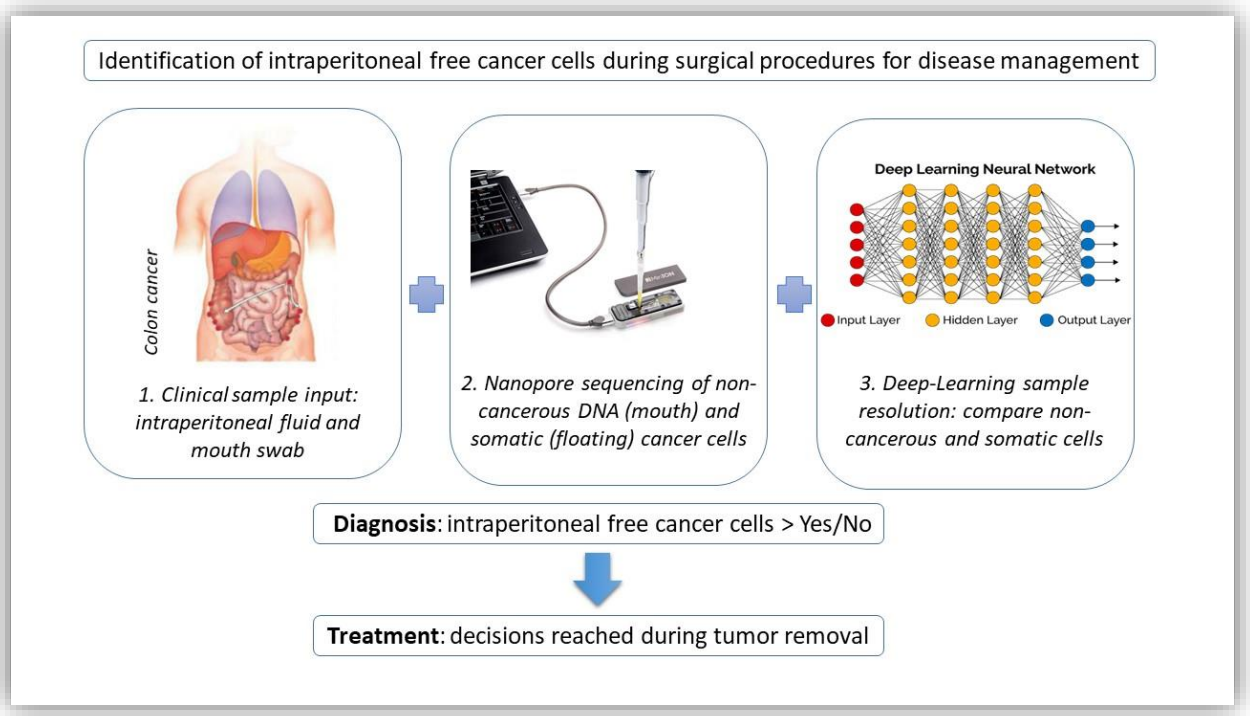
Tel Aviv University, Zimin Institute for Engineering Solutions Advancing Better Lives

Update report – January 2019

Identification of intraperitoneal free cancer cells during surgical procedures for disease management: a multidisciplinary engineering solution using deep learning and real-time-rapid sequencing technology

Professor Noam Shomron, Faculty of Medicine, Tel Aviv University



Why we fit the Zimin Institute ideas

We propose a solution for a pertinent medical question.
We employ deep learning algorithms in our solution.
We use a multi-disciplinary approach in our research.

Abstract

Intra-abdominal malignancies often result in intraperitoneal free cancer cells (IPCCs), which increase the chances of cancer spreading to distal organs, and serve as an important prognostic tool in cancers such as ovarian and gastric cancers. Curative treatments based on intraperitoneal chemotherapy often have effective outcomes, however, the time between IPCC sampling and detection is critical, currently taking weeks to reach a conclusion. The MinION (Oxford Nanopore Technologies) – the first handheld genetic sequencer – is capable of reading long stretches of DNA in a real-time, but has not yet reached the level of accuracy of older, slower technologies. Deep learning techniques mimic the learning process of the human brain in order to recognize patterns in digital representations. In our lab, we use deep learning to circumvent the limitations of Nanopore sequencing by learning the 'signal' rather the 'sequence' of the DNA. The aim of our project is to establish a rapid real-time method for the detection of IPCCs during colorectal resection. We propose a solution that takes advantage of our close collaboration with the Surgery Division at Sourasky Medical Center, and is based on three components: (i) access to clinical samples (abdominal fluid and mouth swabs) collected during resection; (ii) rapid real-time DNA sequencing; and, (iii) deep learning algorithms for the discrimination between somatic versus non-cancerous DNA. We believe the time has come to merge DNA sequencing (as a 'digital signature') with deep learning in order to afford surgeons the opportunity to quickly identify IPCCs during surgical procedures, thereby allowing immediate treatment and decreasing the need for future intervention.

Update

We have established protocols for collection and extraction of DNA from several sample scenarios (swab, fluid with high and low number of cells). We have run some of the experimental protocols and the MinION on human (non-patient) positive and negative control samples in our lab. We have performed Deep Learning on MinION sequencing data we have collected. We have regular meetings and a submitted IRB with Sourasky Medical Center (with Dr. Guy Lahat and Dr. Shelly Loewenstein) which is pending final signatures at the local (Sourasky) committee. Since the National Ethical Committee was non-functional for more than four months, our approval was delayed and therefore we will receive the samples from patients undergoing operations at Sourasky upon approval. This is expected in the coming couple of weeks as they have only asked our team for a small amendment.

In year two

We expect to complete the scanning of the tissue samples, to identify cancer cells in the abdominal fluid, and to prove that this could be done in a timely fashion, namely during (and not after) the operative procedure.

Challenges and Solutions

Intraperitoneal free cancer cells (IPCCs): Intra-abdominal malignancies often result in IPCCs, which increase the chances of cancer spreading to distal organs, and serve as an important prognostic tool in cancers such as ovarian and gastric cancers. Colorectal cancer is one of the most frequent cancers worldwide, with development of peritoneal carcinomatosis in 10-30% of patients. Curative treatments for peritoneal carcinomatosis, such as cytoreductive surgery and intraperitoneal chemotherapy, have been shown to be effective, especially in malignancies of colorectal origin, thus increasing interest in free malignant cell detection. Nanopore sequencing: Oxford Nanopore Technologies (ONT) has developed several devices that use newly developed nanopore technology to read DNA sequences. The smallest model, the MinION, is half the size of a mobile phone and can be directly connected to a laptop or even, making it the most portable genetic sequencing device currently on the market. Deep Learning: In simple terms, deep learning techniques attempt to mimic the learning process of the human brain in order to recognize patterns in digital representations of sounds, images, and other data. We will use deep learning to identify and differentiate digital signatures of non-cancerous DNA (healthy DNA from patient mouth swabs) and DNA from cancer cells (collected from intraperitoneal fluid) that have undergone Nanopore sequencing.

Combining fields

Here we introduce a novel approach for IPCC identification during colorectal cancer surgery. We will use deep learning on signal output from rapid real-time sequencing. Our negative control (baseline measurement) would be non-cancerous DNA derived from patient mouth swabs, which is potentially 'healthy' non-mutated DNA. We will compare this signal to intraperitoneal fluid that might contain cancer cells and will supply mutated cancer DNA. Contrary to the current approach to sequencing, which is only capable of reading DNA sequences, we intend to use this technology to identify multifactorial patterns of DNA characteristics for cancer cell classification. Cancerous and non-cancerous DNA signals will be translated into digital signatures (voltage measurements), which will be processed using deep learning algorithms.

Goals and Milestones

1. Collect clinical samples in the operating room: mouth swabs and intraperitoneal fluid from each colorectal patient. Send to clinical/pathological evalation.

   Todate *we have conducted multiple meetings with the medical team at Sourasky Medical Center, Dr Guy Lahat and Dr Shelly Loewenstein. We have outlined and tested protocols for collection of the samples from the operation room. We are currently waiting for the final IRB committee ethical approvals in order to receive the samples. We have applied severeal DNA extraction protocols on practice samples and evaluated the yeild and purity.*
   In year two *we will run our tested protocols on the sample we will receive from the operation theater.*

2. Train the deep learning algorithm on non-cancerous and cancerous data in order to evaluate its ability to characterize both types of sampes.

   Todate *we have run simulations to demonstrate that the Nanopore machine is productive in our hands. Valuable data was collected on control samples (see the results presented below).*
   In year two *we will apply these algorithms on the clinical samples.*

3. Run MinION experiments on mixed cancer cell line and primary healthy cells at various ratios for testing our deep learning algorithm to seperate the two types of cells.

   Todate *we have run simulations to make sure that the Nanopore analysis gives useful and meangful data in our hands on several non patient samples (see the results presented below).*
   In year two *we will apply these algorithms on the clinical samples.*

4. Extract DNA from the clinical samples and characterize using rapid real-time DNA sequencing. Initial quality control of sequencing data will include elimination of non-informative data, such as non-human derived signals. *(Steps 3-4, expected 6 months)*

   To be *carried out in year two, commencing April 2019 or before.*

5. Apply deep learning techniques to develop effective tests of discriminiation performance. Then, apply the resulting algorithms to collected DNA data from positive/negative controls (from above) and validate accuracy. *(Expected 6 months)*

   Todate *we have run simulations to make sure that the Nanopore analysis gives useful and meangful data in our hands on several non patient samples.*

   In year two *we will run the analysis on the patient data commencing April 2019 or before.*

6. Apply algorithms to rapid real-time data from samples with unidentified cancer cells in order to distinguish between and classify IPCCs for prognostic evaluaiton. *(Expected 6 months)*

***To be** carried out in year two, commencing April 2019 or before.*

Overall, we intend to combine all of the goals listed above for rapid real-time analysis of IPCCs in colorectal cancer resection procedures for better therapeutic outcomes.

Our year two studies are critical for our success as they will transform our simulations to real life data based on the clinical samples. At the end of the second year we expect to be at a position where we will be able to present our findings and to discuss potential applications of our study.

Preliminary results

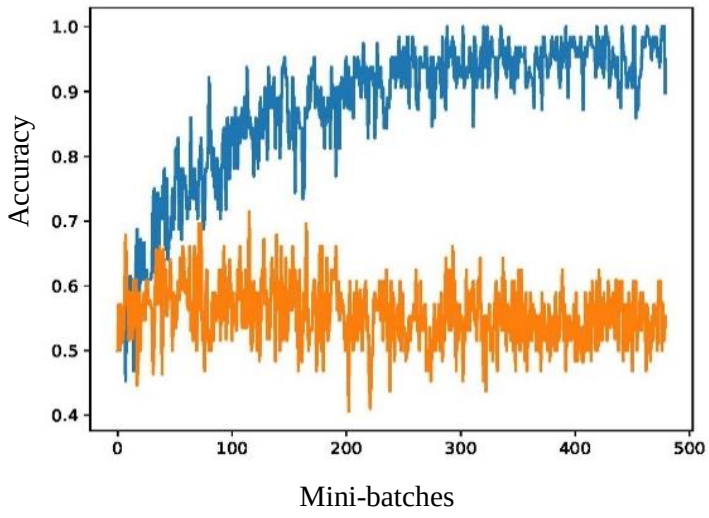


Figure 1: Training process for selective sequencing of a specific gene. Training process of LSTM with recurrent batch normalization model, the figure demonstrates one of the possible problems during deep learning model training. The divergence between accuracy values on the training dataset and test dataset indicate a clear case of overfitting where the model "remembers" the training dataset therefore have a high accuracy of that dataset, but when the model encounters unseen samples from the test dataset the classification accuracy decrease drastically. Accuracy on the training dataset (blue) compared to the accuracy on the validation dataset (orange). The time-span of the X is equivalent to 300 epochs.

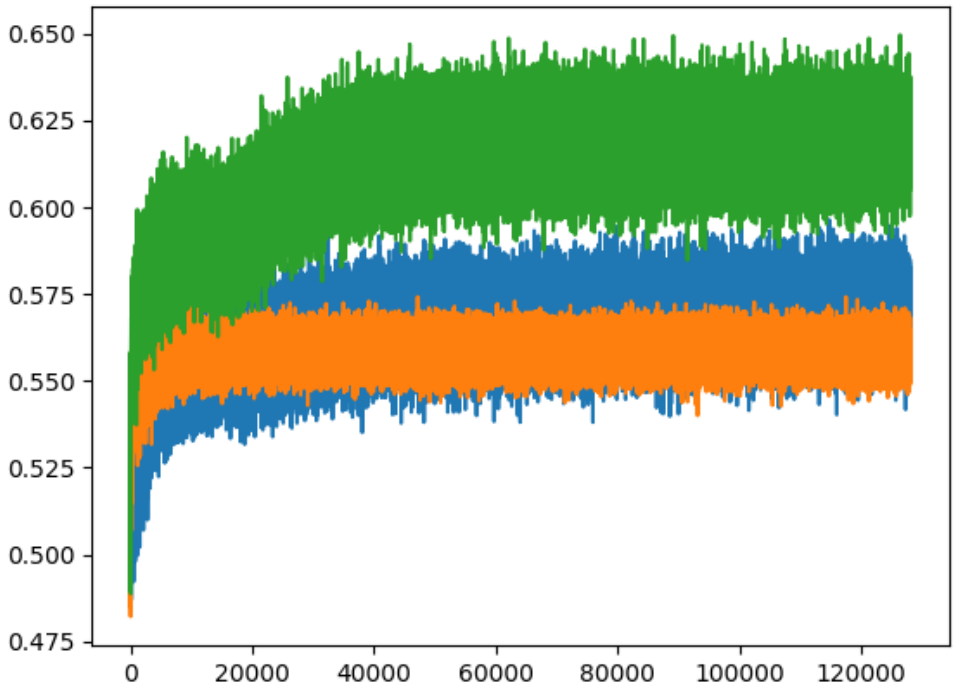


Figure 2: Training dataset accuracy of three separate models. Training dataset accuracy of the three separate models, which were combined into the final model, during the training process. It can be seen that all three models increased their accuracy during training, meaning all inputs were important for the classification of the reads, therefore combining all models into one final model should utilize all the available data to perform the classification. Blue tracks the accuracy of model based on nucleotide sequence. Green tracks the accuracy of the model based on the aligned sequence length parameters. Orange tracks the accuracy of the model based on the first and last positions of the alignment on the chromosome. The Y axis is the accuracy value; X axis is number of mini-batches from the beginning of training.

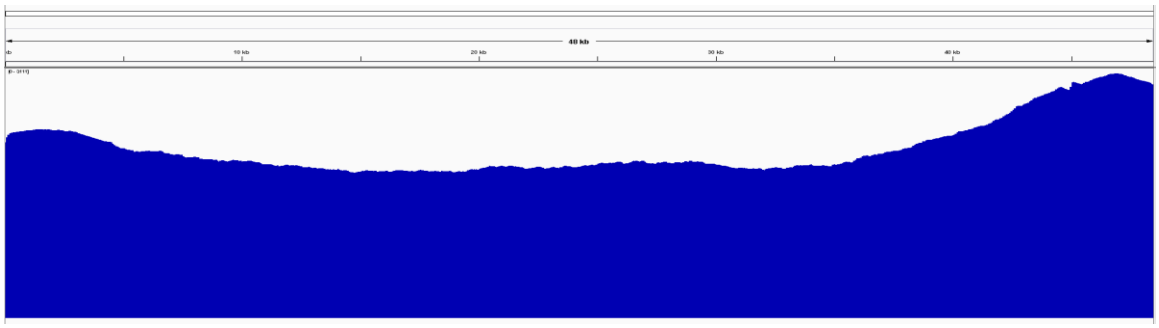


Figure 3: Genome coverage of a virus (Enterobacteria phage λ) as represented by IGV (a visualization software). This genome might represent an anomaly in a genomic sequence from a cancer cell. There is an even coverage on the Labmda phage genome with slight elevation at the last portion of the genome. The increase in coverage is caused by the control DNA added during library preparation steps, the control DNA is identical to this portion of the Lambda phage genome. The x-axis represents the genome of Enterobacteria phage λ in length, the y-axis represents the coverage of each nucleotide of the Enterobacteria phage λ genome.

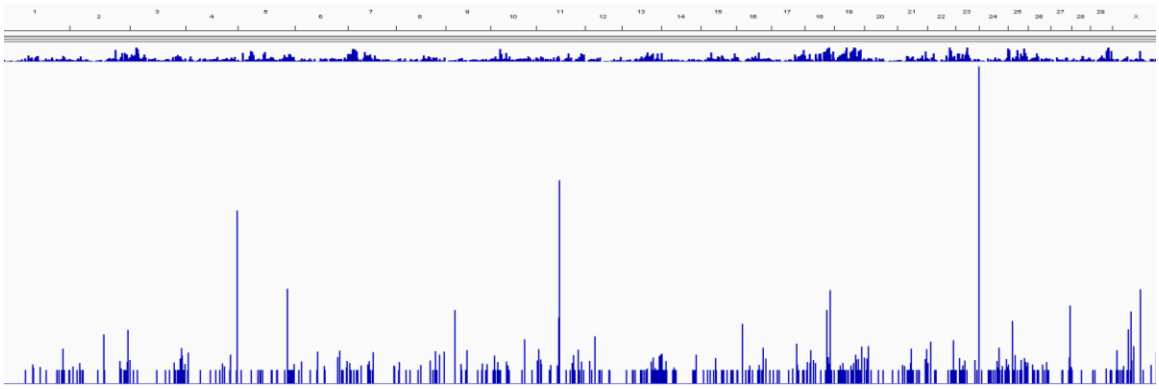


Figure 4: Coverage of human genome as represented in IGV. The coverage of the human genome (which is the background in our experiments) achieved from sequencing a human cell line sample. Although it may appear the reads cover large portions of the genome, in reality the reads aligned sparsely to the genome with large gaps in between. This was the expected result based on the amount of output data from the sequencing. This would be the basis for our future comparison when looking at cancer genomes compared to non-cancerous ones.

Year two is essential for commercialization

Our project is at the verge of its critical experiments. We expect year two to be our turning point in establishing the proof of concept that cancer cells could be told apart from non-cancer ones using a rapid analysis while the patient is still in the operation room. Our preliminary results of year one prove that as a concept it is possible, yet it is critical for us to show it in a 'real-life' setting. Once this is achieved we can plan how to implement it as a standard of care where physicians require rapid feedback during invasive procedures.

Potential commercialization of our project

We believe that rapid feedback during medical procedures is essential for improving clinical intervention. We reached this project after carrying out extensive discussions with physicians from multiple departments. We are aware that there is missing information that the physicians require to carry out knowledgeable decisions and there are very little solutions that can assist them at the moment. We believe that our project should be integrated into one of the departments at a leading hospital. For example, in the Pathology department where samples are received regularly for analysis, or closer to the decision making physician at the Surgical department in the operation room. Once one of these hospital units adopts our setting, a routine protocol could be followed for sample collection, processing, analysis and reporting. Early adopters could be our test case and should work closely with our team to make sure that valuable data is transferred back to the decision makers. We want to make sure that the decision could lead to better treatment, therapeutics or procedures. Given that we focus on the academic – experimental side, any valuable commercial – business guidance would be welcome.