Nano-Sequencing

Matthew Young
Biomedical and Electrical Engineering
University of Rhode Island

The field of medicine is ever advancing and within this seemingly natural progression, DNA sequencing has the potential to exponentially increase its evolution. The ability to map entire genomes efficiently would lead to numerous break throughs in modern medicine from the ability to more accurately and efficiently diagnose diseases to the discovery of new drugs and treatments. The United States understands the opportunities genetic sequencing presents and have set a goal for the research community: To develop a low cost sequencing and resequencing of the entire human genome by 2014, better known as the "The $1000 Genome." The importance of this challenge is described by a ten million dollar reward for completion.

The current method of DNA sequencing is done through a process known as dideoxy chain termination. This process is expensive, inefficient, and vastly time consuming. The first genome ever mapped cost three billion dollars and was conducted over a period of years. One follow-up to this research was conducted in 2008 at a cost of one million dollars, spread across several weeks. Currently at one million dollars it cost one one-hundreth of a cent for each base pair, a huge improvement to the original three dollars per base pair, but a far cry from the proposed one one-hundred thousandth of a cent.

There are many approaches being proposed to more efficiently map the genome. Some of these include capillary electrophoresis on micro-fabricated chips, hybridization on micro-arrays, cyclic array sequencing of amplified DNA, mass spectrometry, and a process that involves processive exo-nuclease. Capillary electrophoresis involves the separation of DNA base pairs on their size to charge ratio within a small capillary filled with an electrolyte. Hybridization involves the application of resonance light scattering. Cyclic array uses fluorescent base incorporated with polymerases and processive exo-nuclease is used to digest a single DNA molecule which can then identified one at a time. These methods are valid, but not as promising as nanotube DNA sequencing, in which a DNA strand is threaded through a nano-pore and its sequence is identified from fluctuations in ionic current.

Nano-sequencing involves the extraction, duplication, splicing, and translocation of DNA to develop single stranded DNA along with its Watson-Crick complement. The DNA is then fed, by manipulating a chemical system and the associated electric field, through a homomeric αHL pore that has a ring of seven ardinines located near the barrel of the pore. To bring the diameter of the conducting pathway closer to the size of the DNA a cyclodextrin adapter is fitted within the pore. At this point amino acids are substituted into the system to deal with charge distribution. Single channel electrical recordings are taken from the trans side of the lipid-bilayer and are detected using a ultra low noise, ultra high gain amplifier that takes a bio-picoamp signal and converts it to a readable voltage – This can be seen on the following page. After the recording is taken computer software begins the analysis.

The algorithm is designed to use the provided data to quickly and efficiently match the base pair moving through the nano-tube to the known current / time data for its complimentary base pair. As simplicity seems inherent there are actually many sources of error that the software must compensate for including read length, complimentarity, orientation, and individual DNA errors. The read length is affected by the physical stress of molecular translocation which causes the DNA to break in several places. It is expected that a common string size would be 10^5 base paitrs buts this number deviates unpredictably. Looking further into the programming, the detection system for the base pairs uses the de-Bruijn method for matching. The de-Bruijn method is a way to identify sequences by looking at the overlap between the known sequences. Within this system it is necessary to identify patterns of overlapping read sizes. The main identifying characteristics of the de Bruijn graph is to find four disjoint paths that are all equal in length and are Watson-Crick compliments or reversals of one another. The problems that arise from this include repeats in DNA sequences, being that they are not going to all be unique. To correct for this, the inputted sequences must be larger than the longest repeatable read in the input DNA. Running duplicate stands of DNA leads to better identifying characteristics and a better final product.

The current program was able to obtain DNA sequences that included 10^5 base pairs in under an hour and the bulk of this time was used to eliminate erroneous regions of DNA. The ultimate goal is to successfully sequence mammalian DNA using parallel processing power for which the explained program compensates for. The major inhibitors in full mammalian genome sequencing include the analysis of 10^9 base pairs and the compensation for its double chromosomal structure.

References:

Astier, Yann, Orit Braha, and Hagan Bayley. "Toward Single Molecule DNA sequencing: Direct Identification of Ribonucleoside and Deoxyribonucleoside 5'-Monophosphates by Using and Engineered Protein nanopore Equipped with a Molecular Adapter." Journal Of The American Chemical Society 128 (2006): 1705-710.

Bokhari, Shahid H., and Jon R. Sauer. "A parallel graph decomposition algorith for DNA sequencing with nanopores." Bioinformatics 21 (2005): 889-96.

"De Bruijn graph." Wikipedia. 7 Oct. 2009. 5 Feb.2009 <www.wikipedia.com>.

Kaptcianos, Jonathan. "A Graph Theoretical Approach to DNA Fragment Assembly." American Journal of Undergraduate Research 7 (2008).

Savage, Niel. "Faster, Cheaper DNA Sequencing." Technology Review. 25 Sept. 2008. MIT. 5 Feb. 2009.