CHEN 4760 Midterm Report

Jonathan Ward

Introduction and Background

The purpose of this project was to see if bacteria could be found in clogged heart arteries. It has been theorized that blood-born bacteria may be a cause of heart disease. Ordinarily, it should be very difficult for bacteria to enter the bloodstream, but certain types of dental work are believed to provide an avenue for entrance of mouth-born bacteria. Many of these types of bacteria are unculturable, making identification and even discovery difficult. For this project, we turned to genetic sequencing to aid in discovering and identifying bacteria in a sample taken from a human heart artery.

Genetic sequencing was historically a slow process because it required many hours of laboratory time that resulted in only a few hundred sequenced base pairs. High Throughput Sequencing (HTS) takes some traditional and new sequencing techniques and parallelizes their operation so that for the same amount of laboratory time, much more DNA can be sequenced.

Without HTS, the human genome project to map the human genome would never have been possible. In that effort, it took 10 years to sequence three billion base pairs. Since then, technology has improved rapidly. It is now possible to map the entire genome in a matter of weeks. By reducing the time required for sequencing, HTS vastly reduces the cost of producing a sequence. This in turn enables many sequencing projects like this one to go forward. To carry out this project with traditional technologies, huge grants and many months of work would be required. Now, it can be carried out on a modest budget with only a few days of an experienced researcher’s time.

Because so much research is enabled by HTS, high throughput sequencing machines are the new standard. Several types exist and more are in development. Currently, the most common type is the high throughput Sanger sequencing machine. This machine takes Sanger sequencing, traditionally done on slab gels, and parallelizes it to run in gel columns. Current machines of this type can sequence 96 columns in parallel. For each column, around 800 base pairs can be accurately sequenced. By running this machine for an entire day, sequencing of up to 490,000 base pairs is possible.

The Illumina DNA analyzer is one of several new types of HTS machines. The Illumina machine uses a massively parallel sequencing process to sequence many small strands, each about 30 bases long. By splitting a long DNA strand into many small chunks which are then sequenced, the sequence of the whole strand can be later assembled using mapping software. This machine is very fast and can sequence 600,000,000 base pairs in a day.

Both types of sequencer have research applications where they perform best. The Illumina DNA analyzer performs best when sequencing large strands very quickly. Its limitation is is that it isn't very good at sequencing completely unknown or *de novo* sequences. This is because each sequence chunk is only 30 bases long. If the sample is completely unknown DNA, it will be very difficult to map to a complete sequence. It is hard to get enough sequence overlap to link the sections together. This is where the Sanger sequencer shines. While slower than the Illumina machine, each read can be as long as 800 bases, providing a lot of room for sequence overlap and eventual mapping of a completely unknown sequence.

Both of these machines can be used to perform the sequencing of microbial DNA required for the project. To identify which microbes are present in the sample, we used Sanger sequencing to analyze the 16S rRNA gene. This gene is useful because it is highly conserved in many species, especially microbes. Analysis of variations in the gene can also be used to accurately perform species identification. A free online database of 16S rRNA sequences exists to help researchers to this end.

Materials and Methods

To amplify any DNA contained in the sample, degenerate PCR was used with a set of universal primers for the 16S rRNA gene. Multiple primers are needed because only amino acid sequences are preserved in genes and there are many DNA sequences that can get translated into the same amino acid sequence.

PCR amplification produces multiple copies of the sequence of interest bounded by the forward and reverse primers. It also creates copies of undesired fragments. Gel electrophoresis was used to get rid of these undesired fragments and isolate the sequences of interest. To extract the target DNA sequences, a section of the gel is cut away to harvest the target band.

In addition to running conventional gel electrophoresis to isolate the target DNA, Agilent's high sensitivity DNA analysis kit was used to quickly double check our results and provide more accurate measurements of sequence length and DNA quantity. The Agilent genechip runs its own electrophoresis on a tiny etched glass slide. The chip can analyze a very small sample and is highly calibrated to provide more accuracy than be achieved by hand with a gel slab.

Once the appropriate band was cut out of the agarose gel slab, the DNA needed to be separated from the gel. This was done with the QIAquick gel extraction protocol. After DNA extraction, we are left with a quantity of the target DNA amplified by PCR step. The DNA strands all have the same length, but because we used degenerate PCR with universal primers and an unknown sample, those strands are probably different types of DNA each with their own sequences. For proper sequencing using Sanger sequencing the DNA sample can only contain one type of DNA.

In order to separate out the different strands of DNA in the sample, a DNA library is created using *in vivo* cloning with TOPO TA Cloning Kit from Invitrogen. Another PCR step is performed to amplify the target DNA and a gel is run to check the results of the cloning procedure. For this PCR step, we avoid degenerate primers by using parts of the cloning vector sequence as forward and reverse primers.

A Sanger sequencing PCR reaction is now performed on the cloned DNA samples. We performed two separate sequencing reactions for each of the sixteen samples, one using the forward primer and one using the reverse primer. Using this method allows us to have better sequencing coverage. The end of a sequence is usually hard to read accurately. The Sanger sequencing reaction takes a sequence and produces fragments of different sizes capped in a terminator dye with a color corresponding to the base type. This is a statistical process and there is a higher probability of producing short fragments near the beginning of the sequence. It produces fewer long fragments with bases near the end of the sequence. By performing the reaction from both directions, we can get good reads at both ends.

After performing the sequencing reaction, the samples were sequenced using an ABI 3730 Sanger sequencing machine. The results were analyzed using SeqMan software.

Results and Discussion

After performing degenerate PCR, electrophoresis was performed to isolate the target DNA. A photograph of this gel can be seen in appendix A. Because I accidentally added sample DNA to both my control and experimental groups, bands appear in both runs. There is significant band blurring, without well defined bands.

The genechip analysis gave the same results, but with more accurate measurements. The genechip software shows a graph of sample quantity versus sequence size. To perform a similar task with an electrophoresis gel would require looking at how bright the bands are and trying to estimate the sequence size of the band with known sample markers and gel translocation properties. The lab-on-chip is highly calibrated to provide good measurement results. The downside to using it is that it is expensive at around $40 per chip and one cannot use it to isolate DNA like one can do in the conventional method by cutting out portions of the agarose gel.

A gel photograph of the direct PCR step performed after cloning can be seen in appendix B. Not all the bands appear bright and crisp. Especially troubling is that it appears that only about half the samples with have both forward and reverse coverage during sequencing. While many reverse bands are bright, many of the forward bands drop out.

Analysis using Seqman software showed no contiguous DNA of significant size and quality. The longest contiguous section was only about thirty base pairs. This sequence and others had only a few reads.

Once contiguous regions were mapped, these DNA sequences were compared to a database to see if they have been previously found and hopefully identify their origin. The Ribosomal Database Project is an online database of 16S rRNA genetic sequences in microorganisms and other species. By uploading our contiguous DNA sections to this database, we could learn what species contain a copy of our discovered DNA sequences.

The short contiguous sections we found did not match up with any known sequence in the ribosomal database. This could be because no DNA was present in the sample or the experimental methods were carried out poorly, resulting in erroneous results.

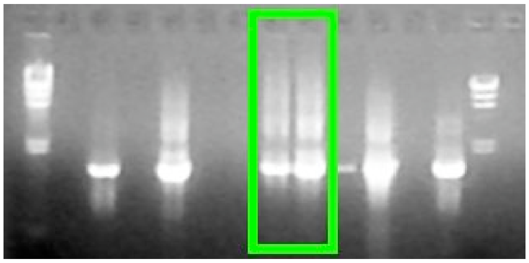
Suggested Future Project Development

We found little evidence of bacterial DNA present in the test sample. This could be because there was no bacteria present, but evidence suggests that the experimental procedure was carried out incorrectly. To carry this project forward, this procedure should be done again using more samples from different test subjects with each sample taken by a different person to remove any bias.

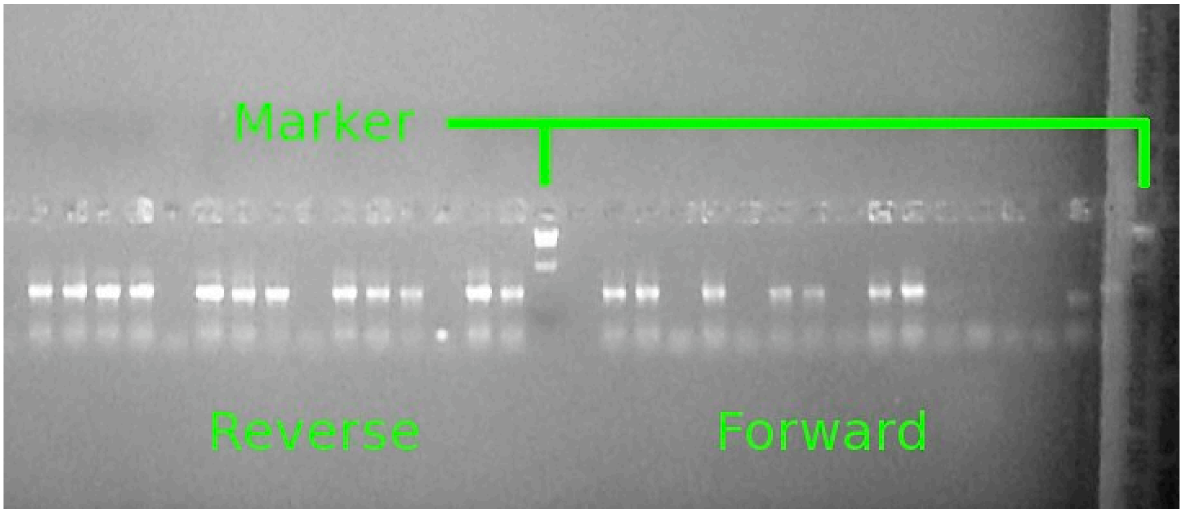
The methods carried out in this project could easily be extended to other projects to find what species' DNA is present in a sample. If a sample contains microorganisms that are not culturable, these methods are especially useful because it would be difficult to identify what species are present using other techniques.

Appendix

A



B



References

"DNA Sequencing." *Wikipedia: The Free Encyclopedia.*Wikimedia Foundation, n.d. Web. 20 March 2010.

"Ribosomal Database Project." Web. *http://rdp.cme.msu.edu*.

W G Weisburg, S M Barns, D A Pelletier and D J Lane. *16S ribosomal DNA amplification for phylogenetic study.* J Bacteriol. 1991 January; 173(2): 697-703Yu, Lin. *The March Toward the $1000 Genome.* E4760 class. Columbia University, New York, NY. Spring 2010.

Acknowledgements

I would like to thank Prof. Ju, Shundi Shi, and everyone who worked hard to make this class possible.