JOINT BASE-CALLING OF TWO DNA SEQUENCES WITH FACTOR GRAPHS

Xiaomeng Shi [†], *Desmond S. Lun* [‡], *Jim Meldrim* [‡], *Ralf Kötter* [§], *Muriel Médard* [†]

[†]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, USA; [‡]Broad Institute of MIT and Harvard, Cambridge, USA; [§]Institute for Communications Engineering, Technische Universität München, D-80290 München, Germany

ABSTRACT

To improve the utility of existing technologies based on Sanger sequencing, this paper examines the possibility of base-calling two superposed DNA sequences jointly. This approach allows a single electrophoresis experiment to process two sequences, using the same quantity of reagents and machine hours as for a single sequence. A practical heuristic is proposed to first estimate the peak parameters, then separate them into two sequences (major/minor) by passing messages on a factor graph. Base-calling on the major alone yields accuracy commensurate with single sequence approaches, and joint base-calling provides results for the minor which, while being of lesser quality, incurs no additional cost and can be ultimately used in the genome assembly process.

Index Terms— DNA sequencing, factor graphs

1. INTRODUCTION

The most widely used chemical experiment for collecting DNA sequencing data is the chain termination method developed by Frederick Sanger in 1977 [1]. In the Sanger experiment, a single strand of DNA is mass replicated starting from a fixed primer, but terminated at random locations by fluorescently-labeled markers. The resulting fragments are electrophoretically separated through gels or capillaries by length, which is inversely proportional to their traveling speed. Optical detection for base-specific dyes then gives rise to a set of time series data in the form of a four-component intensity vector, corresponding to the four base-types (Adenine, Cytosine, Guanine, Thymine). Assuming N points are sampled uniformly in time, a trace in the resulting chromatogram can be written as $y_n(t), 1 \leq n \leq 4, 1 \leq t \leq N$. The raw data is pre-processed to remove background noise, normalize mobility, and color-correct any correlation between components caused by overlapping dye emission spectra.

Base-calling is the process of identifying the order of DNA bases from pre-processed data, into a sequence of the four base types (A, C, G, T). Owing to random motion of the segments as they pass the detection region, the collected data are successive pulses corresponding to the spread of fragment concentrations around their nominal positions. Figure 1(a) shows a pre-processed trace at the beginning and towards the end, with different base types represented by different line styles. A typical run, which requires more than 30 minutes to complete, gives approximately 600 to 800 bases, corresponding to 7000 to 10,000 sample points. Given that genomes easily contain millions of bases and that repetitions are needed to achieve high accuracy in subsequent assembly, a large number of machine days is required to sequence a single genome. In addition, the fixed cost of the machine and variable cost of the reagents sum to thousands of dollars per machine day. To increase the throughput of the overall process while maintaining cost, we mix two DNA segments in electrophoresis, and aim to base call the superposed trace. Figure 1(b) gives a set of sample data, where the average amplitude ratio between the major and minor sequences is close to 2. Here we refer to the sequence with a larger average amplitude as the major, and the other as the minor.

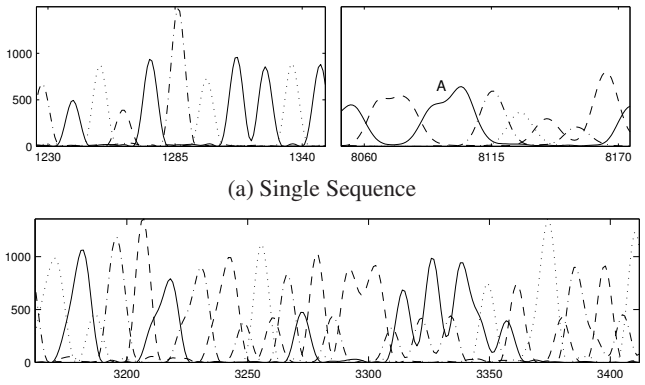

Fig. 1. Sample pre-processed DNA trace data.

To base call a single sequence, an automated sequencer needs to take into account at least three undesirable features of the data: amplitude variation, increasing pulse widths that deteriorate peak resolutions as in Figure 1(a), and jitter in peak spacings. Such timing jitter makes it difficult to apply a dynamic programming algorithm to resolve the intersymbol interference (ISI), because its inherent randomness makes data association with individual peaks no longer uniform, and thus hard to determine.

The most widely used algorithm for base-calling a single sequence is Phred [2], which combines a set of heuristics such

as the running average peak spacing, peak areas and concavity measures to determine the bases. Other approaches include parametric deconvolution [3]; combining Kalman prediction of peak locations with dynamic programming to find the maximum likelihood sequence [4]; and performing Markov Chain Monte Carlo methods with a complete statistical model to estimate the peak parameters [5]. A direct extension of these to sequencing two superposed traces is not trivial, for the major and minor traces are not synchronized in time, nor is separation into two sequences an easy task. The average amplitude ratio is imperfectly related to the reagent concentration. It can only be set to some range instead of a specific value.

In this paper, we first examine the joint base-calling problem with a complete statistical model represented graphically on a factor graph (FG). With this setup, the MAP estimation of individual bases is very computationally expensive. Instead, we propose a two-stage model. By viewing the data as similar to pulse amplitude modulated signals in a communication channel, we first try to find the spike train underlying the mixed sequence data using nonlinear minimum mean square estimation. Next we assign the spikes to the major and minor, to identify the two source sequences. We also present some preliminary results at the end.

2. JOINT BASE-CALLING OF TWO SEQUENCES

2.1. Maximum a Posteriori Base Estimation

Assuming there are N_1 and N_2 peaks in the major and minor sequences respectively, with amplitudes α_{1i}, α_{2j}, peak positions τ_{1i}, τ_{2j}, $1 \le i \le N_1, 1 \le j \le N_2$, and a generic pulse shape $p(t)$, we can write the sampled time series as:

$$\mathbf{y}(t) = \sum_{i=1}^{N_1} \alpha_{1i} p(t-\tau_{1i}) \mathbf{x}_{1i} + \sum_{j=1}^{N_2} \alpha_{2j} p(t-\tau_{2j}) \mathbf{x}_{2j} + \mathbf{e}(t)\,, \tag{1}$$

where $\mathbf{x}_{1i}^{\mathbf{T}}, \mathbf{x}_{2j}^{\mathbf{T}}$ takes on one of the four codewords $\{0001, 0100, 0010, 0001\}$, corresponding to four base types, and $\mathbf{e}$ is the additive noise. Joint base-calling is the process of estimating the parameters $\mathbf{x}_1$ and $\mathbf{x}_2$. Experimental data shows that for each sequence, the peak amplitudes are approximately i.i.d. with a Gamma distribution; the peak timing locations are first-order Markov, i.e., $f(\tau_{l,i+1}|\tau_{l,i})$ satisfies

$$f(\tau_{l,i+1}|\tau_{l,i}) = f_{\Delta\tau}(\tau_{l,i+1} - \tau_{l,i}) \qquad l \in \{1,2\}\,,$$

where $f_{\Delta\tau}$ has its mean equal to the slowly varying average peak spacing, and standard deviation is less than two samples. Also for simplicity, assume the additive noise is white Gaussian, with zero mean and standard deviation σ_e. With this setup, the dependencies between the peak parameters can be represented by an FG, as shown in Figure 2. Circles in the FG represent random variables, while squares represent probability distributions. $\mathbf{y}_k$ denotes all data points associated with the k-th peak. This dependency structure, together with conditionals obtained from training data, allows the Sum-Product

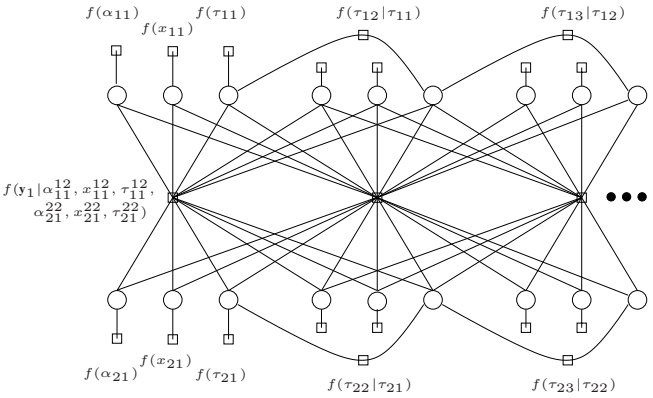

Fig. 2. Factor Graph for MAP estimation of individual bases.

Algorithm [6] to be applied for MAP estimation of individual bases. One simplification we have made in composing this graph is to assume that near uniform alignment between the major and minor exists; hence interference is caused only by adjacent peaks in both sequences. In reality, this assumption is not true, nor is alignment information known a priori. A consequence is that there will be many more edges in the graph, but only a few will carry significant information. The strength of the links can only be determined after at least one iteration of the algorithm. Equivalently, we could view the need for more edges as a difficulty of data association. Clearly this approach for joint base-calling is computationally impractical, albeit being theoretically optimal.

2.2. Two-Stage Base Calling Formulation

Since the MAP base estimation on an FG is very computationally expensive due to random peak timing jitters and difficulties with data association, we develop a two-stage algorithm, where timing recovery and source sequence identification are separately carried out to give a suboptimal solution. Consider Eq.(1). $\mathbf{y}$ can be viewed as the convolution product of the generic pulse $p(t)$ and two superposed spike trains

$$\mathbf{z}(t) = \sum_{i=1}^{N_1} \alpha_{1i} \delta(t-\tau_{1i}) x_{1i} + \sum_{j=1}^{N_2} \alpha_{2j} \delta(t-\tau_{2j}) x_{2j}\,.$$

Using an indicator variable m, $\mathbf{z}$ can be rewritten as

$$\mathbf{z}(t) = \sum_{k=1}^{N} \alpha_k x_k \delta(t-\tau_k) \left[\delta(m_{k,1}) + \delta(m_{k,2}) - \delta(m_{k,1}, m_{k,2})\right]$$

where $N > N_1 + N_2$, and $m_k \in \{(00), (10), (01), (11)\}$, representing whether a spike at time τ_k has originated from noise, the major, the minor, or both. Note that we have dropped the indices distinguishing the two components. If we could estimate values of α, x, τ, m for each base with high accuracy, the two constituent sequences could be identified.

Recovery of peak amplitude, timing, and base type information requires deconvolution of the sample data $\mathbf{y}$. To

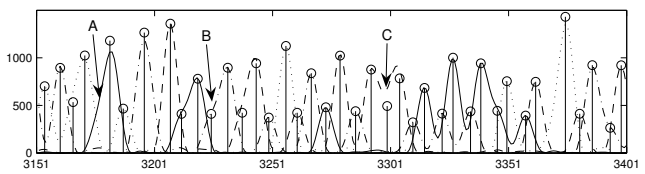

Fig. 3. Sample pre-processed DNA trace data.

reduce the estimation complexity, account for the slowly increasing pulse widths, and reduce edge effects, data is first divided into windows of size 500 samples, with adjacent blocks overlapping by 250 samples. From the earlier part of the trace, a few isolated, well defined pulses are chosen base on curvature, cumulative area, and relative amplitudes, then normalized to be the generic pulse shape $\hat{p}(t)$. For the l-th window, similar computations can be performed to find some well resolved pulses. However, since the ISI effect becomes more severe towards the end of the trace, as illustrated by the overlapping pulses at position A in Figure 1(a), pulse shapes estimated this way are in general far from representative. Instead, the average full width at half maximum (FWHM) are computed, and $\hat{p}(t)$ is scaled accordingly to obtain $\hat{p}_l(t)$.

To deconvolve, first assume the additive noise **e** is white Gaussian. The following mean square minimization can be performed on each of the four base types, yielding the most likely underlying sequence:

$$\underline{\hat{\alpha}}^l, \underline{\hat{\tau}}^l = \arg\min \sum_t \left(y(t) - \sum_{k=1}^{N_l} \alpha_k \hat{p}(t-\tau_k) \right)^2 . \quad (2)$$

Here $\underline{\hat{\alpha}}^l, \underline{\hat{\tau}}^l$ are the set of amplitudes, and locations for each of four base types in the l-th window. Observation of the data shows that the assumption of white noise is not accurate; the baseline noise is colored to have a similar shape as data in the time domain. Nonetheless, the above computation is relatively simple and gives reasonable results. Allowing the values of both $\underline{\hat{\alpha}}^l$ and $\underline{\hat{\tau}}^l$ to be continuous, Figure 3 shows the deconvolved spike train for the data in Figure 1(b). Notice that the hidden pulses at around positions B and C are captured, while the minor peak to the left of position A has been missed. This may or may not contribute to a deletion error in the final base-call, depending on if the spike is counted as a single major peak or overlapped major and minor peaks.

One technical detail worths mentioning is that because there is minimal prior information on N_l, we overestimate its value when minimizing (2). Overfitting will always occur, but the added spikes are either those that overlap the correct results or noise spikes that are very low in amplitude. For the former, we consolidate by combining overlapping spikes that are a distance of less than 1 time unit away. For the latter case, thresholding with the running average of peak amplitudes reduces the problem significantly.

For the second stage, we want to separate the deconvolved peaks into major and minor sequences. For simplicity, assume that there is no prior information about the classification of

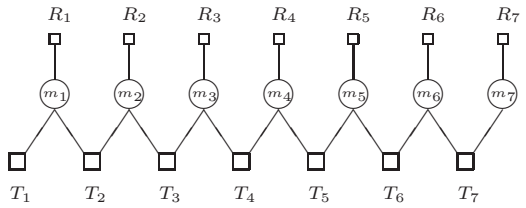

Fig. 4. First order factor graph for separating two sequences.

the two sequences, except that their average amplitudes differ by a multiple between 1.5 and 2.5. Let $\underline{\alpha}, \underline{\tau}, \underline{x}$ be estimated peak parameters for the overall data set, obtained by taking the union of the windowed parameters from stage one. The goal is to find

$$\underline{\hat{m}} = arg \max_{\underline{m}} \log p(\underline{m}|\underline{\alpha}, \underline{\tau}, \underline{x}) .$$

Assuming the distribution of peak amplitudes is independent of the base type and peak location, and the superposition of the two spike trains is dependent on only the peak locations, i.e., $p(\underline{\alpha}|\underline{m}, \underline{\tau}, \underline{x}) = p(\underline{\alpha}|\underline{m})$, $p(\underline{m}|\underline{\tau}, \underline{x}) = p(\underline{m}|\underline{\tau})$, we have

$$\underline{\hat{m}} = arg \max_{\underline{m}} \left\{ \sum_k \log p(\alpha_k|m_k) + \log p(\underline{m}|\underline{\tau}) \right\} .$$

To find the dependence of the indicator variable $\underline{m}$ on $\underline{\tau}$, first we observe that although the major and minor sequences are not synchronized, some uniformity of peak spacing is still maintained within each. A simple model here is to assume $\underline{m}$ is first-order Markov, with transition probabilities determined by the timing difference, i.e., given m_k, we assume the timing difference $\Delta\tau = \tau_{k+1} - \tau_k$ determines the distribution of m_{k+1}:

$$\log p(\underline{m}|\underline{\tau}) = \log p(m_1) + \sum_{k=2}^{N} \log p(m_k|m_{k-1}, \Delta\tau_k) .$$

Let

$$R_k = p(\alpha_k|m_k), T_k = \begin{cases} p(m_k) & k = 1 \\ p(m_k|m_{k-1}, \Delta\tau_k) & k > 1 \end{cases} .$$

The dependency of $\underline{m}$ on the peak parameters can be represented graphically using an FG as in Figure 4. This is the trellis graph of a Markov chain, where the Sum-Product Algorithm [6] can be applied to find the maximum likelihood estimate of $\underline{m}$.

3. RESULTS AND DISCUSSION

We applied the algorithm stated in Section 2.2 to two sample data sets. The first one was shown in Figure 1(b). There were a total of 7000 sampling points, corresponding to about 580 bases starting from sample index 1401 to 8400. The front section was removed because the chemical process begins sporadically but settles after a short time period. Data collected after sample 8400 was not used because ISI compounded with the larger pulse widths make it more difficult to distinguish

the two sequences. The second data set also contained 7000 sample points, although the starting position was adjusted according to the quality of the trace.

For R_k, we assume the conditional distribution of α_k was Gaussian, parameterized by m_k. On the other hand, we approximated $p(m_k|m_{k-1}, \Delta\tau_k)$ by manually examining the first data set. To evaluate the joint base-calling error rate, the sequencing result was compared with reference sets using the *cross_match* program [2]. The dynamic programming based Smith-Waterman algorithm was employed to find the longest lengths of consecutive bases which gave the best local alignment. Results are listed in Table 1. Also given in this table is the performance of a single sequence base caller on the same data sets. This base caller was constructed similar to a phase lock loop, where the envelop of the trace was used to track the peak locations one at a time in the forward direction.

One observation is that mixing two sequences in electrophoresis has little effect on the single sequence base-calling accuracy. In other words, we could use existing techniques for calling the major, while employing the joint base caller for the minor. This corresponds to the shaded cells in Table 1. The first data set was called with higher accuracy, for some over-fitting has occurred when estimating $p(m_k|m_{k-1}, \Delta\tau_k)$. A throughput gain is achieved with the additional minor sequence, which has come at no extra cost. Although of lesser accuracy, this information might be used in the subsequent process to assemble DNA segments to their original order. In particular, to reach the high accuracy needed for genome study, each DNA segment is sequenced multiple times (e.g., 8x) before assembly. The number of repetitions is known as the "depth of coverage." One possibility is to replace some repetitions with the minor base calls. Furthermore, to close gaps between segments in more complex genomes with repeated genes, the major and minor can be set to a known distance away; alignment of the major may help the utilization of the minor in the assembly process.

Although performance of the joint base-calling algorithm is not comparable to that of single sequence base callers, it does have the potential to do better. First, single sequencing results on the major may be used as prior information for initializating the FG in Figure 4. Table 1 shows that performance of joint sequencing on the major does not match that of a single sequencer. Some of the errors affect the accuracy of joint sequencing on the minor: the higher percentage of deletion in the minor is reflected in the higher percentage of insertion in the major. Second, examination of the sequencing results shows that many errors occur near spikes labeled as having originated from neither of the source sequences. These errors can be explained as local disturbances caused by erroneous messages passed through such nodes. Compensation is possible if we increase the complexity of the FG in Figure 4 by linking nodes that are two steps apart. Moreover, deletion errors caused by the deconvolution process may be reduced by iterating between the deconvolution and source sequence separation stages, with missed peaks inserted heuristically based on spacing uniformity.

Table 1. Performance of joint (J) and single (S) sequence base-calling.

	length of best single match	% substitution	% deletion	% insertion
Major(J)	430	2.87	5.50	2.39
Minor(J)	386	2.49	7.73	0.83
Major(S)	582	0.17	0.34	0.00
Major(J)	231	4.31	9.05	1.29
Minor(J)	131	6.01	10.61	0.76
Major(S)	578	2.25	1.21	0.00

4. CONCLUSION

In this paper, we explored the possibility of base-calling two superposed sequences jointly by deconvolution and source sequence identification using factor graphs. Combined with single sequence base-calling, this algorithm enables the sequencing of an additional segment. Although not at the same accuracy as single sequence base-callers, the results are promising. Several venues for further exploration emerges: matching the quality of the major joint calls to that of the major single calls should lead to improvement in that of the minor joint calls. Also, more complex factor graphs, and additional iterations between deconvolution and source sequence identification may lead to improved performance.

5. REFERENCES

[1] F. Sanger, S. Nicklen, and A. Coulson, "DNA sequencing with chain-terminating inhibitors," in *Proc. Natl. Acad. Sci*, 1977, pp. 5463–5467.

[2] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green, "Base-calling of automated sequencer traces using P*hred*. I. accuracy assessment," *Genome Research*, vol. 8, no. 3, pp. 175–185, 1998.

[3] Lei Li and Terence P. Speed, "Parametric deconvolution of positive spike trains," *Ann. Statist.*, , no. 5, pp. 1279–1301, 2000.

[4] Stephen W. Davies, Moshe Eizenman, and Subbarayan Pasupathy, "Optimal structure for automatic processing of DNA sequences," *IEEE Trans. on Biomedical Engineering*, , no. 9, pp. 1044–1056, 1999.

[5] Nicholas M. Haan and Simon J. Godsill, "Modelling electropherogram data for DNA sequencing using variable dimension MCMC," in *2000* IEEE *Intl. Conf. on Acous., Speech, Signal Processing*, 2000, pp. 3542–3545.

[6] F. R. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 498–519, 2001.