Chapter 4
Hidden Markov models for database similarity searches

Ewan Birney
The Sanger Centre, Wellcome Trust Genome Campus, Cambridge, UK.

1 Introduction

Despite the huge number of genes in an organism, the protein coding genes are thought to be made from a limited number of basic protein structures. Evolution has reused these protein structures, combining them to form different proteins, and altering them in different genes to achieve different functions. The diversity of species, each with its own copies of genes made from the limited number of building blocks, means that, for a protein of interest, a number of different related proteins may be found. In this chapter, I will discuss one set of techniques which can be used to take advantage of this diversity of protein sequence. These techniques are all related to the use of *profiles*, which are also discussed in Chapter 5. In this chapter, the emphasis will be on the use of *hidden Markov models* (HMMs) for profile analysis. Some practitioners consider profiles to be a type of HMM.

It is important to realize that a protein might be related to another protein in a variety of different ways. It could be that the entire protein is homologous (that is, derived from a common ancestor) to another, such as human and mouse *src* protein (see *Figure 1*) or the human *src2* protein which is a paralog to the *src* protein. Alternatively only a portion of the protein might be derived from a common ancestor, such as the *fyn* protein, which shares a common C terminal region with a divergent N terminus to the *src* protein. Finally only a small region might be conserved, such as the SH3 domain which is also found in the Grb2 protein (along with many other proteins) with no other organization conserved between the two proteins. This last type of conservation, conservation of a *domain* generally corresponds to a structural domain of the protein which can fold independently and, in most cases, function independently of other regions. Figuring out when you have really defined a domain rather than a more extensive piece of conservation is one of the challenges for a researcher. Profile analysis is useful for all these different types of conservation. It is especially useful for domain analysis as this is the hardest feature to define using other methods.

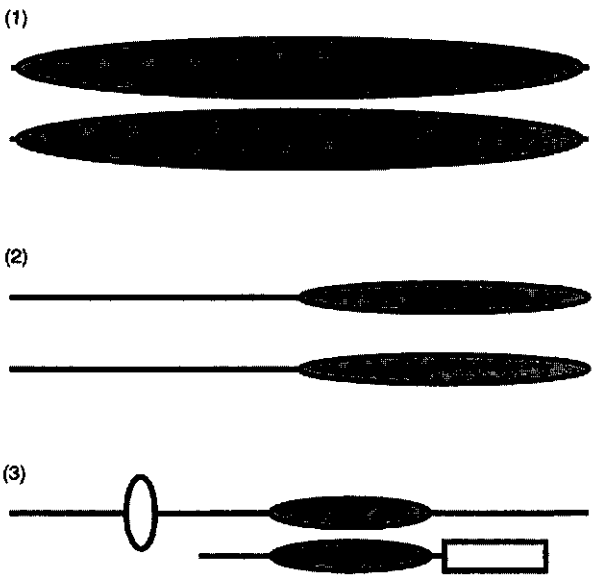

Figure 1 Three different types of relationship are shown. The grey ovals indicate regions which are conserved, whereas the lines and other boxes show regions which are not related. (1) Two very closely related genes, where the entire protein sequence in each gene is conserved. (2) Two genes where the C termini are related but the N termini are unique. (3) Two genes which share one domain but the other regions are entirely different.

2 Overview

For people coming from outside the field, the use of profiles and profile-HMMs can require confronting much confusing jargon and cryptic computer programs. This chapter is meant to demystify this type of analysis. The first point to emphasize is that the programs are, basically, just employing some concept of a 'consensus'. This follows intuitively from the observation that if some sequences have an Aspartate before a critical catalytic residue and others a Glutamate then a new enzyme can be expected to have either an Aspartate or a Glutamate at this position. This sort of simplistic rule is recast into a mathematically convenient form: resulting in some idea of a probability for each possible amino acid at a different position, called a *profile*. The difficulty lies, as in many areas in sequence analysis, that there may be different numbers of amino acids between conserved residues. A consequence of the differing lengths is that there are usually a number of different ways of providing a match to a 'consensus', and some way of choosing the 'best' one must be decided. The variable lengths between conserved residues also makes the statistical behaviour of the technique very hard to handle using conventional statistical analysis.

This chapter will concentrate first on using databases of profile-HMMs through the World Wide Web (WWW), which is by far the easiest way of using them. Then we will concentrate on PSI BLAST (1) which is the easiest do-it-yourself profile method, also available through the Web. The final example will

Table 1 Some useful Web addresses

Protein family sites	
Pfam (Europe)	http://www.sanger.ac.uk/Software/Pfam
Pfam (USA)	http://pfam.wustl.edu/
Prosite Profiles	http://www.isrec.isb-sib.ch/profile/profile.html
Prints	http://www.biochem.ucl.ac.uk/dbbrowser/PRINTS
BLOCKS	http://www.blocks.fhcrc.org/
SMART	http://coot.embl-heidelberg.de/SMART/
Software tools	
PSI-BLAST	http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast
HMMER2	http://hmmer.wustl.edu/

cover the use of the HMMER2 (2) package which I find to be the most effective profile-HMM package available. It is UNIX based and relatively easy to use, though it is not currently available through the Web. Finally I will outline some of the theories behind profile HMMs from the point of view of how it impacts on their practical use.

The reader should be aware that there are many other profile-HMM packages. I would draw your attention in particular to the Meta-MEME package (3) and the PROBE package (4) as well thought out solutions. There are also a number of other profile packages (5, 6) which are more focused on the use of the package by their own groups. Finally, a number of commercial solutions exist (7–9), and you may well have access to them. If you know someone on site who is already skilled in using one of these packages, it is best to use that local knowledge and treat this chapter as more of an introduction to the concepts involved. In addition, it is likely that when you are reading this chapter, that new methods or new presentations of old methods will become available. To keep up to date, use the web, and try the URLs in *Table 1* to find the most up-to-date resources.

Finally, I would like to warn users that I have a strong bias towards using a probabilistic framework to explain and justify the methods: this fits easiest with the HMM formalism and the use of Bayesian statistics (a branch of probability analysis). Other researchers are less zealous about using this sort of framework to explain the results. In either case the most important question is whether these methods are biologically useful, whatever the theories say.

3 Using profile and profile-HMM databases

The starting-point of this sort of analysis is usually a protein sequence which you might have derived from your own sequencing project of a gene of interest. The aim is to use a pre-made profile-HMM, previously constructed by another group to highlight regions of your protein which have homology to already well characterized domains. All these resources focus on domains, being the conserved building blocks of proteins, so the end result of the analysis will be to return which regions of your protein look as if they have particular domains.

Once you have a protein sequence it is probably best to put it into Fasta format (see *Section 8.5*), though many resources will allow you to use other formats. Then connect to one of the resources shown in *Table 1* and find the page for searching with your sequence.

Choose the 'search' page. Then use the 'file-upload' button on the forms to submit your own sequence, and click 'submit' or 'run analysis'. The search against the database will probably take a little over a minute, and should not take more than 10 minutes. Each resource returns its own particular format of results, but what is generally reported is the type of the domain, the position in your protein as a start- and end-point, and some indication of how confident your can be of the hit. They all provide a nice graphical representation of the domain on your sequence as a cartoon of the sequence with different coloured or shaped regions indicating the different domains. Clicking on the graphic will take you usually to an in-depth description of the domain, which in many cases will contain links to other resources and literature references. How to interpret the precise results varies from resource to resource.

3.1 Pfam

Pfam (10) is a database of protein families and corresponding profile-HMMs. Pfam uses the HMMER2 package to provide tools for making the HMMs in the first place and then for searching them. A search against Pfam will provide you with three ways of deciding confidence in the matches. The first is a classical e-value (expectation value) which generally is considered significant if it is below 1.0 for individual searches. The second is the Bits score which is derived from the underlying scoring scheme used to score the match between the sequence and the profile-HMMs. It is related to the Bayesian inference of the probability of the match (for a deeper explanation of the statistics read *Section 8*). A final check is provided by a manually derived cut-off which an 'expert' has chosen to separate the true examples from false examples. These cut-offs are chosen conservatively so that, to the researcher's knowledge, they do not misclassify any protein. This can mean that, in some cases, known trues are missed using this cut-off.

At the time of writing the Pfam database (Version 3.3) had 1344 protein families, which covered 57% of the protein primary sequence database. In new genome projects over one third of proteins had at least one hit to a protein family.

3.2 Prosite profiles

Prosite profiles (11) are an addition to the Prosite resource to define protein domains using profile-HMM technology. Prosite profile reports a classical e-value type statistic which is presented on a log scale. Scores above 5 can be considered significant and scores above 7 very significant. The raw score is not a meaningful number except for different examples of the same domain the better the score the better the match. There are no manually set cut-offs.

At the time of writing, there were 205 prosite profiles. There is no reliable

way of estimating the coverage of prosite profiles. However, a researcher can combine a prosite profile search with a Pfam search, allowing the two resources to be combined in the same submission, with a common output. You are advised to make sure the prosite form is using the most up-to-date Pfam release.

3.3 SMART

SMART (12) is currently based on conventional, non HMM profile technology. The raw score is meaningless, rather you must trust the manually set cut-offs provided internally. At the time of writing there were 302 profiles in SMART. SMART is not focused on coverage but rather on providing very accurate alignments and resources of the domains of interest. It is likely by the time of reading this that SMART has switched to using the HMMER2 package rather than old style profiles.

3.4 Other resources and future directions

There are a number of other resources which provide access to ready made homology databases; in particular PRINTS (13) and BLOCKS (14) (see *Table 1* and Chapter 5). Both these resources are less focused on finding individual domains and instead focus on finding smaller 'motifs'. They may give less clear-cut answers, and for difficult domains may be less sensitive, but come with a number of useful utilities and options.

By the time of reading this chapter, it is likely that a number of resources will be using a common documentation resource (Interpro). This will provide more consistent documentation between the different resources, and is likely to be a forerunner to further integration between the resources.

3.5 Limitations of profile-HMM databases

The obvious limitation of a profile HMM database is that if the domains in your protein are not represented in the database then the databases will (hopefully) return nothing. The only option here is to start your own profile analysis using one of the techniques listed below. In addition, it might be that there is an error in the database, giving the wrong start/end points or misclassifying a region. In all the above databases, the most likely error will be that you miss a true domain in your protein. You can lower the thresholds for determining whether a domain exists or not, but be careful that you do not simply accept a false match due to 'noise'. Use the e-value statistic to decide whether this domain is justified on statistical grounds and read the information in section 7.0 on validating matches.

4 Using PSI-BLAST

PSI-BLAST (1) is a profile building and searching package which is fast, accessible through the Web, and aimed at a less expert audience than the other profile packages. This makes it ideal for occasional use or quick investigations of a particular protein sequence.

In many ways PSI-BLAST follows the same methodology as using the HMMER package below, just that this is done behind the scenes. PSI-BLAST starts from a single sequence, which is then searched against a database using the fast BLAST method. The resulting matches are aligned back to the query sequence, and this derived multiple alignment is used to estimate a profile. The profile is then used to search the database, collect homologues, and align back to the profile, and so the process iterates onwards until it stabilizes or some cut-off is exceeded.

A URL to start the process off is given in *Table 1*. You load in your protein sequence and launch the first search. At the end of each search you have the option of including or rejecting each sequence for the next iteration. This gives you the chance to eliminate potential false positives and include weak but true matches from your knowledge of the biology. An e-value statistic is provided to give an automatic selection of the next round of sequences, which should guide you in your selection.

Many of the problems inherent in using PSI-BLAST are also present when using HMMER, and so I would encourage you to read sections 6.0 and 7.0 carefully. Crucially you must be aware that the statistic to quote for the significance of a match is the first one in which it appears in the profile: once a particular sequence has been included in the set which makes the profile, it will, unsurprisingly, score very well against the resulting profile.

The other problems of PSI-BLAST are less to do with the method and more to how it is used. Because it starts with a single sequence, it is tempting to put in an entire sequence of interest and simply start iterating. If the sequence contains one common domain, although PSI-BLAST will find all the homologues of the sequence, both including the domain and excluding it your results will be dominated by this domain and become unmanageable. As you focus your effort on a particular region, it is better to excise that region and use that as a starting point for further analysis.

5 Using HMMER2

HMMER2 (2) is a package of UNIX command line programs which make and use profile HMMs. If you have no experience of the UNIX command line, then using HMMER2 is going to be a struggle. I suggest taking a short course in UNIX first. In addition to the HMMER2 software, you will need a number of other reasonably standard bioinformatics resources. In particular:

(a) A copy of an up-to-date protein database in fasta format as a single file.

(b) A method of retrieving sequences from this database, preferably with the ability to retrieve only a portion.

(c) A multiple alignment program such as Clustal W (17).

(d) A specialized multiple alignment editor.

It may also be useful to have some experience of a text reformatting language such as Perl or Python, or access to someone who can write small glue programs

for you. Installing the resources is best done with the co-operation of the systems support group for the UNIX machine you are using.

5.1 Overview of using HMMER

Figure 2 gives the basic flow of profile analysis. The main steps are to create a multiple alignment of the region of interest and from this multiple alignment make a profile-HMM. The profile-HMM is searched against a protein database: a number of new protein matches may be identified and these can then be incorporated into the multiple alignment. There are two places where human knowledge can make a large difference in the analysis; firstly manual editing of the multiple alignment can produce dramatically better results, secondly which sequences are included or not into new alignments can be vetted using biological knowledge of the process.

5.2 Making the first alignment

To start the whole procedure off one needs to both identify potential homologues and produce the first multiple alignment. The first potential homologues are usually found using single sequence searches. Then these homologues are aligned using a multiple alignment program such as clustal w. This is discussed in depth in Chapter 3. An important issue to realize is that in many cases you will be attempting to make a multiple alignment of a *domain* common to a number of different proteins. To successfully make such a multiple alignment you will need to determine the rough boundaries of the domain in each protein from the single sequence searches. You will then need to excise the regions of the protein with a small (10 residue or so) leeway on each side: hopefully your database retrieval program will have this ability built in. Once the multiple alignment has been made you will probably want to edit it.

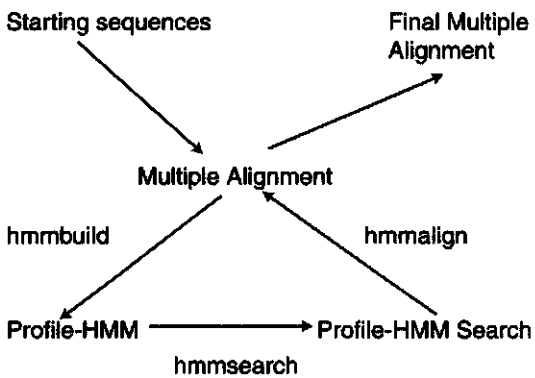

Figure 2 A flow diagram of how profile-HMMs are commonly used. The programs in the HMMer package which are used to provide the different transformations are given beside the arrows. PSI-Blast uses the same principles although much of the mechanics are then hidden from the user

5.3 Making a profile-HMM from an alignment

Thankfully the latest version of HMMER has sensible defaults for making profileHMMs. The program hmmbuild is used to make the HMM, which is run with defaults by typing at the UNIX prompt: 'hmmbuild HMM Alnfile' where alnfile is the name of a file containing a multiple alignment. The defaults which are run are as follows:

(a) **Tree weighting**. A tree is calculated from the multiple alignment and sequences weighted by how many close neighbours they have. This means that overrepresentation by one sub family (for example, many haemoglobin alpha chains for a globin profile-HMM) does not violently bias the profileHMM towards that subfamily.

(b) **Dirchlet mixtures**. A concept of accepted amino acid conservation patterns, such as Valine, Isoleucine, and Leucine being common replacements for each other is provided as Dirchlet mixtures. For more information about Dirchlet mixtures, see *Section 8.4*.

(c) The effective number of sequences is estimated, meaning that alignments which are of very close homologues will be assumed to represent fewer sequences than alignments of more distantly related sequences (see *Section 8.4*).

(d) The placement of which columns in the multiple alignment are match columns as opposed to insert columns.

These defaults are typical for the needs of most HMMs. As you get used to the package, you may want to change them (options are provided on the command line; to get a full list of options type 'hmmbuild -help'), though even experts tend not to deviate too much from these settings.

The only parameter which is worth altering routinely is the local/global mode switch. By default the profile-HMMs are built in global mode, where the entire HMM must be matched against the sequence. An alternative is local mode, by using the -f flag on hmmbuild, which allows only a portion of the HMM to match. If the global mode is used, the HMM becomes far more sensitive for finding distant members, but is unsuitable for finding fragments. If you have misdefined the region of interest, perhaps making it too long at the N or C termini, then the global mode will penalize sequences that do not fit the entire model, preventing them being found. Generally global mode is a better choice, but at the start of the analysis, local mode can be a more sensible choice as one is not always very confident about the locations of the ends of a domain.

Having made the profile-HMM, you need to calibrate it by typing 'hmmcalibrate HMM'. The calibration step calculates statistical parameters for the HMM by generating a random database. This statistical parameterization is crucial for its effective use. Calibration takes around 10 minutes or so.

5.4 Finding homologues and extending the alignment

The HMM can be searched against a database of sequences using the program hmmsearch. This takes the HMM file as the first argument and the database file

as the second argument. The database file is in Fasta format. The results are printed on standard output, so you usually need to redirect the output to save that information in a file. The results give you the following information:

(a) A list of sequences which the HMM hit, ranked from most significant to least.

(b) A list of domains contained in the sequences, ranked from most significant to least.

Notice that a particular sequence can contain more than one domain. In particular, although each of the domain scores might, on their own, not be significant, the combined score of multi-domain match might easily be so.

Both the per-sequence and per-domain matches are provided with two statistics: a bits score and an e-value (see *Section 8.2*). The more reliable score is the e-value; e-values down to 1.0 can be considered significant (an e-value of 1.0 means that, by chance, 1 random sequence is expected to get this score in the database of the size which you used). The bits score is helpful as it is independent of database size.

Having chosen a significance level one would then like to make a new multiple alignment of all the protein sequences found by the HMM. At the moment, this is the most labour intensive step, as the HMMER package does not provide all the functionality for this task. Somehow, one needs to extract all the sequences which are hit and truncate them to the correct start/end points. This is best done by a perl script or similar device. Once you have all the sequences which were hit, as a Fasta file, the program hmmalign will provide a multiple alignment of the proteins on the basis of the HMM. This multiple alignment can then be used to make a new HMM for the next round.

6 False positives

One of the problems inherent to the iterative procedures, both PSI-blast and the use of HMMER outlined above is that if a false positive is added to the alignment, itself and any close relatives will score highly against the profile-HMM. For example, if you inadvertently add a globin sequences to a protein kinase alignment the resulting HMM will match globin sequences surprisingly well.

This ability to start collecting false positives at will means that a researcher should ideally by very vigilant as the iterations progress. Indications that the profile might be picking up noise are:

(a) Low complexity regions occurring in alignment.

(b) A region overlapping a known domain, where it is clear that the multiple alignment is not a divergent subfamily for this domain.

(c) Biological information that indicates that this match is false.

7 Validating a profile-HMM match

Once a researcher has found a suggested domain, how can they validate this? The score of the sequences to the HMM from the final alignment is not the

correct measure of the significance of the match, as it includes all the sequences you wish to score, and they will all score well. In fact the problem of justifying a grouping of sequences is not well handled by the current statistics, in particular when an iterative strategy is used. The following lines of evidence may be used to give a researcher confidence that the similarity they observe is not by chance.

(a) See whether all the sequences can be connected together by significant single sequence scores (e.g. from programs such as BLAST2). Ideally one should be able to show this with the full length proteins (just taking the domain improves the statistics considerably).

(b) Quote the significance of the 'new' sequences for the first time they provided a significant score against the profile-HMM.

(c) To show that A is related B, show that by starting from either A or B one can produce a profile which finds the other sequence using criterion (b).

(d) Provide biological justification that the relationship makes sense (e.g. common mode of enzymatic action). Conversely, biological information which indicates that they should not be related should lessen the researcher's belief in the result.

8 Practical issues of the theories behind profile-HMMs

8.1 Overview of profile-HMMs

A profile is a sequence of conserved positions, each conserved position having a score for each amino acid. For practical purposes, one should neither assume that a particular sequence has all the conserved positions in a profile, nor that a particular sequence will not introduce additional amino acids between two conserved positions in the profile. These two possibilities are the two types of gaps, that is a deletion of part of the profile, or an insertion of residues relative to the profile. A number of *ad hoc* methods have been produced to solve the gap problem in ways that seem biologically sensible.

The *ad hoc* nature of profiles was replaced by a stronger theory based around HMMs which still essentially produced the same sort of profile as before. An HMM is a mathematical model which produces a stream of some observable information in a probablistic manner. In the case of protein sequences, the observable information will be the amino acids. The probablistic nature of the process of producing the amino acids means that a particular HMM can produce more than one set of protein sequences, and also that different sequences are produced with different probabilities. The hidden part of a HMM is that one does not know which model made the particular amino acids one is looking at, and, if one did know the model, which amino acids were made by which part of the model. It is these two problems which one wants to solve, and they correspond to the two questions: 'does this sequence have an example of this HMM in

it' and 'if this sequence does have an example of the HMM, what is the alignment of the HMM to the sequence'.

The HMMs which have been used in this field deliberately mimic the profile model described above. For each conserved column in the multiple alignment, three possible states are permitted: a match state which indicates that a single residue is being aligned to the position, a delete state, which indicates that no amino acid in this protein is present for this position, and an insert state, which allows any number of amino acids to be inserted after the match position. A full-length HMM will have some 500 or so different states, broken down into triplets representing conserved column positions. The behaviour of these states is governed by probabilities for the production of different amino acids from match and insert states and probabilities for the transitions to the neighbouring states. These probabilities are analagous to the scores of the profile and the gap penalities in the profile respectively. Indeed, for practical purposes, the probability representation of a profile-HMM is rarely used. Instead, the probabilities are transformed into sensibly sized integers via a log transformation. In this logged representation, adding the numbers is equivalent to multiplying the underlying probabilities, making the correspondence between profiles and profile-HMMs all the more clear.

Given a particular profile-HMM, the questions 'does this sequence have an example of my HMM' and, given that the last question is true, 'what is the alignment of the sequence to the HMM' can be easily answered using some well known algorithms. These two questions are essentially what hmmsearch and hmmalign provide answers for in the HMMER package.

8.2 Statistics for profile-HMM

Every sequence will match the profile HMM in some manner: some sequences will match the profile HMM better (in the sense that the probability that the HMM would produce such a sequence is higher) than others. How does this statistic (called a likelihood) allow you to say whether this match really is due to the presence of this domain or is it just by chance?

As the underlying basis for HMMs is a probabilistic model, this question is easily answered by Bayesian statistical methods. Non-mathematicians usually have not been exposed to Bayesian statistics previously. Bayesian statisitics try to answer such questions by assigning a probability to its being true or not. In contrast classical statistics answers the question by assuming a particular hypothesis, which is rejected or not by the data. Both approaches are guaranteed to give the same answer when enough data is taken into account, but will often provide different answers for practical problems.

To provide a Bayesian interpretation of the profile-HMM, all possible different models of how the sequence was produced, have to be defined: for profile-HMMs, two models will be considered, the model of the profile-HMM domain and a *null* model of random amino acids drawn from the frequencies found in a large protein database. The probability of seeing the observed sequence under the assumption of the two different models are given, each being a likelihood.

What is quoted is the log-likelihood ratio of the two models: when the base of the log is 2, this statistic (the log-likelihood ratio) is called a *Bits score*. A bits score of 0 is when the likelihood ratio is 1, and hence each model is equally likely to have produced the sequence. Depending on how the ratio is quoted, either more negative or more positive scores indicate that the desired model is more likely. In the HMMER package, the more positive the bits score the better the match to the profile-HMM.

The likelihood ratio does not provide quite enough information to allow an estimate of the probability of the profile-HMM occurring, given the sequence seen. An additional piece of information, being the probability of the profile-HMM occurring without seeing the sequence data needs to be defined. As this information has to be defined without seeing any sequence, it is called *prior* information. Mathematically this is the same idea as the prior information which will be introduced in the next section, but in practice it is used in a very different aspect. Sensible priors include $1/d$ where d is the size of the data base or one could use the probability of a random sequence having this domain in a genome, say 1/10 000. Finally, to be confident of the match, the probability of the profile-HMM occurring should be over 0.95. These two extra manipulations —the prior information and the need for a significant probability translate into a bits cut-off above which one considers matches to be significant. 25 bits translates to sensible choices of prior and significance, and so matches over 25 bits can be considered to be significant.

There is another statistic that can be used to estimate whether the match is significant or not. This is a classical (or frequentist) statistic and is one that most users will be more familiar with. To provide a frequentist statistic, one needs to assume that the match is random, derive the probability that a random match would produce the score, and reject the assumption if this seems very unlikely to have occurred. The problem with this sort of analysis was that it was clear that the distribution of scores of random sequences against a profile-HMM was not normally distributed, and so estimation of probability was very difficult. In recent years the field has produced theoretical and empirical evidence that the distribution is closely related to an Extreme Value Distribution (EVD) (15). PSI-Blast assumes that for a particular way of making a profile, all profiles have the same EVD parameters, regardless of content. PSI-Blast therefore tabulates this information for all possible profile construction mechanisms, and uses the tabulated parameters. HMMER uses a separate calibration step, where the profile-HMM is compared to a large random database, and an EVD is fitted to the resulting distribution. The parameters from this fitting are stored in the HMM so they can be reused for individual sequence searches.

The natural way of reporting the classical statistic is as an expectation value (e-value). This is the number of sequences expected to get this score by chance, and is simply dr where d is data base size and r is the probability that a random sequence will get this score. An e-value of 1.0 is therefore where you expect to start seeing random sequences: e-values less than this are significant.

Which statistic to use: bits score or e-value? It is clear that the e-value statistic

is more robust and more sensitive in the HMMER2 package, and it is what I would recommend. However, the e-value has some less desirable properties, in particular, it changes as the database size changes, unlike the bits score—of course, if you decide the prior on the bits score should be $1/d$ (d is database size), the cut-off for significance of the bits score will change with database size. Quoting both in publications is very sensible.

8.3 Profile-HMM construction

The use of a particular profile-HMM is well understood. A harder problem to solve is 'what profile-HMM best represents my collection of known family members'. By analogy with other fields, such as speech recognition, this problem can be answered by expectation maximization, or similar methods, where an HMM is constructed that maximizes the probability of producing all the sequences known to belong to a certain family. In theory this can work from just the sequences, and no multiple alignment, effectively both aligning the sequences and making the profile-HMM at the same time. In practice, training an HMM from unaligned sequences does not work well, principally because of local minima problems. A better solution is to train the HMM from already aligned sequences, as in *Figure 2*. When already aligned sequences are presented to the training program, the only aspect which the program must estimate is which columns to consider as conserved positions, and which columns should be collapsed into an insert state of the preceding conserved position. Indeed, if you so wish, you can provide this information directly. Once this is known, it is relatively easy to estimate the probabilities for the HMM from the observed sequences using standard theories.

8.4 Priors and evolutionary information

The final problem in HMM construction is that, in general, one is not interested in finding proteins which are only slightly different from the examples one already knows. Rather one wants to find a new subfamily related to the subfamily one has already gathered. This usually means that a sequence from the new subfamily will have some features which are not present in any of the sequences one has already gathered, and yet because of the pattern of conservation and knowledge of behaviour of proteins (for example, a conserved valine position is more likely to have a leucine in a distant subfamily member at this position, than an arginine) one can recognize it as being a related member.

The introduction of extra knowledge into the process of estimating an HMM is called *prior* information, indicating that it is known before any sequence data is seen. Ideally one would like to represent all the knowledge about protein evolution and protein structure in some manner which would allow the profile-HMM construction machinery to use it. In practice a number of assumptions have to be made to allow the mathematics of profile construction to work.

1. Profile-HMM building does not currently work with a concept of the observed sequences being related on an evolutionary tree. Therefore, if you present a

profile-HMM construction method with 10 near-identical alpha globins and one beta globin, the resulting HMM would be predominantly alpha globin. This is solved by weighting the sequences by a tree before applying the profile- building machinery. In the above example, each alpha globin sequence might get a weight of 1/10, and the beta globin sequence a weight of 1.

2. The estimation of amino acid probabilities, taking into account protein evolution, has a stronger theoretical backing. The problem is phrased as an under sampling problem, where although one has a column of, say, 10 amino acids at this position the frequencies of amino acids represented by observation is not an ideal way to estimate the underlying probabilities; clearly not all the amino acids can be represented even once! This problem occurs in many other situations and has been well studied. A good solution is to provide the estimation machinery with prior knowledge of what sort of amino-acid frequencies one expects in columns, for example, one with high leucine, valine, and isoleucine probabilities, and another with a high probability of arginine and lysine, but low probability for hydrophobic amino acids. These distributions are represented in a complicated mathematical form called Dirchlet mixtures (16) and, by using them, the estimation of probabilities for amino acid positions can take into account evolutionary information. A Dirchlet mixture is a just a convenient mathematical form for this information; there is nothing special about them except that they make the downstream mathematics far easier to handle. The Dirchlet mixture can be thought of rather like a protein comparison matrix used in ad hoc profile methods.

3. The final twist to profile-HMM construction is how to balance the information from the Dirchlet mixtures (the prior) with the information from the observed multiple alignment. The problem here is that the aim is to find new sub families so, the fact that one has seen over 1000 different alpha globins, does not mean that the observed amino acid frequencies on their own will make a good HMM for finding beta globin sequences. This is solved in the HMMER package by estimating how many effective sequences one has observed (a collection of divergent sequences will count more than a collection of close relatives).

8.5 Technical issues

In Fasta format the first line starts with the greater-than sign (>) and is followed by the name of the first protein, which should be composed of only characters from the alphabet, or the underscore symbol (_) or numbers. Most programs (though not all) will allow any non space character in the name. Following this line the next lines are the protein sequence in one letter code. The sequence stops at either the end of the file or the next greater-than (>) sign (which marks the start of the next sequence name if there is more than one sequence in the file).

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997). *Nucleic Acids Res.*, **25**, 3389.
2. Eddy, S. R. (1998). *Bioinformatics*, **14**, 755.
3. Grundy, W. N., Bailey, T., Elkan, C. P., and Baker, M. E. (1997). *Comput. Appl. Biosci.*, **13**, 397.
4. Neuwald, A. F., Liu, J. S., Lipman, D. J., and Lawerence, C. E. (1997). *Nucleic Acids Res.*, **25**, 1665.
5. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). *J. Mol. Biol.*, **235**, 1501.
6. Bucher, P., Karplus, K., Moeri, N., and Hoffmann, K. (1996). *Comp. Chem.*, **20**, 3.
7. http://www.gcg.com/
8. http://www.netid.com/
9. http://www.compugen.com/
10. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., and Sonnhammer, E. L. (1999). *Nucleic Acids Res.*, **27**, 260.
11. Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999). *Nucleic Acids Res.*, **27**, 215.
12. Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. (1999). *Nucleic Acids Res.*, **27**, 229.
13. Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., *et al.* (1999). *Nucleic Acids Res.*, **27**, 220.
14. Henikoff, J. G., Henikoff, S., and Pietrokovski, S. (1999). *Nucleic Acids Res.*, **27**, 226.
15. Altschul, S. F. and Gish, W. (1996). In *Methods in enzymology* (ed. R. F. Doolittle), Vol. 266, p. 460. Academic Press.
16. Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjolander, K., and Haussler, D. (1993) In *Proceedings of the First International Conference on Inteligent Systems for Molecular Biology* (ed. L. Hunter, D. Searls, and J. Shavlik), p. 47. AAAI Press, Menlo Park, CA, USA.
17. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). *Nucleic Acids Res.*, **22**, 4673.

This page intentionally left blank