CSC 366 Spring 2011

C-PLOP: Cal Poly Library of Pyroprints

Database Requirements Document
Microbial Source Tracking (MST), a branch of molecular forensics, is a relatively new field of study focused on the tracking of fecal contamination in foods and resource waters. One common approach used for MST investigations is the collection of a database or library of isolates from known fecal sources (e.g., raw sewage, pets, livestock). The database approach requires the isolation of fecal indicator bacteria (FIB), such as *E. coli*, from a variety of fecal sources, followed by database archiving of strain-typing data from these isolates. In this way, environmental isolates can be matched to isolates from the database and an environmental fecal source can be identified.

The project you will be working on in this course is part of a multi-class collaborative to construct an MST database at Cal Poly. In MCRO 224 (General Microbiology I), students isolate, identify and store *E. coli* from fecal samples obtained from a variety of hosts using selective and differential growth media. In BIO 161 (Introduction to Cell & Molecular Biology), these *E. coli* are further characterized at the molecular level to obtain discriminating fingerprints.

The data obtained by students in BIO 161 needs to be stored in a database. The initial design and development of the database prototypes will happen in Spring 2011 CSC 366, Database Design, Modeling and Implementation course. The software produced by you in this course will serve as the basis for the eventual deployed version of the database.

The goals of this project are:

(a) to design the database architecture for C-PLOP (Cal Poly Library of Pyroprints), a database of forensic footprints of the *E.coli* bacterial isolates;

(b) create a prototype database from the data provided to you, and

(c) develop software for communicating with the database, which implements the use cases specified in the supplemental C-PLOP Use Cases document provided to you.

The database and the software you develop will be used by the faculty in the Biology and Chemistry departments, students in a number of BIO and CHEM courses, and will eventually be made publicly available for any research group outside of Cal Poly to submit their information, or to use the database resources to find *E.coli* strain matches.

Short term, students in BIO/CHEM 441 (Bioinformatics Applications) course will evaluate the *E.coli* isolate information and use the software you construct in this course to further populate the database. Additionally, senior project students and Masters students in Biology, Microbiology and Biochemistry will use this database to identify sources of fecal contamination in environmental samples as part of their research. The database design is the key step to building an effective tool for our approach to microbial source tracking.

This document contains some introductory material designed to get you acquainted with the key processes used to produce forensic information on bacterial isolates. It also provides a full specification of the information that needs to be stored in the database.
Microbial Source Tracking and Pyrosequencing

Data to be Stored in the Database

Host Species
In Biology, a species is often defined as a group of organisms capable of interbreeding and producing fertile offspring. Often, a species will harbor other organisms, such as the bacteria that reside in the human digestive tract. From the point of view of the bacteria, the human would be considered a host. If there were more than one species that can serve as a host for a particular type of bacteria, they would be collectively called host species.

In terms of the database, Host Species has two fundamental features that will be used to distinguish between them: a common name, such as “Cat”, “Dog” or “Seagull” and a Latin name, such as “Felis catus”, “Canis lupus familiaris”, “Larus californicus”. Some species of bacteria are found in a large number of host species. For example, E. coli is found in the intestines of most warm-blooded animals. This large number of host species require a unique way of identifying them.

Hosts
A host is a unique individual belonging to one of the host species. The database will contain information about multiple hosts (individual warm-blooded animals) from which E. coli samples were drawn. For example, there may be three different horse hosts: horse1, horse2 and horse3. With each host, we associate the following information: a name of the specific host (e.g., “horse1” or “Mr.Ed”) and the host species.

Samples
Each host present in the database is there because it has provided the research group with a sample. The data from each sample, which is normally a fecal sample (AKA poop), needs to be recorded in the database. Each sample is produced by a specific host. The number of samples from each host can vary widely. Each sample needs a unique identifier and because each sample was collected on a specific date and at a specific site (location) these are the best data for identifying a sample. The location information is informal. Examples of locations include “Pismo pier”, “Cal Poly pig unit”, or “San Luis Obispo residence”.

Isolates and Strains
For each sample recorded in the database, individual isolates of E. coli are obtained and frozen in vials stored in the lab freezers. The number of isolates from a single sample can vary widely. Some samples may provide only one or two isolates, while others may have hundreds. These isolated bacteria may be genetically the same or different, even though they are all considered the same species. To distinguish between those that are identical and those that are different, the genetic material is examined (by making a pyroprint) and the isolates are separated into strains. Strains represent groups of individual isolates that may be distinguished from one another based upon differences in
their genetic material, but they are still considered part of the same species. In bacteria, isolates in the same strain would be like identical twins in animals – individuals with the same genetic material.

Individual isolates are obtained in the lab from the specific collected (poop) samples (see above). The isolates are stored on … plates and are frozen in laboratory freezers to preserve the biological material for further use in the experiments. The collection of isolates located in the labs supervised by the Department of Biology may be occasionally referred to as the “freezer stock”. Information about each isolate from the freezer stock needs to be stored in the database. This information includes the freezer location, the date the isolate was frozen, the person who isolated the E. coli (a student, a lab technician, or a researcher), and information on whether this isolate has already been pyroprinted.

**Pyroprinting**

For this molecular forensics database project to be successfully implemented in classrooms, a rapid, reproducible, discriminating source-typing method is required. Recent advances in DNA sequencing technology have provided us with such a method: pyrosequencing. Pyrosequencing produces a graph, called a *pyrogram*, with a series of light quantities corresponding to the incorporation of nucleotides (individual units that are linked together to make DNA) that are sequentially dispensed during the sequencing process. Using the total light output at each dispensation, one may obtain differences in peaks based upon differences in DNA sequence between strains. The distinguishing features of these pyrograms can be likened to how fingerprints distinguish between humans, thus we are referring to these data as *pyroprints*.

The pyroprints are the main objects to be stored in the database and they consist of pyrograms and the metadata required to identify and produce pyrograms. Pyrograms are obtained by taking one of the isolates from the freezer stock, performing a PCR reaction on it (to “amplify” particular segments of its DNA) and then using pyrosequencing equipment to obtain sequence-based information from the amplified fragments. Currently, there are four different pyrosequencing machines in the Biology department labs, more machines can be used with time. Each pyrosequencing machine can sequence in parallel up to 24 isolates (more can be sequenced by other pyrosequencing equipment). Each amplified product from each isolate is sequenced in a separate well, a location on the sequencing plate, where the strain DNA material is deposited. For a 24-well machine, the wells are arranged in an 8 x 3 pattern and each well is encoded with a combination of a letter (“A”, “B”, “C”) and a number (1-8) (e.g., “A1”, “B4”, “C7”). An isolate from a freezer stock can be pyrosequenced multiple times, producing multiple pyroprints. Each pyroprint needs to be stored separately in the database.

With each pyroprint we associate the freezer stock isolate used in the sequencing process, the date the pyrosequencing took place, the informal name of the region that was amplified and pyrosequenced (e.g., “16S-5S”), the pyrosequencing machine used, and the well where the isolate’s DNA was placed. Additionally, we need to record information about the PCR (date, PCR machine), and the information about the three *primers* used in PCR and sequencing: the forward primer, the reverse primer and the sequencing primer. We also associate the dispensation sequence used in the pyrosequencing and the name of
the person performing the PCR and pyrosequencing.

**Dispensation Sequence**

The result of the pyrosequencing operation is directed by the *dispensation sequence* of nucleotides (DNA building blocks) used in the sequencing process. Dispensation sequence is a particular order that nucleotides are dispensed into the reaction mixture. The light signal generated from each dispensation is proportional to the number of nucleotides incorporated into the new DNA chain. From each nucleotide dispensation, a *peak height for light output* will be recorded as part of the *pyrogram*. To track the unique properties of each pyroprint the database needs a unique way of identifying every dispensation sequence used by our lab.

The pyrosequencing machines store the data for each *pyrogram* in XML format. The database will associate the name of the XML file containing the pyrogram with the pyrogram record. XML files store the data for up to 24 pyrograms (identified by the *well Id*). For each pyrogram and for each dispensation sequence nucleotide, the following information needs to be stored in the database:

- peak height value
- compensated peak height value
- four *compensation slope* values

(see description of the pyrosequencing for detailed explanation of what the quantities are. The specific locations of this information in the XML file will be provided to you in further documentation).

The XML files describing the pyrptints and the information about extracting the pyroprint data and the dispensation sequences will be made available to you.

**Primer**

The enzymes that link nucleotides together to make DNA during the PCR and pyrosequencing reactions require a short string of nucleotides to start the process. These short strings help to prime the enzymatic reaction, so they are referred to as *primers*. Primers also direct the amplification and sequencing to take place at a particular locations in the DNA based upon their ability to bind to specific sequences. PCR uses two primers, called the forward primer and reverse primer, to direct the amplification of a specific segment of the DNA. Pyrosequencing requires only one primer, called the sequencing primer, to direct the DNA synthesis necessary for the light output measured in the pyrograms. Since this database may include pyroprints of more than one segment of DNA, different primers may be used for PCR and pyrosequencing. As a result, there needs to be a unique way of identifying each primer.

**References**

1. [http://www.sciencedaily.com/releases/2003/05/030526103731.htm](http://www.sciencedaily.com/releases/2003/05/030526103731.htm)