5

Profile HMMs for sequence families

So far we have concentrated on the intrinsic properties of single sequences, such as CpG islands in DNA, or on pairwise alignment of sequences. However, functional biological sequences typically come in families, and many of the most powerful sequence analysis methods are based on identifying the relationship of an individual sequence to a sequence family. Sequences in a family will have diverged from each other in their primary sequence during evolution, having separated either by a duplication in the genome, or by speciation giving rise to corresponding sequences in related organisms. In either case they normally maintain the same or a related function. Therefore, identifying that a sequence belongs to a family, and aligning it to the other members, often allows inferences about its function.

If you already have a set of sequences belonging to a family, you can perform a database search for more members using pairwise alignment with one of the known family members as the query sequence. To be more thorough, you could even search with all the known members one by one. However, pairwise searching with any one of the members may not find sequences distantly related to the ones you have already. An alternative approach is to use statistical features of the whole set of sequences in the search. Similarly, even when family membership is clear, accurate alignment can be often be improved significantly by concentrating on features that are conserved in the whole family.

How, in brief, do we identify such features? Just as a pairwise alignment captures much of the relationship between two sequences, a multiple alignment can show how the sequences in a family relate to each other. Figure 5.1 shows a multiple alignment of seven sequences from the large globin family (hundreds of globin sequences are available in the protein sequence databases). The three dimensional structure has been obtained for each protein in the alignment shown, and the sequences have been aligned on the basis of aligning the eight alpha helices of the conserved globin fold, and also on the basis of aligning certain key residues in the sequences, such as two conserved histidines (H) which are the residues which interact with an oxygen-binding heme prosthetic group in the globin active site.

It is clear that some positions in the globin alignment are more conserved than others. In general the helices are more conserved than the loop regions between

```
Helix                AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus            Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE            FFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGGG     HHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus    v.   f  l . ..  ....    f   . aa. k. .     l sky
```

Figure 5.1 *An alignment of seven globins from Bashford, Chothia & Lesk [1987]. To the left is the protein identifier in the* SWISS-PROT *database [Bairoch & Apweiler 1997]. The eight alpha helices are shown as* A–H *above the alignment. A consensus line below the alignment indicates residues that are identical among at least six of the seven sequences in upper case, ones identical in four or five sequences in lower case, and positions where there is a residue identical in three sequences with a dot.*

them, and certain residues are particularly strongly conserved. When identifying a new sequence as a globin, it would be desirable to concentrate on checking that these more conserved features are present. How to obtain and use such information will be the subject of this chapter.

As might be expected, our approach to consensus modelling will be to make a probabilistic model. In particular, we will develop a particular type of hidden Markov model well suited to modelling multiple alignments. We call these *profile HMMs* after standard *profiles*, which are closely related non-probabilistic structures introduced previously for the same purpose by Gribskov, McLachlan & Eisenberg [1987]. Profile HMMs are probably the most popular application of hidden Markov models in molecular biology at the moment [Eddy 1996].

We will assume for the purposes of this chapter that we are given a correct multiple alignment, from which we will build a model that can be used to find and score potential matches to new sequences. The multiple alignment could

be built from structural information, like the globin alignment shown here, or it could come from a sequence-based alignment procedure, such as those discussed in Chapter 6.

Much of this chapter makes use of the theory presented in Chapter 3 for general HMMs. The most important algorithms will be presented again in the specific form relevant to profile HMMs. There is also an extensive discussion of how to estimate optimal probability parameters from multiple sequence alignments.

5.1 Ungapped score matrices

One general feature of protein family multiple alignments, which can be seen in Figure 5.1, is that gaps tend to line up with each other, leaving solid blocks where there are no insertions or deletions in any of the sequences. We will start by considering models for these ungapped regions.

As an example, consider the E helix of Figure 5.1. A natural probabilistic model for such a region would be to specify independent probabilities $e_i(a)$ of observing amino acid a in position i (we use letter e because these will turn out to be the *emission probabilities* of the hidden Markov model when we introduce gaps). The probability of a new sequence x according to this model is then

$$P(x|M) = \prod_{i=1}^{L} e_i(x_i),$$

where L is the length of the block, 21 in this case. As usual, we are in fact more interested in the ratio of this probability to the probability of x under a random model, and so to test for membership in the family we evaluate the log-odds ratio

$$S = \sum_{i=1}^{L} \log \frac{e_i(x_i)}{q_{x_i}}.$$

The values $\log \frac{e_i(a)}{q_a}$ behave like elements in a score matrix $s(a,b)$, where the second index is position i, rather than amino acid b. For this reason, such an approach is known as a *position specific score matrix* (PSSM). A PSSM can be used to search for a match in a longer sequence x of length N by evaluating the score S_j for each starting point j in x from 1 to $N-L+1$, where L is the length of the PSSM.

5.2 Adding insert and delete states to obtain profile HMMs

Although a PSSM captures some conservation information, it is clearly an inadequate representation of all the information in a multiple alignment of a protein

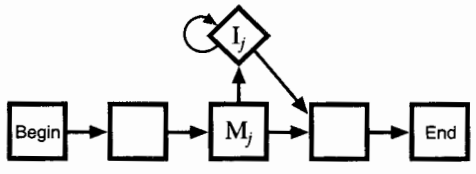
family. We have to find some way to take account of gaps. It is possible to combine the scores of multiple ungapped block models, and this is the approach taken by Henikoff & Henikoff [1991] in the BLOCKS database. However, we will pursue here the aim of developing a single probabilistic model for the whole extent of the alignment.

One approach is to allow gaps at each position in the alignment, using the same gap score $\gamma(g)$ at each position, as in pairwise alignment. However, this is also ignoring information, because the alignment gives us explicit indications of where gaps are more and less likely. We want to capture this information to give us position sensitive gap scores, just as the emission probabilities gave us position sensitive substitution scores.

The approach we take is to build a hidden Markov model (HMM), with a repetitive structure of states, but different probabilities in each position. This will provide a full probabilistic model for sequences in the sequence family. We start off by observing that the PSSM can be viewed as a trivial HMM with a series of identical states that we will call *match* states, separated by transitions of probability 1.

Alignment is trivial because there is no choice of transitions. We rename the emission probabilities for the match states to $e_{M_i}(a)$.

The next step is to deal with gaps. We must treat insertions and deletions separately. To handle insertions, i.e. portions of x that do not match anything in the model, we introduce a set of new states I_i, where I_i will be used to match insertions after the residue matching the ith column of the multiple alignment. The I_i have emission distribution $e_{I_i}(a)$, but these are normally set to the background distribution q_a, just as for seeing an unaligned inserted residue in a pairwise alignment. We need transitions from M_i to I_i, a loop transition from I_i to itself, to accommodate multi-residue insertions, and a transition back from I_i to M_{i+1}. Here is a single insert state of this kind:

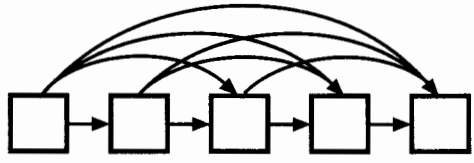


We denote insert states in our diagrams by diamonds. The log-odds cost of an insert is the sum of the costs of the relevant transitions and emissions. Assuming that $e_{I_i}(a) = q_a$ as described above, there is no log-odds contribution from the emission, and the score of a gap of length k is

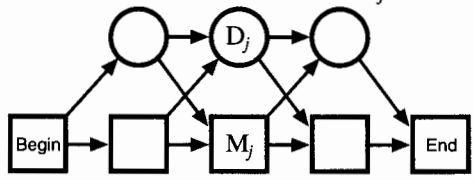
$$\log a_{M_j I_j} + \log a_{I_j M_{j+1}} + (k-1)\log a_{I_j I_j}.$$

From this you can see that the type of insert state shown corresponds to an affine gap scoring model.

Deletions, i.e. segments of the multiple alignment that are not matched by any residue in x, could be handled by forward 'jump' transitions between non-neighbouring match states:



However, to allow arbitrarily long gaps in a long model this way would require a lot of transitions. Instead we introduce silent states D_j as described in Section 3.4:



Because the silent states do not emit any residues, it is possible to use a sequence of them to get from any match state to any later one, between two residues in the sequence. The cost of a deletion will then be the sum of the costs of an M $\rightarrow$ D transition followed by a number of D $\rightarrow$ D transitions, then a D $\rightarrow$ M transition. This is at first sight exactly analogous to the cost of an insert, although the path through the model looks different. In detail, it is possible that the D $\rightarrow$ D transitions will have different probabilities, and hence contribute differently to the score, whereas all the I $\rightarrow$ I transitions for one insert involve the same state, and so are guaranteed to have the same cost.

The full resulting HMM has the structure shown in Figure 5.2. This form of model, which we call a profile HMM, was first introduced in Haussler *et al.* [1993] and Krogh *et al.* [1994]. We have added transitions between insert and delete states, as they did, although these are usually very improbable. Leaving them out has negligible effect on scoring a match, but can create problems when building the model.

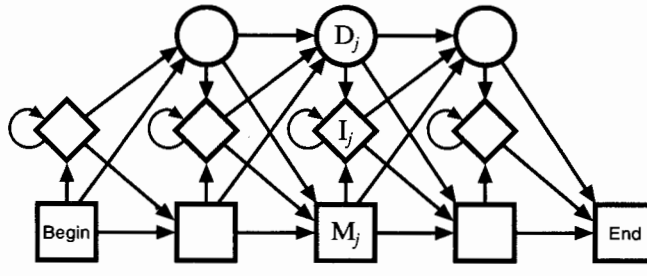


Figure 5.2 *The transition structure of a profile HMM. We use diamonds to indicate the insert states and and circles for the delete states.*

Profile HMMs generalise pairwise alignment

We have seen how the costs of using gap states in a profile HMM mirror those used in pairwise alignment with affine gaps. To help make clear the relationship, it is useful to consider the degenerate case where the multiple alignment from which we build the HMM contains just one sequence.

Let us compare Figure 5.2 with Figure 4.2. If we call the example sequence y, then Figure 5.2 is an unrolled version of Figure 4.2, with the y_j emissions each coming from a separate copy of the pair HMM. The states M_j correspond to a sequence of match states M, the I_j to corresponding incarnations of X, and the D_j to incarnations of Y. To achieve as close a correspondence as possible, the natural values for the match emission probabilities $e_{\mathrm{M}_i}(a)$ are $p_{y_i a}/q_{y_i}$, the conditional probabilities of seeing a given y_i in a pairwise alignment, and for the transition probabilities $a_{\mathrm{M}_i\mathrm{I}_i} = a_{\mathrm{M}_i\mathrm{D}_{i+1}} = \delta$ and $a_{\mathrm{I}_i\mathrm{I}_i} = a_{\mathrm{D}_i\mathrm{D}_{i+1}} = \varepsilon$ for all i.

In formal terms our profile HMM is effectively the hidden Markov model obtained by conditioning the pair HMM of Figure 4.2 on emitting sequence y as one of the sequences in its alignment. Because of this, the Viterbi equations for finding the most probable alignment of x to our profile HMM are essentially the same as those for the most probable alignment of x and y to the pair HMM described in Chapter 4. If we convert them into log-odds ratio form we recover our standard affine gap cost pairwise alignment equations of (2.16), as we will see below. Any differences are due to slightly different Begin and End arrangements.

5.3 Deriving profile HMMs from multiple alignments

Although it is nice to see that the profile HMM is doing the same sort of dynamic programming as we have used before for pairwise alignment, this is not why we introduced them. The key idea behind profile HMMs is that we can use the same structure as shown in Figure 5.2, but set the transition and emission probabilities to capture specific information about each position in the multiple alignment of the whole family. Essentially, we want to build a model representing the consensus sequence for the family, not the sequence of any particular member.

There are a number of different ways to derive the parameter values from a multiple alignment of the sequences in the family. To provide an example for illustrating these methods, Figure 5.3 shows a short section of the globin alignment shown in Figure 5.1.

Non-probabilistic profiles

A model similar to the profile HMM was first introduced by Gribskov, McLachlan & Eisenberg [1987] who coined the name 'profile' (see also Gribskov, Lüthy & Eisenberg [1990]). However, they did not have an underlying probabilistic model,

```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                  ***  *****
```

Figure 5.3 *Ten columns from the multiple alignment of seven globin protein sequences shown in Figure 5.1. The starred columns are ones that will be treated as 'matches' in the profile HMM.*

but rather directly assigned position specific scores for each match state and gap penalty, for use in standard 'best match' dynamic programming. They set the scores for each consensus position to the averages of the standard substitution scores from all the residues seen in the corresponding multiple alignment column. For example, they would set the score for residue a in column 1 of our example to be

$$\tfrac{5}{7}s(\mathrm{V},a)+\tfrac{1}{7}s(\mathrm{F},a)+\tfrac{1}{7}s(\mathrm{I},a)$$

where $s(a,b)$ is the standard substitution matrix. They also set gap penalties for each column using a heuristic equation that decreased the cost of a gap (either insertion or deletion) according to the length of the longest gap observed in the multiple alignment spanning the column.

Although this seems an intuitively obvious way to combine information, and it has been used effectively by many people for finding new members of families, it does produce anomalies. For example, column 1 is much more strongly conserved than column 2 in the example shown in Figure 5.3, but the information in column 1 will be smeared out just as much by the substitution matrix as that in column 2. If we had an alignment with 100 sequences, all with a cysteine (C) at some position, then the implicit probability distribution for that column for an 'average' profile would be exactly the same as would be derived from a single sequence. This does not correspond to our expectation that the likelihood of a cysteine should go up as we see more confirming examples.

In addition to these observations about substitution scores, the scores for gaps do not behave as expected. For example, from the alignment in Figure 5.3 the score for a deletion would be set to be the same in column 2, where there is a deletion in one sequence, HBB_HUMAN, as in column 4, where there is a deletion opening in five of the seven sequences. It would be more reasonable to set the probability of a new gap opening to be higher in column 4.

Changes have been made to non-probabilistic profiles to address these and

other problems [Thompson, Higgins & Gibson 1994b; Gribskov & Veretnik 1996], and we shall return to some of these later.

Basic profile HMM parameterisation

Let us turn back to hidden Markov model profiles. Like all HMMs, these have emission and transition probabilities. Assuming that these probabilities are non-zero, a profile HMM can model any possible sequence of residues from the given alphabet. It therefore defines a probability distribution over the whole space of sequences. The aim of the parameterisation process it to make this distribution peak around members of the family.

The parameters we have available to control the shape of the distribution are the values of the probabilities, and also the length of the model. There is a lot to say about setting these optimally. We give here the basic methods from Krogh *et al.* [1994]. After sections on database searching and variants for local alignment, we will return to an extended discussion of alternative parameter estimation techniques.

The choice of length of the model corresponds more precisely to a decision on which multiple alignment columns to assign to match states, and which to assign to insert states. The profile HMM we derived above from the single sequence y had a match state for each residue y_i. However, looking at Figure 5.3 it seems clear that the consensus sequence for this region should only have eight residues. and that the two non-starred residues in `GLB1_GLYDI` should be treated as an insertion with respect to the consensus. For the time being we will use a heuristic rule to decide which columns should correspond to match states, and which to inserts. A simple rule that works well is that columns that are more than half gap characters should be modelled by inserts.

The second problem is how to assign the probability parameters. We regard the alignment as providing a set of independent samples of alignments of sequences x to our HMM. Since the alignments are given, we can estimate the parameters directly using equations (3.18) from Section 3.3. We just count up the number of times each transition or emission is used, and assign probabilities according to

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad \text{and} \quad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')}$$

where k and l are indices over states, and a_{kl} and e_k are the transition and emission probabilities, and A_{kl} and E_k are the corresponding frequencies.

In the limit of having a very large number of sequences in our training alignment, this will give an accurate and consistent estimate of the probabilities. However, it has problems when there are only a few sequences. A major difficulty is that some transitions or emissions may not be seen in the training alignment, and so would acquire zero probability, which would mean they would never be

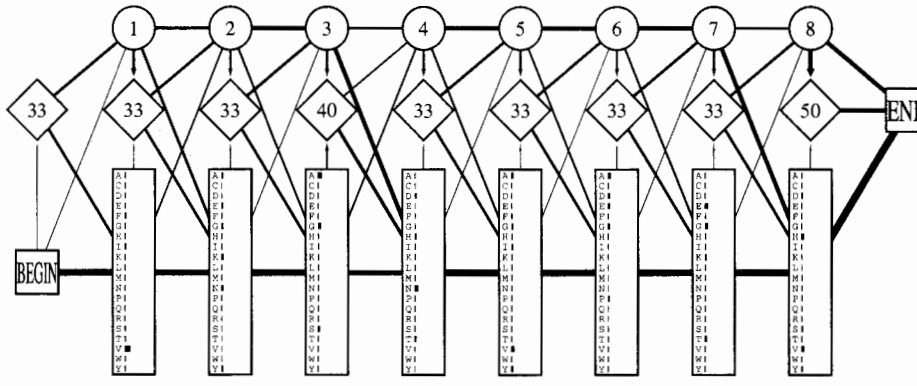


Figure 5.4 *A hidden Markov model derived from the small alignment shown in Figure 5.3 using Laplace's rule. Emission probabilities are shown as bars opposite the different amino acids for each match state, and transition probabilities are indicated by the thickness of the lines. The* $\mathrm{I} \rightarrow \mathrm{I}$ *transition probabilities times 100 are shown in the insert states. (Figure generated automatically using the* SAM *package.)*

allowed in the future. More broadly, we are not using any previous knowledge about protein alignments, as the earlier non-probabilistic methods did implicitly, by using an independently derived substitution matrix. As a minimal approach to avoid zero probabilities, we can add pseudocounts to the observed frequencies (as in Chapters 1 and 3). The simplest pseudocount method is Laplace's rule: to add one to each frequency. We discuss better ways to choose the pseudocount values, and other approaches to estimating the parameters, at greater length below in Section 5.6.

Example: Parameters for an HMM based on Figure 5.3

Let us assume that we use Laplace's rule to obtain parameters for an HMM corresponding to the alignment in Figure 5.3. Then $e_{\mathrm{M}_1}(\mathtt{V}) = 6/27, e_{\mathrm{M}_1}(\mathtt{I}) = e_{\mathrm{M}_1}(\mathtt{F}) = 2/27$, and $e_{\mathrm{M}_1}(a) = 1/27$ for all residue types a other than V, I, F. Similarly, $a_{\mathrm{M}_1\mathrm{M}_2} = 7/10, a_{\mathrm{M}_1\mathrm{D}_2} = 2/10$ and $a_{\mathrm{M}_1\mathrm{I}_1} = 1/10$ (following column 1 there are six transitions from match to match, one transition to a delete state, in HBB_HUMAN, and no insertions). Figure 5.4 shows the complete set of parameters for the HMM in diagrammatic form. □

5.4 Searching with profile HMMs

One of the main purposes of developing profile HMMs is to use them to detect potential membership in a family by obtaining significant matches of a sequence to the profile HMM. We will assume for now that we are looking for global matches.