A TIME-VARYING MODEL FOR DNA SEQUENCING DATA SUBMERGED IN CORRELATED NOISE

Nicholas M. Haan and Simon J. Godsill

Signal Processing Group
Department of Engineering, University of Cambridge, U.K.
email: nmh28@cam.ac.uk

ABSTRACT

Methods for determining the letters of our genetic code, known as DNA sequencing, currently depend on clever use of electrophoresis to generate data sets indicative of the underlying sequence. Typically the subsequent off-line data processing is carried out using a combination of heuristic methods with little mathematical rigour. In this paper, we present a novel model which is able to accurately predict the effect of the many biological processes which are involved, and moreover, which is usable on-line. Off-line methods have been hampered by the need for processing in as little time as possible after the data is generated; performing the processing on-line has enabled a more advanced algorithm to be used with associated improved performance. The algorithm is framed within a Bayesian probabilistic framework, thereby allowing representation of the random nature of the generative process, and relies on new advances in the burgeoning field of Sequential Monte Carlo Methods to perform the required highly non-linear filtering and model selection operations.

1. INTRODUCTION

Deoxyribonucleic acid (DNA) is the molecule used to encode the genetic information within each of us. For our purposes, DNA can be thought of as a sequence of symbols (in reality, chemical bases) taken from a four letter alphabet comprising: A (Adenine), G (Guanine), C (Cytosine), and T (Thymine).

In 1974, Sanger proposed a method for DNA sequencing which, with technical improvements, has since been almost universally accepted [10]. The idea behind the process is simple. Initially, via a process of replication and truncation the DNA sequence of interest is used to form a large population of partial replicas. Each replica is identical to the sequence of interest over a range of bases, always commencing with the first base of the initial sequence, and terminating some random distance down the strand. That is, for the sequence ACGGG the population would contain a number of each of the following: A, AC, ACG, ACGG, and ACGGG. Each fragment is fluorescently labelled according to its terminating base. Subsequently, the entire population is aligned at the start of a large rectangular gel, and an electric field is applied. The fragments progress through the gel at rates approximately inversely proportional to their length, resulting in the various subpopulations arriving at the end of the gel in sequence order. A laser positioned near the end of the gel excites the fluorescent labels, allowing an emission detector to estimate the number of fragments terminated by a given base passing at each time instant.

After some preprocessing, four data sets are obtained (henceforth, "channels"), corresponding to the variation of fragment concentration with time for each of the four terminating bases. This collection of data is known as an electropherogram and is quite clearly indicative of the underlying base sequence. The electropherogram is a mixture of peaks in four channels, with each base in the sequence associated with one major peak in the corresponding channel, and three secondary peaks in the remaining channels resulting from leakage effects; the peaks corresponding to a particular base have common position and shape. An example data set is shown in figure 1.

A range of prior information, mainly detailing the effect of base sequence on the amplitudes and positions of the peaks, is available to constrain the problem; [11] provides a good review. The current state of the art from an off-line signal processing perspective is described in [3], where a combination of heuristic peak detection algorithms is proposed. [5] presents a more advanced algorithm based on statistical modelling of the underlying process, with an associated increase in computational burden.

Here, a model similar to that of [7] [6], which is capable of representing available prior information about the system, is detailed. The model is developed in a framework which allows sequential updating of the required inference. One of the major improvements of the model is that it allows removal of slowly varying background noise, which may be correlated with the desired signal. It is also able to track nonstationarity in the various processes - no previous sequential algorithm has attempted this. The resulting algorithm can be run on-line and has immediate application to all data sets which comprise a series of peaks arriving sequentially in time (for example, some spectroscopy applications). A simple version of the algorithm is presented below, which does not account for background variation, nor (fully) nonstationarity in the processes. The final version of the paper will include modifications for the more advanced case.

2. PROBLEM FORMULATION

2.1. Signal Model

We present a model suited to sequential processing of the system; an alternative block based approach is detailed in [5]. Electro-

pherogram data is well described as the summation of a series of peaks, observed in noise. A general model for electropherogram data in the four channels at time n ($n \in \{1, \ldots, N\}$), $\mathbf{y}_n \triangleq \{y_{n,1}, \ldots, y_{n,4}\}$, is thus:

$$\mathbf{y}_n = \mathbf{e}_n + \sum_{i=1}^{k} \mathbf{b}_i \phi_i(n) \tag{1}$$

where k denotes the total number of bases, $\mathbf{b}_i \triangleq \{b_{i,1}, b_{i,2}, b_{i,3}, b_{i,4}\}$ is a vector defining the amplitudes in the four channels corresponding to base i (and is representative of the number of fragments passing the end of the gel at a given time via the emission spectra of the dyes), and $\phi_i(n)$ defines the peak shape. $\mathbf{e}_n \triangleq \{e_{n,1}, e_{n,2}, e_{n,3}, e_{n,4}\}$ represents the noise in the system at time n. Here, we consider the subset of electropherogram data where the noise can be assumed Gaussian with zero mean and constant variance, σ_e^2 (a slightly more complicated noise model is presented in [5]):

$$\mathbf{e}_n \sim \mathcal{N}\left(\mathbf{e}_n | \mathbf{0}, \sigma_e^2 \mathrm{I}_{4\times 4}\right) \tag{2}$$

where $\mathrm{I}_{4\times 4}$ denotes the identity matrix. In many cases, the peaks can be assumed truncated Gaussian in shape such that:

$$\phi_i(n; p_i, v_i) = (2\pi v_i)^{-.5} \exp\left\{-\frac{1}{2v_i}(n-p_i)^2\right\} \mathbb{I}(|n-p_i| \le \epsilon) \tag{3}$$

where ϵ is defined a-priori to be sufficiently large that for the range of possible variances, the truncation effect is minimal. v_i denotes the variance of the peak.

The "base-state" of the system at base position i, $\mathbf{s}_i \triangleq \{s_{i,-2}, s_{i,-1}, s_{i,0}\}$, is defined as a base triplet, e.g., $\{A, G, C\}$, with the last element corresponding to the current base, and the other elements containing the previous two (the previous two are included to account for sequence dependent effects, e.g. [8]). The following prior structure is proposed for the parameters of base i:

$$s_{i,0} \sim p(s_{i,0}|s_{i-1,0}), \quad \mathbf{s}_{i,-2:-1} = \mathbf{s}_{i-1,0:-1} \tag{4}$$

$$p_i = p_{i-1} + 1 + \Delta_i, \quad \Delta_i \sim \mathcal{G}\left(\Delta_i | \alpha_p(\mathbf{s}_i), \beta_p(\mathbf{s}_i)\right) \tag{5}$$

$$\mathbf{b}_i \sim \prod_{j=1}^{4} \mathcal{G}\left(b_{i,j} | \alpha_{b,j}(\mathbf{s}_i), \beta_{b,j}(\mathbf{s}_i)\right) \tag{6}$$

$$v_i \sim \mathcal{G}\left(v_i | \alpha_v(v_{i-1}), \beta_v(v_{i-1})\right) \tag{7}$$

where $\mathcal{G}(.)$ is the Gamma distribution, and $\mathbf{a}_{i:j} = \{a_i, \ldots, a_j\}$. The functions $\left\{\alpha_{b,j}(\mathbf{s}_i), \beta_{b,j}(\mathbf{s}_i), \alpha_p, \beta_p, \alpha_v, \beta_v\right\}$ are considered time invariant and known *a-priori*. The states evolve according to a Markovian structure. Further, $p_1 \sim \mathcal{N}(p_1|.)$, $v_1 \sim \mathcal{G}(v_1|.)$, and $\mathbf{s}_{1,-2:0} \sim p(\mathbf{s}_{1,-2:0})$, a prespecified initial state distribution. It should be noted that in electropherogram data there is almost never more than one base in a given unit time interval as evidenced by the $+1$ in equation (5); the framework is extensible to the case where there are multiple peaks in a unit interval.

2.2. State-Space Form

The set of peaks affecting the data at time n is given by the index set $\mathcal{I}_n \triangleq \{i : |n - p_i| \le \epsilon, i \in \{1, \ldots, k\}\}$. The data at time n is then completely defined by:

$$\mathbf{y}_n = \sum_{i \in \mathcal{I}_n} \mathbf{b}_i \phi\left(\frac{n-p_i}{\sqrt{v_i}}\right) + \mathbf{e}_n \tag{8}$$

and, therefore, the state of the system at time n can be written $\boldsymbol{\theta}_n \triangleq \{\mathbf{b}_i, p_i, v_i, \mathbf{s}_i : i \in \mathcal{I}_n\}$. The dimension of the state, $k_n = \dim \mathcal{I}_n$, varies with time according to the number of peaks affecting the data. Here, we make the mild (and not strictly necessary) assumption that $k_n > 0$, in order to ensure that, given $\{\boldsymbol{\theta}_{n-1}, k_{n-1}\}$, the prior on $\{\boldsymbol{\theta}_n, k_n\}$ is completely determined. The resulting model is a Hidden Markov Model.

The peak spacing prior of the previous section requires some reformulation to be used within a sequential framework. Given $\{\boldsymbol{\theta}_{n-1}, k_{n-1}\}$, and assuming none of the peaks affecting the data at time $n-1$ become superfluous, the probability of a new peak being introduced is the probability, as defined by the peak spacing prior, of there being a new peak in the interval $(n+\epsilon-.5, n+\epsilon+.5]$, where $\gamma_n \triangleq \max\{\mathcal{I}_n\}$ is introduced as the base index of the last peak affecting the data at time n:

$$p\left(k_n = k_{n-1} + 1 | p_{\gamma_{n-1}}\right) = \frac{\int_{p \in (n+\epsilon-.5, n+\epsilon+.5]} \mathcal{G}\left(p - p_{\gamma_{n-1}} - 1|.\right) dp}{\int_{p \in (n+\epsilon+.5, \infty]} \mathcal{G}\left(p - p_{\gamma_{n-1}} - 1|.\right) dp} \tag{9}$$

Given there is a new peak in this interval, its position *a-priori* is drawn according to $\mathcal{G}\left(p_{\gamma_n} - p_{\gamma_{n-1}} - 1|.\right)$ truncated appropriately. The remaining variance, amplitude, and state parameters of the new peak are then defined according to the priors of the previous section, and the state of the system becomes: $\boldsymbol{\theta}_n = \{\boldsymbol{\theta}_{n-1}, \{\mathbf{b}_{\gamma_n}, p_{\gamma_n}, v_{\gamma_n}, \mathbf{s}_{\gamma_n}\}\}$, with $k_n = k_{n-1} + 1$.

If a state ceases to be relevant at time n, i.e., its position is less than $n - \epsilon$, then the state at time n is deterministically reduced: $\boldsymbol{\theta}_{n,1:k_n} = \boldsymbol{\theta}_{n-1,2:k_{n-1}}$, $k_n = k_{n-1} - 1$. Similarly, if there is no change in the number of peaks affecting the data: $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1}$, $k_n = k_{n-1}$. Effectively, we have defined a sliding parameter window of variable dimension.

2.3. Estimation Objectives

In a Bayesian framework the posterior distribution at time n, $p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n} | \mathbf{y}_{1:n})$, is used for inference, with the expected value of a function of interest $f(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n})$ under this posterior given by $\int\int f(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}) p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n} | \mathbf{y}_{1:n}) d\boldsymbol{\theta}_{1:n} d\mathbf{k}_{1:n}$. In most cases, including ours, the posterior is not amenable to closed form analysis owing to non-linearity in the parameters, and it is necessary to resort to numerical methods. In [5] a batch processing scheme using Markov Chain Monte Carlo (MCMC) methods is successfully used to simulate variates from the posterior, and make associated inference on quantities of interest. These methods are, however, computationally intensive and make little use of sequential-in-time structure in the system. Here, we develop a numerical algorithm to estimate the posterior distribution recursively in time for on-line estimation.

3. SEQUENTIAL SIMULATION

3.1. Ideal Recursion

Since the system is Markovian, a recursion for calculation of the posterior at times $n > 1$ is:

$$p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n}) = \frac{p(\mathbf{y}_n|\boldsymbol{\theta}_n, k_n)\, p(\boldsymbol{\theta}_n, k_n|\boldsymbol{\theta}_{n-1}, k_{n-1})}{p(\mathbf{y}_n|\mathbf{y}_{1:n-1})} \times p(\boldsymbol{\theta}_{1:n-1}, \mathbf{k}_{1:n-1}|\mathbf{y}_{1:n-1}) \quad (10)$$

where at time $n = 1$,

$$p(\boldsymbol{\theta}_1, k_1|\mathbf{y}_1) = \frac{p(\mathbf{y}_1|\boldsymbol{\theta}_1, k_1)\, p(\boldsymbol{\theta}_1, k_1)}{p(\mathbf{y}_1)} \quad (11)$$

We emphasise that, despite the time-dependent notation, all of our parameters are genuinely time varying.

3.2. Sequential Bayesian Computation

The ideal recursion cannot be used directly owing to analytic difficulties. One alternative is to represent the posterior at each time by a set of weighted particles [2, 9]:

$$\hat{p}(d\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n}) = \sum_{i=1}^{P} \tilde{w}_n^{(i)} \delta_{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}}(d\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}) \quad (12)$$

where P denotes the number of particles, $\tilde{w}_n^{(i)}$ denotes the normalised importance weight associated with the particle of value $\left\{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}\right\}$, and $\delta_{\boldsymbol{\theta}_{1:n}^{(i)}, \mathbf{k}_{1:n}^{(i)}}(.)$ is the delta function. An algorithm for updating the particles as time progresses is [2, 9]:

Algorithm 1 - ***Monte Carlo Filter***

For $n = 2, \ldots, N$

For $i = 1, \ldots, P$*:*

- *Draw from the importance distribution*

$$\boldsymbol{\theta}_n^{(i)}, k_n^{(i)} \sim \pi\left(\boldsymbol{\theta}_n, k_n|\boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}, \mathbf{y}_{1:N}\right)$$

- *Evaluate the unnormalised importance weights:*

$$\overline{w}_n^{(i)} = \frac{p\left(\mathbf{y}_n|\boldsymbol{\theta}_n^{(i)}, k_n^{(i)}\right) p\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)}|\boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}\right)}{\pi\left(\boldsymbol{\theta}_n^{(i)}, k_n^{(i)}|\boldsymbol{\theta}_{n-1}^{(i)}, k_{n-1}^{(i)}, \mathbf{y}_{1:N}\right)} \overline{w}_{n-1}^{(i)}$$

- *Normalise the importance weights:*

$$\tilde{w}_n^{(i)} = \left(\sum_{j=1}^{P} \overline{w}_n^{(j)}\right)^{-1} \overline{w}_n^{(i)}$$

- ***Optional:*** *Resample to obtain P samples approximately distributed according to $p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n})$. Set the weights equal.*
- ***Optional:*** *Apply a Markovian transition kernel invariant to the posterior for each particle stream.*

End For

End For

At time $n = 1$ the particles $\left\{\boldsymbol{\theta}_1^{(i)}, k_1^{(i)}\right\}$ are similarly drawn from an initial importance distribution, and weighted appropriately.

3.3. Importance Distribution

At time n the importance distribution predicting whether a new peak has been introduced is chosen to be a mixture of a distribution based on a simple deterministic peak detection algorithm and the prior. Given that a new peak is proposed, the peak position is then generated according to a truncated uniform on the space $(n + \epsilon - .5, n + \epsilon + .5]$. The variance of the peaks changes very slowly with time, and so is quite predictable from one base to the next without reference to the data. The prior is therefore sufficient for an efficient importance distribution.

The importance function for $\mathbf{b}_{\gamma_n}$ (recall γ_n denotes the newest peak affecting the data at time n) is obtained by approximately integrating out the effect of two possible future peaks (in electropherogram data it is very rare to have more than 5 peaks in total affecting a given data point significantly). Details of this step, will appear in the final version. Alternatively, we have found local linearisation of this distribution, or the state space, to be useful in some scenarios (see [2] for a review of these methods). Also, in cases where the prior on the amplitudes can be assumed Gaussian, the amplitudes can be marginalised (with a slight reformulation such Rao-Blackwellization [9] can be performed via the Kalman filter- this too will appear in the final version).

The importance distribution for $\mathbf{s}_{\gamma_n}$ is set to be the full conditional $p\left(\mathbf{s}_{\gamma_n}|\mathbf{s}_{\gamma_{n-1}}, p_{\gamma_{n-1}}, p_{\gamma_n}, \mathbf{b}_{\gamma_n}\right)$ which can be directly evaluated.

Proposing the initial set of particles can be difficult since often parameters for 5 or 6 bases will be required. MCMC methods can be particularly helpful [5].

3.4. Resampling

The resampling step aims to multiply or discard particle trajectories according to how important they are to our approximation of the posterior distribution. When a resampling step is performed we use the standard residual method described in [9]. The resampling schedule must be chosen carefully or degeneracy can result. The problem can be isolated by noticing that the state at time n may include peaks which do not yet significantly affect the likelihood at time n. Therefore, for these peaks a resampling step corresponds to a re-weighting according to the prior; it takes time for new peaks to filter through the likelihood function. If the prior deviates significantly from the likelihood, poor results will follow. However, this is rarely the case, and provided the particle cloud is relatively large, degeneracy is usually not serious.

3.5. Markov Transition

Our model is defined on a variable dimension space, with parameters fixed over moderate time intervals. Degeneracy of the standard particle filter for such systems is commonly known. Here, since the interval of invariance is not too large, MCMC transitions can be used to help replenish the particle set [1]. That is, a kernel invariant to the posterior distribution, $p(\boldsymbol{\theta}_{1:n}, \mathbf{k}_{1:n}|\mathbf{y}_{1:n})$, is applied, the idea being that such a transition can only decrease the difference between the current approximate distribution and the invariant distribution. The kernel used consists of a fixed dimension part based on a modified Gibbs sampler and a variable dimension part based on a birth/death - split/merge Reversible Jump kernel. The reader is referred to the algorithm described in [5] for more details. In order to reduce the computational load, transitions are

applied only on a subset of the total parameter space, corresponding to those peaks centred relatively near the current time.

4. RESULTS

To maintain simplicity, we consider an instance of real data where we may assume there are no complicated sequence dependent effects, and that each of the state transitions is equally likely. The prior on peak spacing was taken to have mean, 11, and variance, 6. The variance of the drift on peak variance with time was set to $.1^2$, which is typical of electropherogram data. The main peak of each quadtuplet was set a-priori to have mean amplitude 100 with a variance of 400 in the channel corresponding to the base in question - a mild assumption - and the remaining peaks of the quadtuplet, to have mean 30 and variance 100. The noise variance was set to 1; ϵ was set to 18. The data set used was approximately 15000 samples long, of which a 70 sample window is considered here. 100 particles were used.

In figures 1 and 2 the data is shown superimposed on the denoised signal as predicted by the two highest probability particles at time 70. It is visually apparent that in both cases the data fit is good. However, though it is quite hard to see by eye, the 7 base interpretation is more strongly supported by the data (likelihood) than the 6 base. The priors on peak spacing and amplitude, however, balance this in favour of the more parsimonious 6 base model. The probabilities of the two models, as given by their frequency in the particle set, were .63 and .32 for the 6 and 7 base systems respectively; other model orders were not strongly supported. The correct interpretation was in fact the 6 base system, with base sequence the same as that shown in figure 2.

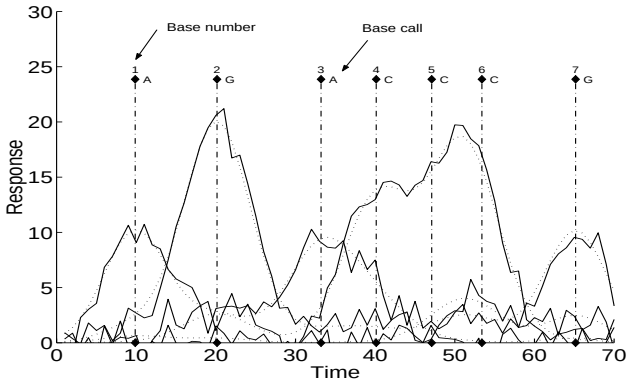

Fig. 1. Data (solid lines) with denoised signal (dotted lines) and base calls for 7 base interpretation. The four channels are superimposed.

5. CONCLUSIONS

We have briefly introduced the DNA sequencing problem, and provided a meaningful statistical framework within which to represent available information. This framework was then translated into one suitable for sequential estimation of the posterior distribution of interest as it evolves in time. Results of the algorithm are promising, indicating that the model selection and fixed-parameter problems are tractable in this framework. A more detailed discussion of more complex models for sequencing data (both in a block based and sequential framework) will be forthcoming in [4]. The algorithm is directly applicable to a broad range of data sets.

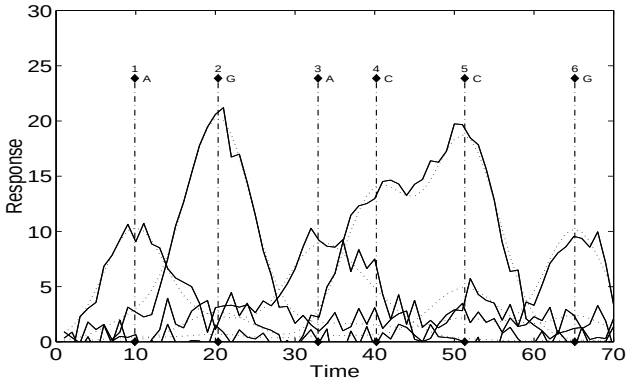

Fig. 2. Data and base calls for 6 base interpretation

6. REFERENCES

[1] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.

[2] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[3] B. Ewing, L. Hillier, M. Wendl, and P. Green. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8:175–185, 1998.

[4] N. Haan. *Applied Statistical Signal Processing for DNA Sequencing.* PhD thesis, University of Cambridge, 2001.

[5] N. Haan and S. Godsill. Modelling electropherogram data for DNA sequencing using MCMC. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000. Paper no. 2573.

[6] N. Haan and S. Godsill. Robust modelling and analysis of DNA sequencing data using MCMC. *Submitted to IEEE Trans. Sig. Proc.*, 2001.

[7] N. Haan and S. Godsill. Sequential methods for DNA sequencing. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001. Paper no. 2227.

[8] R. Lipschutz, F. Taverner, K. Hennessy, G. Hartzell, and R. Davis. DNA sequence confidence estimation. *Genomics*, 19(417-424), 1994.

[9] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *J. Amer. Stat. Assoc.*, 93, 1998.

[10] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci*, 74:5463–5467, 1977.

[11] D. Thornley. *Analysis of Trace Data from Fluorescence Based Sanger Sequencing.* PhD thesis, University of London, Imperial College of Science, Technology, Medicine, Dept. of Computing, 1997.