Single Lane Unlabelled DNA Sequencing – a Bench Top Genomics Unit

Stuart R. Hassard, D.Sideris, J.Harford, J. Hassard **deltaDOT Ltd**, 13 Princes Gardens, London, United Kingdom

ID#: 26843
TA011 Advances in Sensing, Detection, and Integration in Bioanalytical Systems

Introduction

deltaDOT is developing tools that will transform biotechnology. We will change the way biotechnologists, pharmaceutical companies and healthcare workers use genomics and proteomics in basic and applied research. This will make a radical change in the way basic molecular biology research is performed, how diagnosis in healthcare can be achieved and how new drugs could be developed. We are actively seeking to expand the application base for our technology and are in talks with other possible users such as defence companies. Our technological base is highly modular at the component level and as such can be rapidly deployed into new areas.

Our central tool, Label Free Intrinsic Imaging or LFII, requires and has led to a suite of advanced signal processing and pattern recognition tools. The LFII approach allows a one-stop-shop in which the same hardware and methodology can be used in analyzing the information encoded in the genes (for example by DNA sequencing and the analysis of mutations) and expressed in the proteins. The entire suite, however, allows LFII to leverage huge advantages in the areas of throughput, cost, ease of operation and return on investment. Disadvantages of the existing approach to DNA analysis (using chemical markers, or labels) include high cost, the time needed for the analysis and the health & safety implications of using the necessary reagents (handling, storage & disposal).

Key issues that differentiate our technology and make the LFII approach highly suitable for consideration in a wide variety of applications include:

*Miniaturization***:** as an imaging system that uses labels gets smaller, the signal and other capabilities such as resolution scale as the cube of the characteristic dimension, since the signal is proportional to the amount of material being studied. In the LFII approach, we decouple the signal from the volume, and only have a weakened dependence on the first power of the characteristic dimension. LFII therefore lends itself directly to ever-smaller systems. The consequences of this felicitous scaling law affect power consumption, weight and linear dimensions as well as sensitivity and resolving power.

Robustness: with no need for labels or labeling systems, we require fewer parts, moving or otherwise, making the system more robust, operationally, mechanically and in analysis terms.

Analysis Power: unlike with affinity chip technologies, where the discovery chemistry must be devised and applied prior to use, deltaDOT's approach is to assume

nothing. The association of signals with targets, the data-mining and subsequent correlation or anomaly analyses, are all done *in silico* and can be remotely controlled or adapted on-the-fly.

deltaDOT enjoys a close relationship with Proctor and Gamble Pharmaceuticals in the form of investment, advice and a joint development programme to develop our advanced Protein Profiling Platform. This deltaDOT instrument passed a full technical audit by a team of P&G core scientists and two units are being used in discovery programmes in P&G's headquarters in Cincinnati, Ohio. Further sales to P&G and other organisations will follow the development of user-friendly sample handling systems, scheduled for initial assessment in late April 2004.

deltaDOT's technology allowed us to develop the concept of our Sample to Silicon integrated approach. This concept is to treat the analysis in an holistic way, designing systems both upstream and downstream of our central analysis suites. Such systems have the potential to include sample acquisition, milli and microfluidic systems, integrated biochemistry-on-a-chip systems and 'real-time' data analysis coupled to decision-making feedback.

Our core technology will impact most of the biotech spectrum, and can be adapted almost *ad hoc* to many applications.

Single Lane Unlabelled DNA Sequencing

Since the advent of DNA sequencing the focus has tended to be on high throughput discovery genomics. Projects like the Human Genome Project (3 000 000 000 bases - http://www.sanger.ac.uk/HGP/) and the *Plasmodium* project (~30 000 bases Malaria - http://www.sanger.ac.uk/Projects/Pfalciparum/) are two examples of large-scale discovery projects. These, and a wide range of other genome projects, rely on 'production line' sequence data acquisition and processing. Projects like these produced, and in turn were enabled by, the development of a very large market for high through put sequencing machines. This in turn created to a reliance on that technology. This dependence means that the skills and equipment used to do lab based sequencing on short DNA templates have largely been neglected.

The need to sequence short templates of DNA cannot be underestimated. Genetic manipulations of single bases, such as point mutations, or larger DNA fragments, as in gene silencing; are now routine in many labs. This was not the case when the DNA sequencing machines were first launched, and the need to do short run sequencing was not so apparent. Sequencing is the best possible way to check these processes, and generally a template of less than 100 bases needs to be read. The validation of known single nucleotide polymorphisms (SNP) is also becoming increasingly important as their role in genetic disorders become clear. Again this can be achieved by sequencing template only a few 10s of bases long. Only a few years ago this sort of work would be done by isotope labelled Sanger sequencing and large gel electrophoresis. In a high

percentage of modern labs all sequencing now sent out to sequencing services, usually a core facility in a University or research institute. While not very expensive for an individual template (typically ~ £9) this can add up for a large lab, with many workers doing genetic manipulations, and this cost will rapidly become astronomical when thousands of SNPs have to be quickly validated.

Another major problem is time. DNA sequencing services usually take a minimum of three days. For short sequencing that is a bad roadblock and deltaDOT want to enable lab scientists to avoid it. Our Label Free Intrinsic Imaging system coupled to advanced signal processing and microfluidic systems allow us to envisage a DNA sequencer that is cheap and small enough to be an analogue to a PCR machine, fast enough to allow genetic manipulations to be sequenced the day they are done, and modular enough to for part of a larger, integrated sequencing system (ISS). The machine will be a small bench top unit, with a user-friendly operating system and very low running costs, due to the removal of the need to buy dyes or radioisotopes.

The ISS is ultimately designed to be portable and consists of a series of modular units to which this DNA sequencing machine will be central. The concept is 'from sample to silicon' starting with crude sample preparation and pathogen identification by novel pattern recognition software. Identified pathogens, for example *Mycobacterium tuberculosis*, will be harvested, distressed, and their genomic material sequenced by a PCR sequencing unit. Sequence reaction product would then be separated and read on the proposed DNA sequencing unit, with integrated data handling. This unit is designed to provide distributed and automated diagnostic tools based on DNA sequencing. An example of the benefits of this system as a field unit would be the previously mentioned drug resistant pathogen analysis, e.g. *Plasmodium falciparum*, the Malaria parasite. If a field healthcare worker could rapidly find out which resistant strain of parasite was infecting the patient, then an informed and timely drug choice could be made. Having a digital, Internet enabled and possibly wireless output means that the movement of such parasites could also be monitored, allowing "real-time" molecular epidemiology. This in turn means drug therapies could be distributed to remote areas in a more intelligent way allowing the best use of what are usually limited resources.

The key to deltaDOT's increased resolution in an unlabelled system lies in the process of equiphase vertexing (Patents P18499US, P32531US-M and P32533US-M). This involves the de-convolution of signals from a multitude of detectors imaging the same biomolecule. deltaDOT have developed and modeled several signal processing algorithms applicable to sequencing.

The first, MITSO ('Multiple Input Transform for Sequence Ordering') (P32532US-M), uses this process and applies it to the problem of base termination identification. Original deltaDOT sequencing chips had four channels to image each termination reaction. This was clearly unsatisfactory, since there are channel-to-channel variations that result in the sequence getting an unacceptable error rate even with calibration

markers. However, it was realized that the equiphase vertexing algorithm would allow the reactions all to be run in one channel so long as the vertex can be found. This allows us to associate any given band with a given injection point in space (z, along the direction of travel) and time t.

Paradoxically, the more complex the sample, the more bands, the better z and t are defined. Chips were designed to allow multiple sample injection with these constraints. The injection ports are separated on a space and time axis so each termination reaction can be performed simultaneously.

As the site of the injection channel gives the space/time constraint needed, the MITSO algorithm can be applied to identify t=0, and thus the termination reaction, of each band. This allows the concept of 'virtual colour' being assigned to each band in a single channel. The key issue here is the resolution in space z_{inject} and t_{inject}.

In principle, we can set up virtual colour in either or both z and t vertices; in practice we will use both. In operational terms, we simply run the DNA down the same channel.

This offers the possibility of an increase in throughput as reactions, loaded at different times in the same channel can be distinguished. However, it is not clear how long it would take for electrophoresis conditions to degrade sufficiently far that errors creep in and research into this will be necessary. Non-linearities in electrophoretic mobility must be corrected: they are the biggest possible source of error. Other algorithms that can be applied to DNA sequencing include the self-referencing algorithm (SRA), which uses the fact that we only need to exploit the relative lengths of the DNA fragments to acquire the sequence and uses the differential in fragment length to increase the accuracy of the data-unlike protein, DNA is a 'digital' system. Finally, we have a pattern recognition tool in development that will be used to annotate the sequence as it is acquired. It searches for, and interprets restriction enzyme sites and other primary structural motifs in real time. It is envisioned that the software will be fully internet capable.

Results

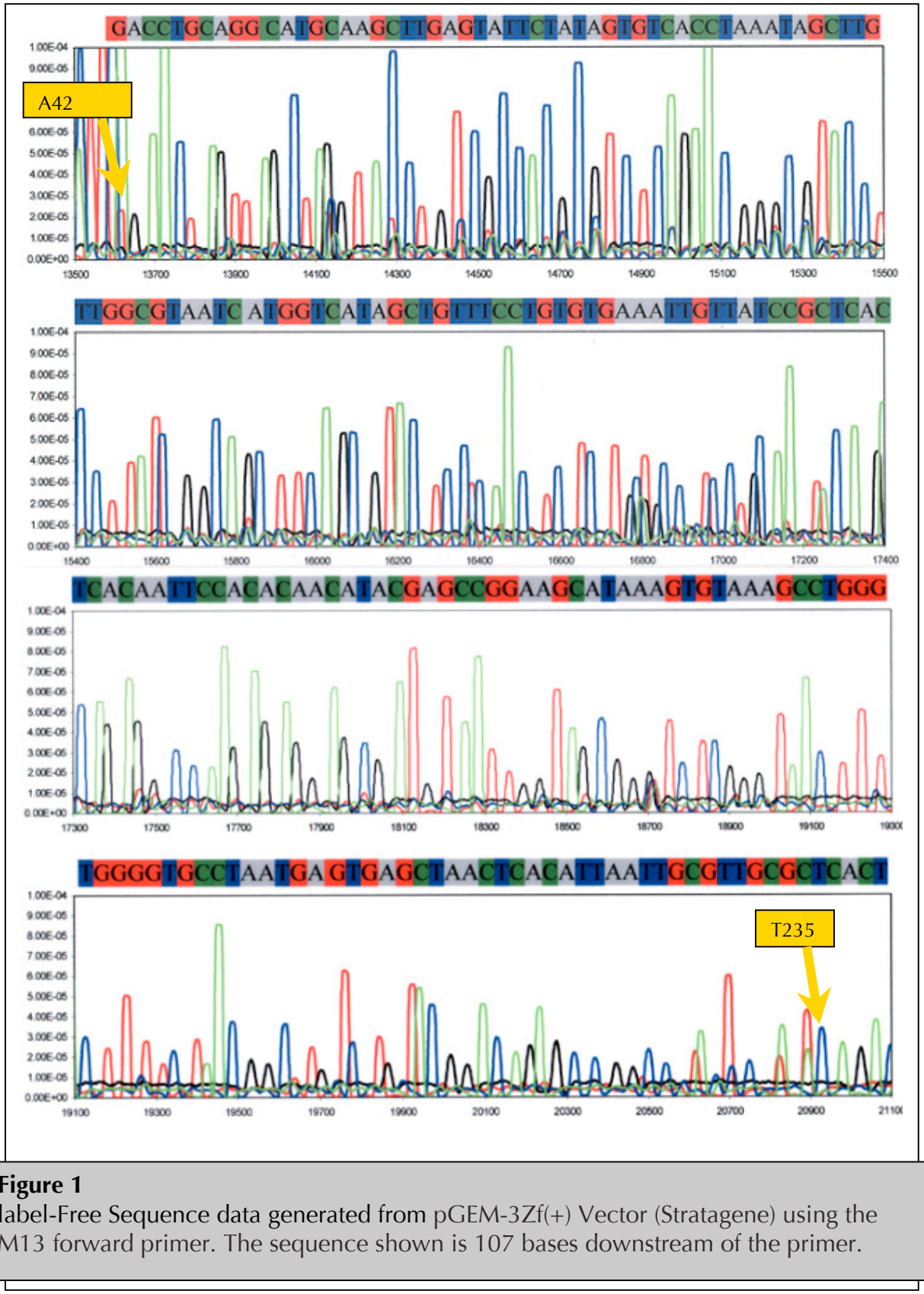

Figure 1
label-Free Sequence data generated from pGEM-3Zf(+) Vector (Stratagene) using the M13 forward primer. The sequence shown is 107 bases downstream of the primer.

Experimental details

A series of capillary electrophoresis runs on x26_11 using a commercial sequencing kit and an in-house gel buffer based on polyethylene oxide were performed. The objective of these runs was to ascertain and optimise the conditions for DNA sequencing on the basis of intrinsic absorption of ultra-violet light. The standard x26_11 rig configuration with a 254nm (20nm bandpass) interference filter. The standard power supply was used at 14kV with a current of approximately 19□A at the longest separation length used.

Electrophoresis configuration - Uncoated fused silica capillaries of 40cm or 70cm separation length (distance to detector; capillary length 56cm and 86cm respectively), 75□m internal diameter were used with Polyethyleneoxide (PEO; molecular weights between 1 and 5 megadaltons; Tris-Taps-Histidine-EDTA buffer, 7M Urea) as the electrophoresis medium. Electrophoresis was carried out at 160V/cm after a stacking injection (2.5 to 5 mins at 25V/cm) employed by dissolving the sequence product in 5 or 10ul of deionised water. The capillary was heated to 36^0C for optimal separation, which was maintained within $\pm 0.1^0$C. The electrophoretic parameters were chosen so as to maximize the resolution within the ability of the current x26_11 system and the standard gel matrix whilst also allowing a reasonable electrophoresis running time. Capillaries were rinsed with deionised water and 2% Polyvinylpyrolidene (PVP) as a dynamic coating between separations and conditioned for 5 minutes at the flow voltage before starting the electrophoretic separation.

DNA Sequencing Sample- Sequencing template was prepared using Promega fmol® DNA cycle Sequencing System (Cat# Q4100). The template of choice was pGEM®3Zf(+) control DNA and pUC/M13 forward primer (24mer). Sequence template was produced on the Robocycler Gradient 40 (STRATAGENE) employing 198 cycles (30 seconds 95°C, 30 seconds 74°C). After 198 cycles the PCR product was cooled (6°C), concentrated and precipitated overnight in ethanol (150% vol) and sodium acetate (10% vol). After centrifugation at 15000 rpm for 20 minutes the supernatant was removed and the DNA air-dried for 15 minutes. This was then dissolved in 10□l of deionised water and denatured for 3 minutes at 95^0C prior to electrokinetic injection into the conditioned capillary.

Eva Processed Track Sequence Data

The mirrorplot shows consecutive runs on X26-11 for one of the sequence tracks in 5 mDa PEO. At smaller fragment lengths as shown in the next figure similar separations can be obtained in lower viscosity separation media such as 1 mDa PEO.

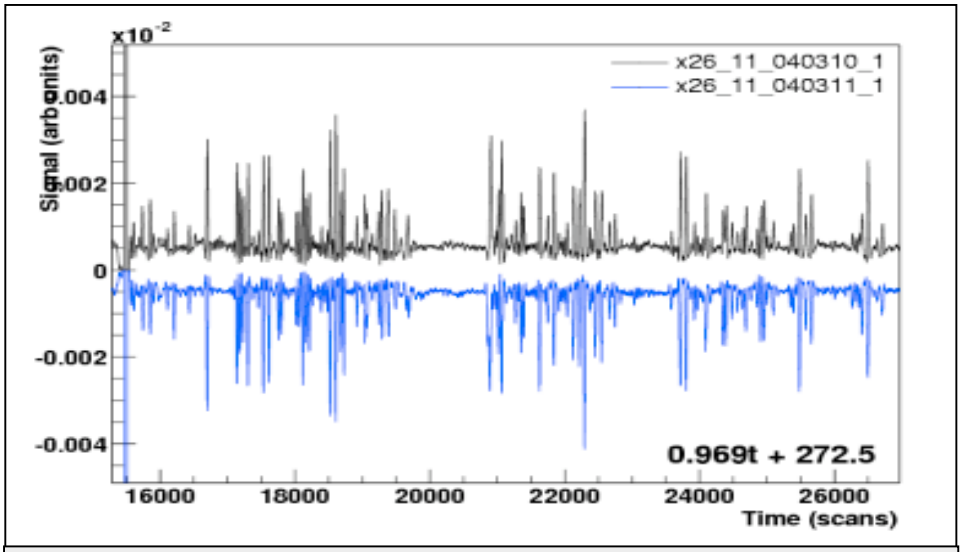

Figure 2
Comparison of Eva processed plots of consecutive runs of A-track sequence from 30bp to 350bp from the same sample batch.

Analysis of the peak-to-peak separation shows very good correlation with the distribution expected. The reproducibility can also be seen to be excellent, which is a vital factor in mutant analysis.

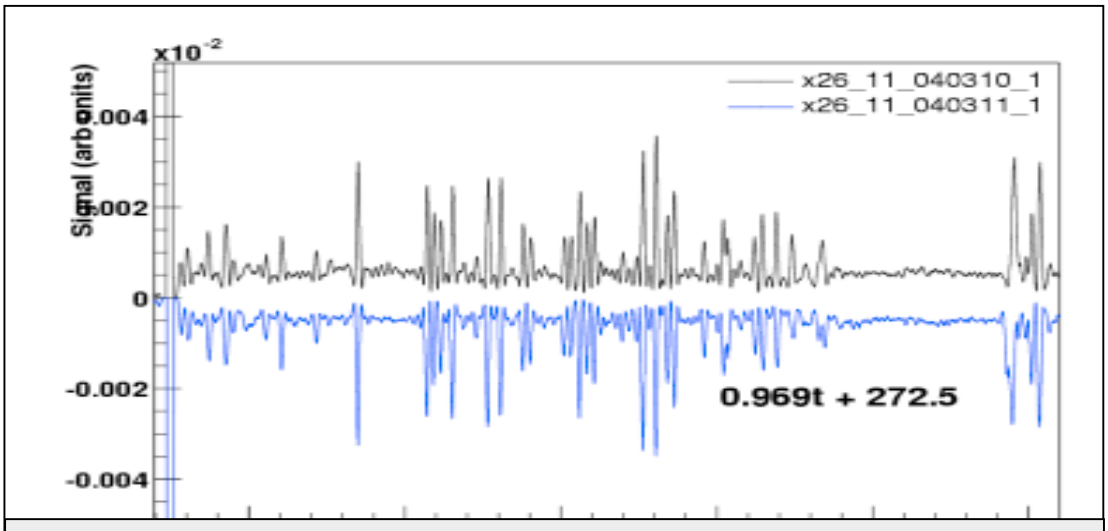

Figure 3.
Comparison of Eva processed plots of consecutive runs of A-track sequence from 30bp to 150bp from the same sample batch.

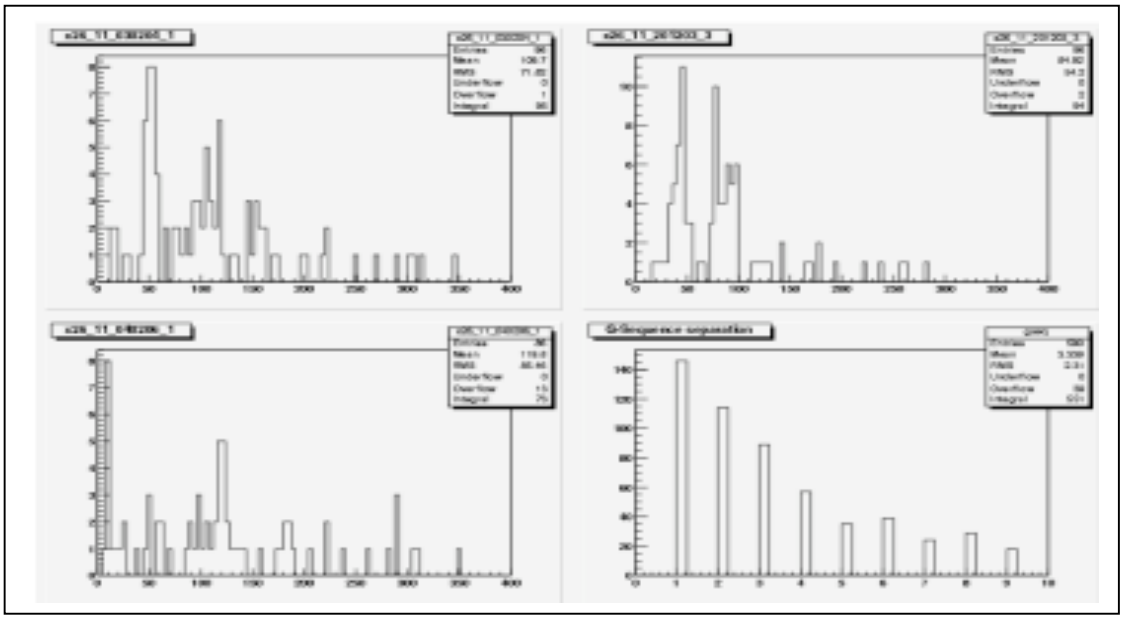

Figure 4

After running a peak finding algorithm over the signal histogram, the plots show the distance between neighbouring peaks. The top two plots are for G-track data run in a 70cm capillary, the bottom left plot is a G-track run with added markers and the bottom right plot shows the separation in base pairs as calculated from the true, known sequence. As can be seen from the true distribution, the most likely separation distance is 1 bp. From the top two plots it would appear that the base pair separation (in time for real data) is about 50 scans.

deltaDOT will have a bench-top DNA sequencer, called Merlin, ready fro beta testing by October 2005, and are actively applying the lessons learned so far to the development of microfluidic systems for fast bench-top DNA sequencing.