BAYESIAN MODELS FOR DNA SEQUENCING

Nicholas M. Haan and Simon J. Godsill

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

ABSTRACT

It is becoming increasingly important to develop novel signal processing and statistical analysis techniques to extract information from biotechnology. This task is complicated by large datasets, intricate physical systems, and the sheer diversity of information that is available. In many systems, classical non-parametric signal processing techniques have been applied with some success. However, where sufficient information is available to construct accurate models, substantial gains can sometimes be derived from a model-based approach. The Bayesian paradigm provides an elegant and mathematically rigorous framework for the objective incorporation of information. In this paper, we develop a Bayesian model for DNA sequencing, with an emphasis on generally relevant Bayesian model selection issues.

1. INTRODUCTION

The aim of Signal Processing is to develop algorithms that use as much information about a system as possible without imposing unfeasible computational requirements. This is particularly true of biotechnology, where there is a wealth of prior information about the chemical and physical processes leading to data generation, but the systems are too complicated for all effects to be included in the analysis process. Further complicating the situation, most biological systems are subject to uncertainty, either deriving from our understanding of the system or genuine random effects (e.g. where quantum effects become significant). In this paper, we discuss the use of the Bayesian paradigm for rigorous inference from DNA sequencing data where there is both uncertainty and a number of different information sources to be included.

The basic form of the DNA sequencing process is common to many systems in biotechnology and is illustrated in figure 1. At the heart is a biological object or system about which we wish to make inference (in this case, the sequence of an unknown piece of DNA). This object is subjected to a physical process, the biotechnology, which outputs a more directly informative quantity, perhaps a physical object or other observable phenomenon. Finally, this output is quantified with some measurement apparatus.

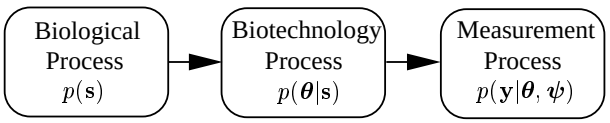

Fig. 1. Modelling Biotechnology experiments.

In the figure, $\mathbf{y}$ denotes the measurements from the system, $\boldsymbol{\psi}$ is a set of parameters used in the description of the measurement process, $\boldsymbol{\theta}$ similarly parameterises the biotechnology process, and $\mathbf{s}$ denotes the underlying biological object or system of interest. The $p(\cdot)$ quantities represent probability densities that describe the stochastic processes in question; for example, $p(\mathbf{s})$ could describe the biologically encoded algorithm used to generate a child's DNA from its parents'.

The aim of biotechnology analysis is to estimate $p(\mathbf{s}|\mathbf{y})$, the posterior distribution of the underlying biological state given the data. Using Bayes's rule, we have:

$$\begin{aligned} p(\mathbf{s}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{s})\, p(\mathbf{s}) \\ &= \left[\int p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\psi})\, p(\boldsymbol{\theta}|\mathbf{s})\, p(\boldsymbol{\psi})\, d\boldsymbol{\theta} d\boldsymbol{\psi}\right] p(\mathbf{s}) \end{aligned}$$

where, owing to the non-linearity of real systems, the integral giving $p(\mathbf{y}|\mathbf{s})$ is rarely available analytically.

The application of the Bayesian paradigm to biotechnology thus necessitates two major tasks:

- Development of suitable models for each of the three stages - biology, biotechnology, and measurement - that capture the essence of the physical processes.
- Numerical estimation of $p(\mathbf{s}|\mathbf{y})$ or a particular feature thereof.

In the following sections we develop these ideas in the context of the DNA sequencing problem, and provide some results to illustrate the potential advantages and disadvantages of the approach. The intention is to give a flavour of the underlying biology, without delving into more esoteric biological issues; a more detailed description can be found in [6, 9].

2. DNA SEQUENCING BACKGROUND

Much, if not all, of the information which determines our physical function is stored within our cells within the double helical chemical structure known as deoxyribonucleic acid (DNA). For our purposes DNA can be thought of as a string of symbols (in reality, chemical bases) taken from a four letter alphabet: A (Adenine), G (Guanine), C (Cytosine), or T (Thymine). In this paper we are concerned with the determination of that sequence.

In 1974, Sanger [8] proposed a method for DNA sequencing which, with technical improvements, has since been almost universally accepted. The idea behind the process is simple. Initially, via a process of replication and termination the DNA sequence of interest, henceforth the *template*, is used to form a large population of partial replicas. Each replica is identical to the template over a range of bases, always commencing with the first base

of the template, and terminating some random distance down the strand. That is, for the template 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, and the entire population is aligned at the start of a gel. An electric field is applied, causing the slightly charged fragments to progress through the gel at rates approximately inversely proportional to their length. The various *subpopulations* thus arrive at the end of the gel in sequence order. A laser positioned near the end of the gel excites the fluorescent labels, allowing a set of four emission detectors to estimate the number of fragments terminated by a each base passing at each time instant. Each emission detector is targeted at the maximum fluorescent emission frequency for one of the four dyes.

Four data sets are obtained (henceforth, *channels*), corresponding measurements of the fluorescence from the changing number of fragments passing the end of the gel. 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 fluorescence. An example data set is shown in figure 3.

When the peaks are approximately of constant amplitude, evenly spaced, well defined, and not obscured by noise, the data is trivial to interpret. Unfortunately, in reality the data is subject to several processes leading to serious degradation of data quality, particularly near the end of the sequence. For example, during a typical sequencing run an electropherogram exhibits variation in the processes which define peak shape, spacing, amplitude, and the nature of the noise.

The current state-of-the-art from a signal processing perspective, Phred, is described in [3], where a combination of peak detection algorithms is proposed. Excellent results are also shown in [2], while [6, 7] presents an algorithm based on statistical modelling of the underlying process that is useful in some scenarios.

3. PROCESS MODELS

We now develop core models for DNA sequencing data that include our understanding of the biology, the biotechnology process, and the observation process. A more detailed discussion of these and peripheral models is supplied in [6].

3.1. Biological Process

The biological process relevant to DNA sequencing is the creation of the DNA sample - usually the intermingling of parental DNA coupled with mutation. Occasionally we may have specific information, perhaps that the sample corresponds to a gene about which certain structural assumptions can be made. Specific structural information is, however, rare, and it is more usual to have non-structural base frequency information, e.g. how much richer is a region in G and C than A and T. Here, we make the commonly made assumption that the i^{th} base in the sequence is randomly selected with probability dependent only on the previous base, i.e. the system is Markovian. This assumption allows incorporation of more basic information, without making overly strong assumptions about DNA structure. Denoting $s_i \in \{A, G, C, T\}$ as the i^{th} base and $\mathbf{s}_{1:i-1} \triangleq \{s_1, \ldots, s_{i-1}\}^{\mathrm{T}}$ we write:

$$p\left(s_i | s_{1:i-1}\right) = p\left(s_i | s_{i-1}\right)$$

3.2. Biotechnology Process

3.2.1. DNA Fragments

Intuitively, the time for each fragment from a particular subpopulation to reach the end of the gel can be modelled as a draw from a probability distribution with mean approximately inversely proportional to the fragment length and variance dependent on the diffusion process induced by the gel. For a given fragment population, the observed distribution of times to traverse the gel can thus be considered as a histogram.

If, as is typical when DNA sequencing, the number of fragments is large, this histogram closely approximates a continuous function. This can be thought of as the product of a continuous unit area function which we will call the *peak shape*, $\phi_i\left(n\right)$, and a scalar quantity denoting the total number of fragments in that subpopulation, a_i. Making the frequently valid assumption of a Gaussian peak shape [1], and denoting p_i to be the mean time for a fragment from subpopulation i to reach the end of the gel and v_i to be its variance, we have:

$$\phi_i\left(n\right) = \frac{1}{\sqrt{2\pi v_i}} \exp\left\{-\frac{\left(n-p_i\right)^2}{2v_i}\right\}$$

The fluorescence emanating from each fragment is dependent on the interaction between the fragment sequence and the attached fluorescent dye. It is therefore reasonable to assume that each fragment from a given population fluoresces in the same way. Using $\boldsymbol{\omega}_i \triangleq \{\omega_{i,A}, \omega_{i,G}, \omega_{i,C}, \omega_{i,T}\}^{\mathrm{T}}$ to denote the fluorescence in each of the four channels associated with each fragment from subpopulation i, the total DNA fragment related fluorescence at time n, $\mathbf{x}_n \triangleq \{x_{n,A}, x_{n,G}, x_{n,C}, x_{n,T}\}^{\mathrm{T}}$ is given by:

$$\mathbf{x}_n = \sum_{i=1}^{N_B} a_i \phi_i\left(n\right) \boldsymbol{\omega}_i$$

where N_B denotes the number of bases in the template.

We now proceed to develop models for the physical processes parameterised above.

a_i : The quantity a_i is used to denote the total number of fragments of length i, and is dependent on the replication and fragment termination processes (see section 2). Here, there are several pertinent pieces of information: the termination process tends to produce more shorter fragments than long; the number of fragments in a given population is directly related to the sequence of that fragment, and particularly strongly to the terminating base and the two bases to either side; in the absence of sequence specific effects, fragments with the same terminating base and similar length will appear in similar number. Sequence dependent termination patterns are discussed in the context of the relevant biology papers in [9].

In reality, a_i is a discrete value. However, since the number of fragments is generally large, the information loss resulting from allowing a_i to be continuous is small, and the range of possible model choices increases. We propose:

$$\mathbf{a}_i \sim \mathcal{G}\left(\alpha_a, \gamma_{a,\mathbf{s}_i} \beta_{a,i}\right)$$

where the choice of a Gamma distribution ensures positivity, constant α_a ensures a constant ratio of expectation to standard deviation, $\beta_{a,i}$ is related to the mean in the absence of sequence dependent effects, $\mathbf{s}_i$ is termed the state and is defined as the terminating base and immediately surrounding bases, and $\gamma_{a,\mathbf{s}_i}$ models the effect of this local surrounding sequence. The model for $\beta_{a,i}$ includes the idea that the number of fragments ending in the same terminating base varies slowly with fragment length. See [6] for a block stationary model.

p_i : The quantity p_i is the mean time taken for fragments of length i to pass the end of the gel and is roughly inversely proportional to fragment mass. The time difference between the arrival of adjacent populations thus tends to decrease with increasing sequence length, although the stochastic nature of the fragments progression through the gel leads to substantial variation, here referred to as *peak jitter*. Peak spacing patterns are also strongly related to the underlying base sequence; local secondary structure can form that leads to a change in fragment mobility [1].

Intuitively, the time taken for a fragment population to reach the end of the gel is equal to the time of the previous population plus an amount reflecting the difference in velocity imparted by the additional base. This velocity differential is dependent on the local sequence in question, and also on the relative lengths of the adjacent fragments. We propose:

$$\begin{aligned} p_1 &\sim \mathcal{N}\left(\mu_{p_1}, v_{p_1}\right) \\ p_i - p_{i-1} &\sim \mathcal{G}\left(\alpha_p, \beta_{p,i}\gamma_{p,\mathbf{s}_i}\right), \quad i > 1 \end{aligned} \tag{1}$$

where α_p is representative of peak jitter and is assumed known, $\beta_{p,i}$ is related to the mean peak spacing in the absence of sequence dependent effects, and $\gamma_{p,\mathbf{s}_i}$ is a sequence dependent modifier. The model for $\beta_{p,i}$ incorporates the slowly varying nature of the process [6].

v_i : The peak width process evolves over time, with a tendency to slowly stochastically increase with time-dependent diffusion effects [3, 4]. To our knowledge, the evolution of width exhibits no noticeable dependence on sequence. The following is proposed:

$$v_i \sim \mathcal{G}\left(\alpha_v, \beta_{v,i}\right)$$

where α_v is assumed known a priori, and $\beta_{v,i}$ is related to the mean of the process. Similarly to $\beta_{a,i}$ and $\beta_{p,i}$, the evolution of $\beta_{v,i}$ is modelled as slowly varying [6].

$\boldsymbol{\omega}_i$: In the absence of interaction between the dye and local sequence, the emission spectra are often accurately known a priori. However, no chemistry is entirely devoid of sequence dependent effects, known or unknown; even fragments with identical local sequence can be expected to produce slightly different emission spectra. Here, the Normal distribution is proposed to incorporate this uncertainty:

$$\omega_{i,j} \sim \mathcal{N}\left(\mu_{\omega,j}\left(\mathbf{s}_i\right), v_{\omega,j}\left(\mathbf{s}_i\right)\right), \quad j \in \{A, G, C, T\}$$

where $\mu_{\omega,j}\left(\mathbf{s}_i\right)$ denotes the expected emission of a fragment with state $\mathbf{s}_i$ in the $j \in \{A, G, C, T\}$ channel . Similarly, $v_{\omega,j}\left(\mathbf{s}_i\right)$ denotes the uncertainty inherent in this expectation. Both are assumed constant and known. The choice of the Normal distribution, as opposed to a distribution ensuring positivity, has been made for mathematical convenience (see section 4).

3.2.2. Background Fluorescence

There are a number of possible sources of additional fluorescence; for example, unincorporated dye or the gel itself. We will call this fluorescence the *background* and denote it at time n by $\mathbf{b}_n \triangleq \{b_{n,A}, b_{n,G}, b_{n,C}, b_{n,T}\}^{\mathrm{T}}$. The precise form of the background is unknown - it depends on a host of conditions unique to the individual experiment in question - but it is usually slowly time-varying in each of the four channels. The specific choice of model is thus relatively arbitrary, other than that it should ensure smoothness and be sufficiently powerful to represent a wide variety of background behaviour. Here, we use cubic B-splines with evenly spaced knots:

$$\mathbf{b}_n = \sum_{j=1}^{N_S} \rho_j \upsilon_j\left(n\right)$$

where N_S denotes the number of splines used, $\upsilon_j\left(n\right)$ denotes the j^{th} cubic spline basis function, and ρ_j is its regression coefficient. A locally linear background can also be used in many scenarios [2, 6], but is inappropriate for some more difficult datasets.

The smoothness of the process can be mathematically incorporated via the model on the regression coefficients using the well known smoothness measure $\int \left|\mathbf{b}_n^{''}\right|^2 dn$. See [6] for more details. The need to separate signal from noise on the basis of smoothness frequently arises in the analysis of biotechnology data.

3.3. Measurement Process

The output of the detectors is given by the sum of the fluorescence and any detector noise. Denoting detector noise as $\mathbf{e}_n \triangleq \{e_{n,A}, e_{n,G}, e_{n,C}, e_{n,T}\}^{\mathrm{T}}$ and the observation process $\mathbf{y}_n$, we have:

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{b}_n + \mathbf{e}_n$$

The noise introduced by each detector, $e_{n,j}$, $j \in \{A, G, C, T\}$, can reasonably be approximated as zero mean Gaussian, independent, and stationary. A model for the variance of this noise can be made on the basis of knowledge of the apparatus in question, or using less informative considerations [6, 7].

4. BAYESIAN INFERENCE

Using Bayes's rule, the individual probabilistic models of the previous section can be combined to yield an expression for the joint posterior distribution:

$$p\left(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\psi}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\psi}\right) p\left(\boldsymbol{\theta}|\mathbf{s}\right) p\left(\mathbf{s}\right) p\left(\boldsymbol{\psi}\right)$$

where $\mathbf{s}$, $\boldsymbol{\theta}$, and $\boldsymbol{\psi}$ respectively denote the unknown parameters introduced in the Biological Process, Biotechnology Process, and Measurement Process sections (see figure 1). The object of our inference, $p\left(\mathbf{s}|\mathbf{y}\right)$, is not analytically available, and must be calculated by numerically integrating the nuisance parameters.

One of the advantages of the Bayesian paradigm is that we can consistently exploit partial linearity and Gaussianity in the system to reduce the space over which numerical integration must be performed. In this case, the posterior is linear and Gaussian in $\boldsymbol{\omega}_{1:N_B}$, leading to

$$p\left(\boldsymbol{\theta}_{-\boldsymbol{\omega}}, \mathbf{s}, \boldsymbol{\psi}|\mathbf{y}\right) = \int p\left(\boldsymbol{\theta}, \mathbf{s}, \boldsymbol{\psi}|\mathbf{y}\right) d\boldsymbol{\omega}_{1:N_B}$$

where, $\boldsymbol{\theta}_{-\boldsymbol{\omega}}$ is simply $\boldsymbol{\theta}$ with $\boldsymbol{\omega}_{1:N_B}$ removed. Inference is then made using the reduced distribution $p\left(\boldsymbol{\theta}_{-\boldsymbol{\omega}}, \mathbf{s}, \boldsymbol{\psi}|\mathbf{y}\right)$. No approximation is introduced.

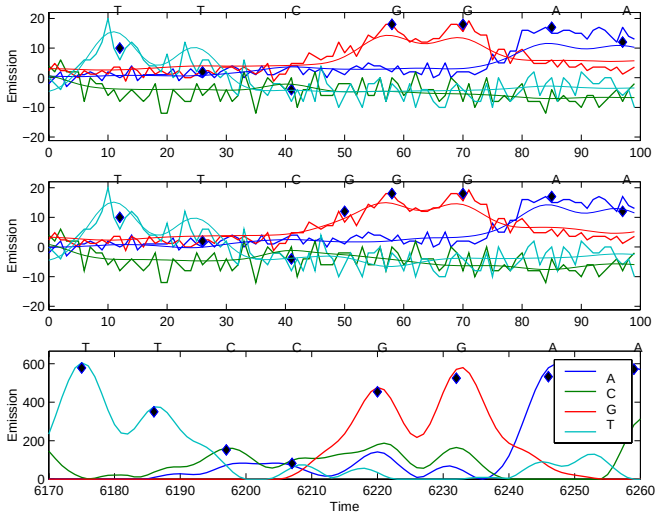

Fig. 2. Example of ambiguity. The top two plots show two predictions from our algorithm superimposed on the raw data, while the bottom plot shows Phred's prediction.

It is also interesting to notice that inherent in the estimation of $\mathbf{s}$ is the estimation of the number of bases present in the sequence. One of the core advantages of the Bayesian paradigm is that it allows this problem of model selection to be tackled within the same framework, without the introduction of less justifiable criteria. It also naturally provides model based error metrics - $p(\mathbf{s} = \mathbf{S}|\mathbf{y})$ gives the probability of $\mathbf{S}$ being the correct sequence interpretation.

The development of appropriate numerical algorithms is a difficult task. Markov Chain Monte Carlo (MCMC) techniques are ideally suited to such statistical problems. These techniques aim to simulate from a distribution of interest through construction of a Markov chain with stationary distribution equal to the required distribution. Quantities of interest can then be examined through histograms of the simulated population; a good review can be found in [5]. A similar algorithm to that used here can be found in [6].

5. RESULTS AND DISCUSSION

We now consider two datasets which demonstrate the ability of the algorithm to improve on the current state-of-the-art, Phred [3], in certain scenarios. The data in figure 2 is ambiguous in that it is difficult to discern the number of bases in the region. Two sequence interpretations are supported by our posterior $p(\mathbf{s}|\mathbf{y})$: TTCGGGAA with probability 60% (corresponding to the true sequence) and TTCGGAA with probability 40%. The preprocessed output of Phred is also shown, and it is seen that an incorrect call is made. The advantage of the model based approach here lies not only in its greater accuracy, but in the provision of a meaningful confidence measure.

Figure 3 illustrates another advantage of the model based approach. Since Phred uses a deterministic peak detection scheme and does not directly model the peak shape or emission spectrum process, it can be susceptible to base calling errors when peak cusps are not distinct or sequence dependent effects are present.

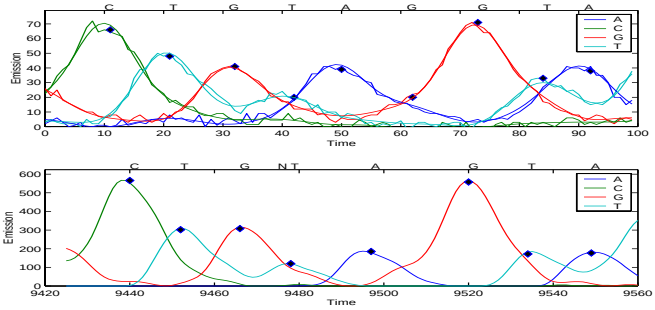

Fig. 3. Another example. The top plot shows our prediction, while the bottom plot shows Phred's. Our prediction corresponds to the true sequence.

We believe the model based approach brings many advantages to DNA sequencing and to biotechnology analysis in general. However, there are certain situations where a non-parametric approach, less sensitive to the precise form of the data, can be preferable. These arise where anomalies in the process lead to a break down of the model assumptions. These issues are discussed in more detail in [6]. Finally, the analysis of real models can be computationally intensive; suitable approximations must be made if fast processing is required.

6. REFERENCES

[1] J. Bowling et al. Neighbouring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Research*, 19(11):3089–3097, 1991.

[2] D. Brady, M. Kocic, A. Miller, and B. Karger. A maximum likelihood base caller for DNA sequencing. *IEEE Trans. Biomed. Eng.*, 47(9):1271–1280, 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] M. Giddings et al. An adaptive object oriented strategy for base calling in DNA sequence analysis. *Nucleic Acids Research*, 21(19):4530–4540, 1993.

[5] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Chapman and Hall, 1996.

[6] N. Haan. *Statistical Models and Algorithms for DNA Sequencing*. PhD thesis, University of Cambridge, Dept. of Engineering, 2001.

[7] 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.

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

[9] 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.