71.30  |    | -13.60 | *  |

Table 2: Results of using BLASTP and HMMs trained by three different methods for detecting similarities to eleven particularly diverged globin sequences in the test set. The percentage identity to the most similar sequence in the 400 training sequences is shown. BLASTP searches against the training set used the BLOSUM62 matrix (Henikoff and Henikoff, 1992); five sequences with significant scores are marked with \*\*. The MSPcrunch BLAST post-processor (Sonnhammer and Durbin, 1994) recovers an additional two sequences with marginal-scoring gapped alignments (marked with \*). HMM log-odds scores (in log base 2, bits) are shown, with a single asterisk if the score is above the highest non-globin score in a search of a database composed of PIR 37 and SwissProt 27 (SWIR4, (Sonnhammer and Durbin, 1994), and two asterisks for scores above zero (a probable match according to the statistics of the model). The highest non-globin scores were -97.06 for GLOB-ML, -50.82 for GLOB-W, and -26.94 for GLOB-MD. See Figure 4 for histograms of these score distributions.

| model     | avg. score | lowest | missed | equiv |
|-----------|------------|--------|--------|-------|
| Glob-ML   | 298        | -103   | 5      | 4     |
| Poison-ML | 275        | -165   | 8      | 4     |
| Glob-W    | 277        | 4      | 2      | 2     |
| Poison-W  | 247        | -52    | 2      | 1     |
| Glob-MD   | 195        | 152    | 1      | 1     |
| Poison-MD | 131        | 53     | 3      | 3     |

Table 3: Results of searching SwissProt 28 with models trained on an alignment containing 20 random sequences in addition to 400 real globins (Poison-ML, Poison-W, Poison-MD), compared to the unpoisoned models. The average log-odds score of the models on their training data and the lowest score in the training data (including scores of the poison, for the poisoned alignments) are shown in the first two columns. MD successfully learns to recognize all 20 random sequences with positive scores, in addition to the 400 globins. The “missed” column shows how many true globins fall below the highest-scoring non-globin in the database. The “equiv” column shows a measure of both sensitivity and specificity called the “equivalence number”, the number of true globins that fall below the threshold defined by the point at which false negatives equal false positives (Bill Pearson, personal communication). For both of these measures, low numbers are better, indicating fewer missed globins. There are 658 real globins in SwissProt 28; a further 10 globin sequence fragments and 2 multidomain globins were ignored in this experiment.

| model     | avg. score | lowest | missed | equiv |
|-----------|------------|--------|--------|-------|
| Kinase-ML | 349        | 16     | 19     | 9     |
| Kinase-W  | 343        | 42     | 17     | 7     |
| Kinase-MD | 297        | 271    | 13     | 4     |

Table 4: Results of searching SwissProt 28 with models trained on an alignment of 261 protein kinases. The columns shown are as described in the legend of Table 3.

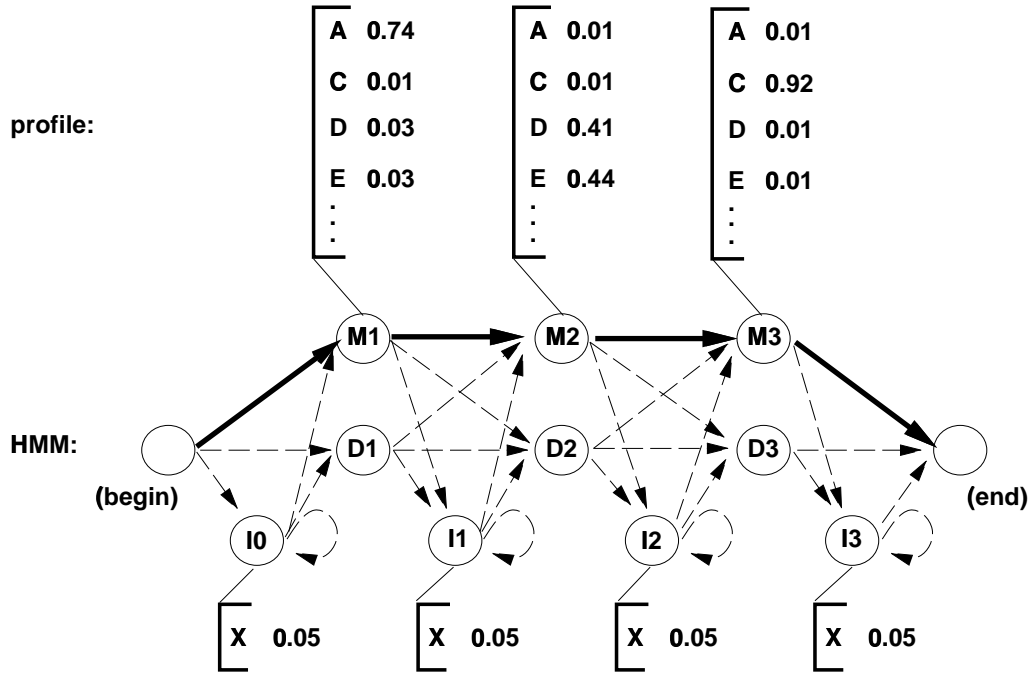
multiple alignment:

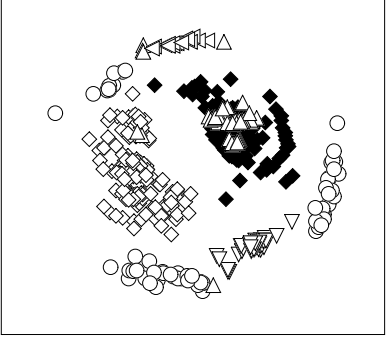
```

      - A D T C
      W A E - C
      - V E - C
      - A D - C
      - A E - C
  
```

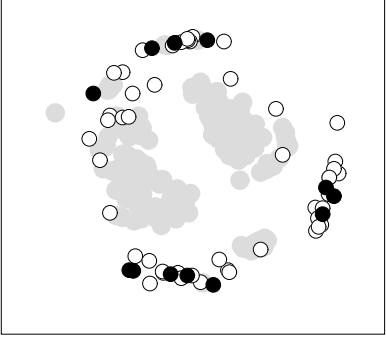
consensus: 

|   |     |   |
|---|-----|---|
| A | D/E | C |
|---|-----|---|

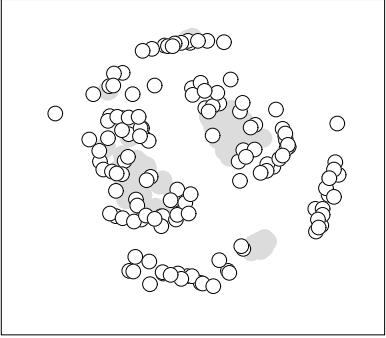




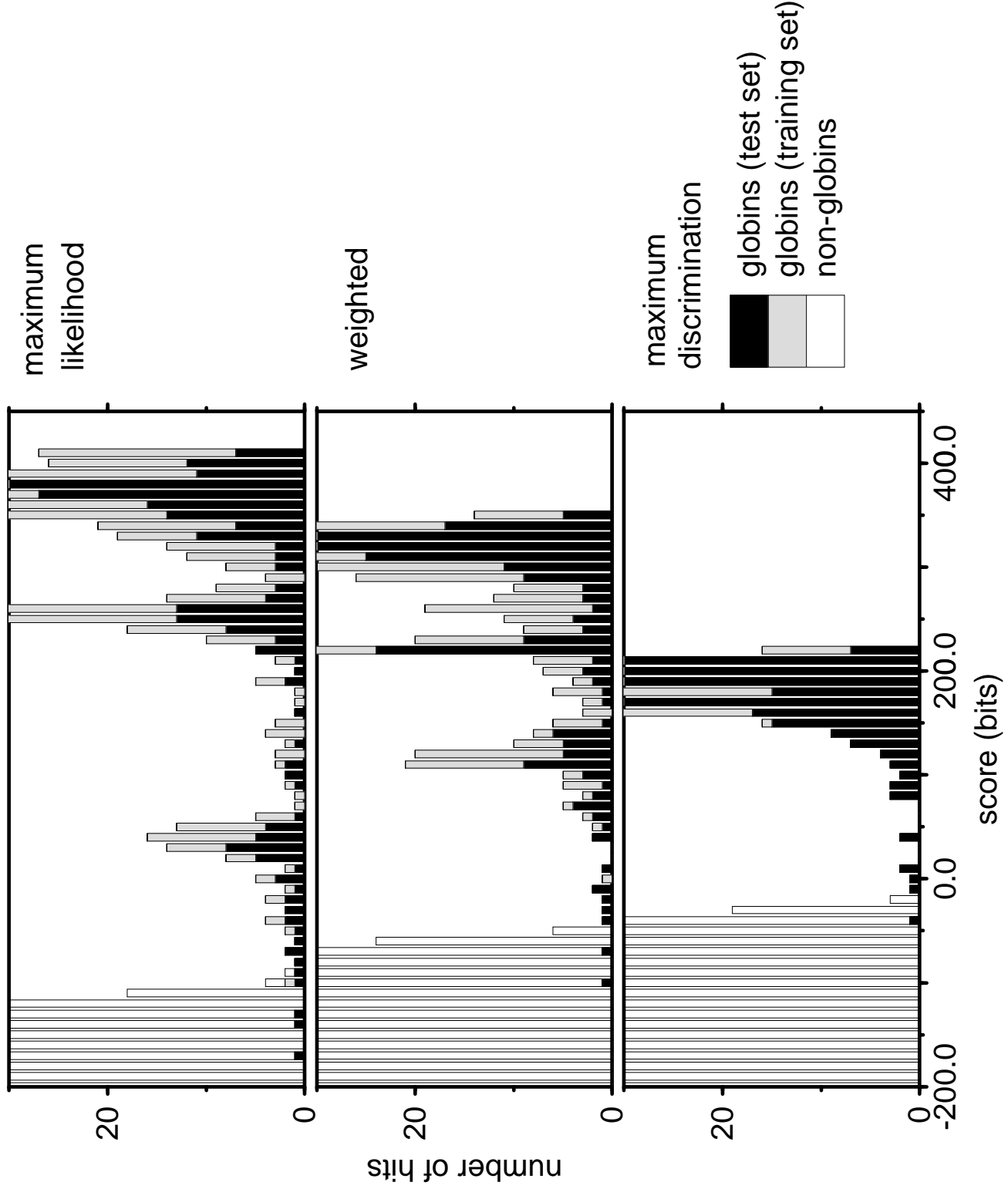
A



B



C



maximum likelihood

weighted

maximum discrimination

- globins (test set)
- globins (training set)
- non-globins