SIMULATION OF “SINGLE MOLECULE REAL TIME” (SMRT) SEQUENCING, AND DATA TRANSFORMATIONS FOR USE WITH EXISTING WORKFLOW TOOLS

Robert M. Horton, Ph. D.

Attotron Biotechnologies Corporation,
Carson City, NV 89706, USA
Email: rmhorton@cybertory.org

Single Molecule Real Time (SMRT) sequencing is an innovative and potentially transformative next-generation approach for determining DNA sequences. It promises high accuracy, high throughput, long read lengths, and the ability to resolve independent alleles from mixed templates. The new type of data produced from this approach requires development of appropriate analytical algorithms and data management tools. Because the technology is based on measurements from single molecules, stochastic effects have great impact on the observed results, and require statistical treatment. Here I present a computer model of this process, written in the R programming language. This makes it possible to generate synthetic data with which to assess various analytical approaches, and to experiment with reaction parameters to observe their expected effects on data quality. I also present methods to transform simulated SMRT data into a format compatible with existing workflow tools, including the Phred base caller and the Staden package for assembly and finishing.

1. INTRODUCTION

SMRT DNA sequencing is an exciting new technology under development at Pacific Biosciences corporation.[1, 2] Relative to competing approaches, it offers long read length, high throughput, small reaction volumes, massive parallelism, and rapid results. It also allows direct distinction of alleles and isomorphic genes in mixtures; this would be quite useful in many genotyping applications, such as tissue typing. Although the fraction of polymerase molecules successfully reading their templates decreases over time, the quality of the surviving reads does not decline with distance from the primer. This makes it possible to generate longer contiguous reads than from approaches that rely on populations of molecules. With these potential advantages, SMRT sequencing may represent a dramatic breakthrough for practical application of individualized genomics.

The SMRT approach is illustrated in Figure 1. A single molecule of polymerase is immobilized within a discernible volume of reaction solution. The polymerase binds a conventional primed template, and extends the primer by incorporation of nucleotides from γ-labelled dNTPs, where each of the four bases is color-coded with one of four distinguishable fluorophores. During incorporation, each successive nucleotide is held in place on the polymerase complex for a period of time before its label is released.

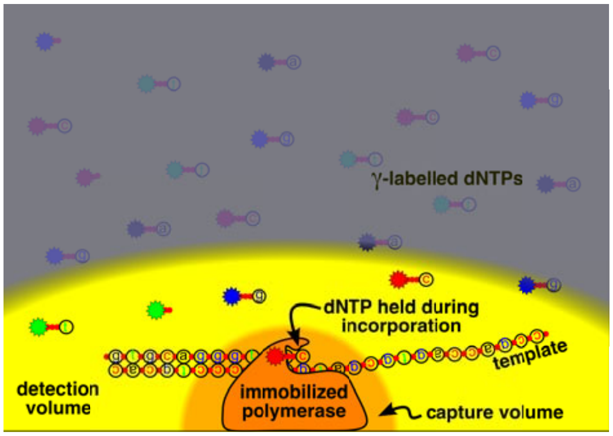


Fig. 1. Overview of the SMRT sequencing approach. A single DNA polymerase molecule is anchored to a substrate. A small volume of solution around the polymerase is illuminated so that fluorophores within this volume can be detected. Fluorescent nucleotides are labelled at the gamma phosphate, so the label is released once the nucleotide is incorporated. While fluorophores free in solution diffuse in and out of the detection volume rapidly, the nucleotide being incorporated is held in the detection volume for a much longer time. Reading sequence with this approach depends on being able to identify the signal emitted by the immobilized base during incorporation above the background fluorescence of unincorporated nucleotides; this can be posed as a statistical problem.

A small amount of solution surrounding the polymerase molecule is exposed to UV light to excite the fluorescent reporters. Unbound fluorophores diffuse in and out of this volume, giving transient fluorescent emissions, while the fluorophore from the nucleotide bound to the polymerase is trapped within the detection volume for a much longer time. The technique relies on the ability to detect the single dNTP being incorporated in the presence of free fluorophores in the detection volume. The number of free fluorphores depends on both the detection volume and on the concentration of dNTPs. In practice, dNTP concentrations must be kept high enough to meet the requirements of the polymerase, which means that the detection volume must be quite small. A small volume is achieved by limiting the illuminated area to an aperture similar in scale to the excitatory wavelength; this limits both the area and the depth of penetration of the excitatory light into the reaction solution.[1]

1.1. Conceptual model

We can model this process in stages. First, consider the "pure signal" function, reflecting which type of flurophore, if any, is bound to the polymerase at any given time. The binding site can hold only one molecule of dNTP, in one of the four colors, so this function will have a value of either 0 or 1 in each channel, and the channels will not overlap. The 'peaks' will be square, with width determined by the amount of time the reporter from the nucleotide being incorporated is trapped on the polymerase. Assuming there is some "refractory period" between releasing the label from one dNTP and being ready to capture the next, the peaks will also be guaranteed to have some minimum space between them.

The spacing between square peaks is determined by the time required for the polymerase to capture a nucleotide to match the next position in the template, plus the refractory period. Assume that the polymerase has access to a certain volume of solution in its immediate neighborhood (its 'capture volume'), and that it can quickly capture any dNTP that diffuses into that volume. Then time can be expressed in terms of the number of samplings it takes to find the required nucleotide. If sampling the dNTPs in the capture volume is a Poisson process, the interval between arrivals of the desired base will follow an exponential distribution.[3]

Two types of variability obscure this signal: *background* from unincorporated nucleotides, and *noise* from measurement variation.

Background is estimated using a simple stochastic model of the number of unincorporated nucleotides in the detection volume. The reaction solution represents a large population of molecules, where the nucleotide concentration describes the average number of dNTPs per volume. Because the detection volume is very small, the number of molecules it contains at any instant is a small sample taken from this population, and can be modeled with a Poisson distribution. Assuming that diffusion in and out of the detection volume is fast relative to the time required to measure the fluorophores, we can treat each measurement as an average of many independent instantaneous samples (*snapshots*), and avoid modeling diffusion.

Finally, the process of measuring the fluorescence intensity is itself stochastic, where the numbers of emitted photons can be modeled using a Poisson distribution, as can *shot noise* from the photon detection apparatus during signal amplification. While the numbers of fluorophores in the detection volume for any given snapshot are whole integers, measured intensities are not. This is because they are averages over many snapshots, and because the intensities of photon and shot noise quanta are on different scales.

2. MATERIALS AND METHODS

This model was implemented in the R programming language, which has excellent facilities for statistical analysis and graphics. Source code is available from the cybertory.org website.[4] The R interpreter is freely available for all major computing platforms.[5]

2.1. Reaction parameters

The general description of the SMRT process, and ballpark estimates of reaction parameters, are based on statements gathered from the Pacific Biosciences website[1]:

- fluorescence detection limits: low nanomolar concentrations
- detection volume: 20 zeptoliters (20e-21 l)
- incorporation time: tens of milliseconds
- dNTP visit time: a few microseconds

2.2. Data structures

Data for each step of the simulation is stored in a *ChannelIntensities* data structure, which is a list of four vectors of floating-point intensity values, one for each color channel (reflecting one of the bases A, C, G, or T). Each indexed position in the vector represents a discrete time point. The model is implemented as a series of transformations applied to ChannelIntensities structures, as are the analyses of the model output.

2.3. Data filtering and smoothing

A filter-style method takes a ChannelIntensities structure as input, modifies it in some way, and returns the modified version. Several of these filters apply a sliding window of fixed width, where the output value of the central point is a function of all the points in the input window. The size of the window is described by its radius, the number of input data points to consider on either side of the central output point. Examples of simple filters include taking the mean of the sliding window to perform simple smoothing, or taking the minimum or median value. Such filters make it easy to experiment by combining them in various ways.

2.4. Statistical model

Plotting the intensities from one channel of a complete read as a histogram shows signal and background peaks. (Four such histograms, from reactions run at four different dNTP concentrations, are shown in Figure 4.) At higher concentrations, the peaks are shifted toward higher average intensities, and they become broader. We can fit Gaussian curves (each described by mean, standard deviation, and height) to the histogram as follows: the background mean is located at the highest point of the histogram. The signal peak is one intensity unit greater than the background, representing the single fluorophore from the dNTP being incorporated. The area under the background peak will be approximately three times greater than that under the signal peak for a given channel, depending on the base composition.

In a Poisson distribution, the mean and the variance are the same (standard deviation is the square root of the variance). The standard deviation of the distribution of sample means is the population mean divided by the square root of the number of samples.[6] The standard deviation of the signal and background peaks is estimated as the square root of the mean (since the underlying distribution is Poisson), divided by the square root of the number of samples, where the number of samples is the product of the number of snapshots per measurement and the number of photons per fluorophore (a model parameter). Since the variances of the two peaks are similar, the relative areas correspond to relative heights.

2.5. Statistical filtering

Given the positions of the signal and background peaks, we can use sliding window filters based on statistical tests. First, we use a one-sample t-test where the output of the sliding window is the p-value for the sample of measurements coming from the signal population. This is divided by the similarly computed probability that the sample comes from the background population, and the logarithm is taken. Positions where the log of the probability ratio is positive are more likely to represent signal than background. This log odds ratio is smoothed by averaging, and only the positive values are retained, showing the areas more likely to come from signal than from background.

2.6. Sequence Chromatogram Format (SCF)

SCF is a well-documented, non-proprietary binary format for storing results of Sanger sequencing experiments.[7] This format can be read and written by most of the software tools commonly used for managing DNA sequencing projects.

Because intensities are recorded as 16-bit integer values in SCF, the floating-point values from a ChannelIntensities structure need to be scaled to fit into a range of values between 0 and 65535. The floating point values represent numbers of fluorphores, plus or minus some measurement variation, and will generally be less than about 20 (depending on the dNTP concentration); multiplying by about 1000 generally produces acceptable scaling.

2.7. Base calling and assembly

Base calling was performed with Phred.[8, 9] This program is customized for different chemistries, dyes, and machines, based on tags it reads from the chromatogram file. The simulated chromatograms leave these tags blank, so a default was set by using the `-process_nomatch` directive and adding the following line to phredpar.dat:

```
""   primer   big-dye   ABI_3700
```

Called SCF files were aligned using programs `pregap4` and `gap4` from the Staden package.[10]

3. RESULTS

The output of two simulated SMRT sequencing reactions are shown in Figures 2 and 3. The top panel in each figure shows the "pure" signal, while the second panel is "raw output", with background and noise added. Further analysis of the high [dNTP] data is shown in Figures 4 and 5. Times given are approximately microseconds, but are labelled *bogoseconds* to reflect the fact that the parameters in this model are estimates taken from general descriptions, and have not been fitted to actual experimental data.

3.1. Effects of nucleotide concentration

Figures 2 and 3 show reactions at low and high dNTP concentration. A symptom of low [dNTP] is stochastic peak spacing, which can complicate interpretation of the experimental data. If [dNTP] is too high, background fluorescence begins to obscure the signal.

Intensities on the y-axes represent measured numbers of fluorophores; values are not whole numbers because measurements are averages of multiple snapshots, and because of stochastic effects of the numbers of photons emitted by the fluorophores, and shot noise during measurement.

3.2. Simple filtering of intensity data

Some simple approaches to signal filtering are shown in the lower panels of Figures 2 and 3. Various transformations of the raw data can reveal the signal peaks. Here, a filter taking the minimum of a sliding window is applied, followed by one taking the mean to smooth out the resulting curve. Excessive smoothing flattens the peaks (not shown). Signal recovery is generally more challenging at higher [dNTP].

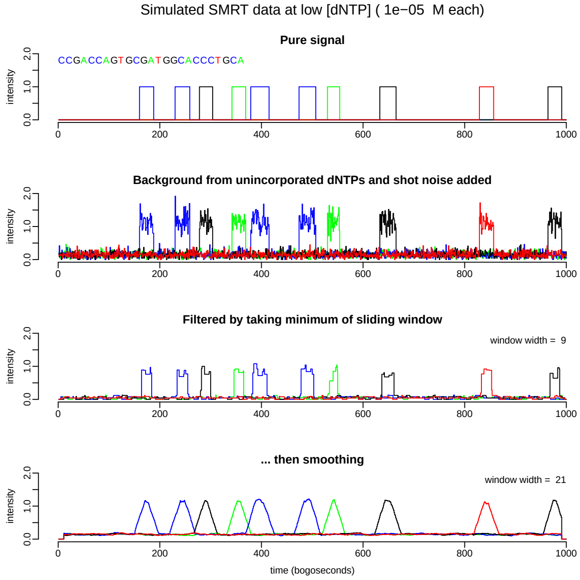


Fig. 2. Simulated SMRT results at low nucleotide concentrations. Note the stochastic peak spacing, and the ease with which the signal can be distinguished from background.

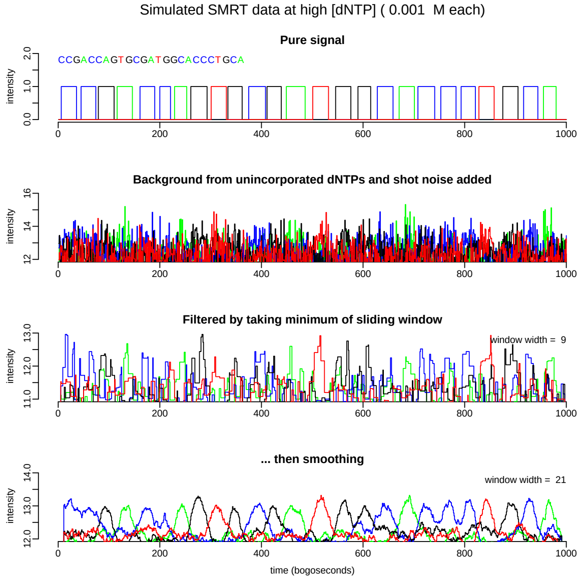


Fig. 3. Simulated SMRT results at high nucleotide concentrations. This figure is like the previous one, but run with a higher [dNTP]. The peaks are much more regularly spaced, but the signal is more difficult to distinguish from the noise. Note the ranges of the y-axes.

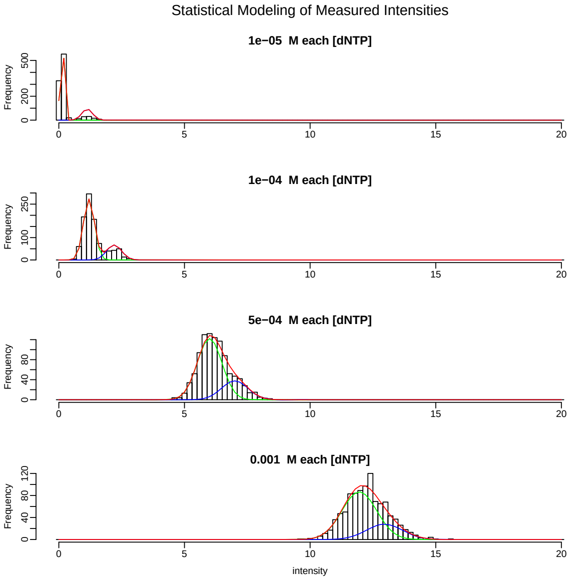


Fig. 4. Fitting probability distributions to measured intensity values. Intensity measurements from a single channel of the high [dNTP] reaction in Figure 3 are plotted as a histogram. Two normal curves, representing signal and background, are fitted to this data. Their sum is shown in red.

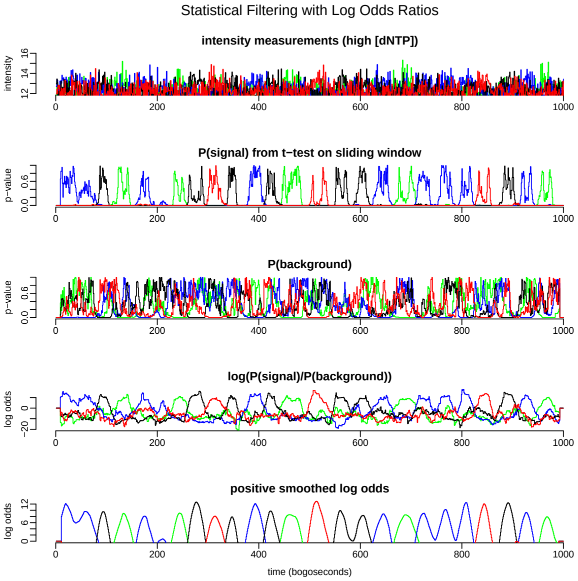


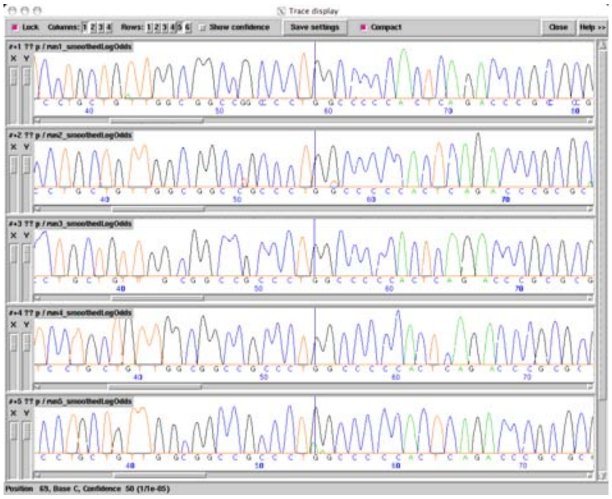
Fig. 5. Statistical filtering. A sliding-window t-test transforms channel intensities into probabilities of the samples in the window coming from the signal population. The probability that the sample comes from the background population is calculated similarly in the third panel. Taking the logarithm of the ratio of the two p-values gives curves that reveal the original signal (fourth panel). In the bottom panel, this curve is smoothed by averaging, and limited to positive values.

3.3. Statistical filtering

As seen in Figure 4, the sum of the two Gaussian curves fitted to the background and signal peaks outlines the histograms quite well. Figure 5 applies the statistical filtering approach to the data from the high [dNTP] reaction. The clean, easily interpreted curves of the bottom panel are reminiscent of Sanger chromatograms; this is the transformation used to generate SCF files.

3.4. Base calling and assembly

Even using data with intentionally high background, with arbitrarily selected default parameters (designed to read Sanger sequencing reactions run with big dye primers on an ABI 3700), the base caller achieves an accuracy on the order of about one miscalled base per hundred.


Trace display
Lock
Columns: 1 2 3 4
Rows: 1 2 3 4 5 6
Show confidence
Save settings
Compact
Close
Help >>
#1 ?? p / run1_smoothedLogOdds
#2 ?? p / run2_smoothedLogOdds
#3 ?? p / run3_smoothedLogOdds
#4 ?? p / run4_smoothedLogOdds
#5 ?? p / run5_smoothedLogOdds
40
50
60
70
80

Position 69, Base C, Confidence 50 (1/1e-05)

Fig. 6. Transformed simulated SMRT data in SCF format can be successfully managed and analyzed using existing software designed for Sanger chromatograms. Data from multiple independent simulated SMRT reads from the same template was transformed by statistical filtering and smoothing, exported to SCF files, run through the Phred base caller, and loaded into the Staden package to align and compare the traces. This illustrates use of existing tools, originally designed for Sanger sequencing, to assemble SMRT results and facilitate examining the data to resolve base calling discrepancies.

Phred adds the called bases to SCF files, which can then be aligned and displayed using Staden, as shown in Figure 6.

4. DISCUSSION

This simulation is helpful for understanding the SMRT sequencing process because it makes it possible to conduct conceptual experiments with model parameters and observe the expected effects on output. It may useful for teaching the principles of the approach, and possibly for training new users.

It would be interesting to see how well the model parameters could be fitted to actual SMRT data, and to evaluate base calling accuracy on real data using the transformation process described. Parameters of Phred (such as the expected background level and peak spacing) could be optimized to SMRT data. Alternatively, the data transformation might be adjusted to better suit the base caller. For example, some level other than zero might be chosen as a cutoff for the log-odds ratio, to adjust the baselines of the curves. As can be seen in the last panel of Figure 5, the bottoms of the peaks extend below zero, indicating that there may be useful information in the negative side of the curves as well.

Converting the results to a form that can be managed and analyzed using conventional tools is potentially useful for two main reasons. First, it may obviate the necessity of developing new data management infrastructure, including software tools and a workforce trained to use them. Second, it may allow analyses to combine sequencing results from different technologies. For example, in a shotgun sequencing project one might use high throughput SMRT sequencing to generate the vast majority of the data, while Sanger sequencing might be used in individual primer walking reactions to resolve ambiguities and connect contigs. This might be particularly useful in the early stages of adoption of the new technology, when types and rates of error are being evaluated.

This model currently ignores differences in nucleotide incorporation rates due to template sequence in the polymerase binding region.[11] This would add variability in peak spacing, though it would tend to be minimized at high dNTP concentrations. Adding such effects to the model would be straightforward.

More powerful data analysis approaches might add experimental flexibility to the SMRT technique. Specifically, better ability to read signals from high-background data would facilitate use of higher dNTP concentrations in reactions. This is desirable for several reasons. First, bases are read more quickly, which could improve throughput, minimize the damage done to the template and polymerase by UV light, and maximize processivity (and thus read length). Also, the more predictable peak spacing expected at high dNTP concentrations would be helpful for base calling, and for comparing data from different reads, as is commonly done when resolving ambiguities in Sanger traces.

Simulated data can be used to optimize analytical approaches for particular variations in experimental design. For example, if the illumination intensity were lowered, fewer photons would be emitted per fluorophore, increasing measurement variance. We can model this effect on the quality of output data. Improved ways of extracting information from this noisier data might make it possible to conduct experiments with lower excitation fluxes, lessening UV damage and further increasing read lengths.

References

1. Pacific Biosciences. `www.pacificbiosciences.com`
2. *Single Molecule Real Time (SMRT) DNA Sequencing Technology Backgrounder.* `http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf` Pacific Biosciences, 2009.
3. Poisson process. *Wikipedia, The Free Encyclopedia.* Available at: `http://en.wikipedia.org/w/index.php?title=Poisson_process&oldid=372071745`. Accessed July 7, 2010.
4. Horton, RM. *Cybertory*$^{\mathrm{TM}}$ *Educational Molecular Biology Simulations.* `http://www.cybertory.org`.
5. *The R Project for Statistical Computing.* `http://www.r-project.org/`
6. Brown B and Hollander M. *Statistics: A Biomedical Introduction.* John Wiley and Sons, 1977, p. 73.
7. Dear S and Staden R. A standard file format for data from DNA sequencing instruments. *DNA Sequence* **3**: 107-110 (1992)
8. Ewing B, Hillier L, Wendl MC, and Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* 1998; **8**:175–185.
9. Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 1998; **8**:186-194.
10. Staden Package `http://staden.sourceforge.net`
11. Turner S. “Single Molecule Real Time DNA Sequencing” (video 12:04) *Sequencing, Finishing, Analysis in the Future* meeting, Santa Fe, NM, May 27, 2009. `http://www.scivee.tv/node/11409`