Hidden Models in Biopolymers

Mor Amitai

The amount of publicly available sequence data (mainly genomic DNA, messenger RNA, and their corresponding protein sequences) is increasing exponentially, primarily as a result of the latest advances in sequencing techniques and the Human Genome Project. Although this sequence data is becoming the most valuable source of information for the understanding of biological processes, the huge amount of data presents many challenges. Two of the most pressing problems are gene-finding and functional annotation. The first problem is simply to find the genes and the translated proteins hidden in the billions of base pairs in the nucleotide databases. Functional annotation involves assigning a function, or at least partial information concerning the function, to proteins (those hidden in nucleotide databases and those already expressed in protein databases).

TECHVIEW
COMPUTATIONAL BIOLOGY
Mailbox: www.sciencemag.org/cgi/dmail?53931

These two problems can be solved in the laboratory, but the amount of lab work needed to accommodate the large and growing amount of sequence data available is daunting. Significant progress in addressing these issues has been made in the past few years through the use of computational techniques that are applicable even for very large databases. One of the most powerful computational techniques is the hidden Markov model (HMM) approach. HMMs have been used in speech recognition over the past 25 years (*1*, *2*) and for the detection of protein domains in the past 6 years or so (*3*–*5*). HMMs are very useful for modeling entities that are composed of a limited set of simple building blocks, such as phonemes in speech and amino acids in proteins. The HMM is used to search for patterns and to detect phenomena in uncharacterized data.

HMMs are also used for other biological problems, such as ab initio gene finding (*6*), detection of unequal evolutionary rates in molecular sequences (*7*), radiation hybrid mapping (*8*), and protein secondary-structure prediction (*9*). These topics, however, are beyond the scope of this review.

The motivation for using the HMM approach for modeling protein domains is similar to that for using motifs (*10*) and profiles (*11*–*14*) to model protein domains. The advantages of HMM are its precise probabilistic modeling and utilization of the experience gained from the use of the same tools in speech.

Figure 1 shows a multiple alignment of the conserved domain of 20 members of the platelet-derived growth factor (PDGF) family (*15*). Some positions (such as 1) are more conserved than others (such as 6). For example, at position 17, one-fourth of the sequences have proline (P, 5 sequences), another one-fourth have arginine (R, 5 sequences) and the remainder do not have an amino acid at this position (often called a deleted position). The basis of the HMM approach is the assumption that a new member of the family is likely to behave similarly in this position. So this position is modeled as having a 25% chance of being a proline, a 25% chance of being an arginine, and a 50% chance of being a deletion.

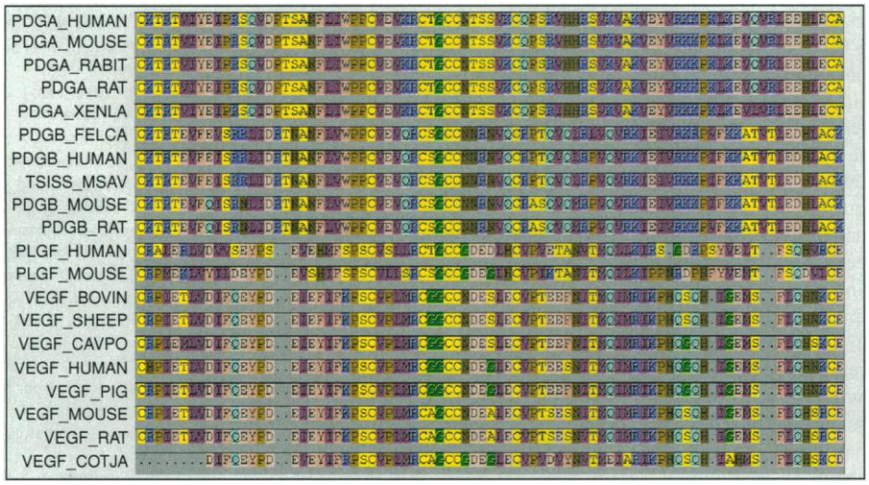

Fig. 1. An example of a multiple alignment (taken from the PDGF family). Different residues are colored differently to facilitate identification of conserved regions.

Of course, because only a small number of sequences have been described, we cannot conclude that there is no chance of any other amino acid. There are well-developed statistical techniques to allot a small but positive probability to other amino acids in this position by use of the knowledge of protein evolution (*16*). The HMM approach models expectations for what unknown members of the family could be through the use of probabilities calculated from the multiple alignment and assuming independence (except within consecutive deletions and insertions) among the amino acids of the protein. Thus, each position is modeled separately; the concatenation of these amino acid probabilistic models is the protein model.

Figure 2 is an extract from an HMM constructed from the multiple alignment in Fig. 1. State M16 is the match state corresponding to position 16 in the multiple alignment. From this state, we can move with a probability of 50% to D17, thereby signifying that position 17 in the multiple alignment is deleted from our query sequence. However, there is a 50% probability that the next state is M17, where with equal probabilities, P and R are emitted (hence, the probability of R is 50% of 50%, which is 25%). Given an unannotated protein, we would like to check whether it contains a domain for which we have developed an HMM. To do this, the protein is aligned to the HMM according to the probabilities (for example, alignment of a proline with a position of high probability for proline rather than one of low probability). This can be thought of as aligning the query protein with the multiple alignment that created the HMM.

Computer programs are used to create and utilize HMMs. Some excellent examples are HMMer (*17*) and the Sequence Alignment and Modeling (SAM) system (*18*). These programs compute the likelihood of the query protein actually containing the functional domain represented by the HMM. The process of modeling and using an HMM to find new family

The author is at Compugen Ltd., Pinchas-Rozen 72, Tel Aviv 69512, Israel. E-mail: mor@compugen.co.il

members involves several steps: (i) align the known members of the family to create a multiple alignment [semi-automatically (*19, 20*), see also (*21*)]; (ii) calculate the derived probabilities and build the HMM (automatically); and (iii) search a protein database with the HMM (automatically). The HMMer and SAM packages address all three steps, as well as providing additional HMM-related tools. A large database of HMMs of protein domains is found in Pfam (*22, 23*). The current version (Pfam 3.2) contains multiple sequence alignments and HMMs of 1344 protein domains, which can annotate more than 50% of the proteins in Swiss-Prot.

The use of the protein HMMs is not limited to searching protein databases, but can also be extended to searching nucleotide databases. In this case, the protein model is automatically translated to an extended model, where amino acids are replaced by codons, and sequencing errors and introns are also modeled. This is done by dividing most of the probability for an amino acid into the probabilities assigned to the appropriate codons, including "erroneous" codons, and by allowing long gaps (introns) in the alignment under certain conditions. For example, most of the probability for histidine at a given location in the original profile will be divided into the probabilities of CAT and CAC, but fractions of it will go to the probabilities of "CA," "CATT," and even "CAG," and so forth, to model various sequencing errors. Long segments of the DNA sequence may be considered introns and left unaligned with the coding part of the new model if they begin with "GT" and end with "AG." The intron models currently used are actually more complex, and some of them even model the poly-pyrimidine tract (which usually occurs just before the end of an intron). The extended HMM can be used to search nucleotide databases in a very similar way to that used in searching protein databases with the original protein HMM (*24, 25*).

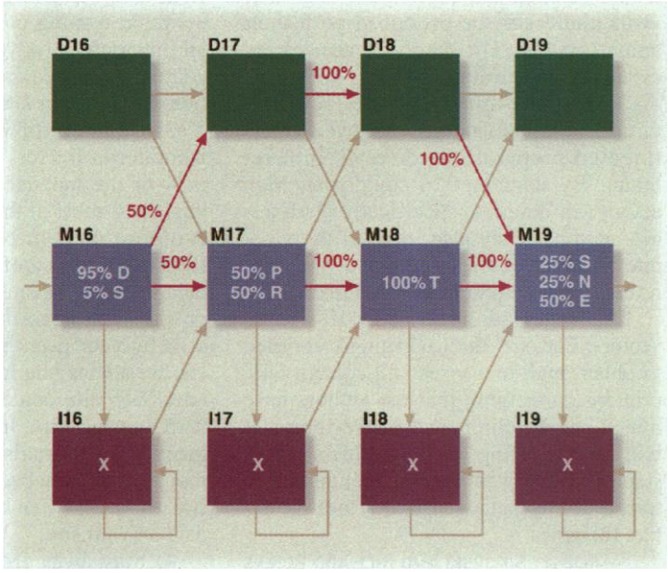

Fig. 2. A part of an HMM built from the multiple alignment from Fig. 1. Each position in the multiple alignment is represented by three states: a match state (M), insert state (I), and delete state (D). The match state contains emission probabilities for each amino acid. The insert state allows insertions of amino acids after the position of the match state (X stands for an arbitrary amino acid). The delete state enables deletion of the amino acid, by emitting no amino acid. Transition probabilities between the states define the probability of each state to be transversed. The HMM actually used will generally be slightly different, in that it will have neither 100% nor 0% transition and emission probabilities, to account for unobserved phenomena.

Currently, experts are annotating genomic data using all the information available (*26*). Even then, such work yields only "best guess" genes that in many cases are incomplete. Novel computational techniques can help find the complete set of human genes. One technique may be direct comparison of pairs of genomic DNA regions that code to homologous genes (either from the same genome or between different species). A comparison of two genomic DNA sequences considering the presence of one or more conserved genes can be naturally phrased using HMMs (*27*).

The HMM approach is very powerful in finding protein domains in peptide chains, as well as nucleotide sequences. However, HMMs are very limited in modeling correlations between distant residues. In secondary-structure prediction, other computational techniques (*28*) perform better than HMMs. In the case of fold recognition problems, encouraging results are obtained through use of secondary-structure prediction and a substantial amount of expert knowledge, in addition to HMMs (*29*). There is still a question whether HMMs are the right tool for incorporating structural information into protein domain models. The availability of HMM-based tools (*17, 18, 25*) and protein domain databases (*23*) enables even scientists who do not specialize in this field to benefit from this powerful technique in their effort to annotate biomolecules. In view of the multitude and nature of the biological data expected to be produced in the coming years, it is likely that HMMs and other probabilistic techniques will prove to be even more important in the near future.

References and Notes

1. L. R. Rabiner and B. H. Juang, *IEEE ASSP Mag.* **3**, 4 (1986).
2. L. R. Rabiner, *Proc. IEEE* **77**, 257 (1989).
3. A. Krogh, M. Brown, I. S. Mian, K. Sjolander, D. Haussler, *J. Mol. Biol.* **235**, 1501 (1994).
4. S. R. Eddy, *Curr. Opin. Struct. Biol.* **6**, 361 (1996).
5. R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, *Biological Sequence Analysis: Probabilistic Models for Protein and Nucleic Acids* (Cambridge University Press, Cambridge, 1998).
6. C. Burge and S. Karlin, *J. Mol. Biol.* **268**, 78 (1997).
7. J. Felsenstein and G. A. Churchill, *Mol. Biol. Evol.* **13**, 93 (1996).
8. D. Slonim, L. Kruglyak, L. Stein, E. Lander, *J. Comput. Biol.* **4**, 487 (1997).
9. N. Goldman, J. L. Thorne, D. T. Jones, *J. Mol. Biol.* **263**, 196 (1996).
10. P. Bucher, K. Karplus, N. Moeri, K. Hofmann, *Comput. Chem.* **20**, 3 (1996).
11. M. Gribskov, R. Luthy, D. Eisenberg, *Methods Enzymol.* **183**, 146 (1990).
12. M. Gribskov, A. D. McLachlan, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4355 (1987).
13. M. Gribskov and S. Veretnik, *Methods Enzymol.* **266**, 198 (1996).
14. J. U. Bowie, R. Luthy, D. Eisenberg, *Science* **253**, 164 (1991).
15. www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00341
16. K. Sjolander *et al.*, *Comput. Appl. Biosci.* **12**, 327 (1996).
17. http://genome.wustl.edu/eddy/hmmer.html
18. www.cse.ucsc.edu/research/compbio/sam.html
19. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
20. In most cases, expert intervention is necessary to improve the multiple alignment automatically generated by the software.
21. S. R. Eddy, in *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, Cambridge, UK, 16 to 19 July 1995, C. Rawlings *et al.*, Eds. (AAAI Press, Menlo Park, CA, 1995), pp. 114–120.
22. E. L. L. Sonnhammer, S. R. Eddy, R. Durbin, *Proteins* **28**, 405 (1997).
23. Available at: www.sanger.ac.uk/Software/Pfam/ or http://pfam.wustl.edu/
24. E. Birney and R. Durbin, in *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Halkidiki, Greece, 21 to 25 July 1997, T. Gaasterland *et al.*, Eds. (AAAI Press, Menlo Park, CA, 1997), pp. 56–64.
25. www.sanger.ac.uk/Software/Wise2/
26. Such techniques include protein, protein domain, and EST database searches, as well as the prediction of ab initio gene-finding programs.
27. Such a comparison is made while comparing pairs of potential codons and detecting introns and sequencing errors on the fly.
28. B. Rost and C. Sander, *J. Mol. Biol.* **232**, 584 (1993).
29. A. G. Murzin and A. Bateman, *Proteins* (Suppl.1) **25**, 105 (1997).
30. I thank E. Birney for insightful discussions and criticism of this manuscript, R. Gill-More and S. R. Pollock for helpful comments and assistance in preparation of the manuscript, R. Durbin for stimulating discussions, and E. Sinyavsky for assistance with the figures.