GENE DETECTION BASED ON HIDDEN MARKOV MODELS

F. Rodolphe[1] and P. Nicolas[1,2]

[1] Laboratoire Mathématique, Informatique et Génome, INRA, Route de Saint-Cyr, F-78026 Versailles cedex, France
[2] Laboratoire de Statistique et Génome, CNRS, Tour Evry 2, 523 place des Terrasses de l'Agora, F-91034 Évry, France

INTRODUCTION

The availability of many entire genome sequences makes gene detection a major challenge. Entire genomes are often being sequenced without much knowledge at the gene level about the organism. So gene detection has to be realized on the DNA sequence as a principal, if not exclusive, source of information. The size of the problem is so large that gene detection must be as far as possible an automatic process.

In this paper, we focus on hidden Markov model (HMM) based methods. A few steps in the short but software productive history of gene detection are briefly recalled in order to give the context and to show the present state of the art. We first consider prokaryotic gene detection, considering later the problem addressed by eukaryotes.

SIGNALS AND TEXTURES

The simplest, and historically the first, idea consists in using stop codons : while reading a gene in the right frame, as a codon sequence, the first stop encountered indicates doubtlessly the gene end. A long segment free of stops containing at least a start codon, called an open reading frame, is a strong indication of the presence of a gene. The method misses many short genes (or detects too much false positives) in A+T poor genomes. Moreover, as starts are not specific, the method does no provide precise indication concerning gene beginnings.

Apart from signals (features related to some short part of the sequence) another kind of information is available. Genes differ from non coding sequences, by statistical properties displayed all along their sequence. Such properties can be qualified as textures, and consist of particular word composition or periodicities. For instance nucleotide frequencies differ between coding and non coding sequences ; furthermore, in coding sequences these frequencies differ according to phase, which gives rise to a statistical 3-periodicity, absent in non coding sequences, a very valuable feature for gene detection. For instance in Bacillus subtilis.genes (A, C, G, T) frequencies are (.30, .19, .34, .17), (.33, .21, .15, .31), (.27, .21, .23, .29), according to phase. Texture was soon incorporated in gene detectors. GENMARK based genome segmentation on likelihood comparisons within a sliding window (Borodovsky and McIninch, 1993). The different models to be compared were Markov chains corresponding to different textures learned before. Glimmer evaluates the coding potential of each ORF using interpolated Markov chains which are more sophisticated texture models (Salzberg *et al.*, 1998).

Hidden Markov models are very well suited to take into account and handle texture information. In these models, a sequence is considered as a series of segments, each one belonging to one out of a set of segment classes. A class (for instance coding or non coding) is

characterised by a typical local composition.

HIDDEN MARKOV MODELS

A hidden Markov model consists of two different stochastic processes.

The first one $U = (U_i , i=1,...n)$ is called the hidden process. It is a first order Markov chain on some alphabet of hidden states (the segment classes). In our setup, index i refers to the position on the sequence and U_i is the class of the segment to which position i belongs. A segment is a run of the same hidden state.

The second process $X = (X_i , i=1,...n)$ is, given the hidden one, a heterogeneous Markov chain, whose transition probability matrix is a function of the actual hidden state. This process takes its values on the four letter alphabet {A,C,G,T} ; it represents the sequence and is called the observed process. Chain order and other characteristics can depend on the segment class ; they usually are specified by the user.

A HMM will be referred to as M1-Mk if the observed process is a conditional Markov Chain of order k. For example, in the M1-M1 model, $L(X) = \Sigma_{U1...Un}\ \alpha(U_1)\beta(X_1;U_1)\ \Pi_{i=1...n-1}\ a(U_i, U_{i+1})b(X_i, X_{i+1} ; U_{i+1})$ is the likelihood of the observed process, where a(u,v), b(x,y;v) stand for hidden (resp. observed) process transition probabilities, $\alpha(u),\beta(x;u)$ for their initial distributions (usually the stationary distribution); $\theta=(a,b)$ is the parameter of the HMM.

The first HMM based algorithms designed to detect genes, and still many others, use training sets to estimate the parameter and consider it as known afterwards. Such training sets consist of carefully annotated sequences of the same organism. Sequence segmentation can be based on Viterbi's algorithm, as in EcoParse (Krogh *et al*., 1994), which consists in computing the most probable hidden sequence (the path which maximizes the likelihood conditional on the observed process). For example, in the M1-M1 model :

$P[X,U]=P[X_1,.....X_n|U_1,.....U_n]P[U_1,.....U_n]=$

$P[X_1,.....X_{n-1}|U_1,.....U_{n-1}]P[U_1,.....U_{n-1}]P[X_n|X_{n-1},U_n]P[U_n|U_{n-1}]$

Let Z(t,u) be the path of hidden states which, among all paths ending at t with u, maximises the likelihood $P_t[X,U]$ of the complete process restricted to [1,t]. The following recurrence holds :

$Z(t,u)=(Z(t-1,\xi_{t-1,u}),u)$ where $\xi_{t-1,u}=\text{argmax}_v P[Z(t-1,v)]a(v,u)b(X_{t-1},X_t;u)$. With a proper initialization ($P[Z(1,u)]=\alpha(u)\beta(X_1;u),u=1,...q$) this relation can be used to compute $Z(n,u),u=1,...q$, and finally the most likely path.

Viterbi's algorithm does not provide any confidence measure of the segmentation it proposes. Baum *et al*. (1970) proposed a procedure called "forward-backward", which computes at each position the probability $P[U_i=u|X;\theta]$ of each hidden state, conditional on the parameter and the entire observed process. Segmentation can then be based on the most probable hidden state, but the probability gives a confidence measure. This algorithm is implemented in GenScan (Burge and Karlin, 1997).

Training sets for gene detection, of a sufficient size, are not always available or can be biased, and constitute a severe limit. Serious efforts were dedicated to the estimation of the parameter, θ, on the data themselves.

ESTIMATING HMMs ON THE DATA

Direct HMM estimation methods are based on likelihood maximization; θ's ML estimator consistency and asymptotic normality were proved for HMMs by Baum and Petrie (1966) and extended by Muri (1997).

Likelihood maximization is not an easy task, due to the large dimension of the parameter, and adapted methods have to be used. In a HMM, hidden states can be considered as missing data. General methods were developed for solving estimation tasks in the presence of missing data, that can be adapted to HMMs. For instance the EM algorithm (Dempster *et al*., 1977) is an iterative procedure that alternates two steps. Given the current value $\theta^{(m)}$, $E[\log P[X,U|\theta]|X,\theta^{(m)}]$, the expectation, is computed during the E-step and maximized over θ during the M-step. Rabiner (1989) gives a detailed description of EM implementation for HMM : E-step consists in computing the probability of two consecutive hidden states $P[U_{i-1}=u,U_i=v|X,\theta^{(m)}]$ using the Baum-Welch forward-backward recurrence, from which follows $P[U_i=v|X,\theta^{(m)}]$. A new value $\theta^{(m+1)}$ is obtained in the M-step. E and M steps are alternated until numerical convergence is achieved.

TEXTURE BASED GENOME SEGMENTATION

Crude HMM can be used for genome segmentation (Churchill, 1989 ; Muri, 1997). Each segment class corresponds to a typical texture; these textures are adjusted to the data. Segmentation results from a compromise between adequacy of the finite number of typical textures to represent the variety of real textures, and the cost of changing the hidden state.

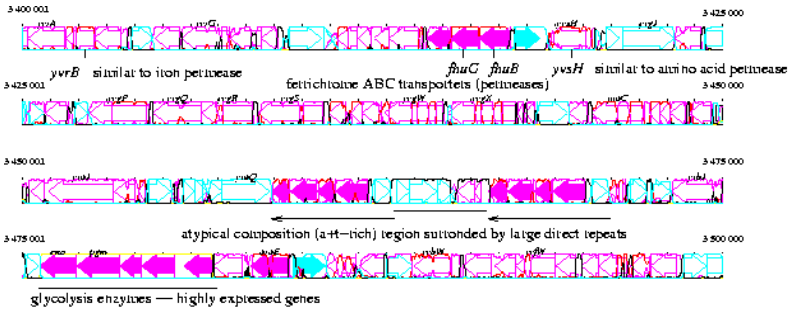

Figure 1. Segmentation of a B. subtilis genome fragment of 100 kbp, based on a crude M1-M2 with 5 hidden states (Nicolas *et al*., 2002). Hidden state probabilities are plotted along the sequence. GenBank annotations are superimposed : arrows stand for genes (filled=known function, empty=unknown). Most of (-) strand genes belong to magenta state; hydrophobic proteins such as permeases, belong to the red one; glycolysis enzymes, highly expressed, to the yellow one; cyan matches with (+) strand genes and black with intergenic regions or atypical genes

Obviously, HMM can capture biologically interesting features. Noteworthy, segmentation is somewhat robust. It does not depend too much on model details. Moreover, while the number

of hidden states is increased, results remain coherent : adding a new hidden state produces the splitting up of a preceeding hidden state, instead of a completely different segmentation.
Beside these interesting properties, it is also evident that gene limits are poorly detected and that crude HMM are unsatisfactory as a gene detection tool.

GENE DETECTION WITH STRUCTURED HMM

Best results are obtained with "good" texture models : second order 3-periodically heterogeneous Markov chains for coding sequences, and homogeneous Markov chains for non-coding sequences. Also, it is well known that genes differ by their codon usage, and therefore their texture. Efficient gene detection must take into account the existence of several gene classes (Borodovsky *et al*., 1995 ; Lukashin and Borodovsky, 1998).
To improve gene limit detection, at least start and stop signals must be taken into account : segments corresponding to coding sequences must begin with a start and end with a stop (Krogh *et al*.; 1994). In a HMM, this can be represented by hidden states dedicated to these signals and a graph connecting the hidden states, which represents structural features of the hidden process. A start leads to a coding state and a stop to a non-coding state : some transition probabilities of the hidden chain equal one, or zero, others remain to be estimated.

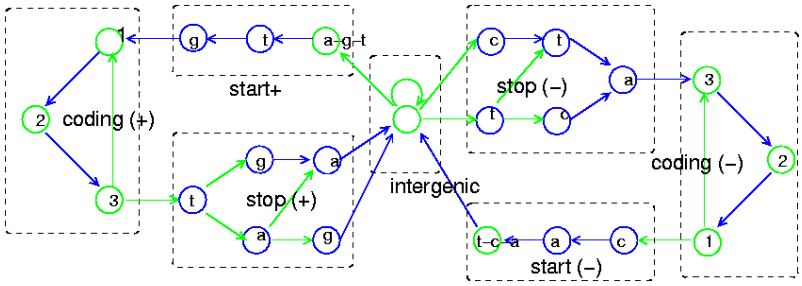

Figure 2. Structure of an elementary gene detector with stop and start, one hidden state for intergenic segments and only one gene class (on both strands) represented by 3 hidden states for statistical periodicity. Letters inside circles indicate which nucleotides can be emitted by the hidden state

SHOW is a software designed for HMM EM estimation and sequence segmentation. It enables one to build up gene detectors which handle textures and signals. The user specifies the structure of the hidden chain (the graph) and the order of the observed process, which can depend on the hidden state. Such gene detectors behave rather efficiently on prokaryotic genomes. GeneMarkS (Besemer *et al*., 2001), uses a different estimation procedure which consists of iterative use of GeneMark.hmm.

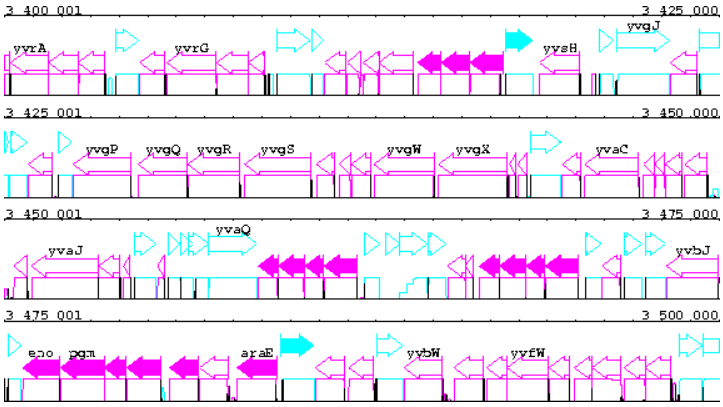

Figure 3. Segmentation of the sequence of figure 1 produced by a SHOW based gene detector. 3 probabilities are plotted along the sequence which correspond to both gene orientations and ribosome binding sites (black curve)

Probably the most serious unsolved problem is a reliable determination of the start codons. No satisfactory solution has yet been proposed, although it certainly exists : bacteria rarely use alternate start sites and make no translation mistakes ! Certainly other signals should be incorporated in the model. In most present day softwares a post-treatment looks for ribosome binding sites around predicted gene beginnings, and eventually modifies start sites (Suzek *et al*., 2001 ; Besemer *et al*., 2001). Similarly promoters and terminators could also be used. Genes on the sequence can no longer be viewed as isolated units : such signals lead one to consider genome structure on a larger scale. In fact things go actually the reverse way : gene detection and annotation precede, and are used for, regulation signal detection and study.

THE PROBLEM OF EUKARYOTIC GENOMES

In eukaryotes, gene detection is a much more difficult problem, which has not yet received a satisfactory response. The reason is mostly that eukaryotic genomes are not compact, they have complex intergenic sequences, and interrupted genes. This causes serious difficulties to all gene detection methods, including those based on HMM.

Intergenic sequences bear several different structures, with many different textures. Good modelling of intergenic sequences is more critical since they are in the majority, and too little is known until now on the biological side. Repeated sequences are an exception. Markov chains are short memory processes and cannot produce neither separated nor tandem repeats, at any scale. Repeated sequences pose therefore a problem to HMM based gene detection methods; fortunately, they are well recognized, and in the most efficient gene detection softwares available today, sequences are preprocessed by special algorithms which detect and remove repeats before gene detection.

Introns interrupt coding sequences without regard to the coding phase. They display a different texture than exons, but the problem they address to texture based methods is hard, since segments become shorter, and sometimes very short. Hence the use of signals flanking exons is highly critical. Unfortunately, donor and acceptor sites are not as well defined or known as

stops and starts are, or not specific enough.
Another difficulty lies in segment length distribution. In a HMM, since the hidden process is a Markov chain, segment lengths should be distributed geometrically. Often this assumption remains acceptable, but sometimes, it is far too unrealistic ; this seems specially true for exons and at least some introns. There is no remedy inside a strict HMM frame, Hidden Semi Markovian Models offer a solution. HSMM, like HMM, are composed of a hidden and an observed process; this last one, conditionally on the hidden one, being a heterogeneous Markov chain. The difference is that the hidden process is semi-Markovian : like in HMM, hidden state changes form a Markov chain (indexed by the jumping instants), residence times within hidden states are independent, their distributions depend on the segment type, but they need not be geometrical. We will not develop these models here; they are implemented in GenScan and GeneMark.hmm; they certainly have a brilliant future.
The best software actually used for gene detection in eukaryotes, basically rely on HMM or HSMM schemes. But owing to peculiarities presented by eukaryotic genomes and the difficulties they induce, some of them use a lot of heuristics and sophisticated tricks in addition to the basic scheme. They do not result from an explicit statistical model and need learning sets. Their performances depend on the organism. For instance, EuGène is designed to use in a flexible way, all kind of information provided by other softwares as splice sites detectors (Schiex *et al*., 2001). It is actually the most efficient gene detector for Arabidopsis thaliana.

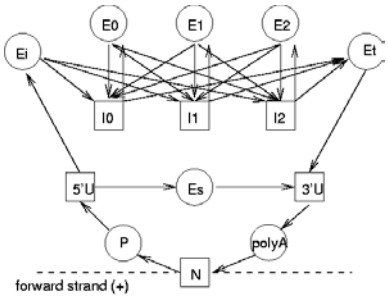

Figure 4. Structure of GenScan.: squares (resp. circles) for hidden states with geometrically (resp. non-geometrically) distributed lengths
E=exon, I=intron, U=UTR, P=Promoter. N=Intergenic. Numbers correspond to the phase: E1=exon beginning on phase 1.
Splice sites which are taken into account in the software, are not displayed here.

GenScan is semi-Markovian and probably the best gene detector for the human genome (Rogic *et al*., 2001). It relies on an exact HSMM scheme. Nevertheless, it also needs a learning set. There are two main reasons for this; any calculation on HSMM, in comparison with HMM, is much more greedy, even a simple Viterbi-like computation. Direct estimation of the parameter in a HSMM is a somewhat formidable task. But also, we still lack information about eukaryotic genomes, and models, which include a lot of parameters, are not very realistic. Estimation on the data would probably, even if feasible, not provide better results.

MAKING PROFIT FROM EXTRA INFORMATION

Until now, we presented gene detection as relying on the sequence to be analysed as an exclusive source of information, with eventually a training set of the same genome. But this was a forced picture ; sequence similarities and expressed sequence tags, for instance, are useful aids to detect small exons, or to make precise exon limits, and have been used for a few

years (Krogh, 2000 ; Schiex *et al*., 2001).
The number of annotated genomes increases very fast, and obviously, the annotated genome of a phylogenetically close relative provides extremely valuable information for gene detection and annotation of a new species. There is no doubt that comparative genomics will provide new research directions for gene detection in eukaryotes.
We tried to give a general overview on gene detection methods based on partially hidden processes. Our leading theme was a classification of the different sources of information at hand for this purpose. Signals and textures are extensively used today. Larger scale genome structure, phylogeny and homologies are key words for the future.

REFERENCES

Baum, L. E. and Petrie, T. (1966) *Ann. Math. Stat*. **37** : 1554-1563.
Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) *Ann. Math. Stat*. **41** : 164-171.
Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) *Nucleic Acids Res*. **29** : 2607-2618.
Borodovsky, M. and McIninch, J. D. (1993) *Comp. Chem*. **17** : 123-133.
Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Médigue, C. and Danchin, A. (1995) *Nucleic Acids Res*. **23** : 3554-3562.
Burge, C. and Karlin, S. (1997) *J. Mol. Biol*. **268** : 78-94.
Churchill, G. A. (1989) *B. Math. Biol*. **51** : 79-94.
Dempster, A., Laird, N. and Rubin, D. (1977) *J. Royal Statist. Soc*. *Ser*. B. **39** : 1-38.
Krogh, A., Mian, I. S. and Haussler, D. (1994) *Nucleic Acids Res*. **22** : 4768-4778.
Krogh, A. (2000) *Genome Res*. **10** : 523-528.
Lukashin, A. V. and Borodovsky, M. (1998) *Nucleic Acids Res*. **26** : 1107-1115.
Muri, F. (1997) PhD thesis, Université René Descartes, Paris V.
Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. and Bessieres, P. (2002) *Nucleic Acids Res*. **30** : 1418-1426.
Rabiner, L. R. A. (1989) *Proc. of the IEEE* **77** : 257-286.
Rogic, S., Mackworth, A. K. and Ouellette, F. B. (2001) *Genome Res*.
Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. (1998) *Nucleic Acids Res*. **26** : 544-548.
Schiex, T., Moisan, A. and Rouzé, P. (2001) “Computational Biology” volume 2066.
Suzek, B. E., Ermolaeva, M. D., Schreiber, M. and Salzberg, S. L. (2001) *Bioinformatics* **17** : 1123-1130.