Appendix D - **UNIVERSITY HONORS PROGRAM**
SENIOR PROJECT - APPROVAL

Name: DAVID A. YAM

College: ARTS AND SCIENCES Department: BIOCHEMISTRY

Faculty Mentor: DR. BRUCE McKEE

PROJECT TITLE: COMPARATIVE DNA SEQUENCING AS A NEW APPLICATION FOR THE ATOMIC FORCE MICROSCOPE

I have reviewed this completed senior honors thesis with this student and certify that it is a project commensurate with honors level undergraduate research in this field.

Signed: Bruce D. McKee, Faculty Mentor

Date: May 9, 2000

Comments (Optional):

Comparative DNA Sequencing as a New Application for the Atomic Force Microscope

David A. Yam, Peter R. Hoyt, Bruce D. McKee, David P. Allison, and Robert J. Warmack

Department of Energy
Energy Research Undergraduate Laboratory Fellowships
University of Tennessee, Knoxville
Oak Ridge National Laboratories

May 8, 2000

Comparative DNA Sequencing as a New Application for the Atomic Force Microscope

David A. Yam, Peter R. Hoyt, Bruce D. McKee, David P. Allison, and Robert J. Warmack
Department of Energy
Energy Research Undergraduate Laboratory Fellowship
University of Tennessee, Knoxville
Oak Ridge National Laboratories
April 12, 2000

Figure 1 - AFM

ABSTRACT

The Atomic Force Microscope (AFM) has proven itself a useful tool for mapping restriction sites on large DNA fragments (Allison et al., 1997). This advance in optical imaging has spawned new ideas about how DNA sequencing may be accomplished. By studying these DNA/Protein interactions, the need for sequencing entire genomes may be eliminated altogether by use of the new technology. One experiment that may help eliminate entire genome sequencing is the detection of mismatch sequences and of insertions and deletions between any heteroduplexed single-stranded DNA fragments. This experiment involves two major phases of study: using MutS, a natural mismatch recognition protein, to detect mismatches in DNA molecules and the simple visualization of insertions and/or deletions relative to restriction sites. The combination of these two phases will potentially allow the sequencing of a genome or genome fragment of unknown sequence by comparing it to a sequenced genome. In conjunction with other AFM mapping techniques, this technology may allow the optical DNA sequencing of molecules up to or exceeding 35-kilobases. The practical use of such a development will allow the rapid, inexpensive, and straightforward sequencing of smaller genomes (viral or bacterial) or fragments of larger ones.

INTRODUCTION

The human genome is one of the millions of genomes yet to be fully sequenced. Currently, the estimated cost of sequencing the genome is approximately one dollar per base. This means that the final cost of the Human Genome Project (HUGO) will be over three billion dollars in addition to the decades-spent sequencing. This is an extremely large amount of money that essentially only sequences the genetic code of one species of the millions of species on the planet. Continued use current techniques and even some of the newest ones will undoubtedly cost a significantly larger amount of money, more time, and an inconceivable number of resources to finish. In an effort to reduce cost and increase the speed of sequencing genomes, scientists at Oak Ridge National Laboratories (ORNL) are researching a new technique that has the potential to shortcut years of work by using a relatively new type of microscope.

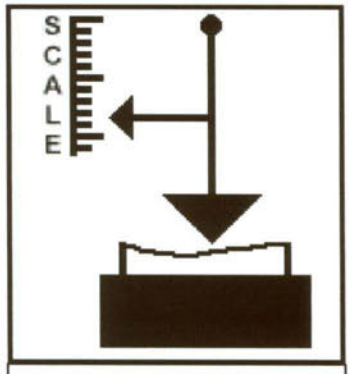

Figure 3 - Profilometer

The AFM was developed in 1986 for the visualization of both conductive and nonconductive samples, and it has just recently become prevalent in the biological sciences (AFM reference). A simple measuring instrument, the profilometer, which measures surface images of large objects, inspired its development. The profilometer operates by moving a styli, or tip, across the surface of an object, essentially measuring the changes in contour

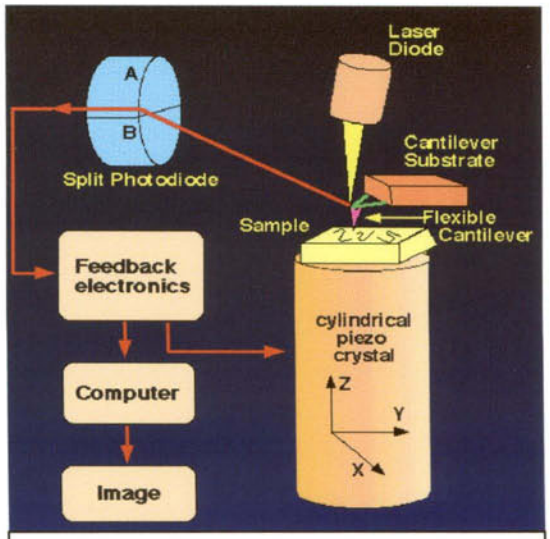

Figure 2 - AFM Schematic

of an object on a two-dimension plane. The AFM operates by moving a very small tip called a cantilever over the surface of a sample that acts much like the styli of the profilometer. As the cantilever moves, repulsive and attractive forces that the tip encounters cause the tip to shift. A laser that is pointed at the tip, and its rays are deflected into a four quad photocell. This photocell measures deflection of the laser, which is directly related to movement of the cantilever tip. The enhanced photocell measurement, the smaller tip, and the low loading force of the horizontal cantilever allow the AFM to measure on the scale of a single nanometer in three-dimensional space. This very small scale allows it to visually measure objects that could previously only be viewed with an electron microscope.

Use of the AFM in the biological sciences introduces non-conventional solutions to everyday problems faced in laboratories. Typically, an electron microscope is unnecessary when working with molecules like DNA and protein because of the expense, availability of electron microscopes, and difficulty of using the microscope. The biologists' other trusted tools: gel electrophoresis and spectrophotometry, do not give three-dimensional views of molecular interactions, but instead just detail attributes like mobility or absorbance that are related to what is actually occurring. This is where the AFM is useful. At a relatively low cost, the AFM creates high-resolution, three-dimensional views of molecular interactions that can be used solely as scientific evidence or in combination with traditional techniques to further elucidate the nature of any particular molecular interaction.

At ORNL, a new non-conventional technique for genome mapping was developed that relies on both conventional techniques and use of the AFM (Genomics, 1997). Allison et al. showed in 1997 that DNA/protein interactions could be visualized using the AFM, which includes binding of restriction enzymes, transcription factors, and other transacting

elements. In this experiment, lambda phage DNA, which contains six EcoRI restriction sites, was bound with a mutated form of EcoRI that binds to but does not cut the specific sequence of GAATTC. The enzyme bound DNA was then visualized under the AFM and it was shown that the restriction enzyme EcoRI bound six times at the six specific loci on the lambda DNA. This experiment showed that without cutting the DNA, the restriction sites of a DNA sequence could be elucidated using the AFM and proved that the tool might be useful in future studies of DNA/protein interactions.

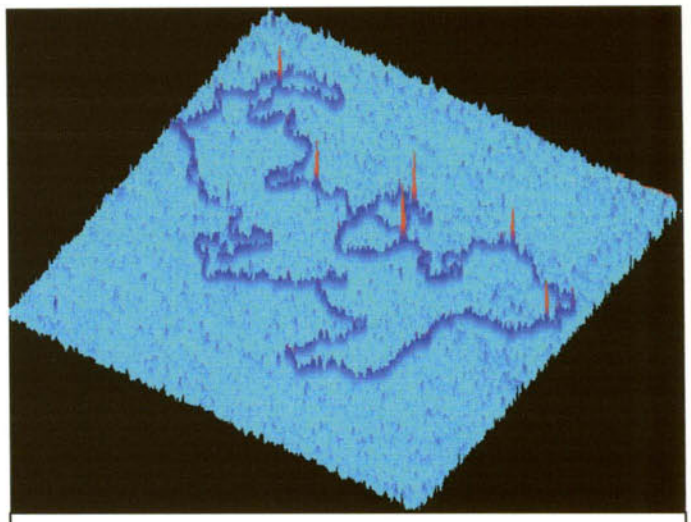

Figure 4 - EcoRI map of Lambda using AFM

The success of the experiment introduced an interesting new way of genome sequencing. Derived from the process of DNA hybridization and the work of Allison et al., a new method of DNA sequencing was proposed. The technique works because of the great deal of sequence homology between organisms. For example, human DNA is more than 99% homologous to that of apes and 95% homologous to mouse DNA. Named comparative sequencing, the technique essentially combines a sequenced DNA strand with a DNA strand of an unknown sequence to form a heteroduplex. This heteroduplex forms normal double-stranded DNA coils where the sequences are homologous, or alike. Where the sequences differ, the DNA develops contortions that can be visualized or tagged with bound proteins. Detection of these differences is the basis of comparative sequencing. Once the differences

are detected, the DNA regions that differ can be sequenced using conventional isolation and sequencing techniques to determine the exact base pair difference.

PURPOSE AND GOALS

The experiment will be to determine if the technique of comparative sequencing is a realistic strategy for the sequencing of DNA. The experiment consists of essentially two phases: 1) detection of base-pair mismatches and 2) visualization of insertion and/or deletion loops that develop between two different DNA sequences. If these regions can be identified, Atomic Force Microscopy can be used to reduce time required to sequence genomes, decrease the number of genes that need to be sequenced, reduce monetary costs involved in complete sequencing in genome projects, and produce accurate maps of genomes.

MATERIALS AND METHODS

In order to determine if the AFM can be used for finding mismatches and insertion/deletions in heteroduplexes, two major experimental phases were performed. The first phase of the project will be to determine if the AFM will work for finding mismatches. This process involves creating mutant plasmids, hybridizing the mutant plasmid with a wildtype plasmid, binding MutS to the heteroduplex, and visualizing the sample under that AFM. The second phase will determine if insertions and deletions in a plasmid can be detected with the AFM and if so what size insertions or deletions can be detected. For both experiments, plasmid DNA instead of full-length bacterial or viral genomes was used because of simplicity of preparation, isolation, and amplification.

Mismatch Plasmid Preparation

The first step in mismatch detection involved the engineering of specific plasmids that could be easily isolated and easily heteroduplexed. A single base pair mutation was

created in the 2961 bp pBluescript II SK+ (pBSSK+) by removing short sections of the plasmid and ligating a synthetic oligonucleotide sequence with a single base pair substitution back into the plasmid. The mutation interrupted a restriction enzyme recognition sequence that would make high-yield isolation possible by simply digesting prepared plasmid to remove any wildtype contaminants. These mutant plasmids were then transformed to DH5α *E. coli* using the heat shock method (heatshock technique). After selection of the appropriate colonies from selective ampillicin-treated LB plates, minipreps and agarose gel electrophoresis of the plasmid DNA were preformed to determine the success of the transformation. Once positive results were confirmed, a Qaigen Midi Prep Kit was used to amplify the plasmid for use creating in heteroduplexes.

Heteroduplexing

Two methods of heteroduplexing were tested for potential use. Both involve mixing ScaI linearized wildtype pBSSK+ and ScaI linearized mutant plasmid in equal concentrations. The plasmid DNA had to be linearized to prevent interfering effects of DNA supercoiling that would prevent double-stranded DNA from separating. The first heteroduplexing technique involves heating the linearized wildtype and mutant plasmids together at 95°C for ten minutes to melt the double-stranded DNA into a single-stranded form. Then the single-stranded DNA (ssDNA) was allowed to cool to room temperature, which formed some heteroduplex plasmids and many other random combinations of the single-stranded DNA. The second technique used formamide to theoretically reduce the number of random recombinations of ssDNA (Koehler, 1978). As little as 0.5 – 1.0 μg of each linearized plasmid were added to a reaction mix of Na_4EDTA and NaOH for 10 minutes, which denatures the double-stranded DNA into ssDNA. Then the reaction is

neutralized with Tris-HCl, pH 8.5, and brought to 50% formamide for at least 1 hour. After this formamide treatment, the DNA is separated from the formamide by gel electrophoresis. The DNA is then visualized by ethidium bromide treatment under UV light and the bands are excised. The excised bands contain both DNA and agarose, the later of which is removed by use of the Geneclean kit (geneclean). The DNA was then concentrated in TE to be used in the MutS binding reaction.

MutS Binding Reaction

The MutS reaction is performed using 0.9 μg of heteroduplex DNA with a 1:5000 dilution of MutS in 20 μL of 50mM HEPES, 100 mM KCl, 1 mM EDTA, 1 mM DTT, 10 mM $Mg(OAc)_2$ at 0°C for 10 minutes as specified by D. J. Allen of the Department of Biochemistry at Duke University Medical Center. Once this is accomplished the DNA/protein is immediately ready to be prepared for use with the AFM.

AFM Preparation and Use

After the binding reaction takes place, the DNA is placed onto freshly cleaved discs of high grade mica, incubated for 5 minutes at 25°C, rinsed in sterilized H_2O, 25% ethanol, 50% ethanol, 75% ethanol, and 100% ethanol, and dried in a critical point dryer (Hansma et al., 1993; and Thundat et al., 1994). The DNA / MutS complex was then visualized using the AFM set on tapping mode in a N_2 Gas chamber to reduce humidity to 20% around the samples.

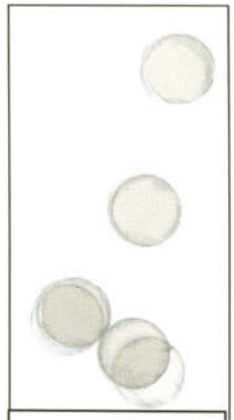

Figure 5 – Mica Discs

Insertion/Deletion Loop Detection

This phase of the project involves producing plasmids with deletions that when heteroduplexed with wildtype plasmid will form deletion loops. Deletions of 77, 156, and

243 base pairs were made in the pSV-β-Galactosidase Control Vector. An additional 180 base pair deletion was added using plasmids provided by Dr. Albrect Von Arnim of the University of Tennessee. The plasmids with 77, 156, 180 and 243 were cloned into HB101 cells by electroporation using a Bio Rad Gene Pulser. These plasmids were then isolated from the HB101 cells using the Qiagen Midi Prep Kit. The successfully isolated samples were then linearize using ScaI, heteroduplexed to the ScaI linearized wildtype plasmid DNA using the formamide technique, isolated, bound to mica, and visualized as described above.

RESULTS

Plasmid Engineering

Preparation of the single base pair mismatch plasmid was a success, but three out of the four deletion loop plasmids were unsuccessful isolated. Thus, the 77, 156, and 243 base pair deletion plasmids were removed from the study leaving only the 180 base pair deletion plasmid. The cause of the failure was the result of an unknown biological contamination either fungi or another bacteria that had ampicillin resistance. The significance of the omitted plasmids will be discussed in the Discussion.

Heteroduplexing

The use of two different heteroduplexing techniques was useful in that it allowed a comparison between the two techniques. Of the heating and formamide treatment techniques, the former took a shorter amount of time, but the latter gave significantly more useful results. Heat treated DNA formed a large number of unwanted incorrectly reannealed DNA strands as shown by the smear of DNA across the electrophoretic gel. The result is undesirable at best because improper reannealing would introduce error in the sequencing phase. Formamide treatment was significantly more successful and greatly reduced the

number of unwanted DNAs. The use of formamide to more accurately select the proper annealing appears to have worked. Efficiency was at 44% for heteroduplex formation which is high considered at most 50% of heteroduplexes where expected to form. The other 50% should be wildtype plasmid and mutated plasmid that resulted from identical strands reanealling. The deviation from 50% was probably caused by failed separation or by mixing unequal concentrations. However, in the final MutS trial, unwanted plasmid (wildtype or mutant without a mismatch) was removed by restriction digestion of the plasmids making the yield close to 100% for heteroduplexes after digestion.

MutS Mismatch Detection

The binding of MutS to the mismatch site yielded excellent results and was highly efficient; meaning that The reannealed wildtype or mutant plasmids did not The visualized DNA/protein molecule clearly labeled the mismatch site located in the middle portion of the plasmid. Using a computer program developed specifically for comparative sequencing, the length from either end of the plasmid (ScaI linearized ends) to the mismatch was within 2 nucleotide base pairs of the actual location of the mismatch. The results are shown in Figures 6 and 7. Under normal circumstances where the location of the mismatch is unknown, binding additional recognition proteins like restriction enzymes or transcription factors will be critical to locating which end of DNA the mutation is on and where positionally the mutation is within the DNA molecule.

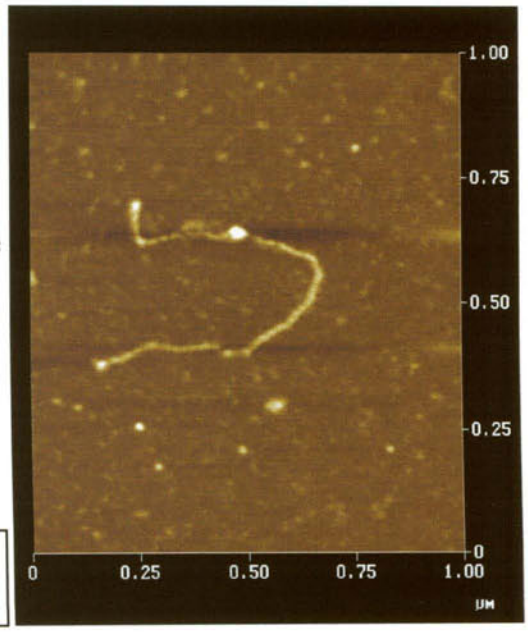

Figure 6 – Mismatch Detection

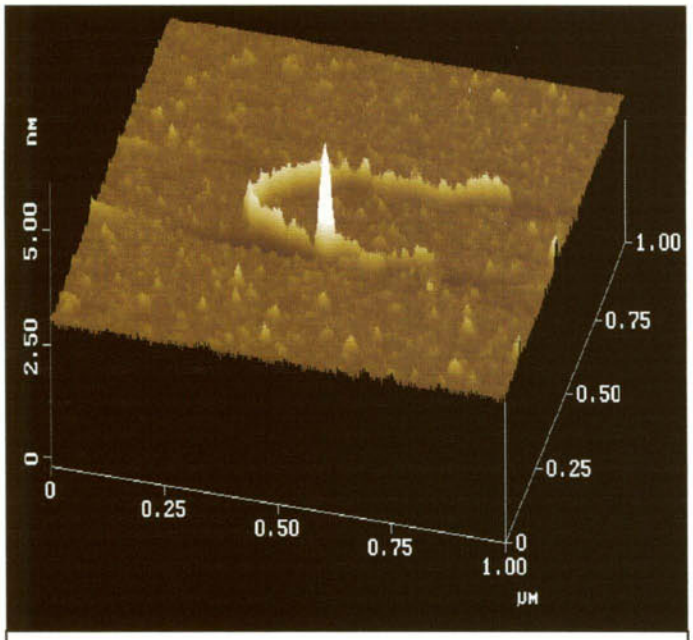

Figure 7 – Mismatch Detection

Insertion / Deletion Loop Detection

Despite the reduced number of samples due to the undesired contamination of cultures, the AFM was proven to be able to visually recognize insertion or deletion loops of at least 180 base pairs. The 180 base pair deletion loop was easily visualized, and its location was identified relative to an EcoRI binding protein. The estimated location of the loop was off by 8 base pairs which was most likely due to the erratic coiling and annealing introduced by the loop and the fact that it was performed manually. The manual measuring process involves tracing the length of the DNA by hand from the high resolution image and making a good estimate of the length by taking the average of several measurements. The results are shown in Figure 8 below.

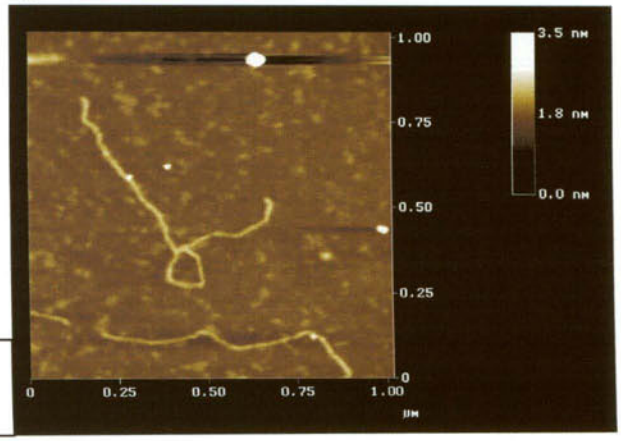

Figure 8 – Mismatch Detection

DISCUSSION

Use of the techniques above has shown that comparative sequencing appears to be a highly feasible means of sequencing an unknown genome that is homologous to a genome of known sequence. The tagging of the mismatches with the protein MutS is highly specific, and the mismatches were shown to be easily identifiable when visualized using the AFM. In addition, the use of the computer has allowed the pinpointing of nearly the exact location of the mismatch. This means that the isolation of the target region containing the mismatch can also be accomplished. Insertion or deletion loops of 180 base pairs or more can also easily be detected using the AFM by simple visualization of the DNA loop. The DNA loop is an obvious structure when viewed with the AFM, and its location was very nearly determined by manual analysis of the DNA, a result that will most likely improve with the development of appropriate software.

Overall, the technique has been proven to be nearly utilizable for use in genome sequencing. However, further study and improvements need to be done in a few crucial areas. First, the failure of the isolation of the 77 and 150 base pair deletion plasmids limited experimental results. Because only a 180 base pair deletion was visualized, it is not clear whether or not a deletion loop as small as 77 base pair can be detected and properly localized. The test of smaller deletion loops will need to be performed to determine the limit of the AFM. Second, the necessary automation of the process will be dependent on the successful update of the computer software to localize deletion loops. This is crucial because the manual process is open to a great amount of human error and is tedious to perform. Third, a more extensive list of reference proteins needs to be established so that more specific locations can be tagged. More restriction enzymes need to be used in the genome restriction

mapping and other groups of DNA binding proteins need to be used. One proposed group of proteins are transcription factors that would bind specifically to promoters. Fourth, a combined study of the mismatch recognition, deletion and insertion loop detection, and restriction mapping needs to be performed to determine if there are interferences that did not arise in each individual study. Finally, the range of homology for the experiment must be determined. The applicability or usefulness of this technology depends very highly on the homology since heteroduplexing, mismatch detection, and insertion/deletion loop detection depends so highly on homology. However, the technology may prove most useful in studies where homology is very high and only a few bases are predicted to differ. An ideal example would be a mutated viral strain or bacteria. Locating the differences in these kinds of situations seems ideal because of the speed at which the mutations could be located, localized by binding restriction enzymes, and excised and sequenced. If these five areas can be improved upon, the technique will be ready for use in and most likely useful to the scientific community.