Joint base-calling of Two DNA Sequences with Factor Graphs

Citation: Lun, D.S. et al., with Xiaomeng Shi. "Joint Base-Calling of Two DNA Sequences With Factor Graphs." Information Theory, IEEE Transactions On 56.2 (2010) : 724-733. Copyright © 2010, IEEE

As Published: http://dx.doi.org/10.1109/TIT.2009.2037029

Publisher: Institute of Electrical and Electronics Engineers

Persistent URL: http://hdl.handle.net/1721.1/62009

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Joint Base-Calling of Two DNA Sequences With Factor Graphs

Xiaomeng Shi, *Student Member, IEEE*, Desmond S. Lun, *Member, IEEE*, Muriel Médard, *Fellow, IEEE*, Ralf Kötter, *Fellow, IEEE*, James C. Meldrim, and Andrew J. Barry

***Abstract*—Automated estimation of DNA base-sequences is an important step in genomics and in many other emerging fields in biological and medical sciences. Current automated sequencers process single strands only. To improve the utility of existing technologies, we propose to mix two independent strands prior to electrophoresis, and base-call jointly by applying the sum-product algorithm on factor graphs. We first present a statistical model for DNA sequencing data and examine the model parameters. A practical heuristic is then proposed to estimate the peaks, which are then separated into two source sequences (Major/Minor) by passing messages on a factor graph. Simulation results show that joint base-calling can provide less accurate but valid results for the minor. The algorithm presented provides a basis for future investigation of joint sequencing techniques.**

***Index Terms*—Base-calling, DNA modeling, DNA sequencing, factor graphs, sum-product algorithm.**

I. Introduction

THE chain termination method developed by Sanger *et al.* in 1977 [1] for collecting DNA sequencing data was the most widely used sequencing technology until recently. One step of the sequencing process is base-calling, where nucleotide-order of a short DNA fragment is determined from data collected through chemical experiments. We propose to jointly base-call two superposed data traces by applying the sum-product algorithm on factor graphs, a method commonly used for iterative decoding in digital communication systems. Our goal is to improve the utility of existing Sanger technologies by mixing two DNA strands together during chemical processing, thus reducing the overall use of reagents and improving the use of sequencing equipment.

As we are addressing a problem in molecular biology with techniques from communication and information theory, in Section I-A we will describe in detail the process of Sanger sequencing, associated terminologies, and related work on base-calling. Readers who are not familiar with these terms may find it helpful to read this part first to gain some understanding of the sequencing problem.

To see why joint base-calling would be beneficial in more quantitative terms, consider the typical run time for shot-gun sequencing on an automated Sanger sequencer from Applied Biosystems (ABI, [2]). Each run often takes more than 30 minutes to complete, and identifies approximately 400 to 600 bases [3]. High throughput sequencing platforms can often read up to 1000 base pairs per run. However, 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. Mixing two DNA segments before electrophoresis and base-calling the superposed traces therefore offer the option of increasing the throughput of the overall process while maintaining cost.

In recent years, the conventional Sanger sequencing method has been rapidly supplanted by next-generation systems. Nonetheless, improvements in Sanger techniques are still of interest, for it is unlikely that Sanger sequencers can be entirely replaced by next-generation systems. The newly developed and commercialized next-generation parallel cyclic-array technologies include pyrosequencing, sequencing with reversible terminators, and sequencing by ligation [4]–[6]. By eliminating chain termination and DNA amplification through bacterial cloning, these techniques can achieve dramatically larger throughputs. In addition, sequence immobilization on arrays reduces the overall usage of reagents, giving rise to further cost reductions. Nonetheless, Sanger technology remains the lead in terms of read-length and accuracy [5], [7]. Typical read lengths for next-generation platforms are 10–40 base pairs, while Sanger sequencing can achieve lengths that are one-order of magnitude longer. The crucial importance of read length can be seen in the sequencing of highly repetitive genomes without known reference sequences [8]. If a repetitive region is longer than the read length, it is almost impossible to assemble accurately since overlaps are critical for bridging gaps among sequenced fragments. Sanger is therefore sometimes the only feasible sequencing method. On the one hand, the competitive environment for commercial sequencer development has accelerated research for data analysis methods and stimulated new applications in genome studies for next-generation technologies; on the other hand, these still have large potential for improvement. Another reason why the study of Sanger sequencing is still of interest is that currently there is a large

Manuscript received May 12, 2009. Current version published February 24, 2010. This material is based upon work supported by the National Science Foundation under Grant CCR-0325496.

X. Shi and M. Médard are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: xshi@mit.edu; medard@mit.edu).

D. S. Lun is with the Phenomics and Bioinformatics Research Centre, the School of Mathematics and Statistics, and the Australian Centre for Plant Functional Genomics, University of South Australia, Mawson Lakes, SA Australia (e-mail: Desmond.Lun@unisa.edu.au).

R. Kötter (deceased) was with the Institute for Communications Engineering, Technical University of Munich, Munich, Germany.

J. C. Meldrim and A. J. Barry are with the DNA Sequencing Operations Group at the Broad Institute of MIT and Harvard, Cambridge, MA 02139 USA (e-mail: meldrim@broadinstitute.org; barry@broadinstitute.org).

Communicated by O. Milenkovic, Associate Guest Editor for the Special Issue on Molecular Biology and Neuroscience.

Digital Object Identifier 10.1109/TIT.2009.2037029

installed base of Sanger sequencers [9]. Given the large costs associated with transition to new technologies and new infrastructures, Sanger sequencers will remain in use in combination with next-generation technologies in the immediate future.

In the remainder of this section, we describe the general principles underlying the Sanger sequencing process, and examine some previous work on base-calling. We also discuss characteristics of the sequencing data that need to be taken into consideration during joint base-calling of two mixed sequences. In Section II, a statistical data model is given and the model parameters discussed in detail. In Section III, we first examine the joint base-calling problem with the complete statistical model represented graphically on a factor graph (FG). With this setup, the maximum *a posteriori* probability (MAP) estimation of the types 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 two source sequences, which are termed *major* and *minor*, respectively, depending on their relative average amplitudes. Results are presented in Section IV, in which we also discuss other issues that need to be considered to improve the general performance of this algorithm. Section V concludes this paper with discussions on possible future work. Data used for analysis are provided courtesy of James Meldrim from the DNA Sequencing Operations Group at the Broad Institute of Massachusetts Institute of Technology and Harvard University. The commercial sequencers employed are from Applied Biosystems, Inc. (ABI)

A. Background on Sanger Sequencing

This section is tutorial in nature, with the aim to introduce some necessary background and terminologies to readers with no direct experience in genome sequencing, and to explain why sequencing is not an easy task. Since this paper is addressed to both the biological and the communications community, the descriptions here are quite extended. What might seem basic to one community may be completely unknown to the other, and readers who are familiar with the sequencing process can safely skip to Section I-B in which we briefly discuss past work on base-calling, and introduce the notion of base-calling two sequences simultaneously.

In this paper, we focus on the technology typical to ABI sequencers: paired-end whole-genome shotgun sequencing based on cycle sequencing and dye-termination capillary electrophoresis. The term *whole-genome shotgun* refers to the random division of an entire genome before fragments are sequenced individually; cycle sequencing is a signal amplification process similar to the better known PCR; the dye-termination method was developed by Frederick Sanger in 1977 [1] and has been the fundamental basis for DNA sequencing ever since. In this section, we limit ourselves to a description of only the general principles underlying signal amplification and dye-termination based electrophoresis. The intention is to establish terminology and notations. For more details on the working chemistry and instrumentation of the ABI sequencers, see the ABI Automated DNA Sequencing Chemistry Guide [3] or other application manuals available on the ABI website [2]. For further descriptions of the Sanger sequencing chemistry, see [10]–[12]. For other sequencing technologies, [13] and [14] give comprehensive reviews of some recent developments, while [15]–[18], and [19], refer to several emerging techniques including pyrosequencing and polony sequencing.

A DNA molecule has the structure of a double helix, where each strand is a chain composed of four types of monomers, identified by their constituting nitrogenous bases. The four base-types are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Genetic information is borne by the ordering of these bases. With an ultimate objective of determining the order of these bases, Sanger sequencing refers to a procedure encompassing five stages: DNA amplification, size separation and fluorescent detection of molecules through electrophoresis, data preprocessing to remove noise and condition the signal, determining base-orders through base-calling, and reassembly of the base-called segments.

1) DNA Amplification: before a piece of DNA can be studied, it first needs to be amplified (i.e., replicated) to reach a significant concentration. One widely used technique for this purpose in molecular biology is polymerase chain reaction (PCR). During replication, free nucleotides in a buffer solution group and extend according to the order of bases in a template sequence. Elongation of a segment terminates when a special monomer, a dideoxynucleotide, is incorporated. The mixture resulting from such amplification processes contains partially extended DNA fragments of different lengths. The terminating dideoxynucleotides are also coupled with fluorescent dyes, one color per base type. These serve as labels during the detection phase. Once amplified, the partially extended DNA strands can be separated by length and each ending base identified fluorescently; thus giving the order of bases along the segment of interest. This is the objective of electrophoresis.

2) Dye-Termination Electrophoresis: in electrophoresis, electrically charged molecules travel across a medium under the influence of an electric field. DNA molecules have a net negative charge. DNA fragments travel at an average velocity inversely proportional to their charges, equivalent to the number of bases they contain. Optical detection is performed at the end of the medium by exciting the fluorescent dyes on the terminator with a laser. A detector periodically records the fluorescence intensity at the end of the medium to give an electropherogram, displaying the ending bases as peaks with different retention times, where the amplitude of the peaks corresponds to fluorescence intensity. The randomness involved in the replication mechanism results in concentration variations among the molecules, leading to random amplitudes for the observed peaks. The raw data stream is in the form of a four-component intensity vector. Assuming K points are sampled uniformly in time, a trace in the resulting chromatogram can be written as $\underline{y}[t], 1 \leq t \leq K$, where t is an integer and for any τ, $\underline{y}[\tau]$ is a length 4 vector.

3) Data Preprocessing: Due to random motion of the segments as they pass the detection region, the collected data are successive pulses corresponding to the spread of DNA fragment concentrations around their nominal positions. Ideally the

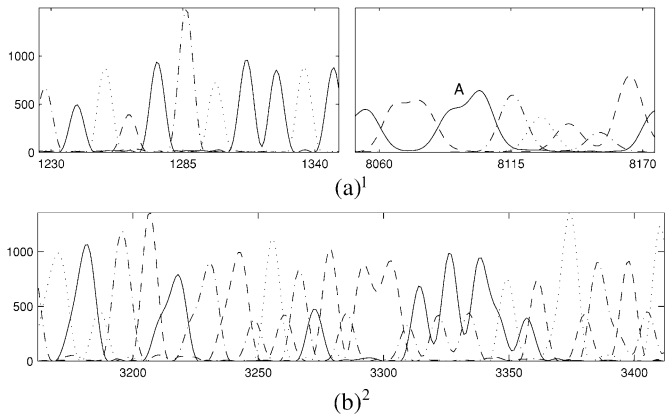

Fig. 1. Sample preprocessed DNA trace data. The four base-types (A, C, G, T) are represented by different line styles. A typical run gives a trace containing approximately 600 to 800 bases, corresponding to 7000 to 10 000 sample points. Peaks are much less resolved towards the end of a trace. When two sequences are mixed prior to electrophoresis, the resulting trace is a superposition of the corresponding individual traces. (a) Single Sequence[1], (b) Two Sequences[2].

acquired pulses would be uniformly spaced, with equal amplitudes, and resolved well enough such that the base types can be identified directly by sampling at the peak locations. Actual experimental data, however, contains not only measurement noise, but distortions inherent from the chemical reaction kinetics [20]. Additional preprocessing of the raw data is thus required before base-calling. There are four functions that are commonly carried out. First, baseline correction removes the slowly varying background fluorescent intensity level. This is achieved through low-frequency noise filtering operations. These noises are slowly drifting, and different for each base type. Second, overlaps among the dye emission spectra leads to correlation among the components, where the presence of a peak in one component may result in peaks in the three others. This crosstalk effect can be eliminated using a linear filtering process called color correction. Third, measurement noise and possible impurities in the reagents lead to additional random noise in the signal. Denoising through a low-pass filter is therefore needed. The last but the most important preprocessing step is electrophoretic mobility correction. The presence of the fluorescent labels affects the mobility of DNA molecules under electrophoresis. This causes compression of the four discrete data series at different rates. Mobility correction aims to scale along the time axis, such that the peak spacings are approximately uniform, and only one of the four concentration levels is dominant around each peak.

4) Base-Calling: Base-calling is the process of identifying the order of DNA bases from preprocessed data, into a sequence of the four base types (A, C, G, T). As stated previously during the discussion of elephophoresis, 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. Fig. 1(a) shows a preprocessed 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 min to complete, gives approximately 600 to 800 bases, corresponding to 7000 to 10 000 sample points. It is intuitive that the well-resolved peaks shown in the first subplot can be called easily by tracking positions of local maximums. However, peaks are much less resolved towards the end, as illustrated by the position marked by A in the second subplot, and such irregularities in data make base-calling difficult. Additionally, even with mobility correction, the stochastic nature of the experiment makes the peaks jitter from uniformly spaced locations. Such timing jitter makes it difficult to apply a dynamic programming algorithm to resolve the interferences among neighboring peaks, because the inherent randomness makes data association with individual peaks no longer uniform, and thus hard to determine.

5) Paired-End Whole-Genome Shotgun Assembly: Since the number of DNA bases that can be read in a single run does not exceed thousands, but genomes are easily millions or billions of bases long, automated mapping strategies are required to cover an entire genome. Shotgun sequencing shears DNA at random locations, performs reads, then assembles fragments on the basis of overlaps. There are two dominant shotgun sequencing techniques: hierarchical, and whole-genome shotgun (WGS) sequencing. In the former, intermediate-sized pieces called contigs are chemically mapped to the original genome, before each is sequenced with the shotgun approach. In WGS sequencing, an entire genome is read at random locations before assembly. Automation is therefore easier to implement, although more complex computations are required, and large-scaled misassembly errors are prone to happen. The choice between these two techniques can often be decided based on the amount of repetitions and complexity of the genomes under study. Furthermore, because overlaps determine the matching of different pieces, but perfect base-calling results are not guaranteed, the same location often need to be sequenced multiple times, where the average number of reads covering a base in the reconstructed sequence is referred to as the depth of coverage. A full coverage often corresponds to 8- or even 12-fold repetition.

B. Related Work on Base-Calling and the Joint Base-Calling Problem

In short, 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 Fig. 1(a), and jitter in peak spacings. The same issues persist when two independent sequences are mixed together prior to electrophoresis. In this case, the resulting trace after preprocessing is a superposition of the corresponding individual traces. Fig. 1(b) gives a set of sample data. Our aim is to base-call both sequences from such superposed trace. Observe that the different average amplitudes and the relative peak spacings are features that can be used to separate mixed sequences. In this example, the average amplitude ratio between the major and minor is close to 2. Here we refer to the sequence with a larger average amplitude as the *major*, and the other as the *minor*. The average amplitude ratio can be controlled by varying the relative reagent concentrations during

[1]Trace file name: $000000000001\_A01\_01.ab1$.

[2]Trace file name: $000016240928\_I07\_024.ab1$.

replication. However, this correlation is imperfect and the average amplitude ratio can only be set to some range instead of a specific value. Another added concern is that the peak locations of the minor sequence do not offset constantly from that of the major, although some regularities in peak spacing can still be observed.

The most widely used algorithm for base-calling a single sequence is Phred [21], 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 [22]; combining Kalman prediction of peak locations with dynamic programming to find the maximum likelihood sequence [23]; and performing Markov Chain Monte Carlo methods with a complete statistical model to estimate the peak parameters [24]. 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 idea of sequencing two or more strands together is not entirely new. However, the emphasis has mostly been on the modification of the underlying chemical processes. References [25] and [26] use two fluorescently labeled primers and requires specialized detection for each. Reference [27] proposes to sequence a strand from both the forward and reverse directions simultaneously in a single reaction. Immobilization of the forward sequencing products allows their separation, through chemical means, from the reverse sequencing products, each of which can then be base-called individually. Reference [28] states two methods for the simultaneous sequencing of multiple genes. The first uses different primers in a single reaction to signal several genes in a serial fashion such that data sets obtained can be analyzed one after another. The second method alternates between amplification and cycle sequencing to obtain both forward and reverse sequencing data. Both methods require additional primer design and can acquire short segments only. Although these techniques all aim to sequence multiple strands simultaneously, the emphasis is on the modification of the underlying chemistry to yield data that can be identified with existing base-callers. In this paper, we propose to collect sequencing data with minimal alteration of the chemical experiments, but base-call superposed data which is often viewed as "contaminated" when they occur in the laboratory.

II. Data Model and Problem Formulation

The output of the preprocessing stage is a set four intensity vectors, each in the form of a pulse train corresponding to a different base type, and spanning over uniformly spaced sampling times. Assuming K points are sampled in time, the sequencing data to be base-called can be written as $\underline{y}[t], 1 \leq t \leq K$, where t is an integer, and for any $\tau, \underline{y}[\tau]$ is a length 4 vector. As discussed in Section I-B, when two sequences are mixed prior to electrophoresis, we define the sequence with a higher average amplitude to be the *major* the other one with a lower average amplitude to be the *minor*. Assuming there are N_1 and N_2 peaks in the major and minor sequences respectively, we can denote the peak amplitudes with α_{1i}, α_{2j} and positions with τ_{1i}, τ_{2j}, where $1 \leq i \leq N_1, 1 \leq j \leq N_2$. The time-varying generic pulse shapes can be represented by $p_i(t), p_j(t)$. Under the assumption of simple superposition, the sampled time series is

$$\underline{y}[t] = \sum_{i=1}^{N_1} \alpha_{1i} p_i(t - \tau_{1i}) \underline{x}_{1i} + \sum_{j=1}^{N_2} \alpha_{2j} p_j(t - \tau_{2j}) \underline{x}_{2j} + \underline{e}(t). \tag{1}$$

Here $\underline{x}_{1i}^{\mathbf{T}}, \underline{x}_{2j}^{\mathbf{T}}$ takes on one of the four codewords $\{0001, 0010, 0100, 1000\}$, corresponding to four base types, and $\underline{e}(t)$ is an additive noise. In writing the above expression, we have assumed that mixing two sequences in one reaction does not alter the underlying chemistry, hence the resulting data is a trivial linear combination of individual traces. Examination of sample data sets such as the one shown in Fig. 1(b) shows that this assumption is empirically valid. Joint base-calling is the process of estimating the parameters $\underline{x}_1$ and $\underline{x}_2$.

To understand qualitatively the time series model in (1), peaks in the data set displayed in Fig. 1 are located manually and the distribution of each model parameter plotted. The experimental data shows that for each sequence, the peak amplitudes are approximately independent and identically distributed (i.i.d.) with a Gamma distribution, where the right tail is larger. Since amplitude distributions of the major and minor sequences overlap, simple thresholding is not sufficient to distinguish these two, although a larger difference in average amplitudes certainly leads to better differentiability. One possible experimental design is to increase the ratio between average amplitudes. It turns out that owing to the nature of the DNA replication process employed in Sanger sequencing, when the amount of reagent used for the major sequence is increased, the corresponding amplitude soon reaches a saturation value. On the other hand, while decreasing the amount of reagent used for the minor, background noise and other non-uniformities from the preprocessing stage soon become significant. In other words, amplitude resolution between the major and minor is limited by the underlying chemical experiment. In this paper, the data sets under consideration have their amplitude ratios set to approximately 2. To put this value into context, in the data set shown in Fig. 1(b), the average amplitude for the major sequence is approximately 750, with a spread of at least 200.

For the base type variable, we do not take into account any gene structures that give rise to possible correlations between neighboring bases or across the two sequences. Also depending on the organism or the gene being sequenced, it is possible to have more prior information on the base type distribution. Instead, we consider them to be uniformly independently distributed among the four base types for each sequence.

As for the peak positions, empirical data shows that for each sequence, the peak timing locations are first-order Markov, i.e., $f(\tau_{l,i+1}|\tau_{l,i})$ satisfies

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

where $f_{\Delta\tau}$ has its mean equal to the slowly varying average peak spacing, and has standard deviation of approximately 0.8 samples. The average peak spacing is approximately $T = 12$, which increases by about half a sample value over 600 pulses. Another way of interpreting (2) is that for each sequence, elements of the peak time series $\{\tau_i\}$ satisfy $\tau_i = \tau_{i-1} + T + \tau_{*i}, i > 1$, where

τ_{*i} is the timing jitter. With a single sequence, such jitter can be tracked easily with an algorithm similar to a phase-lock loop in digital communication systems. A phase-lock loop employs negative feedback mechanisms to track changes in the phase or frequency of a sinusoid signal. In the case of two sequences, cumulations of independent timing jitters in each sequence prevents any synchronization of the major and minor peaks, rendering joint base-calling much more difficult. This is not to say that the locations of neighboring peaks are uncorrelated. For example, two peaks separated 4 to 8 samples apart are very likely to be from different sequences. Such relationships will be explored in our proposed joint base-calling algorithm later.

As discussed in [29] and [23], the additive noise $\underline{e}(t)$ consists of measurement noise, possible contaminants from DNA replication and electrophoresis, and residual errors from the preprocessing stages. It has a non-stationary spectral color as determined by the generic pulse shape. Nevertheless, we assume that $\underline{e}(t)$ is white Gaussian with zero mean and standard deviation σ_e, instead of further investigating its noise spectrum. Such simplifications greatly reduce the complexity of the base-calling process, while simulation results show that even such loose approximations yield acceptable accuracies during base-calling.

The last important characteristic of the electrophoresis data is the generic pulse shape $p(t)$. The model described by (1) contributes approximation errors for the generic pulse shape to the additive noise term. Nelson [29] has proposed to model $p(t)$ as the distribution function of the sum of two independent random variables, one Gaussian with distribution $\mathcal{N}(\mu, \sigma^2)$, and one exponential with parameter λ. The result is an exponentially mediated Gaussian curve, which has tails heavier on the right than on the left. We selected 100 pulses manually from a single sequence to fit this model in MATLAB, 25 from each base type. Each parameter was represented as a function of the peak location b, also manually identified. Generic pulses can then be generated using the fitted model at different peak positions. Although this approach gives very good estimates of $p(t)$ if the parameters have been determined, owing to variations of data quality across different data sets and the impracticality of manually estimating the parameters for each sequence to be base-called, we need to resort to simpler heuristics rather than using this complete model. Davies *et al.* [23] proposed a simpler unit-width pulse shape with a central Gaussian part and uneven exponential tails. The pulse width is set to be linear in time. Since this model is also fitted from observed data and the automated sequencers they have employed are different from our ABI machines, we decided not to use a similar model.

Instead, we propose a heuristic procedure to obtain generic pulse shapes for each two-sequence data set. One observation from the single sequence sample data set is that although the average pulse width increases with time, the increments are always very small, hence the average pulse width can be viewed as being constant locally. Also, smaller pulse widths lead to less interference across neighboring bases at the beginning of a trace. Given a two-sequence data set, we first divide it into windows of 500 samples. Note the number 500 is arbitrary and can be changed if needed. To reduce edge effects, the overlap between neighboring windows is set to be 250 samples. Within each window, we collect pulses of shape which can be deemed typical regarding curvature, cumulative area, relative amplitudes, and size of the tails on both size, to be referred to as *good* pulses. These are then normalized to unit peak amplitude and averaged. As expected, the full widths at half maximum (FWHM) of the average good pulse in each window displays an increasing trend along the data trace. Towards the end of the trace, the number of good pulses becomes small since all have very large tails in comparison to those at the front. To avoid this problem, we take the average good pulse with the smallest FWHM as the unit generic pulse $\hat{\underline{p}}(t)$. For the lth window, the corresponding generic pulse $\hat{\underline{p}}_l(t)$ is $\hat{\underline{p}}(t)$ scaled by the factors found before. The resulting set of generic pulses are very close fits to observed peaks.

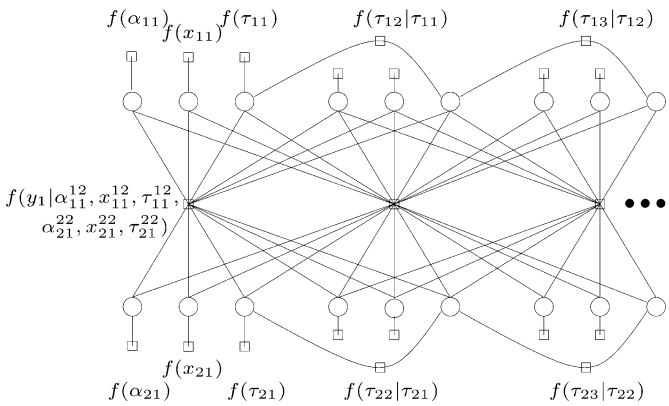

Fig. 2. Factor Graph for maximum *a posteriori* probability (MAP) estimation of individual bases. Circles represent the peak parameter variables given in (1); squares represent conditional probability distributions. Variables on the top row are associated with the major, and those on the bottom row are associated with the minor. The shortened variable notation with both superscripts and subscripts represents pairs or triplets of variables: $\alpha_{11}^{12} = (\alpha_{11}, \alpha_{12}), x_{11}^{12} = (x_{11}, x_{12}), \tau_{11}^{12} = (\tau_{11}, \tau_{12})$.

III. Development of Joint Base-Calling Algorithm for Two Sequences

A. Maximum a Posteriori Base Estimation

According to the data model given by (1), the dependencies between the peak parameters can be represented by a factor graph (FG), as shown in Fig. 2.

Together with Bayesian Networks and Markov Random Fields, factor graphs are graphical models that describe the dependencies among components of complicated functions. A factor graph is an undirected, bipartite graph, in which vertices can be divided into two disjoint sets, where no vertices from the same set are connected. Two types of vertices exist in a factor graph: one type for random variables, and the other for factors, or functions of a subset of the variables. An edge connects a variable node x to a factor node f, if and only if x is an argument of the function $f(\cdot)$. The importance of a factor graph is that under the cycle-free condition, its structure gives not only how a global function factors into local ones, but also a method for computing associated marginal functions for each individual variable node. This method, called the Sum-Product Algorithm [30], exploits the locality properties of the factor graph by passing messages along the edges. In the case where cycles exist, convergence of the Sum-Product Algorithm has

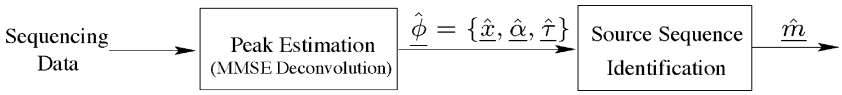

Fig. 3. Block diagram for the two-stage base-calling algorithm. The deconvolution process is followed by source sequence identification through message passing. The sequencing data can be used to estimate the generic pulse shape as well.

not been proven, although it is still routinely used in such cases with plausible results.

In Fig. 2, circles represent the peak parameter variables, while squares represent probability distributions. $\mathbf{y}_i$ denotes all data points associated with the ith peak. This dependency structure, together with conditionals obtained from training data, allows the Sum-Product Algorithm 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 theoretically optimal.

B. 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. Mathematically, let $\underline{\theta} = \{\underline{x}_1, \underline{\alpha}_1, \underline{\tau}_1, \underline{x}_2, \underline{\alpha}_2, \underline{\tau}_2\}$ be the peak parameter vector, then its MAP estimate is

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}} f(\underline{\theta}) = \arg\max_{\underline{\theta}} \{\log f(\underline{y}|\underline{\theta}) + \log f(\underline{\theta})\}. \quad (3)$$

Assuming that the amplitude and type of each individual base is independent of those of the others, the prior distribution of the peak parameters can be written as log-sums

$$\log f(\underline{\theta}) = \sum_{l=1}^{2} \sum_{i=1}^{N_l} \{\log f(\alpha_{li}) + \log f(\underline{x}_{li})\} + \log f(\underline{\tau}_1, \underline{\tau}_2).$$

On the other hand, under the assumption of additive white Gaussian noise with zero mean and variance σ^2, the log likelihood of the observed data is

$$\log f(\underline{y}|\underline{\theta}) = -\frac{1}{2\sigma^2} \sum_t \|\underline{y}[t] - \sum_{i=1}^{N_1} \alpha_{1i} \hat{p}_i(t - \tau_{1i}) \underline{x}_{1i} - \sum_{j=1}^{N_2} \alpha_{2j} \hat{p}_j(t - \tau_{2j}) \underline{x}_{2j}\|^2 + c$$

where c sums terms that do not affect the maximization. Consider a new parameter vector $\underline{\phi} = \{\underline{x}, \underline{\alpha}, \underline{\tau}\}$, where $\underline{x} = \{\underline{x}_1 \cup \underline{x}_2\}, \underline{\alpha} = \{\underline{\alpha}_1 \cup \underline{\alpha}_2\}, \underline{\tau} = \{\underline{\tau}_1 \cup \underline{\tau}_2\}$, and an indicator variable $\underline{m}$ where $m_k \in \{(00), (10), (01), (11)\}$. m_k represents whether a spike at time τ_k has originated from noise, the major, the minor, or both. Although the correspondence between $\underline{\theta}$ and $\{\underline{\phi}, \underline{m}\}$ is not one-to-one, accurate estimates of $\underline{x}$ and $\underline{m}$ leads to identification of the constituent sequences $\underline{x}_1$ and $\underline{x}_2$. In other words, the cost functions above are equivalent to the following:

$$\log f(\underline{\phi}, \underline{m}) = \sum_{k=1}^{N} \{\log f(\alpha_k|m_k) + \log f(\underline{x}_k)\} + \log f(\underline{\tau}|\underline{m}) + \log f(\underline{m}) \quad (4)$$

$$\log f(\underline{y}|\underline{\phi}) = -\frac{1}{2\sigma^2} \sum_t \|\underline{y}[t] - \sum_{k=1}^{N} \alpha_k \hat{p}_k(t - \tau_k) \underline{x}_k\|^2 + c. \quad (5)$$

In (4), it is assumed that the amplitude α_k of each individual base is dependent on m_k only and distributed independently of its neighbors. Also, x_k is assumed to be uniformly independently distributed, also independent of the indicator variable. The dependency relationship for the peak locations is more complex, as will be explained later. In (5), the mean square error for fitting pulse trains to the data is minimized without additional constraints. We have assumed this is independent of the indicator variable associated with each base, although strictly speaking, the minimization should give rise to approximately equal numbers of peaks in $\underline{x}_1$ and $\underline{x}_2$. Taking such independence into account, we can rewrite the maximization in (3) as follows:

$$\hat{\underline{\phi}}, \hat{\underline{m}} = \arg\max_{\underline{m}} \left\{ \arg\max_{\underline{\phi}} [\log f(\underline{y}|\underline{\phi}) + \log f(\underline{\phi}, \underline{m})] \right\}. \quad (6)$$

As a further simplification, we instead consider a suboptimal solution, derived by neglecting the effect of $\log f(\underline{\phi}, \underline{m})$ on the estimate of the peak parameters such that

$$\hat{\underline{\phi}} = \arg\max_{\underline{\phi}} \log f(\underline{y}|\underline{\phi}) \quad (7)$$

$$\hat{\underline{m}} = \arg\max_{\underline{m}} \log f(\hat{\underline{\phi}}, \underline{m}). \quad (8)$$

By expressing the cost function using two components, we obtain a two-stage algorithm for joint base-calling: the first stage is to locate all peaks, and the second stage is to find the best indicator sequence to identify the sources for each peak. Fig. 3 illustrates the two stages involved and the corresponding outputs.

For the first stage, (7) describes a deconvolution process aimed at finding the minimum mean squared error estimate of the peak parameters. Again, we compute such deconvolutions in separate windows. Fig. 4 shows the deconvolved spike train for the data in Fig. 1(b). Note that the hidden pulses at around positions B and C are captured, while the minor peak to the left at position A has been missed. This may or may not contribute

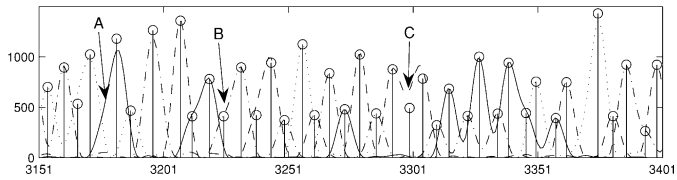

Fig. 4. Output of the MMSE deconvolution process for the data in Fig. 1(b). Spike trains are identified in terms of amplitude, location, and base type. In this example, the minor peak hidden at position A is not captured during deconvolution, while those at around positions B and C are correctly located.

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.

Another issue we have conveniently overlooked during deconvolution is that because there is minimal prior information on N, or N_l for each window, we overestimate its value when maximizing (7). 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 one sample 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. Assumption of independent amplitudes and base types decouples those terms in (4). However, since peak spacings are approximately first-order Markov within a single sequence, once we take the union of $\underline{\tau}_1$ and $\underline{\tau}_2$, the dependency becomes at least second order. More specifically, the offset between the current base and the one next to its immediate neighbor should be approximately 12 samples, which is the average peak spacing in a single sequence. After mixing, if the current base of interest is labeled as major and the previous base is labeled as minor, then the offset between these two peaks can be of any value up to the average peak spacing. As an example, consider position B in Fig. 4. Here the major and minor offset by approximately six samples, which is half the average peak spacing in a single sequence, but B is closer to the pulse on its left than to the one on its right. As a starting point for our algorithm, we approximate the higher order Markov model of the peak spacings with a first-order one. Using c' to denote terms that do no affect the maximization results, m_{k-1}^k to denote $\{m_{k-1}, m_k\}$, and assuming neighboring indicator variables are independent, we have

$$\log f(\hat{\underline{\phi}}, \underline{m}) \simeq c' + \sum_{k=1}^{N} \log f(\hat{\alpha}_k | m_k) + \log f(\tau_1) + \sum_{k=2}^{N} \log f\left(\hat{\tau}_k | \hat{\tau}_{k-1}, m_{k-1}^k\right) + \sum_{k=1}^{N} \log f(m_k).$$

Let

$$R_k = f(\hat{\alpha}_k | m_k) \tag{9}$$

$$T_k = \begin{cases} f(\hat{\tau}_k) f(m_k), & k = 1 \\ f\left(\hat{\tau}_k | \hat{\tau}_{k-1}, m_{k-1}^k\right) f(m_k), & k > 1 \end{cases} \tag{10}$$

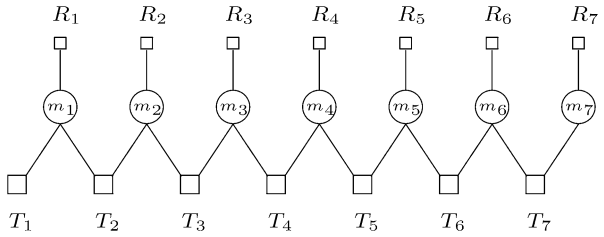

Fig. 5. First-order factor graph for separating two sequences. Expressions associated with the functional nodes are given in (9).

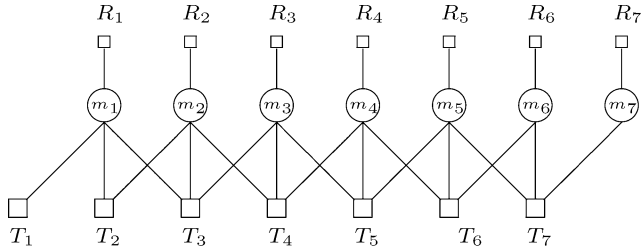

Fig. 6. Second-order factor graph for separating two sequences. Expressions associated with the functional nodes are as given in (11), defined similarly as in the first-order model through (9).

the distribution of $\underline{m}$ parameterized by $\hat{\underline{\phi}}$ can be represented graphically using an FG as in Fig. 5. This is the trellis graph of a hidden Markov model, where the Sum-Product Algorithm [30] can be applied. The cycle-free property of this FG ensures convergence of the algorithm. We can also write a second-order approximation for the log-likelihood of the peak timing locations

$$\log f(\hat{\underline{\tau}} | \underline{m}) = \log f\left(\hat{\tau}_1^2\right) + \log f\left(m_1^2\right) + \sum_{k=3}^{N} \log \left\{ f\left(\hat{\tau}_k | \hat{\tau}_{k-2}^{k-1}, m_{k-2}^k\right) f(m_k) \right\}. \tag{11}$$

The alternative factor graphs is shown in Fig. 6.

IV. Simulation Results and Discussions

We applied the algorithm stated in Section III-B to eight data sets, results are shown for two of those in Table I. The simulation results for different datasets varied significantly depending on the ratio of average amplitudes between the major and the minor. To control the ratio of average amplitudes between the major and minor peaks, the sequencing primer responsible for initiating the replication process was added at limited concentration for the minor relative to the major. To target an average amplitude ratio of approximately 2, the relative concentration of major and minor primers was set to 20 to 1. Even so, the achieved amplitude ratio can greatly exceed 2 or be much smaller, close to 1. In the first case, the minor became less distinguishable; in the second case, neither sequence could be called reliably, especially in parts where bases from the two sequences are approximately aligned. The two data sets used to generate Table I were those that displayed good amplitude ratio characteristics. The sensitivity of our algorithm to large amplitude ratio variation was not quantified at the current exploratory stage, but needs to be studied in detail in the future.

The first dataset used for simulation was the one shown in Fig. 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 peak spacing in this part of the read are not uniform due to artifacts in mobility correction. Data collected after sample 8400 was not used because interferences among neighboring pulses 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, in both the first and second-order models, we assumed the conditional distribution of α_k was Gaussian, parameterized by m_k, instead of a Gamma distribution with uneven tails. More specifically, the mean amplitudes were assumed to be $25, 750, 350$, and 1100, respectively for *noise*, *major*, *minor* and *both* sequences, while the standard deviations were assumed to be $10, 200, 100$, and 224. For the prior distribution of m_k, it was observed from the deconvolution results that only a few were labeled as *noise*. We, therefore, assigned a uniform value of 0.33 for the other three cases.

Unlike in our previous work [31], where the posterior distribution of m_k was approximated by a histogram of values obtained manually from the first data set, the likelihood of τ_k was approximated with a more explicit model. In the first-order case, $f(\hat{\tau}_k|\hat{\tau}_{k-1}, m_{k-1}^k)$ was computed from the following distributions depending on the values of m_{k-1} and m_k. First, if either of two neighboring peaks are labeled as *noise*, i.e., $m_{k-1} = 00$ or $m_k = 00$, it provides no information regarding its neighbor. We therefore assume τ_{k-1} and τ_k to be independent, and $f(\hat{\tau}_k|\hat{\tau}_{k-1}, m_{k-1}^k)$ to take on a uniformly distributed value of $1/12$. Next, if two peaks are labeled as (*major*, *minor*), or (*minor*, *major*), we assume $f(\hat{\tau}_k|\hat{\tau}_{k-1}, m_{k-1}^k)$ is computed from a Gaussian distribution with mean 6, half the average peak spacing, and standard deviation 3, which is larger than 0.8 for the single sequence case. An increased spread in peak location is expected when two sequences are superimposed, for relative positions are random in the inherent absence of synchronization. Finally, for all other values of m_{k-1}^k, we assume $f(\hat{\tau}_k|\hat{\tau}_{k-1}, m_{k-1}^k)$ is computed from a Gaussian distribution with mean 12, and standard deviation 3.

In the second-order case, $f(\hat{\tau}_k|\hat{\tau}_{k-2}^{k-1}, m_{k-2}^k)$ can be computed similarly, starting from the cases where one of m_{k-2}, m_{k-1}, m_k is 00. For example, if $m_{k-2} = 00$, or the $(k-2)$th peak is *noise*, it provides no information regarding its neighbors. Consequently, $f(\hat{\tau}_k|\hat{\tau}_{k-2}^{k-1}, m_{k-2}^k)$ becomes $f(\hat{\tau}_k|\hat{\tau}_{k-1}, m_{k-1}^k)$, which is described in the previous paragraph. If none of the three peaks under consideration is *noise*, $f(\hat{\tau}_k|\hat{\tau}_{k-2}^{k-1}, m_{k-2}^k)$ can be computed from either a Gaussian of mean 12 or 6 depending on the specific value of m_{k-2}^k.

To evaluate the joint base-calling error rate, the sequencing result was compared with reference sets using the *cross_match* program [21]. The dynamic programming based Smith-Waterman algorithm was employed to find the longest lengths of consecutive bases which gave the best local alignment. Given a set of penalties for different error types including mismatch, insertion, and deletion, it finds the best run of matched bases to achieve the smallest cost. Results are listed in Table I. Averaging over the two data sets, the first-order model achieves an overall error probability of 7.5% and 12% for the two sequences respectively, while the second-order model achieves 5% and 12%. Also given in this table is the performance of a single sequence base caller on the same data sets. On the averages it achieves an error probability of 2%. This base caller was constructed similar to a phase-lock-loop, with the peak location and corresponding base-types tracked one at a time in the forward direction. A phase-lock-loop is an algorithm from the communications field. It employs negative feedback mechanisms to track changes in the phase or frequency of a sinusoid signal.

TABLE I
PERFORMANCE OF JOINT (J) AND SINGLE (S) SEQUENCE BASE-CALLING ALGORITHMS FOR TWO DATA SETS (SEPARATED INTO THE TOP AND BOTTOM PORTION OF THE TABLE). JOINT BASE-CALLING IS PERFORMED ACCORDING TO THE ALGORITHM GIVEN IN SECTION III-B, WHERE THE ORDER OF THE FACTOR GRAPH USED IS GIVEN IN THE BRACKETS. THE SAME DATA IS ALSO PROCESSED BY A SINGLE-SEQUENCE BASE-CALLER

	length of best single match	% substitution	% deletion	% insertion
Major (J, 1st)	399	1.00	3.59	0.40
Minor (J, 1st)	262	2.77	3.84	4.26
Major (J, 2nd)	432	0.76	2.67	0.76
Minor (J, 2nd)	260	2.14	4.06	4.49
Major (S)	582	0.17	0.34	0.00
Major (J, 1st)	257	5.83	1.35	2.91
Minor (J, 1st)	60	4.64	0.66	7.95
Major (J, 2nd)	312	1.85	1.85	2.08
Minor (J, 2nd)	57	3.31	0.66	9.27
Major (S)	578	2.25	1.21	0.00

The first observation from the sequencing result 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. A throughput gain is achieved with the additional minor sequences. Although at a lesser accuracy, this information could be useful in several ways. First, recall from Section I-A5 that sequencing can be performed on DNA fragments only, so subsequent assembly by matching overlapped base-calls is required. Since perfect base-calling results are not guaranteed, we often sequence the same location multiple times, where the average number of reads covering a base in the reconstructed sequence is referred to as the the depth of coverage. A full coverage often corresponds to approximately eight-fold repetition [32]. It may be possible to replace some repetitions with the minor base calls, hence increasing the depth of coverage while reducing the overall number of reactions needed. Second, the major and minor can be set to a known distance away such that the presence of the minor facilitates easier sequence alignment in the assembly process, especially for sequences containing multiple repeated genes. For example, paired-end reads are commonly employed to assist merging paired pairs of overlaps. Data are produced by sequencing both ends of a template of known length in separate reactions. The template is chosen to be longer than the corresponding traces. Merged sections around each end-read are then linked through the known distance. If such paired-end reads can be performed within one reaction,

manual handling of the samples can be reduced. A third use of such joint sequencing results is to identify contaminants in naturally occurring mixtures. For example, such hypothesis testing results may be desired in forensic science where we want to determine the presence of a second source of DNA.

The second observation from our sequencing result is that the second-order model does increase the number of bases which can be identified and does slightly reduce the sequencing error, at least for the major sequence. In using factor graphs, the sum-product algorithm always converges for a cycle-free graph. On the other hand, as messages iterate on a cycled factor graph, as in our second-order case, the end-results of the sum-product algorithm are no longer exact marginals, but rather approximations that may or may not be accurate. Intuitively speaking, locality of a graph is better exploited when cycles in the factor graph are larger in radius, i.e., messages passed in one part of the graph does not affect those in other parts immediately. When edges are less sparse and very small cycles exist, on the other hand, we can only hope for the best, as the sum-product algorithm has not been verified to give a good approximation. Our setup leads to very small cycles indeed, but empirically reasonable results are obtained.

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 factor graph in Figs. 5 and 6. Second, examination of the sequencing results shows that many deletion and insertion errors occur not during the second stage of the algorithm, but are caused by the deconvolution process. Such errors not only are problematic on their own, but also weaken the time dependence among neighboring peaks. One possible compensation is to iterate between the deconvolution and source sequence separation stages, where valid peak locations from stage two are set as the initial states for the deconvolution process in stage one, with missed peaks inserted heuristically based on spacing uniformity. The amount of overfit for the total number N of peaks can also be controlled.

V. Conclusion

In this paper, we explored the possibility of base-calling two superposed sequences jointly. Specifically, a two-stage algorithm was developed, where spikes corresponding to different bases are identified through deconvolution first, then assigned to the two source sequences by message passing on a factor graph. The factor graph localizes the statistical dependencies of the overall observed data on the amplitudes, locations, and types of the spikes. Simulation results show that combined with single sequence base-calling, this algorithm enables the sequencing of an additional segment. For the two datasets which we analyzed in detail in this paper, simulations show that using a first-order factor graph achieves an average error probability of 7.5% and 12% for the major and minor sequences respectively, while the second-order model achieves 5% and 12%. A single-sequence base-caller on the same data achieves an average error probability of 2%. Although not at the same accuracy, these results are promising. Several venues for further exploration emerge: sensitivity of the algorithm to variations in average amplitude ratios can be quantified to determine the validity of the base-calls; 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; additional iterations between deconvolution and source sequence identification may lead to improved performance.

Acknowledgment

The authors would like to thank Stuart Licht for helpful discussions during the development of the algorithm, and the reviewers for their valuable comments and suggestions.

References

[1] F. Sanger, S. Nicklen, and A. Coulson, "DNA sequencing with chain-terminating inhibitors," in *Proc. Nat. Acad. Sci.*, 1977, pp. 5463–5467.
[2] Applied Biosystems 2008 [Online]. Available: http://www.appliedbiosystems.com
[3] Applied Biosystems, Automated DNA Sequencing, Chemistry Guide 2000.
[4] N. Rusk and V. Kiermer, "Primer: Sequencing—The next generation," *Nature Meth.*, vol. 5, no. 1, pp. 15–, 2008.
[5] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnol.*, vol. 26, no. 10, pp. 1135–1145, 2008.
[6] S. Schuster, "Next-generation sequencing transforms today's biology," *Nature Meth.*, vol. 5, pp. 16–18, 2008.
[7] D. Hert, C. Fredlake, and A. Barron, "Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods," *Electrophoresis*, vol. 29, no. 23, pp. 4618–4626, 2008.
[8] J. Kidd *et al.*, "Mapping and sequencing of structural variation from eight human genomes," *Nature*, vol. 453, no. 7191, pp. 56–64, 2008.
[9] J. Karow, Large Genome Centers Replace Sanger With Second-Gen Sequencers for Many Applications Oct. 2008 [Online]. Available: http://www.genomeweb.com/sequencing/large-genome-centers-replace-sanger -second-gen-sequencers-many-applications
[10] T. A. Brown, *DNA Sequencing: The Basics*. : Oxford Univeristy Press, 1994.
[11] *DNA Sequencing Protocols*, C. A. Graham and A. J. M. Hill, Eds., 2nd ed. Totowa, NJ: Humana Press, 2001.
[12] , M. D. Adams, C. Fields, and J. C. Venter, Eds., *Automated DNA Sequencing and Analysis*, 1st ed. New York: Academic, 1994.
[13] J. Shendure *et al.*, "Advanced sequencing technologies: Methods and goals," *Nature Rev. Genet.*, vol. 5, no. 5, pp. 335–344, 2004.
[14] N. Hall, "Advanced sequencing technologies and their wider impact in microbiology," *J. Exper. Biol.*, vol. 210, no. 9, pp. 1518–, 2007.
[15] M. Ronaghi, "Pyrosequencing sheds light on DNA sequencing," *Genome Res.*, vol. 11, no. 1, pp. 3–11, 2001.
[16] M. Margulies *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.
[17] M. Levene, J. Korlach, S. Turner, M. Foquet, H. Craighead, and W. Webb, "Zero-mode waveguides for single-molecule analysis at high concentrations," *Science*, vol. 299, no. 5607, pp. 682–686, 2003.
[18] J. Korlach, P. Marks, R. Cicero, J. Gray, D. Murphy, D. Roitman, T. Pham, G. Otto, M. Foquet, and S. Turner, "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures," *Proc. Nat. Acad. Sci.*, vol. 105, no. 4, p. 1176, 2008.
[19] J. Shendure, G. Porreca, N. Reppas, X. Lin, J. McCutcheon, A. Rosenbaum, M. Wang, K. Zhang, R. Mitra, and G. Church, "Accurate multiplex polony sequencing of an evolved bacterial genome," *Science*, vol. 309, no. 5741, pp. 1728–1732, 2005.
[20] M. C. Giddings, R. L. B. Jr., M. Haker, and L. M. Smith, "An adaptive, object oriented strategy for base calling in DNA sequence analysis," *Nucl. Acids Research*, no. 19, pp. 4530–4540, 1993.
[21] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using Phred. I. accuracy assessment," *Genome Res.*, vol. 8, no. 3, pp. 175–185, 1998.
[22] L. Li and T. P. Speed, "Parametric deconvolution of positive spike trains," *Ann. Statist.*, vol. 28, no. 5, pp. 1279–1301, Oct. 2000.
[23] S. W. Davies, M. Eizenman, and S. Pasupathy, "Optimal structure for automatic processing of DNA sequences," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 9, pp. 1044–1056, Sept. 1999.

[24] N. M. Haan and S. J. Godsill, "Modelling electropherogram data for DNA sequencing using variable dimension MCMC," in *Proc. 2000 IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, pp. 3542–3545.
[25] S. Wiemann, J. Stegemann, D. Grothues, A. Bosch, X. Estivill, C. Schwager, J. Zimmermann, H. Voss, and W. Ansorge, "Simultaneous on-line DNA sequencing on both strands with two fluorescent dyes," *Analyt. Biochem.*, vol. 224, no. 1, pp. 117–121, 1995.
[26] S. Wiemann, J. Stegemann, J. Zimmermann, H. Voss, V. Benes, and W. Ansorge, "'Doublex' fluorescent DNA sequencing: Two independent sequences obtained simultaneously in one reaction with internal labeling and unlabeled primers," *Analyt. Biochem.*, vol. 234, no. 2, pp. 166–174, 1996.
[27] D. van den Boom, C. Jurinke, A. Ruppert, and H. Köster, "Forward and reverse DNA sequencing in a single reaction," *Analyt. Biochem.*, vol. 256, no. 1, pp. 127–129, 1998.
[28] K. Murphy and J. Eshleman, "Simultaneous sequencing of multiple polymerase chain reaction products and combined polymerase chain reaction with cycle sequencing in single reactions," *Amer. J. Pathol.*, vol. 161, no. 1, pp. 27–, 2002.
[29] D. O. Nelson, "Improving DNA sequencing accuracy and throughput," in *Genetic Mapping and DNS Sequencing*. New York: Springer, 1996, pp. 183–206.
[30] F. Kschischang, B. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, 2001.
[31] X. Shi, D. Lun, J. Meldrim, R. Kotter, and M. Medard, "Joint base-calling of two DNA sequences with factor graphs," in *Proc. 2008 IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP 2008)*, 2008, pp. 2049–2052.
[32] S. Batzoglou, D. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. Mesirov, and E. Lander, ARACHNE: A Whole-Genome Shotgun Assembler pp. 177–189, 2002.

Xiaomeng Shi (S'05) received the B.Eng. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2005 and the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2008, where she is currently working towards the Ph.D. degree in the Department of Electrical Engineering and Computer Science.

Her research interests include network coding, security, and signal processing.

Desmond S. Lun (S'02–M'06) received Bachelor's degrees in mathematics and computer engineering from the University of Melbourne, Australia, in 2001 and S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002 and 2006, respectively.

He is Director of the Phenomics and Bioinformatics Research Centre (PBRC) and an Associate Professor in the School of Mathematics and Statistics at the University of South Australia, Adelaide, Australia. Prior to his present position, he was a Computational Biologist at the Broad Institute of MIT and Harvard University and a Research Fellow in the Department of Genetics at Harvard Medical School. In 2006, he was a Postdoctoral Research Associate in the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign. His research interests are in synthetic biology, systems biology, and biological signal processing. He is coauthor (with Tracey Ho) of *Network Coding: An Introduction* (Cambridge University Press, 2008).

Ralf Kötter (S'91–M'96–SM'06–F'09) (deceased) received the Diploma degree in electrical engineering from the Technical University Darmstadt, Germany, in 1990 and the Ph.D. degree from the Department of Electrical Engineering at Linköping University, Linköping, Sweden.

From 1996 to 1997, he was a Visiting Scientist at the IBM Almaden Research Laboratory in San Jose, CA. He was a Visiting Assistant Professor at the University of Illinois at Urbana-Champaign and a Visiting Scientist at CNRS in Sophia-Antipolis, France, from 1997 to 1998. From 1999 to 2006, he was member of the faculty of the University of Illinois at Urbana-Champaign, and in 2006 he joined the faculty of the Technische Universität München, Munich, Germany, as the Head of the Institute for Communications Engineering. His research interests included coding and information theory and their application to communication systems.

Prof. Kötter served as an Associate Editor for both the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE TRANSACTIONS ON INFORMATION THEORY. He received an IBM Invention Achievement Award in 1997, an NSF CAREER Award in 2000, an IBM Partnership Award in 2001, and a 2006 XEROX award for faculty research. He was corecipient of the 2004 Information Theory Society Best Paper Award, of the 2004 IEEE SIGNAL PROCESSING MAGAZINE Best Paper Award. He received the Vodafone Innovationspreis in 2008. From 2003 to 2008 he was a member of the Board of Governors of the IEEE Information Theory Society.

Ralf Kötter passed away in February 2009.

Muriel Médard (S'90–M'95–SM'02–F'08) received B.S. degrees in electrical engineering and computer science as well as mathematics in 1989, the B.S. degree in humanities in 1990, the M.S. degree in electrical engineering in 1991, and the Sc.D. degree in electrical engineering in 1995, all from the Massachusetts Institute of Technology (MIT), Cambridge.

She is a Professor of Electrical Engineering and Computer Science at MIT. She was previously an Assistant Professor in the Electrical and Computer Engineering Department and a member of the Coordinated Science Laboratory at the University of Illinois Urbana-Champaign. From 1995 to 1998, she was a Staff Member at MIT Lincoln Laboratory in the Optical Communications and the Advanced Networking Groups. Her research interests are in the areas of network coding and reliable communications, particularly for optical and wireless networks.

Prof. Médard has served as an Associate Editor for the Optical Communications and Networking Series of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, as an Associate Editor in Communications for the IEEE TRANSACTIONS ON INFORMATION THEORY, and as an Associate Editor for the *OSA Journal of Optical Networking*. She has also served as Guest Editor for the IEEE/OSA JOURNAL OF LIGHTWAVE TECHNOLOGY, for the Joint Special Issue of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE/ACM TRANSACTIONS ON NETWORKING on "Networking and Information Theory" and for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY Special Issue on "Statistical Methods for Network Security and Forensics." She serves as an associate editor for the IEEE/OSA JOURNAL OF LIGHTWAVE TECHNOLOGY. She is a recipient of the William R. Bennett Prize in the Field of Communications Networking, the 2002 IEEE Leon K. Kirchmayer Prize Paper Award, and the Best Paper Award at the Fourth International Workshop on the Design of Reliable Communication Networks (DRCN 2003). She received the NSF CAREER Award in 2001 and was corecipient of the 2004 Harold E. Edgerton Faculty Achievement Award at MIT. She was named a 2007 Gilbreth Lecturer by the National Academy of Engineering. She serves as a member of the Board of Governors of the IEEE Information Theory Society.

James C. Meldrim received the B.Sc. degree from Cornell University, Ithaca, NY, in 1994.

He played an integral role in the Whitehead Institute Center for Genome Research's contribution to the Human Genome Project (1998–2004). He is a Manager of Technology Development within the DNA Sequencing Operations group at the Broad Institute of MIT and Harvard, Cambridge, MA, where he continues to provide his leadership and experience to developing next-generation DNA sequencing platforms.

Andrew J. Barry received the B.S. degree in biology from Stonehill College, Easton, MA. He has concurrently pursued graduate studies in biomedical engineering at Tufts University, Medford, MA.

In 2002, he began working at the Whitehead Institute's Center for Genome Research, Cambridge, MA, where he worked on the production floor at the genome center for a year before transitioning to the Technology Development group, which at the time, was called The Broad Institute, founded in 2003. His group worked on the development of the medical-directed sequencing process from transition to production. In 2005, he began working with next-generation sequencing technologies, where he has been working on the scaling-up of the Illumina GA platform and the development of automated platforms for next-generation sample preparation.