C-PLOP
Cal Poly Library of Pyroprints

Use Cases Document
1. Introduction

**C-PLOP**, Cal Poly Library Of Pyroprints consists of a database of *E.coli* isolate **pyroprints** and the Web-based suite of software tools for working with the database. This document presents information about the desired functionality of the C-PLOP software suite.

2. General Requirements

Throughout this document, we refer to the relational database storing the library of pyroprints as the **C-PLOP database**, and to the software designed to work with the database as the **C-PLOP software suite**. We refer to the combination of the C-PLOP database and the software suite as the **C-PLOP system** or the **C-PLOP project**. This section documents some general requirements for both the database and the software suite.

The C-PLOP system has to be designed and developed for deployment on a single computer running a web server (e.g., Apache web server) and a DBMS server that can be accessed through the web applications. It can be assumed that both the web server and the DBMS will be dedicated to supporting solely the C-PLOP system. The development teams are responsible for setting up the appropriate web server and DBMS.

The C-PLOP software suite shall be developed as a collection of web-based applications which run within a modern Web browser. Full browser compatibility (i.e., support for all modern Web browsers) is not required for the prototype software suite, but your software suite shall work properly on at least one major (Internet Explorer, Firefox, Chrome, Safari) web browser. Preference is given to browsers available on multiple platforms.

The language or languages of implementation for the C-PLOP software suite are not specified. Each development team can select the language to their liking subject to the obvious limitation that the development language of choice shall include database connectivity API, i.e., be able to access the C-PLOP database.

Eventually, the C-PLOP system will serve the wide community of biologists, environmental scientists, government agencies and commercial entities with an interest in microbial source tracking. However, the designated users of the prototype C-PLOP system to are faculty, staff and students from Cal Poly's Biology and Biochemistry programs. Further mentions of users in this document refer to representatives of the abovementioned group.

The rest of this document describes the full list of use cases (i.e., tasks to be completed by the users using the C-PLOP software suite) to be implemented. This list is a superset of what each development team will be asked to do. While some use cases must be implemented by each development team, other use cases may be designated for implementation by specific teams at a later stage in the project.
3. Use Cases: Overview

The full functionality of the C-PLOP software suite can be broken roughly into three main categories:

- **Database creation and maintenance.** This group encompasses a set of use cases for adding a variety of information to the C-PLOP database. The two key *dynamically growing* parts of the database are the **freezer stock library of bacterial isolates** and the **bacterial isolate pyroprint library** (see the C-PLOP data description document for more details). This data can be provided to the C-PLOP system in a variety of ways.

- **Database browsing.** Use cases in this category mainly deal with users browsing the information stored in the C-PLOP database, possibly using a variety of filtering conditions on the data to be viewed.

- **Forensic matching.** The key use of the C-PLOP library is source tracking. This is implemented in the C-PLOP system by comparing a pyroprint of an unknown isolate to the pyroprints stored in the database or by comparing library pyroprints to each other.

These use cases are described in more detail in the remainder of the document.

4. Use Cases: Database Creation and Maintenance

**UC-1. Bulk Insertion.** The user provides the C-PLOP system (a) a collection of CSV files containing information about the **freezer stock library** and the pyroprinting activities, and (b) a collection of XML files produced by the Qiagen pyrosequencers which contain the individual programs of the pyro sequenced bacterial isolates.

The C-PLOP system parses the CSV and XML files, and adds the data contained in these files to the C-PLOP database.

This is a **high-priority** and **low-frequency** operation.

This operation assumes that the CSV files provided contain **all** necessary information to correctly populate the C-PLOP database. The XML files may be optional: if they are not provided, the **freezer stock librar**, and the meta-data for the **bacterial isolate pyroprints** (including the information about the **primers** and the **dispensation sequences**) will be stored in the C-PLOP database, while the actual pyroprints (information found in the XML files) – won’t be.

**UI notes.** The C-PLOP system shall provide a UI that allows the user to specify the names of all CSV and XML files containing the information. Please note that while there may be multiple CSV files provided¹, each CSV file will contain information about a different part of the database (e.g., the isolate library, the pyroprint meta-data and so on). At the same time, there may be multiple XML files provided to you in a single run of the use case (the number of these

¹ The raw data archive will be provided to you after the database design stage is complete. This is done so in order to remove database design bias. The CSV file formats for the raw data are mostly accidental, and should not influence your design, yet, should the formats become known in advance, they will exhibit an influence. Because of it, this document is purposefully vague about the specifics of *how many* different CSV files will be provided, and what information will be in each of them.
UC-2-1. **Add Bacterial Isolate to Freezer Stock (isolate library).** The user provides information about a single bacterial isolate which was added to the freezer stock. The C-PLOP system parses the information and adds it to the database.

Note, that this use case may involve adding more than a single record to the list of bacterial isolates in the freezer library. Each bacterial isolate comes from a sample, which in turn comes from a host which comes from a host species. This information may need to be conveyed to the C-PLOP database if necessary.

This is a **high-priority, high-frequency** operation. Any time a new bacterial isolate is obtained and added to the freezer, this information will be passed to the C-PLOP database via this use case.

UC-2-2. **Add Pyrosequencing Data to the Library.** The user provides information about a pyrosequencing event for a bacterial isolate from the freezer stock, and an XML file containing the produced pyroprint.

The C-PLOP system creates database entries to store both the meta-data (who, when, how) and the actual program.

This is a **high-priority, high-frequency** operation. Each time a new pyrosequencing event occurs, the pyrograms and the information about them needs to be added to the database.

Note, that the actual pyrogram (i.e., the peak height values for each dispensation) will always be presented by the user to the C-PLOP system using the XML files produced by the Qiagen pyrosequencers. Meta-data about a pyrosequencing event includes information about various primers (see the C-PLOP data description document for more information) and the applicable dispensation sequence. In this use case it is assumed that all primers and dispensation sequence provided already exist in the database.

UC-2-3. **Add Primer data to the database.** The user provides information about a primer. The information is recorded in the database.

This is a **medium-priority, low-frequency** operation. Primers must be specified for each pyrosequencing event, but they can also be provided in bulk add (see UC-1). There are not too many different primers for each region of the bacterial DNA, so they won’t be added often.

**UI Note.** The actual primer is a relatively short DNA string. It should be possible to paste the actual string text into the UI.
UC-2-4. **Add Dispensation Sequence data to the database.** The user provides information about a dispensation sequence, including the sequence itself. The information is recorded in the database.

A *dispensation sequence* is a sequence of nucleotide letters which specifies in which order the reagents will be dispensed during the pyrosequencing process.

This is a medium-priority, low-frequency operation. Dispensation sequences are prerequisites to the actual pyrosequencing process, but they can also be put into the database via use case UC-1 (bulk insertion). Note that two pyrograms can only be compared to each other if they were built using the same dispensation sequence. Because of this, not too many different dispensation sequences will actually be used for data from the same DNA region.

5. **Use Cases: Database Browsing**

**UC-3-1. Browse freezer stock.** The user selects the view the freezer stock (bacterial isolate) collection. The information about the freezer stock is output in a tabular form. For each isolate, its key features (host species, host, sample, sample date) are reported. The table can be sorted in a number of different ways (e.g., by host species, by sample date, by isolation/freezing date).

This is a medium-priority, high-frequency operation. In general, viewing available DNA material may be useful for various housekeeping activities. However, most of the user interactions with the C-PLOP software suite will be in pursuit of a specific goal.

**UI Note.** The UI implementing this use case can be combined with the UI implementing use case UC-3-2.

The user should have the opportunity to examine the details of each isolate. One way to do it is to provide hyperlink/button/clickable environment for each line in the table of isolates, and provide a new (or a pop-up) page containing the full information about a specific isolate from the freezer library when the hyperlink/button is clicked.

**UC-3-2. Filter freezer stock.** The user selects to view a portion of the freezer stock which matches the criteria specified by the user via the special filtering UI. C-PLOP software suite shall support a wide range of filtering parameters for the freezer stock library. In particular it should be possible to filter the freezer stock list using the following information:

- Host species
- Specific hosts within a host species
- Specific fecal samples
- Date or range of dates for strain isolation/freezing
- Status (pyrosequenced or not)
- Person who isolated the strain
- Freezer location
This is a medium-priority, high-frequency operation. This use case addresses user information needs of the type “find all isolates in the freezer stock such that X”. Examples may include “find all isolates obtained from cats”, “find all isolates prepared by X between January 15 and February 20 of 2011” and “find all isolates obtained by cows in September of 2010, which have not yet been pyrosequenced”.

UI Note. The filtering UI can be combined together with the UI needed for UC-3-1.

UC-3-3. Browse list of pyrograms. The user selects to view the collection of pyrograms stored in the database. The system outputs the information in a tabular form. For each pyrogram, its key features (freezer stock isolate, pyrosequencing date, sequencing region, primers, dispensation sequence) are reported. The table can be sorted in a number of different ways (e.g., by host species, by isolate, by pyrosequencing date, by sequencing machine).

This is a high-priority, high-frequency operation. Most of the searches conducted on the database will be on the program data. This use case and UC-3-4 facilitate that.

UI Note. The UI for this use case can be merged with the UI for UC-3-4.

Individual entries in the list of pyrograms shall be “clickable”. Selecting an entry in the table shall lead to the use case U-4.

UC-3-4. Filter list of pyrograms. The user selects to view a portion of pyrogram library which matches the criteria specified by the user via the special filtering UI. C-PLOP software suite shall support a wide range of filtering parameters for the pyrogram library. In particular it should be possible to filter the pyrogram library using the following information:

- Sequencing region (high priority)
- Primer(s) used (high priority)
- Dispensation sequence used (high priority)
- Freezer stock isolate (high priority)
- Freezer stock isolate sample, host, host species (high-to-medium priority)
- Date of sequencing (high priority)
- Person performing sequencing (medium priority)
- Sequencing equipment, well (medium-to-low priority)
- PCR information (date, machine) (medium-to-low priority)

This is a high-priority and high-frequency operation. A lot of the analytical work with the database will be performed by first accessing the library of pyrograms, and selecting a specific group to observe.

UI Note. The filtering UI can be merged with the UI for use case UC3-3.

Individual entries in the list of pyrograms shall be “clickable” (see UC 3-3). Selecting an entry in the table shall lead to the use case U-3.
**UC-4. View full information about individual pyrogram.** User selects a pyrogram to view. The selection can be made either from the list of available pyrograms displayed as discussed in use cases **UC-3-3** and **UC-3-4** or from a separate “location” in which user provides the system with the information uniquely identifying a specific pyrogram in the pyrogram library. The system displays a web page describing the pyrogram. The web page shall have conveniently located information (meta-data) about the pyrogram. It shall also contain a graphical depiction of the pyrogram itself in a form of a bar chart representing the **peak heights** of the pyrogram.

This is a **high-priority** and **high-frequency** operation.

**UI Note.** A dedicated HTML page is required for the display of information about an individual pyrogram.

**UC-5. Search inside pyrograms.** User specifies a substring of a dispensation sequence of length \( N \) and a list of \( N \) numbers representing peak heights. The system searches for pyrograms whose dispensation sequence contains the specified substring, and where the peak heights for the positions represented by the characters from the substring are as specified in the input. A margin of error can also be provided by the user.

For example, if a user provides a substring “ATGT” and a list of numbers “250 150 0 75”, the system will find all dispensation sequences that contain a substring “ATGT”, and for each occurrence of this substring in the dispensation sequence and for each pyrogram built using the sequence, compare the four peak height values with the “250 150 0 75” quadruple. There are two types of comparisons that can be implemented. If both are implemented, the UI shall contain the appropriate “switch” functionality.

**Comparison 1: direct compare.** The system will compare the absolute values of the peak heights in the pyrogram with the specified values. If all values fall within the margin of error range around the user-specified values, a match is declared, and the pyrogram is retrieved.

**Comparison 2: Pearson Correlation.** Just as one can use Pearson correlation to compare full pyrograms, it can also be used to compare a pyrogram fragment to the input sequence of numbers. The system computes the **Pearson correlation coefficient** between the user-specified sequence of numbers and peak heights from the pyrogram. If the computed coefficient is within the margin of error from 1.0, the pyrogram is retrieved. (See **UC-6-1** and **UC-6-2** for more information on the use of Pearson correlation in the project).

The pyrograms shall be displayed in a tabular fashion similar to the ones described in **UC-3-3** and **UC-3-4**. The location of the match shall be shown in one or more columns of the table (as needed).

This is a **low-priority, low-frequency** operation.
6. Use Cases: Forensic Matching

UC-6-1. Find matches for a single pyrogram in the entire library. The user provides the system with information about one pyrogram. The system searches the library of the pyrograms and retrieves and displays, in the descending order of the similarity/confidence score a list of pyrograms that are found to be the closest matches.

Information specification. There are three ways in which the user can provide information about a query pyrogram:

- **New pyrogram data from the sequencer.** User provides the name of an XML file that contains the query pyrogram, and the well number. The system shall extract the data about the pyrogram from the XML file and use it in the rest of the use case. Note, that this method supplies only the pyrogram – no origin/freezer stock isolate information is provided beyond what can be extracted from the XML file itself.

- **New manually entered pyrogram.** User provides the following inputs to the system:
  - the DNA region under study;
  - a string in the {"A", "T", "C", "G"} alphabet representing the dispensation sequence of the query pyrogram, OR the name of an existing dispensation sequence.
  - a sequence of numbers representing the peak heights in the query pyrogram.

- **A pyrogram selected from the library.** User browses/filters the list of pyrograms (see use cases U-3-3 and U-3-4) and selects a pyrogram. The user then initiates search.

Matching pyrograms. Each pyrogram stored in the database has two sets of numbers that can be used for comparison: the original peak heights obtained from the pyrosequencer and the compensated values computed by the analytical software supplied by the pyrosequencer vendor. Both can be used to match pyrograms. Each time a matching is performed, the system shall provide the user with the ability to select the set of numbers to be used in the operation. The default is to use the raw peak heights. The descriptions below apply to both raw peak heights and compensated scores.

There are different ways in which similarity between two lists of numbers can be established. The most important of these is **Pearson correlation**.

Let \( X = \{x_1, \ldots, x_n\} \) and \( Y = \{y_1, \ldots, y_n\} \) are two sequences of numbers (in the context of this project: two pyrograms obtained for two isolates from the same region and using the same dispensation sequence and primers). The **Pearson correlation coefficient** between \( X \) and \( Y \) is computed as follows:

\[
 r(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

Here, \( \bar{x} \) and \( \bar{y} \) are mean values of sequences \( X \) and \( Y \) respectively.
Pearson correlation coefficient ranges from -1 to 1, with -1 being interpreted as the two sequences being “direct opposites” of each other, 0 being interpreted as the two sequences being independent of each other, and 1 being interpreted as the two sequences representing exactly the same behavior.

Two pyrograms *match* if their Pearson correlation coefficient is greater than or equal to a special pre-defined value $r_{\text{match}}$ sometimes referred to as the *match threshold*. The actual value of $r_{\text{match}}$ may vary and may have to be specified explicitly. At present, a value $r_{\text{match}} = 0.997$ is used, however this may change in the future, the software suite shall support user-specified match threshold values.

The system shall provide the user with a variety of choices on what to report. The choices are:

- report only the matches with scores equal to or above the match threshold;
- ask the user for a *match threshold* and a second number: *interestingness threshold* (usually, a smaller than the *match threshold* number). Output all pyrograms whose similarity to the query pyrogram is above the *interestingness threshold*, mark all *matches* among the output.
- Output top $N$ ($N$ supplied by the user) pyrograms ranked/ordered by the similarity score, regardless of the actual similarity values. (Mark all *matches* if the *match threshold* is provided).

This is a **high-priority** and **high-frequency operation**.

**UC-6-2. Find matches for multiple pyrograms in the entire library.** The user provides the system with information about multiple pyrograms. The system searches the entire library of the pyrograms and retrieves and displays, for each input pyrogram, in the *descending order of the similarity/confidence score* a list of pyrograms that are found to be the closest matches.

**Information specification.** There are two ways in which the information can be specified.

- **XML File(s).** The user provides the system with one or more file names for the XML files produced by the Qiagen pyrosequencing equipment, and specifies, where necessary the wells to be used from each file (if no well are specifies, *all wells will be used*).
- **Pyrograms selected from the library.** The user browses/filters the library of pyrograms and selects and adds to the query list the desired pyrograms from the library. (Notice that this may cause some changes in how you approach UI development for use cases UC-3-3 and UC-3-4).

**Matching.** See UC-6-1 for detailed description of the matching process and possible outputs.

**UI Notes.** The system shall report the list of closest matches for each of the pyrograms in the query list.

This is a **medium-priority** and **medium-frequency** operation. Generally speaking, it is subsumed by a number of consecutive runs of UC-6-1. From this perspective, it can be viewed as purely a convenience feature – the one that will allow for faster turnaround time in getting data from the C-PLOP software suite.

**UC-6-3. Find matches for a single pyrogram in a subset of the library.** The user provides the system with information about a single pyrogram, and specifies the subset of the pyrogram library in which to search for matches. The system compares the input pyrogram to the pyrograms in the
selected subset of the library and retrieves and displays, in the descending order of the similarity/confidence score a list of pyrograms that are found to be the closest matches.

For information on how the input pyrogram is provided see UC-6-1. For information on selecting the subset of the library, see UC-3-4. For information on matching and output see UC-6-1.

This is a high-priority and high-frequency operation.

**UC-6-4. Find matches within a group of pyrograms.** The user provides the system with information about multiple pyrograms. The system compares the pyrograms in the group to each other. It then displays all matches found.

*Note:* this use case is NOT an extension of UC-6-3 onto multiple pyrograms. Rather, here, a single set of pyrograms is specified. Pyrograms in the set are only compared to other pyrograms in it!

The input pyrogram set can be specified via either of the means specified in UC-6-2.

**UI notes.** For the output, the user shall be able to provide one of two requests: “succinct output” or “verbose output”. In the former case, the system shall only report all observed matches. In the latter case, the system shall report the matrix of pyrogram similarity scores (e.g., Pearson correlation coefficients) for each pair of pyrograms considered.

This is a high-priority and high-frequency operation.

**UC-6-5. Export match data.** Upon examining the information provided in the output for the use cases UC-6-1, UC-6-2, UC-6-3 and UC-6-4, the user requests (via some UI interaction with the system) that the results are exported. The system exports the results to the file whose name, location on disk and format are specified by the user.

**Export data formats.** The main export data format is CSV – comma-separated values files. The specific organization of data in a CSV file is left underspecified at this moment and will be subject to further clarification.

This is a medium-priority and high-frequency operation.

**General Note 1.** Please note, that in all use cases related to matching pyrograms, the following needs to be observed. It makes sense to compare two pyrograms to each other only if, (a) the two pyrograms represent the same DNA region and (b) both pyrograms were produced using the same dispensation sequence. When primers are known/supplied, they must coincide too, however, in some use cases above information about primers may be not included in the input.

**General Note 2.** Please note that the filtering mechanism (described in UC-3-4 and used in various other use cases) should include the opportunity to select all pyrograms in the library except a specified group. For example, it should be possible to specify a filter for “all pyrograms except those coming from seagull hosts”.