FEATURE EXTRACTION FOR DNA BASE-CALLING USING NNLS

Lucio Andrade-Cetto and Elias S. Manolakos

Communications and Digital Signal Processing Center,
Center for Research and Graduate Studies
Electrical and Computer Engineering Department,
Northeastern University, Boston MA 02115

ABSTRACT

We present new algorithms that can be used to extract features from a DNA chromatogram prior to base-calling. The algorithms assume that the inter-base distance has already been equalized using methods such as those presented in [1]. We show first how a good estimate of the peak diffusion (spread) can be calculated from the raw trace and without having to known the underlying base sequence. Using the estimated inter-peak distance and peak spread parameters a non-negative least squares problem can be formulated in order to find the weight factors of the multiple shapes immersed in broad peaks, typically found towards the end of the chromatogram. The two algorithms combined provide peak hypotheses that are tested by the subsequent base decisions and scoring stage of the base-caller using probabilistic methods.

1. INTRODUCTION

DNA sequencing is the experimental process of determining the unknown nucleotide sequence of a DNA sample. The sample under analysis is used as a template to generate a large number of partial copies of all possible lengths (fragments). The terminal nucleotide, (A, C, G or T) of each fragment is labeled with either a radioactive or fluorescent marker (dye) which can is used for detection at a later step [2]. Labeled DNA fragments in the mixture are separated into sub-populations according to their length by gel or capillary electrophoresis, because shorter fragments travel faster than longer ones due to their difference in molecular weight. Labeling allows the identification and abundance characterization of each sub-population of fragments. By combining both operations (labeling and electrophoresis) in a sequencing experiment it becomes possible to infer the unknown sequence of bases present in the original DNA sample.

The analysis of the four signal traces generated by a sequencing experiment is performed by software programs, known as DNA *base-callers* , whose purpose is to estimate the time location and abundance of fragments in sub-populations manifested as peaks in the resulting DNA chromatogram. After merging the information from four channels (one channel for each nucleotide type) the original sequence of bases can be deduced. However, a sequencing experiment is prone to several effects that may distort the signal [3, 4] making accurate DNA base-calling more challenging than just a peak ordering procedure. DNA sequencing is so widely used in biology so it is desirable to improve the number of bases that can be accurately and routinely deciphered in every run. Even a 10% improvement in average readlengths (number of bases that can be read with a specified accuracy e.g. 99% in every run) can reduce significantly the cost of genome sequencing projects.

Our research group has introduced a *divide-and-conquer* top-down processing strategy which aims at the accurate interpretation of the raw DNA sequencing traces. It encompasses several signal pre-processing algorithms [1, 3, 4] that try to improve the SNR and undo distortions before attempting to base-call [5]. In [6] we have presented a first attempt to formulate the base-calling problem within a Bayesian probabilistic framework. We have shown that using probabilistic modeling and unsupervised learning methods allows base-callers to adapt online to different experimental conditions, sequencing technologies and chromatogram quality, without the need for recalibration. Furthermore, probabilistic modeling allows using posterior probabilities (of a detected peak being a true base given the data) as base confidence levels.

In this paper, we present the front-end stage of a new statistical base-caller based on a less restrictive probabilistic graphical model [7]. Specifically we discuss here the main elements of the feature extraction stage, that tries to determine potential landmarks in the traces which may represent true symbols of the DNA sequence. Our approach can address effectively the challenging problem of unmixing poorly resolved merged peaks appearing towards the end of the chromatogram due to the inevitable resolution loss. In Section 3 we present an algorithm based on Non-Negative Least Squares (NNLS) employed to unmix kernels representative of peaks. To improve the robustness of this algorithm two key quantities must be estimated from the raw trace: the expected distance between peaks representing true base symbols (a problem addressed in [1]) and the expected diffusion (spread) of peaks, to be discussed here in Section 2. In Section 4 we present examples of how our approach can unmix potential peak alleles and discuss preliminary results on resulting base-calling performance improvements. Discussing the new probabilistic model for base-calling (the system's back-end) is beyond the scope of this paper and will appear elsewhere.

2. PEAK SPREAD ESTIMATION

In a typical DNA denaturing experiment the number of members in a sub-population of same-length fragments is usually large, therefore the detected signal peak for every allele reassembles to the probability density function of the fragments. Let us model an *ideal* signal peak generated by a subpopulation of same-length fragments with a Gaussian shape:

$$\phi_j(i) = x_j \, e^{-\alpha_j (i - t_j)^2}, \tag{1}$$

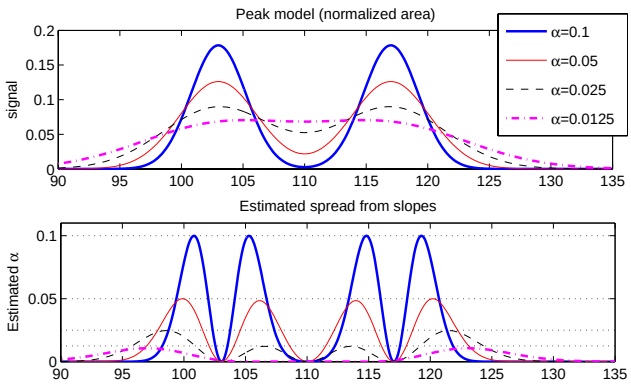

Fig. 1. Interaction between alleles. Top panel: Model signals for two close-by alleles. In all four cases considered there is the same quantity of strands in a subpopulation (same area) but differently distributed (different height and spread). Bottom panel: Plot of $\breve{\alpha}(i)$ for every case. The dashed lines show the values of α used to generate the model signals. Observe that by using the maximum values of the derivative it is possible to estimate α even for severely merged peaks.

where $[x_j\ \alpha_j\ t_j\ ]^T\ \in \mathcal{R}^3_+$ is the vector of unknown parameters. For every peak j, i is the sample index, the weight x_j is related to the abundance of fragments in the subpopulation, t_j is related to the expected value of their time-of-arrival to the photo-detector, and α_j models the degree of uncertainty about this location.

Estimating all elements of the parameters vector simultaneously for all peaks in a lane of the chromatogram (or even for a medium size region) is prone to fail due to several expected deviations from the the ideal situation modeled by Eq. (1) attributed, among other reasons, to: overlapping peaks, low SNR, non regular mobility patterns, chimeras, compressions, dependencies of peaks to the underlying sequence etc. This is possibly the reason why base-calling strategies grounded on sound statistical methods (e.g. [8, 9, 10, 11, 12]), have yet to provide evidence of accuracy superiority relatively to *Phred* [13], the most popular but heuristically built base-caller.

We follow a *divide and conquer* approach and propose first a method for the accurate estimation of the uncertainty of the fragment arrival time, (α_j), regardless of the knowledge of t_j and x_j. Let us define *peak spread* ($\frac{1}{\alpha}$) as the variance of the arrival time (or sample location in the chromatogram) of DNA fragments of a given length. It is well known that this parameter increases with length because large fragments in a sub-population interact more than short ones while traveling in the separation media. As a result, we propose to model the spread as a monotonically increasing function of fragments length. Four different curves will be regressed independently, one for each lane. (To avoid notational complexities we will concentrate on a single lane for the remaining of the section.)

Two approaches come immediate to mind for calculating the observed spread (the response variable in the regression): (i) search for representative peaks and measure the width at a certain height relatively to their maxima, or, (ii) compute a histogram count of the sampled data. However, they may both usually fail in the presence of overlapping peaks, alleles with tails, and low SNR, which are problems often appearing in typical DNA chromatograms. Recall that at this stage in the processing we have not yet formed a hypothesis on which peaks truly represent one base symbol.

To compute the peak spread (α) from the observed data we

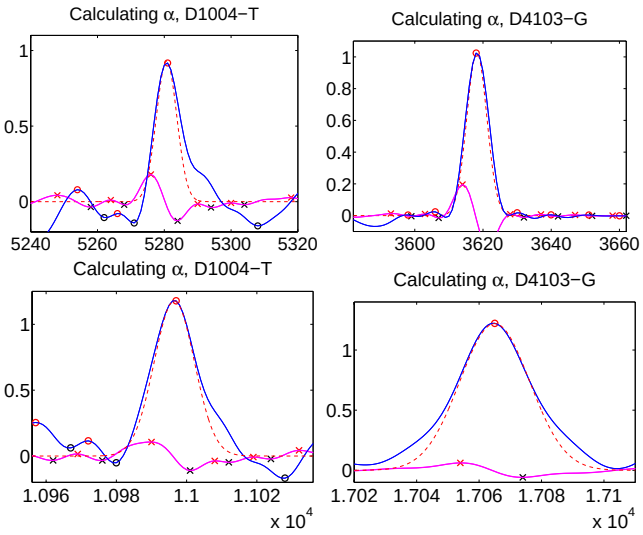

Fig. 2. Four examples where the Gaussian model (dashed line) is recovered from the derivative of the observed trace (continuous thick line). The continuous thin line is the original sampled chromatogram. α is computed from the values of the derivative at the MSP's (crosses) and the peak of the observed signal (circle).

perform the following operations: (i) We take the second derivative of Eq.(1) and by equating to zero we find the maximum slope point locations (MSP's) ; (ii) by substituting these locations into the derivative of Eq.(1) we can solve for α and obtain,

$$\alpha_j = \frac{e \cdot \max_i(\phi'_j(i)^2)}{2 \max_i(\phi_j(i))^2}. \qquad (2)$$

This expression forms the basis for estimating the spread of a peak by using its observed value and an estimate of the signal's derivative.

In the top panel of Figure 1 we see four signals modeling two close alleles but with different spread. In all four cases it is assumed that the quantity of fragments (area) is the same but their distribution different. The bottom panel shows the trace of $\alpha_j(i)$ for every case. The response variables will be picked at the location of the MSP's as suggested by expression (2). It can be observed that in all cases these maxima approximate very well the dashed lines which correspond to the values of α used to generate the model signals in the top panel. If the chromatogram is smoothed with a low pass filter and only MSP's belonging to considerable area alleles are considered, then the maximal values of $\breve{\alpha}(i)$ can only correspond to sides of peaks (going uphill or downhill) which are not interacting with other artifacts and can be used to estimate the trend of the peaks spread.

Figure 2 shows four examples using experimental traces where the Gaussian model (dashed line) is recovered from the derivative of the observed trace (continuous thick line). Early (top panels) and late (bottom panels) peaks of two typical chromatograms are shown, one generated by gel electrophoresis and primer dyes (D1004) and another by capillary electrophoresis and terminator dyes (D4103). The continuous thin line is the original sampled chromatogram. The spread $\breve{\alpha}$ is computed from the values of the derivative at the MSP's (x symbols) and the peak of the observed signal (o symbols). The model is centered and scaled to the crest of the allele for demonstration purposes, the complete method show-

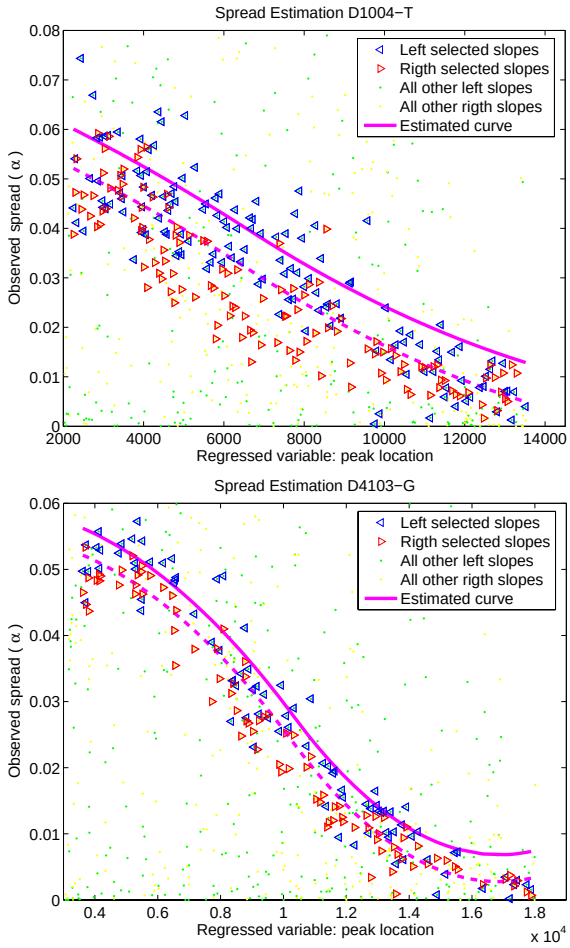

Fig. 3. Spline regressed (dashed line) for one channel of two data sets sequenced with different chemistries; primer dyes and gel electrophoresis (PD) on the top, terminator dyes and capillary electrophoresis (TD) in the bottom. The continuous line is the final estimate for α after the adjustment of Eq. (7). The bottom row of scatter plots show the residual analysis for the respective data sets.

ing also how to also estimate parameters x and t is provided later in Section 3. Using an alternative approach, such as measuring the width of the peak, would have produced an incorrect estimation of $\check{\alpha}$, due to the tails that appear after some peaks in these examples.

We now formulate the *spread* estimation algorithm. Let the vector $\{\check{d}_i\}$ be one lane of the sampled chromatogram, $i \in \mathcal{N}$ is the sample index [1]:

1. Select peaks and valleys: $\mathcal{P} = \{(p,q,r) \in \mathcal{N}^3\}$, such that p is the sample index of a local maximum in $\check{d}$ and q (r) is the index of the first local minimum to the left (right) of p.

2. By using Eq. (2) we can find two estimates for $\check{\alpha}$, when using the left and right MSP's of every peak:

[1]Notation: A ˇ marks a value computed from observations or directly observed, a ˆ denotes an estimate after regression or probabilistic modeling, and, a ˜ denotes an intermediate value.

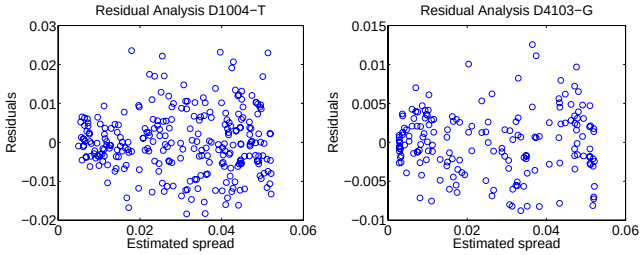

Fig. 4. Scatter plots with the residual analysis respective to the regressed data sets shown in Figure 3.

$$\begin{aligned}\mathcal{A}_L &= \left\{(\check{\alpha},p) : \check{\alpha} = \tfrac{e}{2\,\check{d}_p^2} \max_{i\in q:p-1}(\check{d}_i - \check{d}_{i+1})^2\right\},\\ \mathcal{A}_R &= \left\{(\check{\alpha},p) : \check{\alpha} = \tfrac{e}{2\,\check{d}_p^2} \max_{i\in p:r-1}(\check{d}_i - \check{d}_{i+1})^2\right\},\end{aligned} \tag{3}$$

$$\mathcal{A} = \mathcal{A}_L \cup \mathcal{A}_R. \tag{4}$$

Set $\mathcal{A}$ now contains the response and predictor variables used to regress a functional of $\check{\alpha}$ versus the sample index.

3. Remove possible outliers using[2]:

$$\mathcal{A} = \left\{ (\check{\alpha},p) \in \mathcal{A} \;\middle|\; \begin{array}{l} \check{d}_p > 4\,\check{d}_q \\ \check{d}_p > 4\,\check{d}_r \\ 0.5 < \check{d}_p < 1.5 \end{array} \right\} \tag{5}$$

4. Regress using two second order splines, with smooth joints at β_2 ($\beta_1 = -\inf$) [14]:

$$\tilde{\alpha}_i = \sum_{k=1}^{2}\sum_{j=0}^{2} b_{kj}(p_i - \beta_k)_+^j + \epsilon_i \tag{6}$$

where $\bullet_+ = \max(0,\bullet)$. For simplification we assume at this point that $E[\epsilon_i] = 0$, which is not accurate since we know that the MSP's will seldom lead to overshooting values for α, while it is common to find smaller estimates as explained earlier.

5. Adjust α_i to minimize overshooting by using the expected value of the residual error:

$$\hat{\alpha}_i = \tilde{\alpha}_i + \sqrt{E[\epsilon_i^2]}. \tag{7}$$

The two panels of Figure 3 show the spline regressed ($\tilde{\alpha}(i)$) using a dashed line for a channel of two different data sets sequenced using different chemistries and electrophoresis technologies; primer dyes and gel electrophoresis (PD) on the top, terminator dyes and capillary electrophoresis (TD) at the bottom. Small dots represent MSPs that have not passed the tests noted in Eq. (5). Left (right) pointing triangles are signal values at left (right) MSPs. The data confirms two typical characteristics of DNA chromatograms: (i) peak shapes are slightly skewed to the left, and (ii) tails are usually expected after the allele. The continuous line shows the final estimate used for $\hat{\alpha}(i)$ after the correction of Eq. (7). Our interpretation of these curves is the following: Let k be the base symbol index. Generally, we can assume that the spread is $\propto k$ except[3]: (i) For short fragments ($k < 100\ bp$) the uncertainty is driven mostly by imperfections in the loading phase of the

[2]Note that when this step is applied the height of the chromatogram has already been normalized.

[3]We can interchange i with k without any loss of generality since estimating the spread is applied after the chromatogram has been preprocessed to have constant inter-base distance for the complete run.

experiment rather than on the dispersion of the molecules while traveling in the media [15]. (ii) Also, for very large fragments ($k > 1000\ bp$) the loss of resolution is so severe that it is difficult to find slopes of peaks non interacting with other alleles or artifacts. The scatter plots shown in Figure 4 summarize the residual analysis for the respective data sets. In both cases the residuals profile indicates an acceptable regression operation.

As a corollary of this section; *Peak resolution* (R), a typical parameter used by the analytical chemistry community to measure the quality of electrophoresis data, can be computed using $\hat{\alpha}_i$ as follows:

$$R(i) = \frac{\text{pulse width}}{t_j - t_{j-1}} = \frac{2}{\hat{\delta}}\sqrt{\frac{\ln(2)}{\hat{\alpha}_i}}. \quad (8)$$

This is true because the expected inter-base space has already been normalized to $\hat{\delta}$, and the pulse width, by definition, is the width of a single base peak at half-height. For the examples shown in Figure 3 the unity resolution threshold occurs at $\alpha = 0.018$ for the PD data set and $\alpha = 0.012$ for the TD data set.

3. FEATURE EXTRACTION

In this section we describe a method for extracting possible features of the chromatogram that might represent an allele, and therefore a base symbol of the underlying DNA sequence. Such feature vectors can be statistically characterized to decide if they correspond to base symbols or not at a later stage of the processing, as we show in [5].

Note that we do expect the method to also extract small peaks of the signal due to: (1) additive noise introduced by the measurement process, 2) presence of small true conglomerations of fragments caused by chemical contamination or observed because of the overlapping spectra of the photo detectors used for the different channels. The extraction of these features at this point does not represent a problem, because a subsequent pattern recognition base-calling stage will evaluate all possible combinations of features which most likely reassemble the correct underlying sequence [5]. During the preprocessing stage we prefer not to exclude features related to artifacts, to ensure that we feed enough information for evaluation to the subsequent decision stage. This action leads to minimizing the probability of introducing deletions (under-calls) at the early stage of the processing.

On the other hand, large conglomerations of fragments, appearing as a single broad peak, present a problem because they may correspond to more than one base symbol. In long read-length DNA chromatograms it is not uncommon to find five or more base symbols forming a very wide peak. In [6] we tried to tackle this challenging problem by letting the statistical characterization of peak events to cover also the case where a single peak may represent two, or more, base calls. However, we have realized that in the presence of severe peak overlapping there is not enough evidence to support learning accurately the parameters of models for large peaks, even with the aid of a *Bayesian* approach. Keep in mind that we employ an unsupervised learning framework where model parameters are estimated from the data set exclusively, so that the base-caller can adapt to different types of chemistries and sequencing protocols on the fly and without the need for extensive re-calibration (as is the case with *Phred*). In addition, under low SNR conditions, broad peaks tend to get over-segmented; e.g. a large bump with three immersed base symbols and some noise on the flattened top may lead our segmentation algorithm to extract

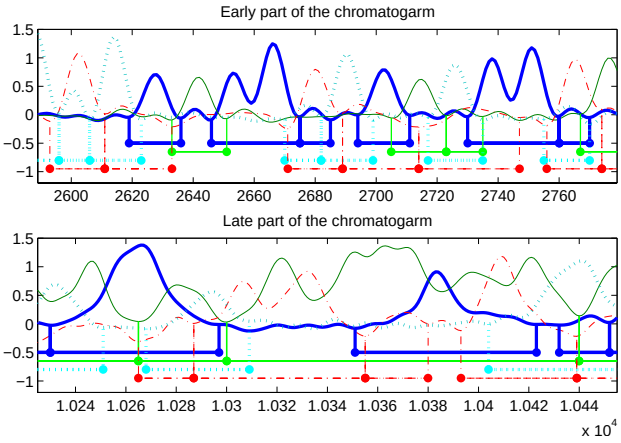

Fig. 5. Example of an early (top) and late (bottom) part of a typical DNA chromatogram with the extracted *segments* shown for the four channels.

only two peaks of unexpected size for further statistical characterization. Small conglomerations of fragments, caused by contamination, tails, and cross-correlation, can also be attached to broad problematic bumps.

Following the divide and conquer strategy we will attempt to unmix any wide peak that might represent more than one base symbols. For this purpose, we will take advantage of two facts: (1) we have at this point already re-sampled the chromatogram and expect evenly spaced base symbols [1, 5], (2) we have estimated the trend of the spread (as discussed in the previous section).

Let us define a *segment* as a small continuous section of a channel in the chromatogram which may contain: (1) a small artifact, (2) a single DNA allele, (3) two or more partially mixed alleles, (4) two or more completely overlapped alleles, and, (5) a combination of partially and completely overlapped alleles. Figure 5 shows examples of all previous cases in two sections (early (top) and late (bottom)) parts of a typical DNA chromatogram. A segment representing one or more alleles must contain all samples of the signal which correspond to sensed fragments of the corresponding lengths, but must not contain samples due to other length strands, i.e. a segment must be completely unmixed from other segments. Observe that the two alleles around sample 2680 are considered as a single segment; although it appears evident that two base symbols are present, we will still feed such segments into the un-mixing processing block in order to properly estimate the area (also called *weight*) of each one as required for accurate base-calling. In less evident cases, part of a segment could be a small artifact or a tail.

To select the segments in a lane (channel) of the chromatogram we first find all valleys

$$\mathcal{V} = \left\{ i \in \mathcal{N} \;\middle|\; \begin{array}{l} \check{d}_{i-1} > \check{d}_i \\ \check{d}_{i+1} > \check{d}_i \end{array} \right\}. \quad (9)$$

A segment is delineated by a pair of indices $(q, r) \in \mathcal{V}^2$ such that q (r) is the first valley to the left (right) of p whose amplitude satisfies $\check{d}_{\tilde{q}} < 0.1\check{d}_p$ ($\check{d}_{\tilde{r}} < 0.1\check{d}_p$), where p is the index to maximum value within the given segment. Some segments can be immediately rejected since they do not contain true alleles representing a base symbol. Those that are are retained should are in the set:

$$\mathcal{S} = \left\{ (q, r) \in \mathcal{S} \,\middle|\, \check{d}_p > 0.2 \right\}. \quad (10)$$

We now use Eq. (1) to model the mixed signal of alleles in a segment. We impose our notion of evenly spaced symbols and constant spread:

$$\hat{d}_i = \sum_{j=0}^{N_S-1} x_j e^{-\hat{\alpha}(i-(j\hat{\delta}-\tau_0))^2} + \epsilon_i \qquad \forall\, i = [q:r] \tag{11}$$

or in matrix form,

$$\hat{\mathbf{d}} = \mathbf{W}(\tau_0)\,\mathbf{x} + \epsilon, \tag{12}$$

where $\frac{1}{\hat{\alpha}}$ is the already estimated spread (assumed constant in a segment), $\hat{\delta}$ is the already estimated *inter-symbol* distance (assumed constant in a segment), $\epsilon_i \sim N(0, \sigma_\epsilon)$ as supported by our previous studies [5, 1], $N_S = \left\lfloor \frac{r-q}{\hat{\delta}} \right\rfloor$ is the number of peak shapes to be fitted in the segment. Eq. (11) represents a sequence of peak hypothesis that might be active within a given segment, note that there might be a peak shape hypothesis in a region where there is not an evident allele. The unknown parameters are $(\mathbf{x}, \tau)$. x_j are the unknown peak shape *weights* , which will later allow us to statistically characterize every peak hypothesis. To avoid overfitting we impose the realistic restriction $x_j \geq 0, \forall j$, i.e. we prevent hypotheses which are realistically not very possible to take negative weight values and disturb the correct estimation of x_j for more likely hypotheses. τ is the location of the center of the left-most peak shape ($j = 0$) in a segment.

To find the unknowns $(\mathbf{x}, \tau)$ given the observed signal $(\check{\mathbf{d}})$ in the least square sense we need to solve

$$\begin{gathered}(\hat{\mathbf{x}}, \hat{\tau}) = \underset{\mathbf{x},\tau}{\operatorname{argmin}} \left(\left|\left|\check{\mathbf{d}} - \mathbf{W}(\tau)\,\mathbf{x}\right|\right|\right) \\ \text{subject to } \mathbf{x} \geq \mathbf{0}, \tau \in [q : q + \hat{\delta}].\end{gathered} \tag{13}$$

If τ can be accurately pre-estimated, Eq. (13) becomes:

$$\begin{gathered}\hat{\mathbf{x}}(\tilde{\tau}) = \underset{\mathbf{x}}{\operatorname{argmin}} \left(\left|\left|\check{\mathbf{d}} - \mathbf{W}(\tilde{\tau})\,\mathbf{x}\right|\right|\right) \\ \text{subject to } \mathbf{x} \geq \mathbf{0},\end{gathered} \tag{14}$$

which is a typical *Non-Negative Least Squares* (NNLS) formulation [16]. Therefore, to solve Eq. (13) we propose a directed exhaustive search for τ performed as follows:

Let k be the index of the current segment (q, r), and $\hat{\delta}$ be the expected inter-base distance.

1. Predict an initial estimate of the location of the peaks (τ) using the peaks observed in the M previous segments in the same channel[4], in practice we use $M = 5$:

$$\tau_k = \frac{1}{M} \sum_{l=k-1}^{k-M} \left\{ \left\lfloor \frac{q - \tau_l}{\hat{\delta}} \right\rfloor \hat{\delta} + \tau_l \right\}. \tag{15}$$

2. Find $\hat{\mathbf{x}}$ for τ_k using the NNLS algorithm presented in [16].

3. Calculate the search direction and step for τ_k by looking at the derivative of the least squares error with respect to τ_0:

$$\begin{gathered}\nabla = \tfrac{\hat{\delta}}{15} \text{sign}\left(\hat{\mathbf{x}}^T \tfrac{\partial \mathbf{W}(\tau_k)}{\partial \tau_0}^T \left[\check{\mathbf{d}} - \mathbf{W}(\tau_k)\hat{\mathbf{x}}\right]\right), \\ \text{where sign}(\bullet) = \begin{cases} 1 & \bullet > 0 \\ 0 & \bullet = 0 \\ -1 & \bullet < 0 \end{cases}.\end{gathered} \tag{16}$$

4. Compute $\tau_k^{new} = \tau_k + \nabla$,

if $\tau_k^{new} < q \rightarrow \tau_k^{new} = q$ else if $\tau_k^{new} > q + \hat{\delta} \rightarrow \tau_k^{new} = \tau_k^{new} + \hat{\delta}$.

5. Find $\hat{\mathbf{x}}^{new}$ for τ_k^{new} using the NNLS algorithm [16].

6. If $\left|\left|\check{\mathbf{d}} - \mathbf{W}(\tau_k^{new})\,\hat{\mathbf{x}}^{new}\right|\right| < \left|\left|\check{\mathbf{d}} - \mathbf{W}(\tau_k)\,\hat{\mathbf{x}}\right|\right|$

then $\tau_k = \tau_k^{new}$, $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{new}$, and goto step 4.

7. The computation is completed. Output estimates $\hat{\mathbf{x}}$ and τ_k.

The iterative search for τ_k may give to the reader the false impression of a compute intensive algorithm, since NNLS is solved for every τ_k. However, there are several factors leading to acceptable performance in our case[5]: (1) Only a fraction of the segments have mixed alleles, (2) Since inter-base distance has been normalized, the expected jitter is small and thus the initially predicted τ_k is close to the optimal, (3) low pass filtering in the pre-processing stage reduces the chances of getting trapped in a local minimum, and, (4) QR decomposition can be implemented in an efficient manner, since only one column vector is added or deleted at every iteration of the NNLS algorithm [16].

4. EXPERIMENTAL RESULTS

Figure 6 demonstrates the unmixing capabilities of the proposed algorithm with three different segments of a primer-dye gel electrophoresis chromatogram (it is well known that resolution is worse in primer-dye chemistry than terminator-dye chemistry). The top left panel shows a single allele with a tail in the early part of the trace. The top right panel shows two partially overlapped alleles followed by a third conglomeration of fragments which could be a tail or even a manifestation of a heterozygous situation; a final decision will be made after analyzing the other channels. Two additional allele hypotheses are also extracted at the extremes of the segment, but their weight values are close to zero, so even if they make their way to the base-caller stage, most likely they will be rejected. The bottom panel shows a broad peak at the late part of the chromatogram containing six base symbols. The resolution in this window is $R(i) = 1.36$. Two more hypotheses have been extracted by the algorithm at the ends of the segment which were, however, rejected by the later decision stage of the base-caller (non-active hypotheses). A ninth peak location (around the sample number 12760) is considered by the NNLS algorithm but will not be passed to the base-caller stage because its weight was exactly equal to zero.

The final set of extracted features is the union of all the unmixed allele weights in all four channels. This list of *events* is further analyzed by the base-caller in order to estimate if each one of them represents an actual base symbol or not. We anticipate that, in addition to our in-house FOVD base-caller that is ground on probabilistic graph theory [7], other base-callers using statistical methods proposed in the literature [8, 12] may benefit from the availability of such a list of potential bases.

The FOVD base-caller has already shown very promising results using the proposed feature extraction method. After testing our end-to-end pre-processing, feature extraction, FOVD base-calling approach we found that it can achieve 25% less errors than *Phred* [13] for long read-length sequences (> 800bp). In the evaluation we used a large and representative pool of M13mp18 chro-

[4] In the absence of M segments we can use the local maxima, recall that at the beginning of the chromatogram peaks usually have good resolution

[5] Segmentation of a data set lasts approximately 15 sec. on a 3GHz Pentium machine using a non-optimized Matlab implementation

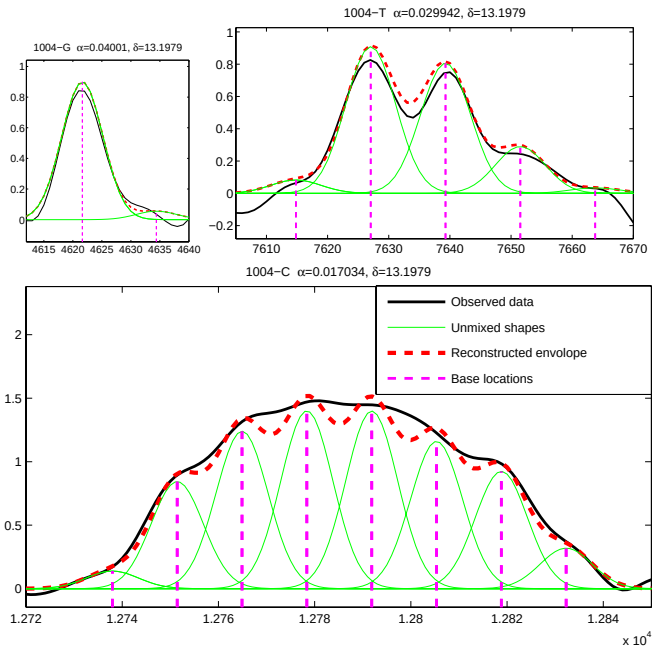

Fig. 6. Unmixing alleles with NNLS. Top left panel: single allele with tail. Top right panel: Two partially overlapped alleles with a third small artifact. Bottom panel: Six alleles immersed in single broad peak.

matograms. Due to space limitations we omitted further discussion on this comparative evaluation, but some details have already appeared in [4, 17].

5. CONCLUSIONS

We have presented a novel method for extracting potential landmarks in a DNA chromatogram which can be used for base-calling purposes. A decision algorithm will subsequently need to disambiguate between different combinations of extracted peak hypothesis. The presented feature extraction strategy assumes that the inter-peak distance has already been equalized using methods as those discussed in [1]. We first show that it is possible to estimate the diffusion of the peak shapes (spread) towards the end of the chromatogram with good accuracy and without knowledge of the underlying DNA sequence. With this information available only the actual quantity of DNA fragments, in a subpopulation of fragments with fixed length, remains unknown. This quantity can be estimated using an iterative NNLS algorithm that computes peak shape weights (normalized amplitude) and refines the peak locations. The extracted feature vectors are fed to a statistical base-caller (FOVD) which provides decisions and confidence values for the generated base symbols by employing a generalized forward-backward probability propagation method. The resulting end-to-end base-calling strategy can routinely exceed Phred's accuracy considerably especially at long readlengths.

6. REFERENCES

[1] L. Andrade-Cetto and E. Manolakos, "Inter-peak distance equalization for DNA sequencing," 2005, in preparation.

[2] L. Alphey, *DNA Sequencing: From Experimental Methods to Bioinformatics*, Springer-Verlag, 1997.

[3] L. Andrade and E. Manolakos, "Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 35, no. 3, pp. 229–243, Nov. 2003, *special Issue on Signal Processing and Neural Networks for Bioinformatics*.

[4] L. Andrade and E. Manolakos, "Robust Normalization of DNA Chromatograms by Regression for improved Base-calling," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 3–22, 2004, *special Issue on Genomic Signal Processing*.

[5] L. Andrade-Cetto, *Analysis of DNA Chromatograms using Unsupervised Statistical Learning Methods.*, ECE Department, Northeastern University, 2005, Ph.D. dissertation in preparation.

[6] M. Pereira, L. Andrade, S. El-Difrawy, B. Karger, and E. Manolakos, "Statistical learning formulation of the DNA base-calling problem and its solution using a Bayesian EM framework," *Discrete Applied Mathematics*, vol. 104, no. 1–3, pp. 229–258, 2000.

[7] L. Andrade-Cetto and E. Manolakos, "A Graphical Model formulation of the DNA Base-calling problem," in *IEEE Workshop on Machine Learning for Signal Processing*, 2005, *submitted*.

[8] D. Brady, M. Kocic, A. Miller, and B. Karger, "Maximum Likelihood Base-Calling for DNA sequencing," *IEEE Trans. on Biomedical Engineering*, vol. 47, no. 9, pp. 1271–1280, 2000.

[9] S.W. Davies, M. Eizenman, and S. Pasuphaty, "Optimal structure for automating processing of DNA sequences," *IEEE Trans. on Biomedical Engineering*, vol. 46, no. 9, pp. 1044–1056, 1999.

[10] N.M. Haan and S.J. Godsill, "Modeling electropherogram data for DNA sequencing using variable dimension MCMC," in *ICASSP 2000, Istanbul*. 2000, pp. 3542–3545, IEEE.

[11] N.M. Haan and S.J. Godsill, "Bayesian Models for DNA sequencing," in *ICASSP 2002, Orlando, Florida*. 2002, pp. 4020–4023, IEEE.

[12] S. El-Difrawy P. Boufounos and D. Ehrlich, "Hidden Markov Models for DNA sequencing," in *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, 2002, pp. CP1–16:1–4.

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

[14] C.L. Schumaker, *Spline functions : basic theory*, John Wiley and Sons, 1981.

[15] E. Yarmola, H. Sokolof, and A. Chrambach, "The relative contribution of dispersion and diffusion to band spreading (resolution) in gel electrophoresis," *Electrophoresis*, vol. 17, pp. 1416–1419, 1996.

[16] C.L. Lawson and R.J. Handson, *Solving Least Squares Problems*, Prentice-Hall, 1974.

[17] L. Andrade and E. Manolakos, "Automatic Estimation of Mobility Shift Coefficients in DNA chromatograms," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, Sept. 2003.