Papillomavirus Alignments and Structure Predictions

Andrew Farmer and Gerald Myers

MS K710, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Analyses of papillomaviruses that take as their point of departure an alignment, either of the nucleic acid or amino acid sequences, will only be as sound as the alignment, which is itself an hypothesis. For this reason, many sequence analyses are conducted over only "unambiguously alignable" stretches of sequence, typically stretches for which the similarities are 50% or greater, and the information in the difficult regions (similarities less than say 30%) is lost to the analysis. In view of the enormous diversity of papillomavirus sequences, the alignments in Part II of the database publications up to this year were focused of necessity upon groups of viral types, which were then "apposed" by eye. These restricted alignments could not support analyses based upon the entire set of PV sequences. This year, in Parts II and III, we have brought into play a new alignment strategy that holds some promise for simultaneously aligning all members of the papillomaviral family.

Most alignment strategies are "progressive", which is to say that the alignment unfolds from the pairs of most similar sequences to the pairs of most dissimilar sequences; the essence of this approach is captured by Doolittle's dictum—"once a gap always a gap" [1]. McClure and coworkers critique twelve different alignment methods, most of which are progressive, according to their abilities to correctly identify ordered series of motifs in highly divergent proteins that have been experimentally studied [2]. Some of the newer multiple alignment programs are intentionally not progressive, partly for the reason that progressive alignments may be trapped by local optima. The Hidden Markov Method (HMM) approach, which we utilize and describe herein, emphasizes position-specific probability distributions of character states, hence a gap in one portion of the alignment may be scored differently than a gap in another portion of the aligment; most alignment programs have position-independent scoring schemes. As HMM is centered upon the columns of information in an unfolding alignment, this approach, sometimes referred to as a "generalized profile", is indifferent to the relatedness of pairs of sequences [3,4]; in order to achieve satisfactory results, HMM should be employed with large sets (40 or more taxa) of highly divergent sequences such as is seen with papillomaviruses.

The HMM approach leads to a model for the sequence set that has been analyzed. Hence with subsequent database searches, this model—sometimes in the form of a "most likely" sequence, a consensus-like sequence—embodies all of the information contained in the data set, not merely one particular sequence. We have also exploited the HMM-generated model for purposes of protein structure prediction, using an array of contemporary algorithms (see Part III sections concerned with the E2 and E4 proteins). Eventually, the sequence alignment and the structure prediction will become intertwined in an effort to optimize the alignment.

In the following text, we first describe in some detail the HMM approach as we have applied it in this compendium, especially in Part II, then we turn to a discussion of our attempts at protein structure prediction.

A. MULTIPLE SEQUENCE ALIGNMENT USING HMM

The Hidden Markov Model (HMM), as it has been applied to sequence analysis, has many similarities to what is called a "profile" [5,6] in terms of the information that it captures concerning a set of related sequences. In a sense, each can be thought of as an extended consensus sequence in which the information retained at each position includes the frequency with which each possible base or amino acid residue is seen in the sequence set at that position. The HMM is constructed from a number of successive nodes generally corresponding to the columns of positional homology of an alignment; each of these nodes contains a match state, an insert state and a delete state. Associated with each of the states in the model is a vector of probabilities that specify the likelihood with which the system might pass to each member of the set of next possible states. Also associated with match states and insert states are vectors of probability specifying the likelihood that the system will generate or “emit” each possible amino acid or nucleotide when in that state (delete states allow for the possibility that a sequence not have a character in a certain column).

The resultant architecture of the HMM allows one to establish a correspondence between the characters of a given sequence and the states of the model. The succession of the characters in the sequence will thus determine a path through the states of the model, and associated with this path will be a likelihood determined both by the probabilities of transition between successive states and the probability that each state has for generating the character that has been assigned to it. Provided that all the probabilities in the model, including both transition and emission probabilities, are non-zero, then each path through the model that is permissible according to the rules governing transitions from one node of the model to the next will have a non-zero probability of generating the given sequence. The task of finding the optimal path through the model for a given sequence, i.e. the path with the highest likelihood, can be thought of as aligning the sequence to the model, and may be solved using dynamic programming techniques.

The most important differences between the profile and the HMM lie not in the resultant information structures, but in the means by which these structures are generated from the sequence data. As with a consensus sequence, the profile is generated from a set of sequences whose alignment has been determined by some independent means. The parameters for describing an HMM can also be derived from a given alignment in this manner. More importantly, however, there exists an algorithm for HMMs that allows one to determine the parameters of the model having the highest likelihood (at least within the neighborhood of the initial model) given a set of unaligned sequences. This approach is quite similar to certain techniques used in connection with artificial intelligence applications, and is known as “training” the model.

The algorithm used in training the parameters of an HMM involves an iterative approach that uses an initial model to estimate an alignment of the given set of sequences, then uses this alignment to re-estimate the model, and so on until the estimates converge to an optimum. For example, if we are given a set of protein sequences that are thought to be related, a good estimate for an initial model can be made by using the frequency distribution of amino acids in the unaligned set as a vector of probabilities assigned to all the match states and insert states of the model; transition probabilities between the states of the initial model can be assigned arbitrarily, or using a prior assessment of the relative frequencies of indel events. All of the sequences in the given set will now be aligned in turn to the model, finding the path through the model that maximizes the likelihood for the given sequence; by aligning all the sequences in the set to the model in this pairwise fashion, one transitively defines a multiple sequence alignment of the sequences to one another. The multiple sequence alignment thus created can be used for an estimate of the parameters of the HMM, by counting the frequency of occurrence of each amino acid at each position of the alignment and the frequency of indel events across the alignment. This adjusted HMM then serves as a model for another round of alignment, and so on. It can be shown that this process is guaranteed to converge to a local maximum of the likelihood function.

To address the problem of guaranteeing convergence to a global maximum for this function, a variation of the simulated annealing algorithm can be applied at each step of the iterative algorithm; this basically allows a stochastically generated sub-optimal alignment to be chosen for the re-estimation of the model’s parameters, where the sub-optimality of the alignment decreases to zero with successive iterations of the re-estimation procedure.

As should be clear from the preceding discussion, the model can be used to generate a multiple sequence alignment of sequences, including sequences not belonging to the set used to train the parameters of the model. One advantage to using the HMM over the standard dynamic programming algorithm for multiple sequence alignments is that since one is really performing a set of pairwise comparisons of the sequences to the model, the time and memory requirements increase only linearly with the number of sequences, as opposed to exponentially with dynamic programming. Further, most algorithms for sequence alignment require position-independent gap penalties, which is unrealistic in the case of most proteins, which are composed of both conserved regions and indel-rich variable regions. The Hidden

Markov Model, on the other hand has parameters for the likelihood of introducing an insertion or deletion that may vary freely from position to position across the model.

An important application of the model is in the discrimination of related sequences from non-related sequences. This is especially useful in connection with database searching. Associated with each sequence in the database is a probability with which the sequence could be generated by the given model. The distribution of likelihood scores for all the sequences in the database will provide a measure of discrimination between similar and non-similar sequence. Using the HMM for database searching has the advantage of utilising a great deal more of the information available for a family of sequences than can be captured by query techniques that force one to use only one sequence from the family or, at best, a standard consensus sequence as a query against the database. We are currently running database searches with HMM-generated models and comparing the results to what might be obtained by other methods (*e.g.*, the pattern approach described in III-91-123 of the 1995 compendium).

We have employed the HMMER implementation that is publicly available (*eddy@genetics.wustl.edu*). Another HMM suite that can be obtained is SAM (*http://www.cse.ucsc.edu/research/compbio/sam.html*). These programs were originally applied to highly studied data sets (globins, EF-hand proteins, etc.) for which some experimentally-based data were available to help assess the alignment results [3,4]. With PV sequences, the results of the approach must be critiqued by scutinizing motifs—do E2 binding sites align, for example? This is problematic as it is not preordained that all motifs (E2 binding sites) need align. Another difficulty encountered by the HMM (and every) alignment method is large indels; we have the least confidence in those. To the extent that these stretches may have arisen through acquisition of genetic material, they may not be intrinsically alignable as they may not be homologous (see E4 alignments in Parts II and III, for example). In short, the alignments in Parts II and III of the compendium are uncertain; alignments in previous publications may be “safer”—although limited—because they were executed over relatively highly related sequences. Both are available on the Web site (*http://hpv-web.lanl.gov*).

We shall now turn to structure-prediction and the potential interplay between primary sequence alignment and structure-based alignment. In a later edition of the compendium, we will resume our survey of PV similarities as uncovered by various methods, including HMM.

B. PV PROTEIN STRUCTURE PREDICTION

A promising starting point for predicting a structure for a given amino acid sequence is to determine whether that sequence is sufficiently similar to any other sequence for which biophysical data, ideally X-ray crystallographic data, is available. Sequences that are 50% are more similar will have similar structures, and less similar sequences can have similar folds over core regions. Our focus herein will be upon weakly similar sequences for which little or no biophysical data is available.

The earliest structure prediction algorithms, such as the Chou-Fasman algorithm, possess a predictive accuracy of no better than about 55%, partly due to the small set of known structures upon which they depend and partly due to their assumptions. Three-state predictions—helix (*H*), sheet (*E*) and coil (*C*)—are more accurate than four-state predictions that include turns (*T*), and the accuracy is poorest at the ends of polypeptides and best in the core regions. Secondary stucture prediction in general is most reliable for transmembrane helices. With the build-up of the protein database and the development of more powerful algorithms, which especially take into account multiple sequence alignments, the predictive accuracy for secondary structure can now reach better than 70%.

SOPM (self-optimised prediction method) is an example of a recent approach to protein secondary structure prediction [7,8]. When applied to 239 dissimilar proteins of known structure, this algorithm yields three-state prediction accuracies of 69% to 73%. Because it involves sizeable subdatabases of sequences and their known structures, it will take longer to run than the older, less accurate algorithms. The basic ideas used in the SOPM are as follows.

First, a sliding window of a fixed size is applied to the protein sequence of unknown secondary structure to define a set of overlapping peptides. For example, suppose we are given the sequence KPQRNSKSTAAL . . . with a window whose size is eight amino acids long and which is moved one amino acid over at each step. Then the resultant set of peptides will be KPQRNSKS, PQRNSKST, QRNSKSTA, RNSKSTAA, NSKSTAAL . . . . Note that most of the amino acids of the original sequence will belong to eight successive peptides, each differing from the previous peptide by the removal of an amino acid from one end and the addition of an amino acid to the other.

Next, each of the peptides thus derived from the original sequence is now compared to a database of peptides that has been created by similar means from a database of proteins of known secondary structure. If the peptide from the query sequence matches a peptide from the database above a certain threshold of similarity, then the similarity score is added to the conformational scores for each of the amino acids in the peptide. In our example, suppose that the first peptide KPQRNSKS matches a peptide in the database RPQRDTKS whose known structure is *HHCCCEEE*, and that the similarity score between these two peptides is 30. If this score is above the threshold parameter, then 30 will be added to the first two amino acids' helical conformational scores, to the next three amino acids' coil conformational scores and to the last three amino acids' sheet conformational scores. There may be other peptides in the database matching the query peptide with alternative predictions for the secondary structure of each of the amino acids, and all these predictive scores will be added together in each of the conformational categories, resulting in a distribution of scores over the possible secondary structure conformations. After the first query peptide has been compared, the process will continue for each of the remaining peptides in the query set. The final scores for an amino acid belonging to eight successive query peptides will thus include the scores for the comparisons of all eight of these peptides against the entire database of peptides of known structure.

After all comparisons have been made, each amino acid in the original protein will have values associated with its propensity to adopt a conformation in each of the secondary structure classes. From the method of calculation detailed above it is clear that the empirical evidence for the prediction of the secondary structure of the amino acid weighs most heavily for that class with the highest score.

However, there are two additional statistics concerning the distribution of the scores over all the classes that can be revealing of the predictive power of this approach.

The first is the actual magnitudes of the scores for any given amino acid. If these are small relative to the cumulative scores for other amino acids, it may indicate a lack of information for the prediction of the secondary structure conformation of that amino acid. This could happen for two reasons: first, if the amino acid is within the window size to either terminus of the original protein, it will belong to proportionately fewer query peptides and have fewer comparisons with the database that could add to its score; second, the amino acid could belong to a series of peptides that for some reason are poorly represented in the database of known structures, and could thus have few comparisons to the database having a large enough similarity score to be added to the conformational scores for the amino acid. In either case, values that are low in magnitude indicate a lack of information in the database for the amino acid in its given environment.

The second statistic that is pertinent to the predictive value of a set of scores for a given amino acid is the difference between the scores of the highest and next-highest scoring classes of secondary structure. If this difference is small, it may be inferred that the information in the database for the amino acid in this particular environment is conflicting. For example, suppose that approximately half of the peptides contributing to a given amino acid's conformational scores support a helical structure, while the other half support its being classed as an element of a beta sheet. In this scenario, it is likely that the cumulative scores for helix and sheet for this amino acid would be nearly equal, and thus the difference between them would be near zero. In order to make this statistic independent of the magnitudes of the scores (which were accounted for in the former statistic), one may normalize the values by dividing the difference between the highest and next-highest scores by the magnitude of the highest score.

It has been shown that the secondary structure of proteins changes much more slowly over time than their primary structure, i.e. mutations in the sequence of amino acids comprising the sequence often do not alter the secondary structure conformations adopted by the amino acids at these positions. Therefore, much more information concerning the secondary structure of a class of related proteins can be obtained from a set of these proteins whose primary sequences may have diverged considerably, but which are not so evolutionarily distant to have diverged to any great extent from a structural standpoint.

In order to make use of this information, one must first make structural predictions for each of the protein sequences in the set, then align the proteins so that structurally homologous residues share identical positions. Having done so, the SOPM algorithm may be easily extended to cover the entire alignment, by simply adding up the conformational scores for each residue in a column to obtain the overall prediction for that position in that class of proteins.

This is the approach that is taken by the SOPMA server (*http://www.ibcp.fr/predict.html*), which accepts a single protein sequence as input, then does a database search for homologous sequences, and computes both the secondary structure prediction for each of these sequences and the alignment of the set. Finally, it computes a consensus secondary structure by averaging over the conformational scores for the amino acids at each column of the alignment, using values of zero when a sequence has a gap at a particular position. We have implemented the final step of this approach (see the E2 and E4 sections in Part III) to look at alignments generated locally by the HMM methods. We have also classified a position's scores as strongly predictive (upper case letters) if they meet the two following criteria: the score for the highest scoring class must be greater than the median value for the score for all positions that were similarly classified, e.g. as helices; the normalized difference between the scores for the highest and next-highest scoring classes must be in the upper quartile of these values for all positions that were similarly classified. Thus in theory, a maximum of one-fourth of the positions classified by the algorithm into each of the possible secondary structural conformations will be considered strongly predictive, and in practice even fewer meet both criteria.

It is the widespread wisdom at this time to evaluate sequences, whenever possible, by more than one algorithmic approach. The SOPMA server, therefore, submits a sequence to alternative methods of structure prediction—Gibrat, Levin, DPM [9–11]—and also generates a consensus over those and the SOPM itself. Thus we have submitted the HMM-generated "most likely" sequence to the SOPMA suite, which produces a prediction using for the four individual algorithms and a consensus prediction. We have also gained individual SOPM predictions for the various PV sequences and constituted a consensus structure prediction as described in the previous paragraph; in the analyses presented for the E2 and E4 protein sequences, the latter prediction always appears at the top where it becomes the reference for the alignment of structures.

References

[1] Doolittle, R.F. (1987) *Of URFS and ORFS: A Primer on How to Analyze Derived Amino Acid Sequence* University Science Books, Mill Valley, California.

[2] McClure, M.A., Vasi, T.K., and Fitch, W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11**: 571–592.

[3] Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M.A. (1994) Hidden Markov models of biological primary sequence information. *Proc. Nat. Acad. Sci. U.S.A.* **91**: 1059–1063.

[4] Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994) Hidden Markov methods in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**: 1501–1531.

[5] Luthy, R. and Eisenberg, D. (1992). Protein. In *Sequence Analysis Primer* (eds. M. Gribskov and J. Devereux), pp. 78–82. W.H. Freeman and Company, New York.

[6] Gribskov, M. and Veretnik, S. (1996). Identification of sequence patterns with profile analysis. In *Computer Methods for Macromolecular Sequence Analysis* (ed. R.F. Doolittle). pp. 146–159, Academic Press, Inc., San Diego.

[7] Geourjon, C. and Deleage, G. (1994) SOPM: a self-optimised prediction method for protein secondary structure prediction. *Prot. Eng.* **7**: 157–164.

[8] Geourjon, C. and Deleage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* **11**: 681–684.

[9] Gibrat, J.-F., Garnier, J., and Robson, B. (1987) Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**: 425–443.

[10] Levin, J.M., Robson, B., and Garnier, J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS* **205**: 303–308.

[11] Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function and Genetics* **19**: 55–72.