BIOINFORMATICS

Vol. 15 no. 12 1999
Pages 980–986

Segmentation of yeast DNA using hidden Markov models

Leonid Peshkin[1] and Mikhail S. Gelfand[2]

[1]Computer Science, Brown University, Providence, RI 02912, USA and [2]State Scientific Center for Biotechnology, NIIGenetika, Moscow 113545, Russia

Received on September 9, 1998; revised on June 9, 1999; accepted on June 23, 1999

Abstract

Motivation: *Compositionally homogeneous segments of genomic DNA often correspond to meaningful biological units. Simple sliding window analysis is usually insufficient for compositional segmentation of natural sequences. Hidden Markov models (HMM) with a small number of states are a natural language for description of compositional properties of chromosome-size DNA sequences.*

Results: *The algorithms were applied to yeast* Saccharomyces cerevisiae *chromosomes (YC) I, III, IV, VI and IX. The optimal number of HMM states is found to be four. The optimal four-state HMMs for all chromosomes are very similar, as well as the reconstructed segmentations. In most cases the models with $k+1$ states are obtained by 'splitting' one of the states in the model with k states, and the corresponding increase of the level of detail in segmentation. The high AT states usually correspond to intergenic regions. We also explore the model's likelihood landscape and analyze the dynamics of the optimization process, thus addressing the problem of reliability of the obtained optima and efficiency of the algorithms.*

Availability: *The system is available on request from the first author.*

Contact: *ldp@cs.brown.edu*

Introduction

Developed initially for speech recognition (Rabiner, 1989), hidden Markov models (HMM) are now widely used in computational molecular biology. The usual applications of HMMs include multiple alignment and functional classification of proteins (Krogh *et al.*, 1994), prediction of protein folding (Di Francesco *et al.*, 1997), recognition of genes in bacterial and human genomes (Burge and Karlin, 1997; Kulp *et al.*, 1996; Henderson *et al.*, 1997), analysis and prediction of DNA functional sites (Crowley *et al.*, 1997), and identification of nucleosomal DNA periodical patterns (Baldi *et al.*, 1996).

However, the first application of HMMs to the analysis of genetic data, done by Churchill (1989), was to analyze the compositional heterogeneity of natural DNA sequences. Although it is the simplest level of DNA statistics, it still is not well understood. Base composition of natural DNA sequences is clearly non-uniform, and HMMs are a convenient language for its description. Despite the importance of the pioneering work by Churchill, a number of questions has remained unresolved. First, computational problems precluded analysis of long sequences, and allowed only the use of models with a small number of states (two or three). Second, no analysis of consistency of the generated segmentation or statistical confidence of a model choice was undertaken, and the features of the model parameter space were not analyzed. Finally, new data have become available recently, especially complete genomes and chromosomes that provide an interesting subject for the analysis.

The present work aims to fill some of these gaps and addresses general properties of HMMs in the domain of DNA sequences. We systematically investigate the computational aspects such as the structure of the parameter space, the model likelihood landscapes and robustness of the obtained data segmentation. The algorithms are applied both to natural and simulated DNA sequences.

In addition to the standard nucleotide sequences, it is sometimes of interest to consider binary representations. In this paper we address the strong–weak hydrogen bonding alphabet, and thus decompose the DNA sequences into segment of homogeneous GC-content. This segmentation is invariant relative to the DNA strand and is meaningful both physically and biologically (Bird, 1986; Churchill, 1989; Goodall and Filipowicz, 1989; Aissani *et al.*, 1991; Fickett *et al.*, 1992).

Materials

Yeast *Saccharomyces cerevisiae* chromosomes (YC) I, III, VI, IX and a fragment of chromosome IV (Goffeau *et al.*, 1996) were downloaded from GenBank (Benson *et al.*, 1999). Lengths and GC-contents of these sequences are given in Table 1.

Table 1. Optimal models (BIC and GC observation probabilities). The one-state model shows GC-contents.

HMM	YC I	YC III	YC VI	YC IX	YC IV (300Kbp)
length	230 203	315 339	270 148	439 885	300 000
one-state	.392	.386	.387	.389	0.384
BIC	−154 226	−210 232	−180 340	−293 984	−199 832
two-state	.354	.347	.355	.357	.351
	.456	.450	.446	.448	.425
BIC	−153 510	−209 345	−179 825	−293 138	−199 531
three-state	.293	.309	289	.269	.276
	.382	.389	.378	.381	.378
	.466	.477	.455	.461	.439
BIC	−153 414	−209 188	−179 749	−292 943	−199 479
four-state	.288	.178	.238	.249	.274
	.372	.336	.352	.356	.371
	.440	.402	.409	.399	.414
	.522	.488	.472	.472	.471
BIC	−153 376	−209 163	−179 744	−292 927	−199 472
five-state	.043	.151	.080	0.193	.06
	.301	.320	.294	0.313	.301
	.372	.371	.352	0.358	.356
	.441	.423	.407	0.402	.401
	.524	.498	.472	0.472	.469
BIC	−153 410	−209 176	−179 779	−292 962	−199 511

Algorithms

HMM

We are not giving a detailed description of HMMs and related algorithms (see Rabiner (1989) and Churchill (1989) for extensive coverage of this subject).

The HMM (sometimes called hidden Markov chains) is a well-known technique of stochastic modeling. Within this model, the observations are considered to be a stochastic function of a Markov process whose states are 'hidden', i.e. not directly observable. An ensemble of observation parameters O, state transition parameters T, and prior distribution over initial state constitutes a Hidden Markov Model. There are standard ways of estimating the model's parameters from data, as well as evaluating the probability of test data being generated by the model. The model produced can be used to impose segmentation of the data.

There are two aspects of this optimization problem: (i) assuming some N, find the best HMM among HMMs with N states; (ii) among best models with different number of states, find the best one.

We use the Baum–Welch (BW) algorithm which, given a sequence S and some initial model $\mathcal{M}$ with a given number of states N, adjusts the parameters (the transition matrix T and the observation matrix O) as to maximize likelihood $Pr(S|\mathcal{M})$. The BW algorithm is a version of expectation-maximization technique proved to converge to a local optimum. We run the BW algorithm starting from models with random initial parameters and keep the best *local optimum* until we agree to stop having found 'satisfactory' good local optimum. An improvement to this would be to draw initial models from some distribution, which is suggested by the problem nature and therefore is more likely to fall close to an optimal model. These questions are addressed in detail in the next two sections.

Segment-biased models

Consider a state transition matrix for a HMM. An assumption that there are long segments with homogeneous composition corresponds to the diagonal values of the transition matrix being close to 1. For example for a HMM with two states: $T = \begin{bmatrix} 1-\varepsilon_1 & \varepsilon_1 \\ \varepsilon_2 & 1-\varepsilon_2 \end{bmatrix}$. If ε values are small, the system remains in the state for a long time (expected $1/\varepsilon_1$ or $1/\varepsilon_2$ correspondingly) before it

makes a transition out of the state. Accordingly, to create a random model with transition matrix 'biased' towards segmented structure having uniform distribution over segments length, one should for each row i of transition matrix: (i) Choose a length of segment L_i between some small value (say, 10) and some fraction of the sequence length (say, half), according to the exponential distribution (so that the segments of length $10p$ are equally likely for all p) and set $\varepsilon_i = 1/L_i$. (ii) Using a uniform distribution spread the probability mass ε_i over off-diagonal elements of the row.

Global optimization and stopping criteria

Theoretically there is no way to determine whether we have found the optimal model with the given number of states for the given sequence. We accumulate the statistics over local optima, in order to obtain some estimate of how 'good' is the best local optimum found. Below we introduce the procedure of 'likelihood landscape exploration' which is an initial stage of sequence analysis: (i) generate a large number of random models and collect statistics of those models' likelihood: minimal and maximal values, mean and standard deviation. (ii) Run the local optimization procedure on a smaller number of random initial models, keeping track of how long it takes for each run to converge, thereby getting an estimate of how many different 'valleys' are there and how 'wide' each of them is. (iii) Collect statistics over the points of convergence. It is often the case that we converge to the same point starting from many different initial models (e.g. Figure 1).

Models comparison

Generally, when we are trying to describe some data, the more degrees of freedom we allow to the model, the better fit is obtained and therefore the higher likelihood for an optimal set of parameters is reached. To compare models with the same number of parameters we can simply use likelihoods, otherwise some form of penalty for extra degrees of freedom is necessary.

We use the Bayesian information criterion (BIC) (Schwarz, 1978). Given several candidates, the model with the maximum value of BIG $argmax_{\mathcal{M}}(l(\mathcal{M}) - \frac{1}{2}k \log L)$, is taken to be the best model. Here $l(M) = \log P(D|\mathcal{M})$ is the log-likelihood, k is the number of degrees of freedom in the model and L is the sequence length. Thus, the BIC value is a penalized form of the log-likelihood with a penalty that increases linearly with the number of parameters in the model.

Along with the theoretical criterion we also use an empirical one. Having found the best models of different orders, we compare them by observing to what extent an extra state is being used for description of the data and segmentation (see e.g. Figure 3). We found that for the genomic sequences this method gives results similar to BIG.

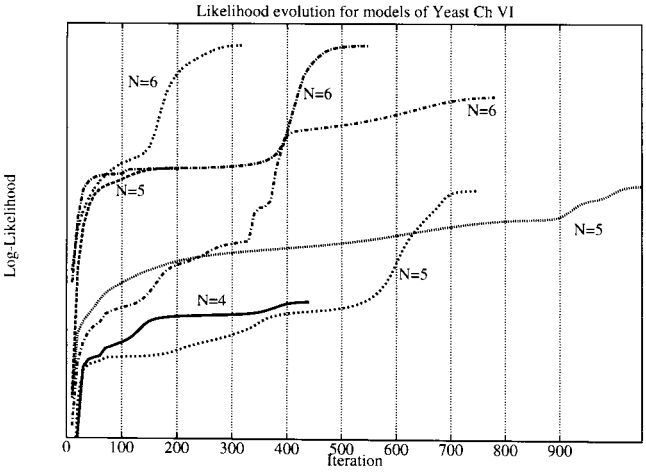

Fig. 1. The evolution of the likelihood for numerous runs of HMM optimization of yeast chromosome VI. Runs significantly vary in length and are often of sigmoid rather than convex shape. Notably, there are a few defining values of the likelihood between which the model makes a fast jump. A model with more states ($N = 5$) often first converges to a likelihood level of the best smaller model ($N = 4$), than jumps to the next level, discovering how to take advantage of extra degrees of freedom.

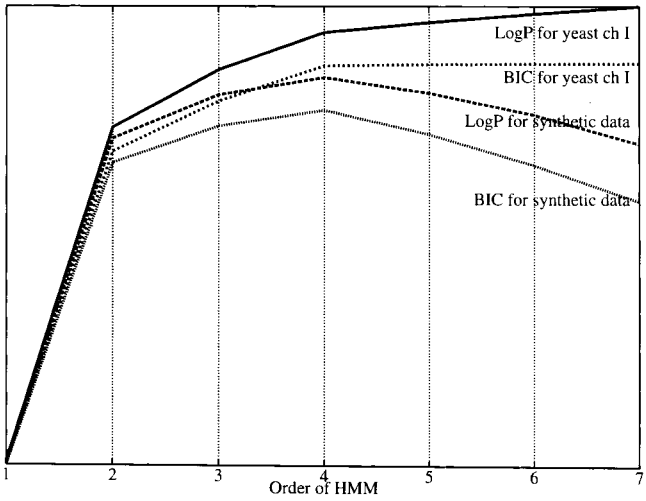

Fig. 2. Plots of log-likelihood and BIC for YCI and a synthetic sequence generated using the best four-state model for YCI.

Implementation

The BW implementation in a modified version of software from package described in Shatkay (1998). The random generator was taken from Cassandra (1997).

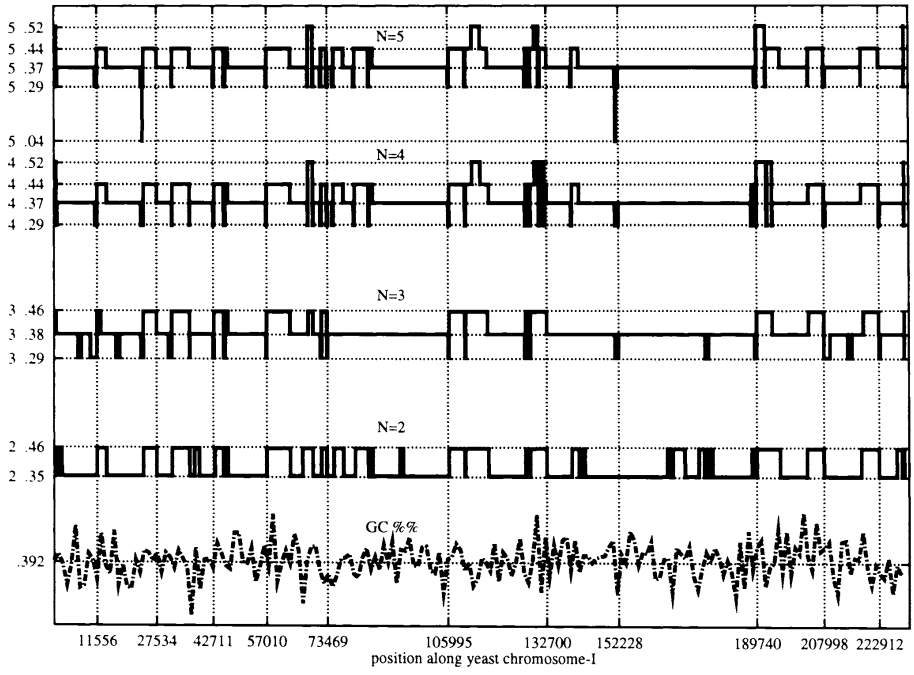

Fig. 3. Segmentation of YCI by HMMs with different number of states. Numbers on the left indicate the order of the model and the GC probability corresponding to each state. The lower stripe corresponds to the two-state HMM segmentation, the upper, to the five-state HMM segmentation. The plot at the bottom represents variation of GC-content for YCI within sliding window of 1150 bp.

Results

Robustness

We have explored the robustness of solutions and effectiveness of HMM algorithms using simulated data to do a reverse computation — keep the HMM and the sequence of hidden states used to generate a sequence of observations, and see if we are able to recover them starting from a distorted initial model.

We performed two series of model restoration experiments: (i) the initial model had the correct state transition matrix T (taken from the actual model used to generate the data), but a randomly distorted observation distribution matrix O; (ii) on the contrary, the observation matrix O was taken from the correct model, whereas the state transition matrix T was altered.

It turned out that distortion of the observation probabilities strongly influences the behavior of the algorithm. This means that estimating an unknown HMM, we should accurately sample the observation parameter space by trying models with initial setting on a regular grid or generated from the uniform distribution. On the other hand, the 'segment biased' models are very robust towards distortion of state transition probabilities in the initial model: the BW algorithm properly converges back to the original model in most cases. Thus the distribution of the initial settings for state transitions could be skewed towards more likely parameters and explored with less care and computational costs.

The Viterbi algorithm seeks the most *likely* hidden state sequence which does not necessarily coincide at every position with a real sequence recorded while generating sample data. To analyze the robustness of the Viterbi algorithm, we have applied it to a number of observation sequences generated by a given model with a known hidden state sequence. The disagreement between the initial and reconstructed sequences of hidden states was less than 0.1%. Thus having found a correct model we are certain to arrive to the correct segmentation.

Finally, we note that BIC correctly estimates the number of hidden states in synthetic data (Figure 2).

Estimation of HMM

Table 1 presents the optimal models with one through five states for each chromosome. Note that in our case the ensemble of observation probabilities O has only two components. Hereinafter we present the GC-content values when the observation probabilities are discussed. The optimal models in all cases have four states, although the difference between BIC for four and five state models for YCIII and the three and four state models for YCVI and YCIV is not large.

Table 2. Comparison of a chromosome segmentations obtained using three different three state models. Each large cell (K, L) $(K, L \in \{\text{I, III, IV, VI, IX}\})$ describes correlation between the segmentation of chromosome K by the model estimated on this chromosome and the model estimated on chromosome L. Within this cell element (i, j) gives the percentage of length of the chromosome described by state i of model K and by state j of model L.

Object sequence	Sequence used to estimate the model														
	YC I			YC III			YC VI			YC IX			YC IV		
YC-I				18.5	3.6	.1	22.0	.1	0	22.0	.1	0	22.0	.1	0
				.46	66.7	5.2	3.0	69.0	.3	1.2	71.0	.1	5.3	66.6	.4
				0	.5	5.1	0	2.3	3.2	0	2.5	3.0	.1	2.8	2.7
YC-III	12.2	.1	0				12.1	.1	0	12.1	.1	0	12.1	.1	0
	4.1	70.4	.3				4.1	70.5	.2	4.3	70.1	.4	9.1	65.4	.2
	.1	6.3	6.6				.2	8.5	4.3	.2	9.3	3.5	.3	9.9	2.7
YC-VI	13.9	1.7	.3	9.5	6.3	.1				14.0	1.8	0	15.5	.3	0
	0	80.1	1.1	.1	76.4	4.7				0	81.1	.1	3.9	77.2	.1
	0	0	2.9	0	.6	2.3				0	1.23	1.6	.3	.8	1.9
YC-IX	13.2	1.0	0	10.7	.1	0	14.2	2.3	0				14.2	7.0	0
	.5	81.7	1.8	3.5	80.5	.4	0	80.5	.2				0	75.9	.2
	0	.1	1.7	.1	3.3	1.4	0	1.1	1.6				0	1.0	1.6
YC-IV	5.7	5.9	.1	3.8	7.6	.1	8.6	2.9	.1	6.0	5.6	0			
	0	86.1	.8	.1	84.2	2.5	.1	86.1	.6	.1	86.5	.2			
	0	.3	1.2	0	.6	.9	0	.5	1.0	0	.3	1.2			

Figure 2 presents the comparison of the best models of different order using log-likelihood and BIC for YCI and a synthetic sequence generated using the best four-state model for YCI. The optimal model for the synthetic data is clearly the four-state one, and additional states provide only minor increase of the likelihood and some of them are never visited. Although the log-likelihood for YCI continues to grow for models with more than four states, the plot of BIC has a marked local maximum at four states. Moreover, the overall shapes of the BIC plots for real and synthetic data, and in particular the relative heights of the local maxima, are very similar. This observation provides additional support for the four-state model.

It is interesting to trace the birth of new states (see Table 1). Normally one of the states for the model with n states would split into two states for the model with $n + 1$ states, leaving the rest of the states almost unchanged. Remarkably, the three-state models have very similar observation probabilities for different chromosomes, meaning that the main features of all three chromosomes are the same. The four-state models thus capture the individual properties of the chromosomes.

As an additional control we have also compared the distribution of segment lengths among the yeast chromosomes, the synthetic data and a fragment of the *E. coli* genome. The distributions of segment lengths for yeast chromosomes are very close to each other and significantly different from the other two sequences (data are not shown).

Segmentation

Figure 3 presents the segmentation of YCI by the best models with two through five states. The patterns of segmentation by the two- and three-state models are rather different. The change is smaller as we go up to four-state HMM. As we further increase the number of states there is no significant change in the segmentation. The fifth state with GC-content .043 is only visited twice very briefly along the whole sequence of 230 Kbp. As a matter of fact this state covers only .06% of the sequence length, whereas the next rarest states with GC-contents .29 and .52 are used for approximately 3% of the sequence, the state with .44 GC for 24%, and state .37 GC for 69%.

Experiments with other yeast chromosomes produced similar outcomes and for brevity we do not present these segmentation results here.

Since the three-state models of the yeast chromosomes are similar, it is interesting to analyze segmentation of a chromosome by the optimal model obtained for a different chromosome. The agreement between different segmentations is illustrated by Table 2. Indeed, the diagonal elements, corresponding to regions consistently described by different models, dominate in all cells.

Finally, it should be noted that positions of transitions between states produced by different models (either trained on different chromosomes or with different number of states) often exactly coincide (compare different panels on Figure 3).

Table 3. Percentage of occupancy of every state (GC-content increases left-to-right) of a four-state model by protein-coding (bottom line) and non-coding (top line) fragments of chromosomes.

CH I	3.45	27.77	5.07	1.02
	0.27	39.90	18.91	3.61
CH III	0.62	14.11	15.18	1.42
	0.03	13.63	45.99	9.01
CH IV	1.60	20.38	3.76	0.61
	0.23	54.82	16.09	2.50
CH VI	1.14	18.35	10.85	2.32
	0.15	32.65	28.91	5.64
CH IX	1.57	11.70	15.11	1.41
	0.22	27.69	33.84	8.44

HMM and functional segmentation

As expected, the HMM segmentation correlates with a natural segmentation of chromosomes into protein-coding and non-coding regions (Table 3). Indeed, the AT-richest state (the leftmost) almost exclusively occurs in non-coding regions, whereas the GC-richest state (the rightmost) occurs mostly in coding regions. The middle states (for four-state models) also are correlated with genes, but to a lesser extent.

Discussion

Dividing genomes into compositionally homogeneous regions is important for a number of reasons. First, these segments are compositionally meaningful. Indeed, we have observed that AT-rich states usually occur in intergenic regions. This agrees with the observation that the GC-content of protein-coding regions is higher than that of non-coding regions in mammals (Aissani *et al.*, 1991) and plants (Goodall and Filipowicz, 1989), but contradicts the observation that GC-rich segments often occur in regulatory regions of vertebrate genes (Bird, 1986). We have also observed that subtelomeric regions of all yeast chromosomes form separate segments. It is noteworthy that some homogeneous segments are rather long (e.g. the region 152,228–189,740 on Figure 3), whereas in other cases the segmentation pattern is rather complex (e.g. the region around pos. 73,469). Further analysis is required to assess the biological meaning of this observation.

Second, it is well known that performance of many algorithms for functional mapping of genomic DNA crucially depends on homogeneity of the sequences and it deteriorates when the sequences are not homogeneous. The compositional segmentation of genomes allows for use of different sets of parameters fine-tuned for each composition and thus to improve performance (Burge and Karlin, 1997).

Finally, the HMM segmentation says something about the general statistical properties of natural DNA.

We conclude that HMM are a useful and convenient tool for statistical analysis of the genomic data. They provide detailed yet robust descriptions that can be used for further structural analysis or direct biological interpretation.

Acknowledgements

We are grateful to Leslie Pack Kaelbling and David Haussler for useful comments, and to Hagit Shatkay who provided an early version of the BW algorithm implementation. M.G. was partially supported by the Russian Fund of Basic Research and Russian State Program ‘Human Genome’.

References

Aissani,B. *et al.* (1991) The compositional genes properties of human genes. *J. Mol. Evol.*, **32**, 493–503.

Baldi,P., Brunak,S., Chauvin,Y. and Krogh,A. (1996) Naturally occurring nucleosome positioning signals in human exons andintrons. *J. Mol. Biol.*, **263**, 503–510.

Benson,D., Boguski,M., Lipman,D., Ostell,J., Guellette,B., Rapp,B. and Wheeler,D. (1999) Genbank. *Nucleic Acids Res.*, **1**, 12–17.

Bird,A. (1986) CG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Cassandra,A. (1997) Exact and approximate algorithms for partially observable Markov decision processes, PhD thesis, Brown University.

Churchill,G. (1989) Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, **51**, 79–94.

Crowley,F., Roeder,K. and Bina,M. (1997) A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.*, **268**, 8–14.

Di Francesco,V. *et al.* (1997) Incorporating global information into secondary structure prediction with hidden Markov models of protein folds. *5th Int. Conf. on Intelligent Systems for Mol. Biology* ISMB-97, pp. 100–103.

Fickett,J.W., Torney,D.C. and Wolf,D.R. (1992) Base compositional structure of genomes. *Genomics*, **3**, 1056–1064.

Goffeau,A. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.

Goodall,G. and Filipowicz,W. (1989) The AU-rich sequences present in the introns of plant nuclear pre-mrnas are required for splicing. *Cell*, **58**, 473–483.

Henderson,J., Salzberg,S. and Fasman,K. (1997) Finding genes in DNA with an HMM. *J. Comput. Biol.*, **4**, 127–141.

Krogh,A., Brown,M., Mian,I., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology: application to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Kulp,D., Haussler,D., Reese,M. and Beckman,F. (1996) A generalized hidden Markov model for the recognition of human genes

in DNA. *4th Int. Conf. on Intelligent Systems for Mol. Biology* ISMB-96, pp. 134–141.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Schwarz,G. (1978) Estimating the dimension of a model. **6**, 461–464.

Shatkay,H. (1998) Learning models for robot navigation, PhD thesis, Brown University.