Hidden Markov models in biological sequence analysis

by E. Birney

The vast increase of data in biology has meant that many aspects of computational science have been drawn into the field. Two areas of crucial importance are large-scale data management and machine learning. The field between computational science and biology is varyingly described as "computational biology" or "bioinformatics." This paper reviews machine learning techniques based on the use of hidden Markov models (HMMs) for investigating biomolecular sequences. The approach is illustrated with brief descriptions of gene-prediction HMMs and protein family HMMs.

Introduction

There has been a revolution in molecular biology over the last decade due to a simple economic fact: The price of data gathering has fallen drastically. Nowhere is this better illustrated than in large-scale DNA sequencing. At current costs, it is economical to determine the DNA sequence of the entire genome of a species (the genome is all of the DNA sequence passed from one generation to the next), even for species with large genomes, such as humans.

The basic information of interest in bioinformatics pertains to DNA, RNA, and proteins. Molecules of DNA are usually designated by different sequences of the letters A, T, G, and C, representing their four different types of bases. RNA molecules are usually designated by similar sequences, but with the Ts replaced by Us, representing a different type of base. Proteins are represented by 20 letters, corresponding to the 20 amino acids of which they are composed. A one-to-one letter mapping occurs between a DNA molecule and its associated RNA molecule; and a three-to-one letter mapping occurs between the RNA molecule and its associated protein molecule. A protein sequence folds in a defined three-dimensional structure, for which, in a small number of cases, the coordinates are known. The defined structure is what actually provides the molecular function of the protein sequence.

The basic paradigm of biology is shown graphically in **Figure 1**. Depicted in the figure is a region of DNA that produces a single RNA molecule, which subsequently produces a single protein having a well-defined biological function.

Roughly speaking, the time and cost of determining information increases from the top of the diagram to the bottom. Determining DNA and RNA sequences is relatively cheap; determining protein sequences and protein

©**Copyright** 2001 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/01/$5.00 © 2001 IBM

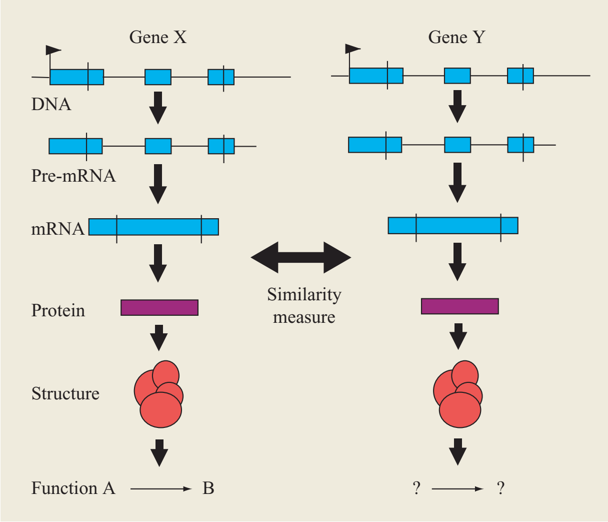

Figure 1

The basic paradigm of biology (DNA produces RNA, which produces protein) and how this relates to bioinformatics data processing. The two separate panels represent different genes. The top line is genomic DNA: Going from the top to the bottom of the diagram are the primary processes which transform the information in the genomic DNA to the functional aspects of the organism. The exons are shown in blue, and the introns and intergenic DNA appear as thin lines. The start and stop codons are represented as thin vertical lines. The magenta rectangle represents a linear protein sequence, and the red circles its three-dimensional structure. The process of comparing two genes at any one of these levels provides evidence for homology between the two genes.

structures is far more expensive; many person-years can be spent trying to elucidate the function of a single protein. A clear goal for bioinformatics is to provide a way to convert the cheaper information at the top to the more valuable information at the bottom. Two steps have proven to be difficult. For unknown reasons, large organisms deliberately process the RNA sequence that is derived from the DNA sequence by a method known as pre-mRNA splicing. This removes specific pieces of the RNA (called introns) and fuses the remaining pieces (called exons). The exons remain collinear with their original layout in the DNA sequence. The ratio of exon sequence to intron sequence is around 1:50 in human DNA, and the intron sequence appears to be extremely "random" in nature, making effective discrimination difficult. Despite this challenge, bioinformatics has developed a reasonably successful solution using HMMs (see below). The second problem is deducing protein structure from a linear protein sequence. This "folding problem" has resisted concerted attack from researchers over the last twenty years. Although there have been many exciting advances in the area of protein folding, it seems likely that there will not be a solution to this problem in the next five or more years.

Bioinformatics can thankfully sidestep both of these problems by using arguments of evolution. Imagine the proto-rodent that represents the common ancestor between mouse and human. This creature had a region of its DNA sequence which made a protein with a specific function (for example, catalyzing the reduction of ethanol to acetaldehyde). At some point there was a speciation event which led eventually to man and mouse. In the two lineages, the DNA sequences were maintained from generation to generation, sometimes suffering a mutation that changed the DNA sequence. As long as the mutation did not disadvantage the individual, in general preserving the function of the protein, the mutation would be passed on to its descendants. In the extant species of man and mouse, one ends up with two similar but not identical regions of DNA sequence which form two similar proteins with similar structures and functions.

This argument of common ancestry, or *homology*, is illustrated pictorially in Figure 1 by the horizontal arrows. Arguments of homology are the bedrock of bioinformatics. It is relatively easy to find a clearly homologous DNA sequence presupposed to exist at the first cellular organism and observable in all living organisms—for example, the DNA sequence which produces the proteins found in the ribosome. This conservation in the face of potentially billions of random mutations in the DNA sequence shows how much selection (i.e., an individual with a deleterious mutation is unlikely to pass on this mutation) occurs in biology.

Given that two proteins are homologous, one can deduce that, at the very least, portions of the 3D structure of the two proteins are similar, if not some functional aspects of the protein. Since there exist large databases of known proteins with known functions, a considerable amount of bioinformatics pertains to transferring knowledge from known to unknown proteins using arguments of homology. This process is very efficient. Despite the millions of organisms, each containing thousands of genes, researchers have estimated that there are only around 4000 unique protein parts which have been reused by evolution over time (although these 4000 unique protein parts form many more than 4000 molecular functions).

The degree of homology is determined by calculating some metric indicating how similar two sequences are. The observed similarity can be due either to homology between the two sequences or simply the "by chance" score created by matching two unrelated sequences. In more advanced formalisms, a single sequence is scored against a mathematical model of a particular type of conserved sequence region, again using a hidden Markov model, as discussed later. In general, the farther down the information flow of Figure 1, the better the measure of similarity, because it is easier to deduce that two protein sequences are homologous than two DNA sequences. The end result is that there are two everyday tasks for bioinformatics: deducing the protein sequence from the DNA sequence and comparing protein sequences to an existing database of protein sequences, both of which utilize hidden Markov models.

The rest of this paper describes these two methods in broad detail. The reader should be aware that the author deliberately ignores three large areas of probabilistic models that are being used in biology. One is the use of probabilistic models to represent the alignment of two sequences (often proteins). The basic HMM used here is quite rare in other fields but ubiquitous in bioinformatics; some recent papers are a fully Bayesian approach to sequence alignment [1] and a novel accuracy-based *a posteriori* decoding method [2]. The second area is the use of probabilistic models for evolutionary tree analysis, concerning which there is a long-established research interest; some recent papers include the integration of an HMM with tree methods [3]. The final area is the use of stochastic-context-free grammars (SCFGs) for RNA analysis. An SCFG is to *yacc* what a hidden Markov model is to *lex*, and they are ideally suited to RNA analysis, since RNA forms stem-loop structures analogous to the nested-bracket structure found in context-free grammars. Some recent papers discuss the exciting ability to push into more-context-dependent grammars [4]. Readers should also be aware that many of these probabilistic methods have nonprobabilistic parameterized counterparts, in some cases predating the probabilistic method by more than a decade and providing very effective techniques. The author's own prejudice is to view nonprobabilistic parameterized systems as being interpretable as some sort of probabilistic model.

Hidden Markov models

An HMM is a graph of connected states, each state potentially able to "emit" a series of observations. The process evolves in some dimension, often time, though not necessarily. The model is parameterized with probabilities governing the state at a time $t + 1$, given that one knows the previous states. Markov assumptions are used to truncate the dependency of having to know the entire history of states up to this point in order to assess the next state: Instead, only one step back is required. As the process evolves in time through the states, each state can potentially emit observations, which are regarded as a stream of observations over time. These models are often illustrated graphically as shown in **Figure 2**, with the states being circles and transitions as arrows between the states.

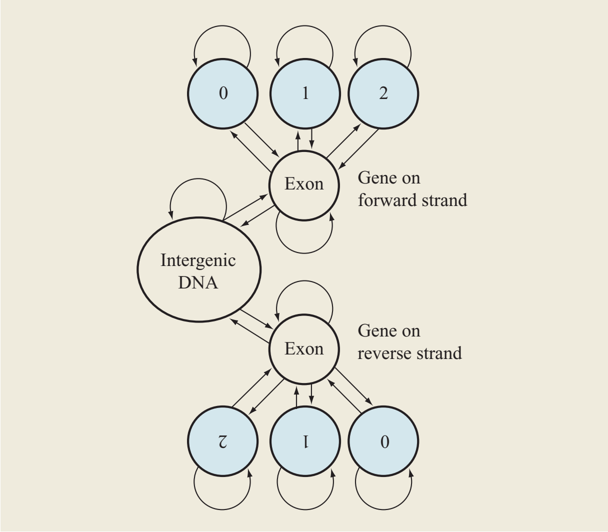

Figure 2

Abbreviated gene HMM model. The HMM is split into two symmetrical parts: genes on the forward or reverse strand of the DNA sequence (DNA sequence can be read in two directions). Each gene model contains a central exon state which has an emission of nucleotides tuned to recognize protein coding regions. Interrupting the exons are introns; three intron states are used, since there are three relative positions at which an intron can interrupt a coding triplet of DNA bases. These introns are distinguished by their "phase" — 0, 1, or 2.

Given a particular set of parameterized models, two questions can be answered: For a given observed sequence, which model is the most likely to explain this data, and for a given sequence and a given model, what is the most likely reconstruction of the path through the states. In addition, models can be learned from the data, in which parameters are estimated by expectation-maximization techniques. It is very natural to use a Bayesian statistical framework with HMMs. This is because the likelihood of, for example, observing a sequence given a particular model is a natural calculation for HMMs. Bayesian statistics provide a framework for converting this likelihood into an *a posteriori* probability (the probability of the model, given the observed sequence) that includes the ability to integrate prior knowledge about, for example, the way in which proteins evolve.

For biological sequences, the "time" dimension is replaced by the position in the sequence. Hidden Markov models prove so successful in this field because they can naturally accommodate variable-length models of regions of sequence. This is generally achieved by having a state which has a transition back to itself. Because most biological data has variable-length properties, machine learning techniques which require a fixed-length input, such as neural networks or support vector machines, are less successful in biological sequence analysis.

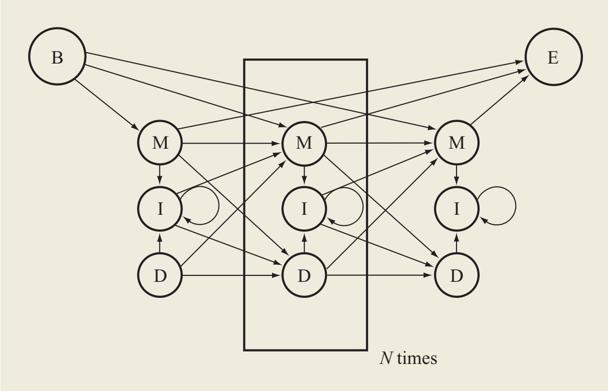

Figure 3

A profile HMM, which has a repetitive structure of three states (M, I, and D). Each set of three states represents a single column in the alignment of protein sequences.

Gene-prediction HMMs

Gene-prediction HMMs model the process of pre-mRNA splicing followed by protein translation. The input of this process is the genomic DNA sequence and the output is the parse tree of exons and introns on the DNA sequence, from which the protein sequence of the gene prediction can be predicted.

The gene-prediction HMMs are relatively standard: An abbreviated HMM is shown in Figure 2. There are states representing exons and introns, with specific states to model aspects of the gene parse; in particular, the crossover points between exons and introns (denoted as the 5′ and 3′ splice sites) have strong sequence biases. The exemplar program for the field is Genscan by Chris Burge and Samuel Karlin [5], with other good examples being Genie by David Kulp and colleagues [6] and HMMGene by Anders Krogh [7]. Depending on one's outlook, these programs either do well, with base-pair specificity in the 80% range in well-defined test sets, or badly, in the sense that 50% gene predictions appear to be completely wrong in large-scale genomic tests.

Many of the new approaches are hoping to integrate additional information from similar sequences at the RNA or protein level. All of the authors mentioned above are integrating this information, and there are other approaches, such as the present author's own work, which provide a formal integration of protein similarity with gene prediction [8, 9].

Profile HMMs

Anders Krogh and colleagues have developed a hidden Markov model equivalent of profile analysis for investigating protein families [10]. Profile analysis provided an *ad hoc* way to represent the "consensus profile" of amino acids for a set of protein sequences belonging to the same family. The hidden Markov model applied was deliberately modeled on this successful technique, but introduced the notion of using probability-based parameterization, allowing both a principled way of setting the gap penalty scores and also more novel techniques such as expectation maximization to learn parameters from unaligned data.

The architecture of the HMM is shown in **Figure 3**. It has a simple left-to-right structure in which there is a repetitive set of three states, designated as *match*, *delete*, and *insert* (M, D, and I). The match state represents a consensus amino acid for this position in the protein family. The delete state is a non-emitting state, and represents skipping this consensus position in the multiple alignment. Finally, the insert state models the insertion of any number of residues after this consensus position. This type of repetitive HMM is also common in speech recognition, where it is sometimes called a "time-dependent" HMM or "time-parameterized" HMM.

The use of profile HMMs was greatly enhanced in the HMMER package by Sean Eddy [11]. HMMER provided a free, stable, and effective software package to build, manipulate, and use HMMs, as well as a number of important improvements to the use of HMMs. First, HMMER provided log-odds likelihood of the model compared to a random model to indicate the relative likelihood that a new sequence belongs to this family. In the second iteration of the package, HMMER2, the HMM architecture was improved, in particular reducing the number of parameters to learn and in addition deliberately modeling repeated occurrences of a single protein "domain" in one protein sequence. HMMER2 also introduced a frequentist interpretation of the log-odds likelihood statistic by providing the ability to calibrate an HMM against a random distribution of sequences and fitting a distribution under the assumption that it was an extreme value distribution. This calibration and curve-fitting approach produced a statistic that is far more powerful than, but still as accurate as, that produced by the Bayesian *a posteriori* probability approach.

The success of HMMER in providing a stable, robust way to analyze protein families gave rise to a number of databases of hidden Markov models. Such databases are similar in many ways to the databases of phonemes and longer words used in speech recognition: Since biology has a limited number of protein families in existence, sheer enumeration of these protein domains is achievable. Despite the early promise of using unsupervised training approaches to derive these HMMs, highly supervised approaches by bioinformatics experts have always outperformed the more automatic approaches. The databases of profile HMMs are therefore focused around manual adjustment of the profile HMMs followed by an automatic gathering of complete datasets from large protein sets, as illustrated by the use of the Pfam (protein family database) [12] and SMART (simple modular architecture research tool) [13] approaches. Pfam in particular now possesses coverage significant enough that 67% of proteins contain at least one Pfam profile HMM and 45% of residues in the protein database are covered in total by the HMMs. This extent of coverage, coupled with the good statistical behavior of the profile HMMs, has made Pfam an automatic protein classification system without peer.

Theoretical contributions from bioinformatics

It is easy to think that much of bioinformatics is the rather mundane application of existing machine learning techniques to yet another set of data. However, like any real-life problem, bioinformatics stretches the methods in ways in which other datasets have not. This has led to a number of advances in machine learning from bioinformatics. Here are some selected highlights:

- *Small dataset usage*. Bioinformatics, like some other fields, uses small sets of data but comprises a large body of theory about how certain distributions are presumed to behave. By integrating some of this theory into more standard prior distribution style methods, a number of novel methods have been developed, such as applying multiple Dirichlet priors [14] and maximizing the use of small datasets [15].
- *Novel decoding methods*. Much of bioinformatics is less interested in the question of which model a particular set of observations come from than in the path taken through a particular model. The standard maximum-likelihood path (also called the Viterbi path) is not always the best path for a particular problem. Novel methods include *a posteriori* decoding to maximize the accuracy of the path [2] and decoding methods for integrated probabilistic methods [7].
- *General extensions of techniques*. Some interesting work has occurred in bioinformatics in the integration of machine learning techniques. In the author's own work with Richard Durbin, we were led to derive a formal process to combine two separate HMMs into one [9]. David Haussler and Tommi Jaakkola provided a way of combining a discriminant method (support vector machines) with a generative HMM for providing better performance in a stringent class-distinction test [16].

Open areas for research in hidden Markov models in biology

Open areas for research in HMMs in biology include the following:

- *Integration of structural information into profile HMMs*. Despite the almost obvious application of using structural information on a member protein family when one exists to better the parameterization of the HMM, this has been extremely hard to achieve in practice.
- *Model architecture*. The architectures of HMMs have largely been chosen to be the simplest architectures that can fit the observed data. Is this the best architecture to use? Can one use protein structure knowledge to make better architecture decisions, or, in limited regions, to learn the architecture directly from the data? Will these implied architectures have implications for our structural understanding?
- *Biological mechanism*. In gene prediction, the HMMs may be getting close to replicating the same sort of accuracy as the biological machine (the HMMs have the additional task of finding the gene in the genomic DNA context, which is not handled by the biological machine that processes the RNA). What constraints does our statistical model place on the biological mechanism—in particular, can we consider a biological mechanism that could use the same information as the HMM?

There are many other topics, both in probabilistic modeling and more generally in bioinformatics as a discipline waiting for enthusiastic machine-learning researchers. The author looks forward to the field growing over the coming decade.

References

1. J. Zhu, J. S. Liu, and C. E. Lawrence, "Bayesian Adaptive Sequence Alignment Algorithms," *Bioinformatics* **14,** 25–39 (1998).
2. I. Holmes and R. Durbin, "Dynamic Programming Alignment Accuracy," *J. Comput. Biol.* **5,** 493–504 (1998).
3. P. Lio, J. L. Thorne, N. Goldman, and D. T. Jones, "Passml: Combining Evolutionary Inference and Protein Secondary Structure Prediction," *Bioinformatics* **14,** 726–733 (1999).
4. Eleanor Rivas and Sean Eddy, "A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots," *J. Mol. Biol.* **285,** 2053–2068 (1999).
5. C. Burge and S. Karlin, "Prediction of Complete Gene Structures in Human Genomic DNA," *J. Mol. Biol.* **268,** 78–94 (1997).
6. D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman, "A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA," *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. F. Smith, Eds., AAAI Press, Menlo Park, CA, 1996, pp. 134–142.
7. A. Krogh, "Two Methods for Improving Performance of a HMM and Their Application for Gene Finding," *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, Eds., AAAI Press, Menlo Park, CA, 1997, pp. 179–186.
8. E. Birney and R. Durbin, "Dynamite: A Flexible Code Generating Language for Dynamic Programming Methods

Used in Sequence Comparison," *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, 1997, pp. 56–64.

9. E. Birney, "Sequence Alignment in Bioinformatics," Ph.D. thesis, The Sanger Centre, Cambridge, U.K., 2000; available from *ftp://ftp.sanger.ac.uk/pub/birney/thesis*.
10. A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *J. Mol. Biol.* **235,** 1501–1531 (1994).
11. S. R. Eddy, "HMMER: A Profile Hidden Markov Modelling Package," available from *http://hmmer.wustl.edu/*.
12. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer, "The Pfam Protein Families Database," *Nucleic Acids Res.* **28,** 263–266 (2000).
13. J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork, "SMART: A Web-Based Tool for the Study of Genetically Mobile Domains," *Nucleic Acids Res.* **28,** 231–234 (2000).
14. K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology," *Comput. Appl. Biosci.* **12,** 327–345 (1996).
15. S. R. Eddy, G. J. Mitchison, and R. Durbin, "Maximum Discrimination Hidden Markov Models of Sequence Consensus," *J. Comput. Biol.* **2,** 9–23 (1995).
16. T. Jaakkola, M. Diekhans, and D. Haussler, "A Discriminative Framework for Detecting Remote Protein Homologies," *Comput. Biol.* **7,** 95–114 (2000).

Received August 2, 2000; accepted for publication December 19, 2000

Ewan Birney *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, U.K. CB10 1SD (birney@ebi.ac.uk).* Dr. Birney is the Team Leader for Genomic Annotation at the European Bioinformatics Institute (EBI). He received a B.A. degree in biochemistry from Oxford University in 1996 and a Ph.D. degree in genetics from Cambridge University in 2000. Dr. Birney has pursued a career in bioinformatics since his undergraduate days, writing the "GeneWise" software package and becoming the Bioperl coordinator in 1999. He is now one of the leaders for the "Ensembl" project, which is providing an open annotation of the human genome (*http://www.ensembl.org/*).