

COVENANT UNIVERSITY

**INTERNATIONAL TRAINING
WORKSHOP**

**ON PATTERN DISCOVERY
IN BIOLOGY**

18-27 APRIL, 2005

MOTIVATION
ON
MOLECULAR SEQUENCE ANALYSIS

BY

ANUKAM KINGSLEY C

(B.MLS, M.Sc, MHPM, PhD-in-view)

DEPARTMENT OF PHARMACEUTICAL MICROBIOLOGY, FACULTY OF
PHARMACY, UNIVERSITY OF BENIN, NIGERIA

A PROBLEM SOLVED BY BIOINFORMATICS, USING BLAST HOSTED BY NCBI

**IDENTIFICATION OF VAGINAL
LACTOBACILLUS AS DETERMINED BY
PCR-DGGE AND 16S rRNA GENE
SEQUENCING USING BLAST AND
PHYLOGENETIC ALGORITHMS**

Some problems in Lactobacillus identification as it relates to my research

- There is paucity of information on the Lactobacillus species colonizing the vagina of African women
- Previous research is usually based on about 17 Phenotypic fermentation tests
- Yet identification to the strain level remains difficult.

Molecular Methods for bacterial species

- There are various techniques but the most widely used is the V2-V3 region of the 16S rRNA gene
- Nucleotide-Nucleotide base sequences provide an accurate basis for phylogenetic analysis and identification.
- The sequence obtained from an isolate can be compared to those of lactobacillus species held in data banks such as NCBI using the BLAST algorithm.

How do you obtain the sequence from micro-organisms or any cell?

1. **Extraction of bacterial DNA using Instagene Martix**
2. **Amplification of the DNA template using PCR with species specific primers or eubacterial primers**
3. **Denaturing Gradient Gel Electrophoresis (DGGE) with Dcode mutation system**
4. **Band Excision from the DGGE**
5. **PCR re-amplification with the cut bands.**
6. **Purification of the re-amplified PCR products.**
7. **Sequencing of the V2-V3 region of the 16S rRNA gene**

POLYMERASE CHAIN REACTION (PCR)

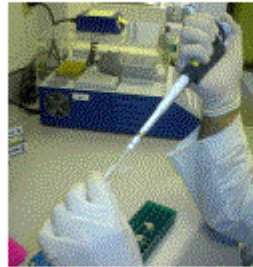
- The principle of PCR is amazingly simple. It is closely patterned after the natural replication of the genetic material, which occurs whenever a cell divides to form two new cells.
- In the first step of the polymerase chain reaction, *denaturation*, the two strands are separated by heating the extracted DNA to 94°C
- In the second step, known as *annealing*, two primers attach to the single strands. These primers are small synthetic stretches of single-stranded DNA
- In the third step, called *extension*, these short stretches serve as starting-blocks for the enzyme *Taq* polymerase. *Taq* polymerase adds the nucleotides complementary to the template at about 72°C, linking them together.
- This three cycle has been fully automated by the use of thermal cyclers. This piece of equipment repeats the same temperature program (94°C, 40-60°C and 72°C) over thirty times to make a billion copy of a given sequence.

PRINCIPLE OF DGGE

- Denaturing gradient gel electrophoresis has been shown to detect differences in the melting behavior of small DNA fragments (200-700base pairs) that differ by as little as a single base substitution
- The rate of mobility of DNA fragments in acrylamide gels changes as a consequence of the physical shape of the fragment
- The position of the gradient where a domain of a DNA fragment melts and thus nearly stops migrating is dependent on the nucleotide sequence in the melted region.
- Sequence differences on otherwise identical fragments often cause them to partially melt at different positions in the gradient and therefore stop at different positions in the gel.

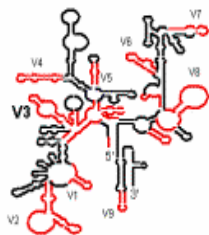
PCR-DGGE SUMMARY

Step 1.



Total DNA is extracted and purified

Step 2.



16S gene are targeted with specific primers and amplified

Step 3.



DGGE gel profile

Step 4.

→ GCCAGCCGTGTTTACAGA...
→ GCCTTAAGCAGGCCCTTCG...
→ GCCAGCCGTGTTTACAGA...
→ GCCGGCTCTAGCTTCCGT...

Prominent gel bands are excised and sequenced

Step 5.

Acinetobacter sp.
Geobacter sp.
Unknown sp X.
Methanosacina sp.

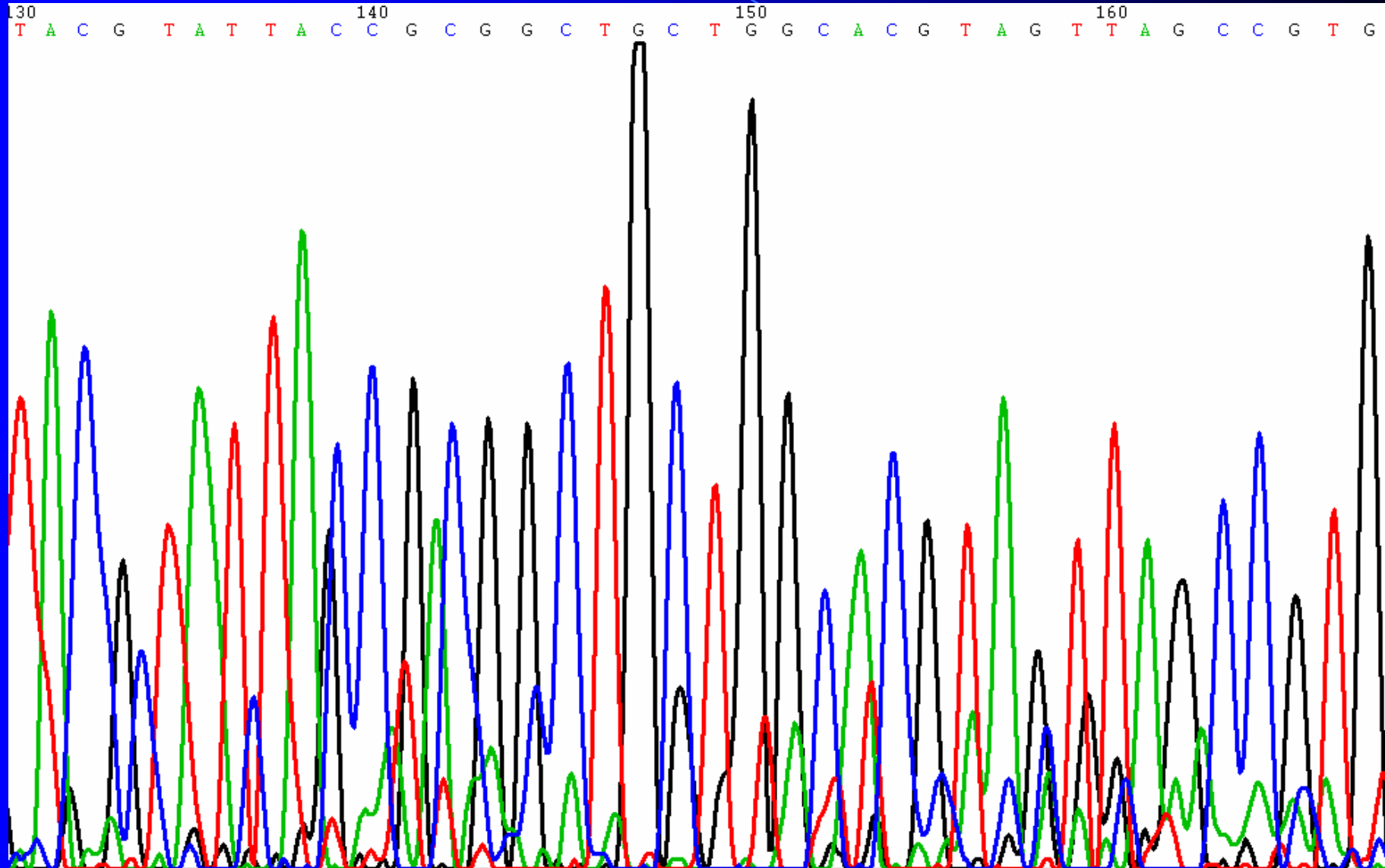
Phylogenetic analysis identifies organisms

SEQUENCING

- Gene sequencing techniques is based on the electrophoretic procedures using the high resolution denaturing polyacrylamide (sequencing) gel (e.g POP 7 TM)
- The key to determining the sequence of deoxynucleotides is to generate separate enzymatic or chemical reactions mixed with Big-Dyes that terminate at the variable end in Adenine (A), Thymine (T), Guanine (G) or Cytosine (C).
- After electrophoresis using the most recent ABI Prism terminators, laser beam is then channeled through the electrophoretic pathway of the micro-capillary array
- The four dyes that couples to each of the nucleotides fluoresces and a computer soft ware captures the electrophoretic patterns as shown.

Chromatogram showing the nucleotide sequence from the ABI Prism 3730x1 DNA Analyzer.

Red = Thymine, Green = Adenine, Blue = Cytosine, Black = Guanine.



Nucleotide-Nucleotide sequence results of the V2-V3 region of the 16S rRNA gene of Lactobacillus species as revealed by BLAST Gene Bank data base and manually corrected with soft-ware Chromas version 2.3.

```
AAAGACTCTGTTGTTGGTGAAGAAGGACAGGG  
GTAGTAACTGACCTTTGTTTGACGGTAATCAA  
TTAGAAAGTCACGGCTAACTACGTGCCAGCAG  
CCGCGGTAATACGTAGGTGGCAAGCGTTGTCC  
GGATTTATTGGGCGTAAAGCGAGTGCAGGCGG  
CTCGATAAGTCTGATGTGAAAGCCTTCGGCTC  
AACCGGAGAATTGCATCAGAAACTGTCGAGCT  
TGAGTACAGAAGAGGAGAGTGGAACTCCATG  
TGTACCGGTGAAATAAACCTATGTGTAT
```

How do you make meaning out of the sequences generated?

- This is where bioinformatics programs come in either with BLAST or BCB algorithm to identify which organism has a similar gene sequence from the gene data bank
- The genomic era has seen a massive explosion in the amount of biological information available due to huge advances in the fields of molecular biology and genomics.
- Bioinformatics is the application of computer technology to the management and analysis of biological data. The result is that computers are being used to gather, store, analyse and merge biological data
- Bioinformatics is an interdisciplinary research area that is the interface between the biological and computational sciences.
- The ultimate goal of bioinformatics is to uncover the wealth of biological information hidden in the mass of data and obtain a clearer insight into the fundamental biology of organisms.
- This new knowledge could have profound impacts on fields as varied as human health, agriculture, the environment, energy and biotechnology.

BLAST ALGORITHM

- BLAST (Basic Local Alignment Search Tool) [Altschul et al 1990; Karlin & Altschul 1993] is one of the most widely used similarity search tools available to today's computational biologist.
- It rapidly identifies statistically significant matches between newly sequenced segments of genetic material (as shown in table 1 for V2-V3 region of the 16S rRNA gene of Lactobacillus species) or proteins and databases of known nucleotide or amino acid sequences
- Such searches allow the scientist to make inferences about the structure and function of their discoveries or to screen new sequences for further investigations.
- BLAST has undergone nearly continuous

For Example

- The sequence result from my research on *Lactobacillus* determined by the automatic sequencer *ABI Prism 3730xl* is the query
- The sequence is entered in the BLAST search engine hosted by NCBI or **BCB search site** as the case may be.
- The program will search for similarities in the data bank and displays the organism's identity with the highest gene score accession number, and percentage similarity, etc.

Phylogenetic analysis

- A phylogenetic tree is a graphical representation of the evolutionary relationship between taxonomic groups.
- The term phylogeny refers to the evolution or historical development of a plant or animal species, or even a human tribe or similar group.
- A phylogenetic tree is a specific type of cladogram where the branch lengths are proportional to the predicted or hypothetical evolutionary time between organisms or sequences.
- Cladograms are branched diagrams, similar in appearance to family trees, that illustrate patterns of relatedness where the branch lengths are not necessarily proportional to the evolutionary time between related organisms or sequences.
- Bioinformaticians produce cladograms representing relationships between sequences, either DNA sequences or amino acid sequences

- Cladograms can rely on many types of data to show the relatedness of species.
- In addition to sequence homology information, comparative embryology, fossil records and comparative anatomy are all examples of the types of data used to classify species into phylogenetic taxa.
- So, it is important to understand that the cladograms generated by bioinformatics tools are primarily based on sequence data alone.
- The cladogram only illustrates the probability that two organisms, or sequences, are more closely related to each other than to a third organism, it does not necessarily clarify the pathway that created the existing relationships

Experience and Recommendation

- My research was done with the same method used in the Northern hemisphere, the finding shows that after all Black women harbour the same Lactobacilli, notably *Lactobacillus iners*.
- With the PHYLIP program, phylogenetic analysis was done showing the genetic relatedness of the various Lactobacillus species present in Nigerian women.
- Besides some strains exhibited **probiotic (Live micro-organisms, which when administered in adequate amounts, confers health benefits on the host)** potentials which are undergoing further studies.
- We envisage that BCB algorithms will be useful in future probiotic research with good collaboration.