2005

Imposed Constraints on the Smith-Waterman Alignment Algorithm for Enhanced Modeling of a Single-Molecule DNA Sequencer

Patrick G. Humphrey
LICOR, Inc.

Gregory R. Bashford
University of Nebraska - Lincoln, gbashford2@unl.edu

Imposed Constraints on the Smith-Waterman Alignment Algorithm for Enhanced Modeling of a Single-Molecule DNA Sequencer

Patrick G. Humphrey[1] and Gregory R. Bashford[2,*]
[1]*LICOR, Inc., Lincoln, NE*
[2]*Department of Biological Systems Engineering, University of Nebraska-Lincoln*
**corresponding author: gbashford2@unl.edu*

Abstract

An effort has been underway to develop a system for de novo sequencing of single DNA molecules with very long reads. The system operates by optically detecting the passage of fluorescently tagged DNA bases through a detection zone. A successful system would be revolutionary with respect to speed, read length, cost and minimized laboratory infrastructure. An important part of system development is modeling of the detection process. In particular, predicting the expected error from a set of sequencing parameters is helpful in system design. This paper describes variations on the Smith-Waterman algorithm for subsequence alignment used in a single-molecule detection model. The alignment algorithm is used to check the modeled output sequence generated from a known input sequence. Variations based on reasonable assumptions led to over an order of magnitude improvement in alignment speed.

1. Introduction

We have been developing a system for real-time single molecule DNA sequencing. A key advantage of single-molecule sequencing is the elimination of cloning and the laboratory infrastructure associated with high-throughput operations. This, along with reduced reagent consumption, results in sequencing costs orders of magnitude lower compared to existing methods. Read lengths will be tens of kilobases to simplify shotgun sequence assembly and preserve haplotype information. Applications include whole-genome sequencing, SNPs, haplotyping, genotype-trait associations, long-read SAGE for expression profiling, analysis of alternative mRNA splicing patterns, and comparative genomics within or between species. Fields-of-use include research, diagnostics and personalized medicine.

System components include novel "charge-switched" nucleotides, an adapted DNA polymerase, a method for isolating and handling single DNA molecules, a microfluidics flowcell for sorting molecules by charge, a TIR (total internal reflection) optical system for single-molecule detection in four spectral channels, and software algorithms for single molecule detection and system control.

A typical difficulty in single-molecule detection is sensitivity to the very low level of signal available. Normally single-molecule detection is performed by conjugating a fluorescent tag to a biomolecule of interest. Point detection of single molecules is well established [1, 2] and several groups have imaged single molecules in two dimensions [3, 4]. Attempts have been made to track single molecules frame-to-frame with less success. The main problem is diffusion of molecules out of the depth of field [5] and within the image plane [6], both of which limit the tracking time. This has restricted tracking to larger molecules such as proteins and viruses that diffuse more slowly (on the order of 1 μ^2/s) [7, 8]. Restricting the volume that the molecule may diffuse in increases the observation time (by decreasing the effective diffusion coefficient), which has been accomplished by point detection in drawn capillaries [9].

By using submicron channels with a flat optical surface suitable for TIR optics, we have extended these results to the two-dimensional imaging and tracking of single-fluorophore molecules (bulk diffusion ~ 200 μ^2/s) in free solution. The channels are on the order of 2 microns wide by 0.5 microns deep. A laser is used to excite a small area (a few hundred microns wide). As the fluorophores travel through this excitation site, they emit photons which are captured by a highly sensitive, cooled CCD camera. Each fluorophore generates a small spot in an image (Figure 1) according to the PSF (point spread function) of the imaging system. If the fluorophore is traveling quickly compared to the shutter period, a "streak" is formed on the image. In our sequencing method, each streak in an image represents a nucleotide in a DNA strand. Accurate sequencing or "base calling" depends on the successful detection of these streaks.

2005 IEEE International Conference on Electro Information Technology,
DOI: 10.1109/EIT.2005.1627002

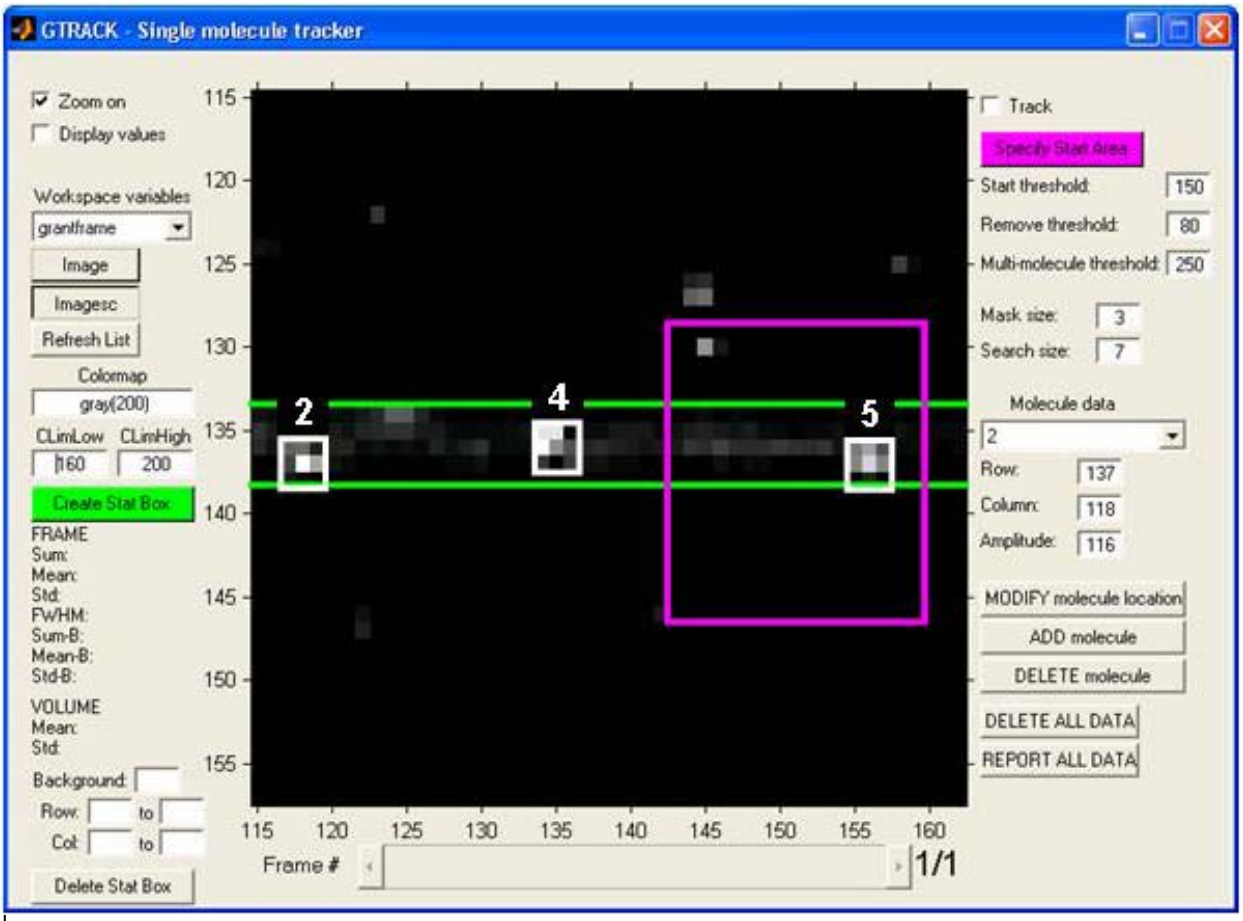

Figure 1. Real data. In-house single molecule tracking software. Fluorescent TAMRA dye molecules are detected as they originate at the lower right of the image (magenta box). Subsequent to detection they are tracked as they traverse through the flowcell; in this case a right-to-left movement. Three molecules can be seen here, in white boxes and labeled "2", "4", and "5". The green lines show microchannel boundaries. The flowcell channels are not seen in the figure, as all background has been subtracted out for a better signal-to-noise ratio.

As there are many parameters that affect the formation of streaks in an image, we have found that an imaging model is very useful for predicting the ability of a detection device to correctly identify molecules traveling through the detection site. This paper describes data analysis techniques and algorithms to perform single molecule base calling, in an automated fashion, from CCD images. In particular, an alignment algorithm was developed based on the well-known Smith-Waterman algorithm [10] with improved speed. This was done by imposing constraints on the brute-force search that is normally performed on the score value matrix. The alignment algorithm quantitatively judges the simulation by reporting errors encountered during sequencing.

2. Methods

2.1 Single Molecule Detection Model

A program was developed using MATLAB to simulate the fluorescence image formation of single-nucleotide incorporation events. Physically, these events are recorded on an image by a CCD camera in the form of a "spot" or "streak" (Figure 2) from accumulated fluorescence photons. The features of this streak (e.g. length, brightness, etc.) are governed by system parameters (e.g. microchannel dimensions, frame rate, integration time, detection zone size, flow rate, diffusion, etc.).

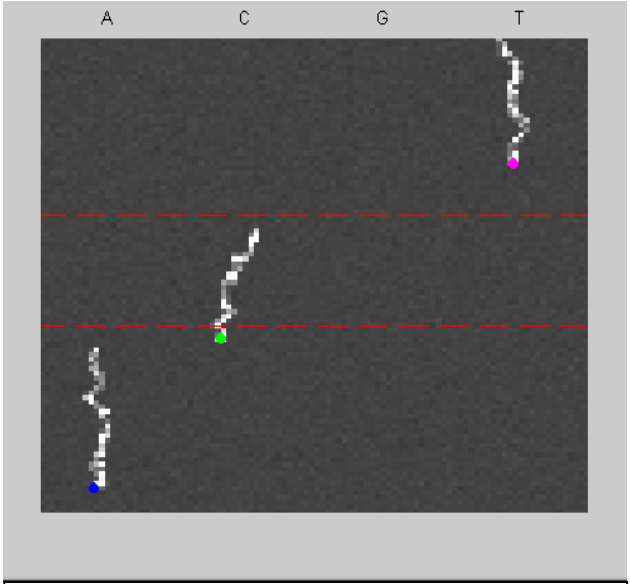

Figure 2. Imaging model. Simulated streaks are shown of fluorescent molecules traversing through an excitation laser.

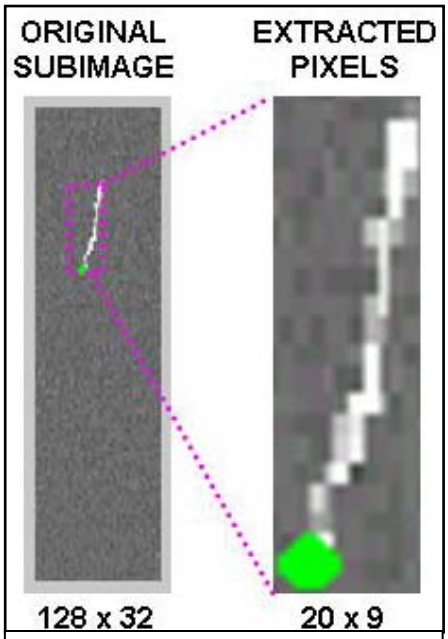

Figure 3. Channel extraction. Pixels at the expected signal location (20x9=180 bytes) are extracted out of a subimage of data (128x32=4096 bytes) for a >22X reduction in data. The green dot represents actual pyrophosphate location at the end of the frame.

2.2 Automated Signal Discrimination

An algorithm was developed to detect the presence of one or more pyrophosphate molecules in an image. The data were reduced by first extracting the image streaks (Figure 3). The extracted subimage was converted to a stream of accumulated counts by binning the subimage both across the rows and down a variable number of columns (Figure 4).

Signal and noise statistics were characterized and utilized in the calculation of an optimal signal detection threshold. These signal and noise characterizations were also utilized in the formulation of potential quality parameters for future research. However, the primary output of this algorithm was a signal-minus-noise data stream and a binary value depicting the presence (1) or absence (0) of signal.

The system background noise was characterized for different system operating conditions enabling the differentiation of the pyrophosphate signals from the noise. The mean and variance of the noise was measured by imaging microchannels devoid of fluorescence molecules. A histogram was constructed and scaled to form a probability density function (PDF) of noise counts received in a given frame period. This PDF was used along with the signal PDF to calculate an optimal threshold for signal yes/no decisions.

The morphology of the single-pyrophosphate signals ("streaks") was characterized (modeled) over various system operating conditions. This signal shape characterization or feature extraction enabled the ability to determine the signal location within the image. This is important in performing data reduction and signal count determination. The signal characteristics or features change as a function of the system parameters (micro-channel dimensions, flow

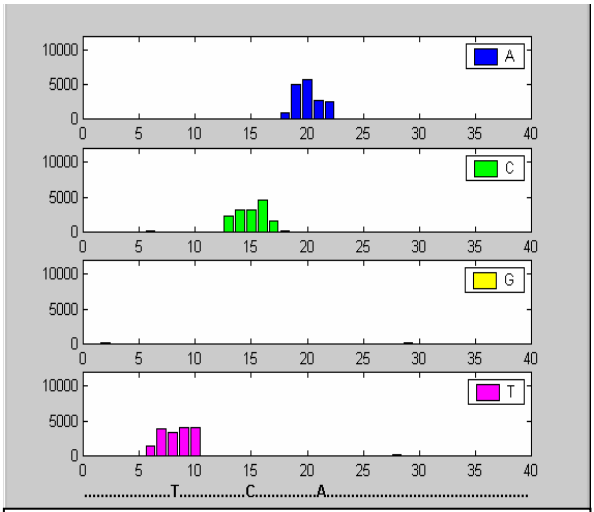

Figure 4. Data reduction. The image data in Figure 2 converted to a stream of binned counts.

rate, diffusion, laser power, frame rate, pixel size, nucleotide concentration, enzyme catalytic rate, dye photophysical properties, and optical point spread function) and the algorithm was designed to account for these changes in an optimal fashion. For example, increasing the flow rate results in longer signal streaks for a given frame period; the detection algorithm adjusts for this. Similar to the noise characterization, a PDF of photon counts was created from a histogram of counts received from the signal pixels in a frame.

The "detection zone" is the region within an imaged area of the flow cell under which the pyrophosphate molecules will pass and be analyzed. This region of interest is extracted from the image and analyzed for detection (Figure 3). The detection zone length is set to match the expected streak length, which is predicted from the flow rate and CCD frame period time. At a later stage the extracted zone is subdivided into smaller sub-zones for more detailed discrimination. The developed algorithm is flexible and adaptive in that the detection zone size is adjusted or adapted optimally to the specified input parameters.

Additional algorithms (not described) are responsible for taking the detected signals and forming an estimated sequence, i.e., "calling" bases. The model is sensitive to errors that may result from real-life situations. Examples are: 1) "swapping" of bases due to molecular diffusion, leading to incorrect calls, 2) false-negatives due to insufficient photon collection, leading to sequence deletion, and 3) false-positives due to excessive noise, leading to sequence insertion. An important task is for the model to quantitatively measure the sequence error. Originally, the Smith-Waterman algorithm [10] was used to assign an error based on alignment score. When modeling hundreds of bases, it was quickly seen that the time to align large sequences was too long to be practical. This motivated the development of algorithm variations in an attempt to minimize the computation time.

2.3 Smith-Waterman Variations

A customized version of the Smith-Waterman sequence alignment algorithm was developed and implemented to perform the comparison of the (random) input and output sequences in providing sequence accuracy statistics. The algorithm was modified principally to provide an increase in execution speed as well as to force a beginning-to-end comparison of the sequence results. The execution speed for the original Smith-Waterman algorithm is $O(n^3)$ where n=sequence length. Two speed enhancements were developed and combined to reduce the number of required computations to compare sequences.

In the original Smith-Waterman algorithm, a similarity matrix is formed by comparing each position of one sequence to positions in the other. A scoring function is defined to "encourage" matching positions (positive scores) while "discouraging" mismatches or induced gaps (negative scores). Table 1 shows a completed similarity matrix for the sequences 'ACGTTGCGA' and 'ATGTCGA'. The output of the original alignment algorithm is:

```
ACGTTGCGA
ATGT__CGA
```

Consider the dark-shaded comparison of 'T' to 'C' in positions 4 and 5 of the sequences respectively. The best score is calculated by comparing the penalty of a mismatch (diagonal move) to that of introducing gaps (insertions/deletions) in the sequence (horizontal or vertical move). To determine the optimal score, all entries above and to the left of the position under consideration are queried. These positions are shown shaded in Table 1.

Table 1. Similarity matrix with full comparisons.

		A	T	G	T	C	G	A
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
C	0.00	0.00	0.67	0.00	0.00	1.00	0.00	0.00
G	0.00	0.00	0.00	1.67	0.33	0.00	2.00	0.67
T	0.00	0.00	1.00	0.33	2.67	1.33	1.00	1.67
T	0.00	0.00	1.00	0.67	1.33	2.33	1.00	0.67
G	0.00	0.00	0.00	2.00	1.00	1.00	3.33	2.00
C	0.00	0.00	0.00	0.67	1.67	2.00	2.00	3.00
G	0.00	0.00	0.00	1.00	0.33	1.33	3.00	1.67
A	0.00	1.00	0.00	0.00	0.67	0.33	1.67	4.00

Table 2. Similarity matrix, variation #1.

		A	T	G	T	C	G	A
	0.00	0.00	0.00	0.00	0.00			
A	0.00	1.00	0.00	0.00	0.00	0.00		
C	0.00	0.00	0.67	0.00	0.00	1.00	0.00	
G	0.00	0.00	0.00	1.67	0.33	0.00	2.00	0.67
T	0.00	0.00	1.00	0.33	2.67	1.33	1.00	1.67
T	0.00	0.00	1.00	0.67	1.33	2.33	1.00	0.67
G		0.00	0.00	2.00	1.00	1.00	3.33	2.00
C			0.00	0.67	1.67	2.00	2.00	3.00
G				1.00	0.33	1.33	3.00	1.67
A					0.67	0.33	1.67	4.00

Two assumptions were formed in order to allow for improved calculation efficiency by coding variations of the original algorithm. The first assumption is that there will be no less than a 50% overlap in the two sequences to be compared. This is a conservative assumption, since with even moderate sequencing

error, the input and output sequences should be approximately the same length. The corresponding similarity matrix adjustments are shown in Table 2. Essentially, matrix positions corresponding to >50% overlap are not considered. These positions are the upper right and lower left diagonals of the matrix. As in Table 1, shaded boxes show the search positions for optimal score. Note that it is possible to search into a region which is not scored.

The second assumption is that a bound may be placed on the maximum gap length allowed. This is reasonable for the sequencer simulation because DNA sequencers with too large of a sequencing error (e.g., >10%) are not marketable. Therefore, the model only needs to report that the error has exceeded acceptable bounds. In the similarity matrix, this corresponds to searching in a limited space when looking above and to the left of a position. For example, Table 3 shows the similarity matrix for a maximum allowed gap length of 3. In the simulations performed for our sequencing scheme, we allowed a maximum gap length of 11.

Table 3. Similarity matrix, variation #2.

		A	T	G	T	C	G	A
	0.00	0.00	0.00	0.00	0.00			
A	0.00	1.00	0.00	0.00	0.00	0.00		
C	0.00	0.00	0.67	0.00	0.00	1.00	0.00	
G	0.00	0.00	0.00	1.67	0.33	0.00	2.00	0.67
T	0.00	0.00	1.00	0.33	2.67	1.33	1.00	1.67
T	0.00	0.00	1.00	0.67	1.33	2.33	1.00	0.67
G		0.00	0.00	2.00	1.00	1.00	3.33	2.00
C			0.00	0.67	1.67	2.00	2.00	3.00
G				1.00	0.33	1.33	3.00	1.67
A					0.67	0.33	1.67	4.00

3. Results

From the variation schemes, the number of required computations can be derived by examination of the similarity matrix search positions. The results are:

Smith-Waterman: $$Comps = \frac{n^2}{2}(n+1)$$

Variation #1: $$Comps = \frac{n}{6}\left(7\left(\frac{n}{2}\right)^2 - 1\right)$$

Variation #2: $$Comps = \frac{nc}{2}(2n - c + 1)$$

Variation#(1+2): $$Comps = \frac{nc}{4}(3n - 2c)$$

Where

Comps = number of computations
n = sequence length
c = constant (maximum allowed gap)

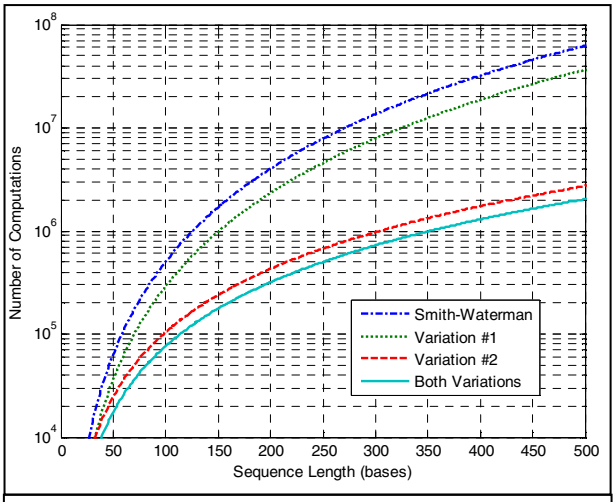

Figure 5. Alignment computations vs. sequence length.

These equations are plotted for the number of computations versus sequence length in Figure 5 (with *c*=11). Each of these four alignment methods was implemented and tested for performance (speed and accuracy). The accuracy results were identical; however, the execution times were markedly different (not shown, but correlate with the number of computations). The most significant decrease in execution times occurred for the second algorithm variation. Both variations (1+2) were implemented into the system simulation program and were a key factor in obtaining reasonable execution times, especially for long sequence reads, which were required. The improvement in calculation time between the original algorithm and variations (1+2) is between 1 and 2 orders of magnitude, for sequence lengths between 500 and 1000.

4. Discussion and Future Work

The developed system simulation program was run many times over multiple system input parameter configurations under which the sequence performance (accuracy) results were recorded and evaluated. No problematic simulation results were encountered.

The results indicate that significant improvements in sequence alignment speed can be obtained by very reasonable constraints. The improvements described are applicable when two long sequences of similar length are compared. Different constraints may be called for when searching sequences of markedly

different length. One example occurs when searching a long sequence (e.g. chromosome) for inclusive similarity to a shorter subsequence (e.g. gene). Alignment improvements for these cases are the subject of future work.

5. References

[1] Peck K., Stryer L., Glazer A., and R. Mathies, "Single-molecule fluorescence detection: Autocorrelation criterion and experimental realization with phycoerythrin", *PNAS*, 86 (1989), pp. 4087-4091.

[2] Nie S., Chiu D., and R. Zare, "Probing individual molecules with confocal fluorescence microscopy", *Science*, 266 (1994), pp. 1018ff.

[3] Macklin J., Trautman J., Harris T., and L. Brus, ""Imaging and time-resolved spectroscopy of single molecules at an interface", *Science*, 272 (1996), pp. 255ff.

[4] Schmidt, T., Schutz G., Baumgartner W., Gruber H., and H. Schindler, "Imaging of single molecule diffusion", *PNAS*, 93 (1996), pp. 2926ff

[5] Kues T. Dickmanns A., Luhrmann R., Peters R., and U. Kubitscheck, "High intranuclear mobility and dynamic clustering of the splicing factor U1 snRNP observed by single particle tracking", *PNAS*, vol. 98, no. 21 (2001), pp. 12021-12026.

[6] Xu X. and E. Yeung, "Direct measurement of single-molecule diffusion and photodecomposition in free solution", *Science*, vol. 275 (1997) pg. 1106ff.

[7] Seisenberger G., Ried M., Endress T., Buning H., Hallek M., and C. Brauchle, "Real-time single-molecule imaging of the infection pathway of an adeno-associated virus", *Science*, vol. 294 (2001) pg. 1929-1932.

[8] Kubitscheck U., Kuckmann O., Kues T., and R. Peters, "Imaging and tracking of single GFP molecules in solution", *Biophys. J.*, vol. 78 (2000), pp. 2170-2179.

[9] Lyon, W. and S. Nie, "Confinement and detection of single molecules in submicrometer channels", *Anal. Chem*., 69 (1997) pp. 3400-3405.

[10] Smith T. and M. Waterman, “Identification of common molecular subsequences”, J Mol. Biol., 25;147(1), (1981) pp. 195-7.