# Parameterization studies for the SAM and HMMER methods of hidden Markov model generation

**Marcella A. McClure**
Department of Biological Sciences
UNLV, Las Vegas, NV 89129
phone: 702-895-4471
fax: 702-895-3956
mars@parvati.lv-whi.nevada.edu

**Chris Smith**
Department of Biological Sciences
UNLV, Las Vegas, NV 89129
phone: 702-895-1551
fax: 702-895-3956
csmith@parvati.lv-whi.nevada.edu

**Pete Elton**
Department of Biological Sciences
UNLV, Las Vegas, NV 89129
phone: 702-895-1551
fax: 702-895-3956
elton@parvati.lv-whi.nevada.edu

**Abstract**

Multiple sequence alignment of distantly related viral proteins remains a challenge to all currently available alignment methods. The hidden Markov model approach offers a new, flexible method for the generation of multiple sequence alignments. The results of studies attempting to infer appropriate parameter constraints for the generation of *de novo* HMMs for globin, kinase, aspartic acid protease, and ribonuclease H sequences by both the SAM and HMMER methods are described.

## Introduction

The earliest application of Markov processes to the analysis of biological data was in area of ecological modeling (Baum and Eagon, 1967). In the last few years a plethora of Markov modeling and Expectation-Maximization (EM) algorithms have been applied to a variety of molecular computational biology problems [Churchill, 1989; Lawrence, 1990; Thorne, 1991 ; Cardon, 1992; Stultz, 1993]. Recently, hidden Markov model (HMMs) methods, based on a human speech recognition approach (Rabiner, 1989), have been applied to secondary structure prediction, multiple sequence alignment, motif identification, and database searching (Asai, 1993; Baldi, et al., 1994; Fujiwara, et al., 1994; Krogh, et al., 1994; Eddy, 1995). These methods essentially create a stochastic production model representing the sequences used to train the model. In its simplest form a model is initialized *a priori* for; 1) the transition into a match, deletion or insertion state, and 2) the occurrence of a given amino acid (or nucleotide) in a match or insert state. Using the initial model and all training sequences, all possible paths for each sequence through the model are evaluated to obtain new estimates of the parameters that will increase the likelihood of the model. This process is repeated until the model converges. A multiple alignment is generated by computing the negative logarithm of the probability of the single most likely path through the model for a particular sequence given all the possible paths generated by the training sequences.

The HMM approach has a variety of advantages over more classical multiple alignment methods: 1) it is grounded in probability theory; 2) knowledge of phylogenetic history or pairwise ordering is not required; 3) insertions/deletions (indels) are treated probabilistically in a variable, position dependent manner; 4) experimentally derived information can be incorporated into the model *a priori*; 5) the model can provide information regarding both the stochastic and selected features of a protein family; and 6) the computational cost of aligning a set of sequences to an HMM is directly proportional to the number of sequences to be aligned. The computational cost increases exponentially with the number of sequences to be aligned in most other existing multiple alignment methods. To date there are no other methods with the flexibility of design allowed by the HMM approach.

It should be recalled that RNA virus genomes accumulate mutations 1-10 million times faster than DNA-based lifeforms. RNA sequences, therefore, provide us with a highly divergent, yet related, set of co-linear genes encoding a variety of enzymatic and structural proteins. Many of these sequence relationships share far less than 25% amino acid identity although the proteins perform the

same functions. We are interested in constructing HMMs which adequately represent homologous proteins, and their functionally equivalent distant relatives, to provide a global representation of sequences that can be used to address a variety of evolutionary questions. Functionally equivalent relatives are those sequences which cannot be shown to be homologous by Monte Carlo simulations under classical criteria (usually those with less that 25% identity). Nonetheless, all biological and biochemical data support a common ancestry for such proteins (see McClure, 1993 for an example).

Although HMM methods allow the use of pre-existing alignments as "hints" for model building there are sets of divergent viral protein sequences for which adequate multiple alignments do not yet exist. To derive an adequate alignment for even a representative subset of these sequences by more classical methods still requires the time consuming effort of manual refinement. It is of interest, therefore, to determine appropriate parameter constraints for the generation of biologically informative *de novo* HMMs representing given sets of proteins. While empirical testing of these methods may be considered anecdotal by some individuals, it is important to remember that it is customary in science that the limits of analytical procedures be defined in some manner before using them to derive new information. This is usually done by defining a set of standards (McClure, et al., 1994) and some form of assessment criteria (McClure and Raman, 1995). It is our position that a method which cannot find the known ordered series of motifs conferring function of a given protein family will be of little value in finding patterns in new protein sequences.

The two HMM methods available in academia, Sequence Alignment and Modeling Software System, SAM, (Krogh, et al., 1994; Hughey and Krogh, 1996) and HMMER (Eddy, 1995) are evaluated. The SAM method implements the full Baum-Welch (BW) algorithm with the injection of noise by simulated annealing to avoid local optima. The HMMER method allows a choice of approaches, simulated annealing, the Viterbi approximation of the BW, and full BW implementations. In agreement with Eddy's demonstration (Eddy, 1995), our pilot studies (unpublished observations) also indicate that the simulated annealing approach performs better than either the Viterbi approximation to the BW or the full BW implementation available in the HMMER method. All subsequent tests of the HMMER method utilize the simulated annealing approach.

The four protein families, globin, kinase, aspartic acid protease (aaPR), and ribonuclease H (RH) sequences, employed to evaluate more classical methods of multiple sequence alignment (McClure, 1994), and the assessment criteria developed in an initial study of the SAM method (McClure and Raman, 1995) are used to test various aspects of both SAM and HMMER. The results of a study comparing the default parameter settings, and initializations of the match and insert states with more data specific settings are presented. Additional evaluation of the HMMER code for the effects of model length and training set size on model generation bring the evaluation of this method to the same status of SAM (McClure and Raman, 1995). New studies on training set similarity distributions are in process for both the SAM and HMMER methods.

## Methodology

All analyses were conducted on SPARCstations 10/514MP, 2, or LX running the Solaris 5.4 operating system. The globin, aspartic acid protease (aaPR) and the ribonuclease H (RH) sequences were extracted from the non-redundant database composed of PIR, 34.0, SWISS-PROT, 23.0 and GenPept, (translated GenBank, 73.0). The kinase catalytic domain database was provided by Salk Institute. Version 1.1 of SAM was used in all studies. The hmmt program of version 1.8 of HMMER was used in the default parameter studies. Based on these results (tables 1 and 2) hmmt was modified as follows for all other studies. A -m option was added which allows the user to provide hmmt with the normalized frequencies of the amino acids in the training set for the initial match state emissions. The default hmmt provides these values for both the match and insert states based on the global estimates of amino acid frequencies in the PIR database. The insert state emissions are hard wired into hmmt. In our modified version the insert state emission probabilities are uniformly initialized, 1/20.

### Training sets

The training sets are similar to those used in the studies described previously, (McClure, 1994; McClure and Raman, 1995) except that sequence fragments have been removed. Only the viral aaPR sequences are used for training in protease model generation. Previous analysis using SAM indicates that a model capable of generating a correct alignment of the ordered series of motifs conferring aaPR catalysis could not be generated using either cellular and viral sequences, or cellular sequences alone (McClure and Raman, 1995). The same test conducted with HMMER confirmed this result (unpublished observation). The RH sequences are the RH domain of the RNA-directed DNA polymerase or reverse transcriptase (RT). This domain, averaging 146 residues, is not found in all RT proteins and therefore is analyzed as a separate sequence. The RH sequences of bacteria are not found as part of larger molecules.

### Validation sets

The validation sets, 12 sequences each, covering the phylogenetic distribution of each family, were used to test the quality of alignments generated by HMMs. The globin sequences, ranging from 10-70% pairwise identity, include alpha- and beta-globins from mammals and birds; myoglobins from mammals; and hemoglobins from insects, plants and bacteria. Five regions of the alignment

were arbitrarily designated to serve as the ordered series of motifs defining the globin family thereby providing a consistent test similar to those described below. The globins are highly conserved with few indels. The five motifs range in size from three to seven amino acids.

The kinase validation set includes serine/threonine, tyrosine and dual specificity kinases from mammals, birds, fungi, retroviruses and herpes viruses. Crystallographic studies of the cyclic adenosine monophosphate-dependent protein kinase confirm that the conserved motifs of the kinase protein core do indeed cluster into the regions of the protein involved in nucleotide binding and catalysis (Knighton, et al., 1991). The kinase family has well defined indel regions interspersed among eight highly conserved motifs each varying in size from one to nine amino acid residues .

The aaPR set includes the amino-terminal domain of pepsin from mammals, birds and fungi, and representative members of the retroid family: retroviruses, caulimoviruses and retrotransposons. As predicted by primary sequence analysis, crystallographic studies confirm the dimeric structure and catalytic residues of the retroviral protease (Miller, et al., 1989). The aaPR family has three motifs, ranging in size from three to five residues with an indel pattern which varies among different lineages.

The RH validation set is comprised of sequences from *Escherichia coli* and the retroid family; retroviruses, caulimoviruses, hepadnaviruses, retrotransposons, retroposons and group II plasmids of the filamentous ascomycete mitochondria. Crystallographic studies of the *E. coli* RH protein (Katayanagi, et al., 1990) and HIV-I RH domain (Davies, et al., 1991) confirm the ordered series of motifs common among these very distantly related sequences as the catalytic residues of the protein as predicted from primary sequence analysis. The RH family has five motifs, ranging in size from one to five residues with an indel pattern which varies among different lineages.

The amino acid identity of all three enzymatic validation sets ranges from 8-30%. This low level of identity along with the significant amount of indels, places the aaPR and RH sequences among the most difficult of tests developed to date. It should be noted that these two datasets are not the most divergent data for which we have developed multiple alignments. Validation sequence sets are available through EMBL (identification no. DS16117).

## Evaluation Criteria

Currently we use three measures of "goodness" in determining a "best" HMM. Optimal models are generated by both SAM and HMMER by slightly different methods. In the SAM method the user can specify the number of models to be generated at a specific length or within a give model length range. The negative log likelihood (NLL) is calculated for all training sequences for each model thereby determining the optimal model. The optimal model then undergoes the model surgery procedure (MSP) which deletes or adds states after training. The HMMER method does not provide an option for generating more than one model at a specific length or within a give model length range. HMMER finds the optimal model by calculating the average log-odds probability of the data fitting the model at each iteration. A new model is determined by fitting the training sequences to the current model until the model score converges. Towards the end of the iteration a maximum *a posterior*, MAP, procedure (S.R. Eddy and R. Durbin, in preparation) is invoked to revise the model length.

The average entropy/position content (A.E.C.) reflects the amount of conservation in a model. The entropy is calculated by

$$H_{avg} = -\left(\frac{1}{M}\right)\sum_{j=1}^{M}\sum_{q\in(d,i,m)}\sum_{x\in\{a.acids\}} P_{jx}^{q}\log\left(P_{jx}^{q}\right),$$

where M is the length of the model and $P_{jx}^{q}$ is the probability of being in the state $q$ (delete, $d$, insert, $i$, and match $m$) in position $j$ of the model and emitting residue $x$. The biologically informative (B.I.) model is defined as the one which produces a multiple alignment of the validation sequences which captures a particular set of the biological features, in this case, the ordered series of motifs common to each protein family. We are investigating the correlation between the optimal model, the highest A.E.C. model, and the B.I. model. Our preliminary studies using SAM indicate that the "best" model is among those with high A.E.C. values, but not necessarily the highest (McClure and Raman, 1995).

## Evaluation Strategy

The initial studies compare the ability of SAM and HMMER generated models to correctly identify the ordered series of motifs in the four validation sequence sets using the default versus user specified parameters (table 1). The default settings for both the initial match and insert state amino acid emission probabilities are globally estimated from the protein database. The derived emission frequencies are estimated from all the sequences of each family for each test. Pilot studies indicate that both the match and insert initializations have an effect on generating a B.I. model. We assessed models generated with derived emission frequencies and uniform probabilities for both the match and insert states. In general, models generated by initializing the match state emissions with the data specific amino acid frequencies and the insert state emissions with uniform probabilities out performed models generated by all other combinations for these two parameters (unpublished observations). All subsequent tests, use the data specific emission frequencies for the initial match states, while the insert state emissions are uniformly initialized, 1/20.

The SAM code also provides various options for manipulating the regularizer, which is provided to keep estimates from over fitting the data. By trial and error we have found that setting the match, insert and delete jump confidences to 50 versus the default of 1 produces models

| | SAM | | HMMER | |
|---|---|---|---|---|
| parameters | Default | Derived | Default | Derived |
| match state | global | data specific | global | data specific |
| insert state | global | uniform | global | uniform |
| del_jump_conf | 1 | 50 | - | - |
| ins_jump_conf | 1 | 50 | - | - |
| match_jump_conf | 1 | 50 | - | - |
| reestimates | 30 | 25 | - | - |
| stopcriterion | 0.1 | 0.05 | - | - |
| model length | | | | |
| globin | 148 | 147 | 165 | 159 |
| kinase | 253 | 255 | 272 | 279 |
| aaPR | 99 | 102 | 109 | 109 |
| RH | 128 | 129 | 154 | 171 |

**Table 1. Parameters and model lengths.** The default match and insert state amino acid emission frequencies are estimated from the entire protein database. The derived match state emission frequencies are estimated from each of the four protein families in each test. The derived insert state emission probabilities are uniformly initialized, 1/20. Changes to other SAM parameters were determined by trial and error. Dashes indicate options not available in the HMMER code. The model lengths were determined as described in the Methods.

| | SAM | | HMMER | |
|---|---|---|---|---|
| | Default | Derived | Default | Derived |
| Globin I(7) | 41 | 100 | 100 | 100 |
| II(5) | 83 | 100 | 100 | 100 |
| III(5) | 58 | 100 | 83 | 100 |
| IV(5) | 75 | 100 | 100 | 100 |
| V(3) | 16 | 100 | 100 | 83 |
| Kinase I(6) | 100 | 100 | 100 | 100 |
| II(1) | 100 | 100 | 100 | 100 |
| III(1) | 100 | 100 | 100 | 100 |
| IV(9) | 83 | 100 | 100 | 100 |
| V(3) | 83 | 100 | 100 | 100 |
| VI(3) | 83 | 100 | 100 | 100 |
| VII(8) | 83 | 100 | 100 | 100 |
| VIII(1) | 16 | 100 | 100 | 100 |
| aaPR I(3) | 100 | 100 | 100 | 100 |
| II(5) | 42 | 92 | 42 | 83 |
| III(5) | 67 | 100 | 92 | 83 |
| RH I(3) | 83 | 100 | 75 | 92 |
| II(1) | 83 | 83 | 75 | 100 |
| III(3) | 66 | 75 | 75 | 83 |
| IV(5) | 58 | 83 | 75 | 66 |

**Table 2. Results of models generated by default and derived parameters.** The first column indicates each protein test, roman numerals each motif, and the values in parentheses the number of residues in each motif. In subsequent columns the correct identification of all residues comprising the motif is indicated as the percentage of sequences in which the motif was identified in the validation sets. Subsequent columns indicate the results of comparing the SAM and HMMER defaults and derived settings (table 1), respectively . The training set sizes are: globins = 873, kinase = 225, aaPR = 160, and RH = 201.

that find motifs more often, but more rigorous tests of varying these parameters are needed. HMMER does not provide such options. The model lengths for the SAM studies are: 1) the default length, randomly selected within plus/minus 10% of the average length of each training set, with subsequent MSP, and 2) the derived length is the MSP length of an optimal model selected by generating 10 models with lengths randomly chosen between the minimal and maximal training sequence length. In the HMMER studies the default MAP length is assessed under the

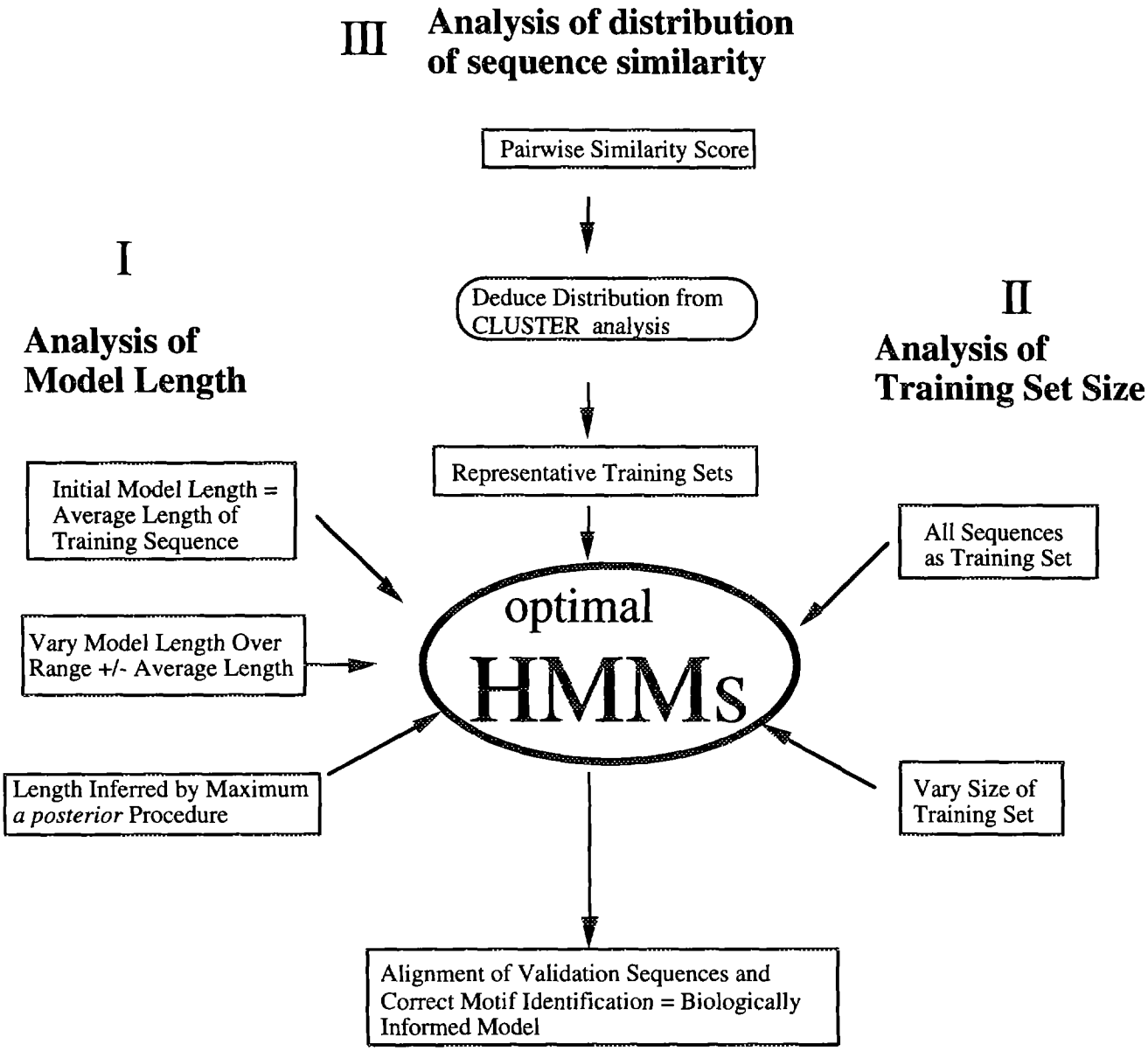


Figure 1. Flow diagram of the strategy for model length (I), training set size (II) and training set similarity distribution (III) studies.

default and data specific match emission frequencies and uniform insert probabilities.

Figure 1 displays a flow diagram of our approach to model length, training set size, and training set similarity distribution studies. In an earlier study of the SAM method the effects of model length and training set size indicates that: 1) generating 10 models, selecting the optimal model from this set, and then allowing the MSP

## Effect of model length on model quality for HMMER method

| lengths (range) | length 1 | length 2 | I(7) 1 | I(7) 2 | II(5) 1 | II(5) 2 | III(5) 1 | III(5) 2 | IV(5) 1 | IV(5) 2 | V(3) 1 | V(3) 2 | A.E.C. 1 | A.E.C. 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (100-120) | 119 | 120 | 92 | 100 | 100 | 100 | 100 | 100 | 50 | 50 | 75 | 83 | 0.986 | 0.985 |
| (121-140) | 139 | 140 | 92 | 100 | 100 | 100 | 92 | 100 | 100 | 100 | 83 | 100 | 1.043 | 1.075 |
| (141-160) | 157 | 150 | 100 | 92 | 100 | 100 | 92 | 100 | 75 | 100 | 75 | 75 | 1.072 | 1.070 |
| (161-180) | 174 | 178 | 67 | 92 | 100 | 100 | 75 | 67 | 100 | 58 | 100 | 75 | 1.128 | 1.079 |
| (181-200) | 197 | 192 | 92 | 92 | 100 | 100 | 75 | 67 | 50 | 83 | 100 | 83 | 1.124 | 1.091 |
| ave. len. | 148 | | 92 | | 100 | | 100 | | 75 | | 75 | | 1.084 | |
| MAP | 159 | | 100 | | 100 | | 100 | | 100 | | 83 | | 1.122 | |

**Table 3. Globins**, training set size = 873

| length (range) | length 1 | length 2 | I(6) 1 | I(6) 2 | II(1) 1 | II(1) 2 | III(1) 1 | III(1) 2 | IV(9) 1 | IV(9) 2 | V(3) 1 | V(3) 2 | VI(3) 1 | VI(3) 2 | VII(8) 1 | VII(8) 2 | VIII(1) 1 | VIII(1) 2 | A.E.C. 1 | A.E.C. 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (231-250) | 248 | 244 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1.735 | 1.718 |
| (251-270 | 269 | 254 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1.737 | 1.739 |
| (271-290) | 278 | 273 | 75 | 100 | 100 | 92 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 92 | 83 | 100 | 100 | 1.732 | 1.743 |
| (291-310) | 295 | 299 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1.745 | 1.751 |
| (311-330) | 311 | 314 | 100 | 100 | 100 | 100 | 92 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 83 | 100 | 100 | 100 | 1.734 | 1.726 |
| ave. len. | 283 | | 100 | | 92 | | 100 | | 100 | | 100 | | 100 | | 100 | | 100 | | 1.745 | |
| MAP | 279 | | 100 | | 100 | | 100 | | 100 | | 100 | | 100 | | 100 | | 100 | | 1.700 | |

**Table 4. Kinase**, training set size = 225

| length (range) | length 1 | length 2 | I(3) 1 | I(3) 2 | II(5) 1 | II(5) 2 | III(5) 1 | III(5) 2 | A.E.C. 1 | A.E.C. 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| (70-90) | 88 | 89 | 100 | 100 | 50 | 58 | 83 | 83 | 1.616 | 1.596 |
| (91-110) | 107 | 108 | 100 | 100 | 50 | 42 | 83 | 92 | 1.661 | 1.645 |
| (111-130) | 128 | 121 | 100 | 100 | 50 | 58 | 75 | 58 | 1.667 | 1.659 |
| (131-150) | 145 | 143 | 100 | 100 | 50 | 42 | 83 | 92 | 1.607 | 1.629 |
| ave. len. | 107 | | 100 | | 66 | | 83 | | 1.639 | |
| MAP | 109 | | 100 | | 83 | | 83 | | 1.656 | |

**Table 5. aaPR**, training set size = 160

| length (range) | length 1 | length 2 | I(3) 1 | I(3) 2 | II(1) 1 | II(1) 2 | III(3) 1 | III(3) 2 | IV(5) 1 | IV(5) 2 | A.E.C. 1 | A.E.C. 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (100-120) | 120 | 118 | 92 | 92 | 92 | 92 | 83 | 83 | 83 | 75 | 1.721 | 1.713 |
| (121-140) | 135 | 136 | 83 | 92 | 92 | 92 | 83 | 75 | 75 | 75 | 1.689 | 1.696 |
| (141-160) | 160 | 155 | 83 | 66 | 92 | 92 | 83 | 83 | 75 | 66 | 1.661 | 1.639 |
| (161-180) | 174 | 177 | 66 | 83 | 92 | 66 | 83 | 75 | 66 | 83 | 1.612 | 1.655 |
| ave. len. | 146 | | 92 | | 83 | | 66 | | 92 | | 1.696 | |
| MAP | 171 | | 92 | | 100 | | 83 | | 66 | | 1.655 | |

**Table 6. RH**, training set size = 201

The first column of each table indicates the length range or procedure for obtaining the model length. The second column indicates the length of the two highest scoring models from each length range (labeled 1 and 2), the average length, and the MAP procedure of HMMER. The columns with roman numerals indicate each motif and the values in parentheses indicate the number of residues in each motif. Correct motif identification is indicated as the percentage of sequences in which the motif was found in the validation sets. The results for each of the optimal models are labeled 1 and 2, respectively. The last column reports the average entropy/position (A.E.C.) of each model. The higher the value, the lower the actual entropy.

out performs other attempts to infer an adequate model length; and 2) there is a lack of correlation between increasing the number of sequences in the training set and performance based on our criteria (McClure and Raman, 1995). We have conducted similar tests on the HMMER method. The first test is designed to determine if the default MAP procedure also proves superior over other attempts to infer an appropriate model length. In these studies all optimal models within each model length range (tables 3-6) are generated. The A.E.C. is calculated for the two models within each range with the highest average log-odds probability score. The validation sequences are aligned to these models and the resulting alignments are assessed for correct identification of the ordered series of motifs thereby selecting the B.I. model.

A similar approach in studying the effect of training set size on the generation of the best model is conducted by decreasing the training size by a specific number of sequences (tables 7-9). In these studies the model length is determined by the MAP procedure. In all tests the smallest training set size is 50 sequences which is considered to be sufficient to generate an adequate model by both the SAM and HMMER methods.

The third phase of these studies, assessing the effects of the distribution of pairwise sequence similarities on model generation, is on going and will not be described further at this time.

## Results

### Parameters and initializations

Table 1 lists the changes to the defaults settings and table 2 displays the results of using these parameters for the initial SAM and HMMER evaluation. In the SAM studies a significant difference is observed between models generated by the default and derived parameter settings. This is illustrated by the increase in motif identification in the four validation sets when these sequences are aligned by the models generated by the derived parameter settings (table 2). Models were also generated under the default match and insert state initializations, with the other parameters set to the derived values (table 1). The results are similar or worse than those obtained under all defaults settings (unpublished observations).

In the test of the HMMER method an insignificant difference in motif identification for the globins is observed, while either set of parameters produces models which correctly identify the eight kinase motifs. As in the SAM study, an improvement in motif identification for the more difficult tests of the aaPR and RH sequences is evident in the HMMER generated models (table 2). A comparison of the performance of the SAM and HMMER methods using the derived settings indicates that the SAM method performs better than HMMER for the globin and aaPR tests, there was no difference in the kinase test, and minor differences for the RH test (table 2).

### Model length studies for HMMER

Models were generated for each training set for: 1) each length within each model range; 2) the average length of the training set sequences; and 3) MAP length procedure (figure 1). The match state emissions were set to the normalized amino acid frequencies of each training set, and the uniform insert state emissions probabilities were used in each test. Two models within each length range for each test were selected as described in the Methods. In the globin test the second most optimal model, with a model length of 140, and an A.E.C. of 1.075 is the B.I. model. The model produced by the MAP procedure has a higher A.E.C., but fails to identify all five motifs (table 3). In the kinase study there are four models with shorter model lengths and three models with longer lengths than the MAP produced model which qualify as B.I. models. All of these models have A.E.C. values greater than the MAP model (table 4). In both the aaPR and RH studies the MAP procedure models are better than all other models. The A.E.C. of the aaPR MAP model is among those models with the highest value, while the A.E.C. RH MAP model is among those with the lowest values (tables 5 and 6). In general, the use of the average length of the training sequences as a model length did not perform as will as the other approaches (tables 3-6).

### Training set size studies for HMMER

Surprisingly, although the MAP procedure fails to find the B.I. model when the training set is all globin sequences (873), two different B.I. models are found by the MAP procedure with smaller numbers of training sequences (table 7). In the kinase study the MAP procedure found B.I. models for the three smallest training sequences sets (table 8). Likewise, in the aaPR test there is a lack of correlation between increased training set size and B.I. model generation or increase in correct motif identification (table 9). In the RH study, however, the MAP procedure produced two models from smaller training set sizes which identified a higher proportion of the four motifs (table 10, training size 100 and 160) than the model generated from all the training sequences (table 6, training size 201). Each training set size in each test was randomly sampled five times with similar results.

## Conclusions and future studies

The objective of multiple sequence alignment of highly divergent, but clearly related protein sequences, is to identify the regions of the sequences common to a

## Effect of training set size on model quality for HMMER method

| training size | model length | I(7) | II(5) | III(5) | IV(5) | V(3 |
|---|---|---|---|---|---|---|
| 50 | 148 | 92 | 100 | 100 | 100 | 100 |
| 100 | 157 | 100 | 100 | 100 | 100 | 100 |
| 300 | 162 | 100 | 100 | 100 | 100 | 100 |
| 500 | 164 | 100 | 100 | 100 | 58 | 75 |
| 800 | 160 | 100 | 100 | 100 | 58 | 83 |

**Table 7. Globins,** total number of sequences 873

| training size | model length | I(6) | II(1) | III(1) | IV(9) | V(3) | VI(3) | VII(8) | VIII(1) |
|---|---|---|---|---|---|---|---|---|---|
| 50 | 275 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 100 | 288 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 140 | 269 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 180 | 273 | 100 | 100 | 100 | 100 | 100 | 100 | 92 | 100 |
| 220 | 272 | 83 | 100 | 92 | 100 | 100 | 100 | 100 | 100 |

**Table 8. Kinase,** total number of sequences 225

| training size | model length | I(3) | II(5) | III(5) |
|---|---|---|---|---|
| 50 | 117 | 92 | 42 | 50 |
| 80 | 108 | 100 | 58 | 75 |
| 100 | 107 | 100 | 42 | 83 |
| 125 | 107 | 100 | 58 | 75 |
| 150 | 111 | 100 | 42 | 75 |

**Table 9. aaPR,** total number of sequences 160

| training size | model length | I(3) | II(1) | III(3) | IV(5) |
|---|---|---|---|---|---|
| 50 | 154 | 83 | 100 | 75 | 66 |
| 75 | 146 | 92 | 75 | 83 | 50 |
| 100 | 169 | 100 | 100 | 83 | 75 |
| 130 | 167 | 83 | 100 | 83 | 83 |
| 160 | 170 | 92 | 100 | 83 | 75 |

**Table 10. RH,** total number of sequences, 201

The first column of each table indicates the number of randomly selected sequences in each training set. The second column indicates the model length. The columns with roman numerals indicate each motif and the values in parentheses indicate the number of residues in each motif. Correct motif identification is indicated as the percentage of sequences in which the motif was found in the validation sets.

specific protein family, as well as the regions which are specific to various subsets of the family. Given that RNA genomes mutate significantly faster than DNA genomes, and there is no way to estimate the actual number of mutations that have occurred, the principle of maximum parsimony is applied in manual refinement of alignments. Basically indels are adjusted to maximize matches and conservative substitutions by invoking the fewest number of evolutionary events which account for the observed differences/similarities among the sequences. Our goal is to derive parameter constraints for the generation of *de novo* HMMs which reduce the amount of manual

refinement required to produce an initial multiple alignment, and establish guidelines for the use of these new methods in inferring patterns among new protein sequences.

Based on the results of the studies presented in table 2 it appears that the initialization of the match state emission frequencies has a significant effect on models produced by the SAM method. Although this effect is less pronounced in the HMMER models the use of data specific match emission frequencies and uniform insert probabilities does increase motif identification for distantly related sequences. It is recommended that users explore the use of data specific match emission frequencies when using either of these methods for modeling highly divergent protein sequences.

In sharp contrast to our previous analysis of the effect of model length on model quality of the SAM method (McClure and Raman, 1995), the HMMER MAP procedure of adjusting the model length does not always out perform a more empirical approach to finding an adequate model length. Only in the case of the aaPR model is the MAP procedure better. The results of the range study indicate that several B.I. models for both the globin and kinase tests can be found at various model lengths. Two RH models capable of better motif identification than the MAP procedure model can be found by the model range approach. The HMMER method provides an experimental option for controlling the model length. Given the variability of models lengths which can produce B.I. models, we will explore the use of this option in generation of a model that can produce a more compact alignment. Such a model may provide a better approximation to multiple alignments that are manually refined.

Although our previous SAM studies indicated that B.I. models are among those with the highest A.E.C. values, this is not the case in the HMMER studies. It should be pointed out that in both studies the A.E.C. values indicate the quality of the model, while the B.I. criterion is used to access the biological quality of the validation sequence alignment generated by the model thereby determining the B.I. model. Neither methods align the residues between the ordered series of motifs for the more divergent tests (aaPR and RH) very well when compared to manually refined alignments. The variability of correlation between either the optimal model or the B.I. model with the highest A.E.C. model may reflect the inadequate alignment of regions between the ordered series of motifs. Once the stability of the model length is achieved for the B.I. models, studies addressing the appropriate alignment of regions between the ordered series of motifs will be initiated. We will explore the use of the experimental PAM matrix option provided by the HMMER method which may prove useful in addressing this problem

The effect of training set size on the generation of HMMER generated models confirms our earlier finding, that contrary to expectation, there is little correlation between the size of the training set and the quality of the model (McClure and Raman, 1995). It is evident that available sequences for various protein families are biased in their distribution. In the studies presented here it is the validation sequences with the least representation in the training set which are mis-aligned. While there is extensive interest in the development of various weighting schemes to correct this problem, we are in the process of developing training sets which more accurately represent the distribution of observed relationships. Figure 1 outlines our approach in generating representative training sets. While these studies are underway for both the globin and kinase families, we are updating both the aaPR and RH datasets to increase the representation of distantly related members. We are investigating the effects of the distribution of sequence similarity in the training sequences on the generation models by both the SAM and HMMER methods.

While the HMM approach to multiple sequence alignment remains the most flexible method available to date, the recent development of these methods mandates a considerable research effort to determine adequate parameter constraints for the application of these methods to real data for *de novo* model generation. The foresight of the developers in providing the ability of model acquisition from initial alignment, or the MAFIA approach of working the problem in reverse so to speak, will certainly accelerate the pace of multiple sequence analysis. In addition to the ongoing studies described above we have used the MAFIA approach to build a model for dUTPase sequences from eukaryotes, prokaryotes, several different DNA viruses, and a subset of the retroviruses in which this gene is found (manuscript in preparation). We are also constructing MAFIA models for over 500 RT sequences, as well as all the other proteins of the retroid family (McClure, 1996), and all of the proteins of the order Mononegavirales, which includes Ebola, rabies, and measles viruses. One of the long-term goals of these studies is to generate robust models representing the natural variation of viral proteins sequences on a per residue basis. Such models will set the stage for comparing the natural variation within a given protein population with the rapidly accumulating sequences from chemotheraputically resistant mutants. Such comparisons will not only provide information relevant to the future design of anti-viral agents, but this approach will also provide data to address issues pertaining to the nature of the selection process itself.

## References

1. Asai, K.; Hayamizu, S. Handa, K. 1993. Secondary structure prediction by hidden Markov model. CABIOS, Vol.9 No.2, 141-146.

2. Baldi, P., Y. Chauvin, T. Hunkapiller, and M.A. McClure. 1994. Hidden Markov models of biological primary sequence information. Proc. Natl. Acad. Sci., USA 91:1059-1063.

3. Baum, L.E. and J.A. Eagon. 1967. An inequaility with applications to statistical estimation of probabilistic functions of Markov processes and to a model for ecology. Bull. Ame. Math. Soc. 70:360-363.
4. Cardon, L.R. and G.D. Stormo. 1992. Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. J. Mole. Biol. 223:159-170.
5. Churchill, G.A. 1989. Stochastic models for heterogenous DNA sequences. Bull. of Math. Biol. 51:79-94.
6. Davies, J.F., Z. Hostomska, Z. Hostomsky, S.R. Jordan, and D.A. Matthews. 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. Science 252:88-95.
7. Eddy, S. *Multiple alignment using hidden Markov models*. p. 114-120 in *Third International Conference on Intelligent Systems for Molecular Biology*. 1995. Cambridge, England: AAAI Press.
8. Fujiwara, Y., M. Asogawa, and A. Konagaya. *Stochastic motif extraction using hidden Markov model*. p. 121-129 in *Second International Conference on Intelligent Systems for Molecular Biology*. 1994. Stanford University, Stanford, CA: AAAI.
9. Hughey, R. and A. Krogh. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. CABIOS in press.
10. Katayanagi, K., M. Miyagawa, M. Matsushima, M. Ishikawa, S. Kanaya, M. Ikehara, T. Matsuzaki, and K. Morikawa. 1990. Three-dimensional structure of ribonuclease H from *E. coli*. Nature 347:306-309.
11. Knighton, D.R., J. Zheng, L.F. Ten Eyck, V.A. Ashford, N.-H. Xuong, S.S. Taylor, and J.M. Sowadski. 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science 254:407-414.
12. Krogh, A., M. Brown, I.S. Mian, K. Sjolander, and D. Haussler. 1994. Hidden Markov models in computational biology: applications to protein modeling. J. Mole. Biol. 235:1501-1531.
13. Lawrence, C.E. and A.A. Reilly. 1990. An expectation maximization (EM) algorithm for the identification and characterization of common sites of unaligned biopolymer sequences. Proteins: Struct. Funct. Genet. 7:41-51.
14. McClure, M., A. 1993. Evolutionary history of reverse transcriptase, p. 425-444. in A.M. Skalka and S.P. Goff (eds.), Reverse Transcriptase, Cold Spring Harbor Laboratory Press.
15. McClure, M.A. 1996. The molecular evolution of the retroid family, p. 404-415. in A. Gibbs, C.H. Calisher, and F. Garcia-Arenal (eds.), *Molecular Basis of Virus Evolution*, Cambridge University Press: Cambridge, England.
16. McClure, M.A. and R. Raman. *Parameterization studies of hidden Markov models representing highly divergent protein sequences*. p. 184-193 in *Proceedings of the 28th Annual Hawaii Conference on System Science*. 1995. Maui, Hawaii: IEEE Computer Society Press.
17. McClure, M.A., T.K. Vasi, and W.M. Fitch. 1994. Comparative analysis of multiple protein-sequence alignment methods. Mole. Biol. Evol. 11:571-592.
18. Miller, M., M. Jaskolski, J.K. Mohana Rao, J. Leis, and A. Wlodawer. 1989. Crystal structure of a retroviral protease proves relationship to aspartic protease family. Nature 337:576-579.
19. Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition p. 257-285in Proceedings of the IEEE 77:257-285.
20. Stultz, C.M., J.V. White, and T.F. Smith. 1993. Structural analysis based on state-space modeling. Protein Science 2:305-14.
21. Thorne, J.L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33:114-124.