CLiC Microscopic Data Organization Database (Group 68)

Stone Chen, Yunwen Ji, Wenquan Li, Yuki Zhang

February 4, 2019

1 Requirement Analysis

1.1 Introduction

1.1.1 Purpose

The database revolves around the CLiC raw data in building its connections with three most relevant components: the lab member, the experiment, and the analysis results. This architecture would benefit all users (most likely the lab members) in efficiently retrieving pertinent information, such as the individual responsible for a dataset, the experimental procedures, the file PATHs, etc.

1.1.2 Scope and Special Design Choice

We limit the sphere of our interest to only the biomedical imaging component of the lab, as there is huge need for safe and efficient data management, particularly for this branch of the lab. Note that this database is not responsible for storing multidimensional arrays such as videos, images, and analysis related information; instead, those information are backed up in the lab cloud drive, of which the respective PATHs are saved as attributes to corresponding entities.

1.1.3 Terminology

CLiC Microscopy: Convex Lense induced Confinement Microscopy

1.1.4 Resources

https://pubs.acs.org/doi/10.1021/ac101041s. Please also see the appendix for a brief overview of this method.

1.2 Database Description

1.2.1 Entities and Attributes

- **Lab Members**: Those who are currently working or have worked in the lab, including research assistants and researchers. They both share the attribute name, contact information, and status. Name is key attribute as there is very low chance of duplication. Status can have value of active/inactive, indicating whether the individual still works here.
- **Research Assistant**: A research assistant (undergraduates) is a subclass of lab members, it has no additional attribute but unique relationships.
- **Researcher**: A researcher is a subclass of lab members with a special attribute named level, which indicates the degree/job level in the lab, for example, PhD. It also has unique relationship compared to research assistant
- **CLiC Raw Data**: The raw data is generated using a microscope installed with CLiC in the format of TIFF time series. This multidimensional array will not be saved directly in the database but rather backed up in the cloud drive, where the unique PATHs are captured by our database. This entity has an

artificial ID as key as well as an additional attribute date. Furthermore, since there are different types of data generated, which will subsequently be studied differently, this raw data contains two subclasses depending on the type of data.

- **Nanoparticle Data**: This subclass of raw data represent the data used in the studies of nanoparticle. It has one unique relationship compared to the other subclass.

- **DNA Unwinding Data**: This subclass of raw data represent the data used in the studies of DNA unwinding. It has one unique relationship compared to the other subclass.

- **Binding Analysis Result**: This entity represents the result analyzed by research assistant based on DNA unwinding data. These results are different from nanoparticle analysis results in the procedures performed as well as the information saved, which includes binding count, diffusion coefficient etc. Those results are not saved directly as attributes but backed up in the cloud drive, of which the PATH is stored as an attribute of this entity. In addition, this entity contains two other attribute: date and analysis type. Date is just the date those results are generated, while analysis type can have values of 'rough','decent' or 'finalized', which represents the degree of confidence in the accuracy of a particular result. We decide to save this as an attribute because this can help users decide whether or not they would like to get the PATH and check this result. For example, if a result is of analysis type 'rough', the user may not consider it valuable to its purposes. Finally, the key of this entity is an artificial ID.

- **Nanoparticle Analysis Result**: This entity represents the result analyzed by research assistant based on nanoparticle data. These results are different from DNA binding analysis results in the procedures performed as well as the information saved, which includes mean square displacement, number of particles etc. Those results are not saved directly as attributes but backed up in the cloud drive, of which the PATH is stored as an attribute of this entity. In addition, this entity contains two other attribute: date and MSD (Mean Square Displacement). Date is just the date those results are generated, while MSD can assume certain numerical values. We decided to save MSD as an attribute because again it is an important analysis result that a researcher would like to know before taking a closer look at the whole dataset. For instance, if a researcher decides to review all results with MSD larger than k, then this setup makes it convenient to do so. Finally, the key of this entity is an artificial ID.

- **CLiC Dashboard**: The entity represent the quality control result of CLiC raw data, which is performed before all analysis and presented in format of a dashboard. It includes general data information and specific quality information that includes the focus, SNR (signal-to-Noise Ratio) and contaminant level. All those information will be backed up in cloud drive, where the corresponding PATH is saved as an attribute of this entity. In addition, since quality information are the strict requirements to deem a raw data as valuable for analysis, we decided to save them as attributes so that users can quickly obtain these information before proceeding to analysis. Last but not least, this entity also has an attribute named date to store the date these results are generated. Finally, the key of this entity is an artificial ID.

- **Protocol**: This is the outline of experimental procedures, assays and safety information etc. Researchers formulate protocols before performing experiments so that they ensure they have made comprehensive planning. Each researcher has their own ways of naming their protocols, including typically their initials and some keywords regarding their experiment, which is always unique. Thus we use those protocol IDs as key of protocol. Since those protocols are usually saved and backed up somewhere, an attribute of their PATHs is also saved.

- **Microscope**: This entity represents the microscopes, which are used to perform experiments and generate raw data. There are currently two sets of microscopes in the lab and a third one is underway. Each device has a unique ID that starts with 'Nikon' and a number, for example 'Nikon 1'. Therefore, those unique microscope IDs are used as the key of this entity. Furthermore, those microscopes can be on temporary maintenance sometimes so an attribute named status is used; it assumes values of either 'active' or 'inactive'.

- **Experiment**: This weak entity represents an experiment that follows a certain protocol with a particular set of microscope. Often times, researchers replicate experiments using the same protocol to collect more data or verify their results. To represent this in ER model, experiment is designed to be a weak entity that is dependent on a protocol; it has a partial key named 'try number' that indicates the replicate number. For example, if try number is equal to 3, then this particular experiment is the third replicate of a certain protocol. Obviously this try number is not unique unless it is tailored to a specific protocol that has a unique protocol ID.

1.2.2 Relationships

- **is a (from lab members)**: *Lab members* have two different subclasses: *researchers* and *research assistants*, who conduct distinctive tasks in the lab.

- **supervise**: A *researcher* supervises a *research assistant*. This is a many-to-many relationship because a researcher can supervise more than one research assistant and one research assistant can be supervised by more than one researcher. There is a participation constraint on the research assistant, which requires that every research assistant must be supervised by at least one researcher.

- **formulate**: A *researcher* formulates a *protocol*. This is a many-to-one relationship because a researcher can formulate many protocols but a protocol can only be formulated by one researcher. There is also a participation constraint on the protocol because every protocol must be formulated by at least one researcher, indicating that every protocol must be formulated by exactly one researcher incorporated with the key constraint.

- **generate**: A *researcher* generates a set of *CliC raw data* through an *experiment*. This is a many-one-many relationship because an experiment can only be conducted by one researcher but a researcher can conduct many experiments. In addition, a researcher can generate many sets of data and a set of data can be generated by multiple researchers.

- **using**: The *experiment* uses the *protocol*. This is a one-to-many relationship since one experiment can only apply one protocol while one protocol can be utilized by replicate experiments. There is a participation constraint on the experiment because every experiment must be conducted according to at least one protocol, indicating that every experiment uses exactly one protocol incorporated with the key constraint.

- **with**: An *experiment* is conducted with a *microscope*. This is a one-to-many relationship because an experiment can only use one microscope while one microscope can be used by plenty of experiments. There is also a participation constraint on the experiment because every experiment must be conducted with at least one microscope, indicating that every experiment must use exactly one microscope incorporated with the key constraint.

- **is a (Clic Raw Data)**: The CliC new data has two subclasses: nanoparticle data and DNA unwinding data.

- **Assess**: The *research assistant* assesses the *CliC new data* to generate the *CliC dashboard*. This is a many-one-many relationship since one research assistant can assess more than one set of CliC raw data to get more than one CliC dashboard. One set of CliC raw data or one dashboard can only be assessed by one research assistant and one dashboard can only be generated from one set of raw data. There is a participation constraint on the CliC dashboard because every dashboard must be generated by at least one research assistant and at least one set of CliC raw data.

- **Analyze (DNA unwinding data or nanoparticle data)**: The *research assistant* analyzes the *DNA unwinding data (nanoparticle data)* to generate *binding analysis results (nanoparticle analysis results)*. This is a many-one-many relationship since one research assistant can analyze more than one set of DNA unwinding data to generate more than one set of binding analysis results. One set of DNA unwinding data or binding analysis results can only be analyzed or obtained by one research assistant. One set of binding analysis results can only be produced by one set of DNA unwinding data. There is a participation

constraint on the binding analysis results since every set of binding analysis results must be achieved by at least one research assistant and at least one set of DNA unwinding data.

2 Functional Analysis

2.1 Overview

2.1.1 Why do we need to store the CLiC data and its derivatives?

The lab members perform their research utilizing this microscopy technique extensively, which means that a huge amount of data is generated every day. Those data play a critical role in testing hypothesis and supporting conclusions in the day-to-day work of every researcher. Therefore, the lab has already set up cloud and local drives for backup purposes as well as a set of strict rules to follow in storing those data.

2.1.2 Why do need this database?

Despite those data are safely stored, researchers in the lab still complains about the difficulty to retrieve those data and their relevant information efficiently. For example, a researcher could not recall whether a dataset has been analyzed or not and spent a whole day looking for it on the cloud. He ended up finding nothing so he asked a research assistant to analyze it again which took him another whole day. This example can be even worse if that data has indeed been analyzed but just no where to find. Therefore, if there can be a magic button that one can simply press to find and present all they want in front of them, that will be exactly what we long to have. In practice, we can have something very similar which only requires more than one simple press — the database that store the links between CLiC data and all its relevant aspects!

2.2 Application Description

The central component of this database is the CLiC raw data and the application seeks to link all relevant components, which includes three major components: the lab members, the experiments, and the analysis results. Currently, those that need to be backed up on the hard drive are consisted of research protocol, CLiC raw data, CLiC dashboard, and CLiC data analysis results (in order of when they are generated). As soon as those are backed up on the cloud, their corresponding component in the database is expected to be updated accordingly. In this way, if one needs to retrieve relevant information on one dataset, all relationship that stems from the raw data entity will offer immediate feedback. Potentially, we would like the database to have the following functions:

2.2.1 Experimental details? Is it legit?

Researchers and their PI usually review the experimental details before looking at the results to ensure the data is valuable to answer some questions. Given a particular dataset, this database can report on where the protocol is, which microscope was used as well as the number of replicates (try number) etc.

2.2.2 Data analyzed? I need it NOW!!

The main intersection between the work of a researcher and a researcher assistant is one ultimate question — Is this dataset analyzed? This database will help them interact better and more efficiently since this question can be answered by checking whether there is a analysis result entity with date specified linked to the particular set of raw data. Researchers will be able to know whether their assistants are idling, while research assistants can have a clearer idea of the upcoming work to do.

2.2.3 This dataset is of such poor quality!! Prove it.

From time to time, some dataset generated can be of poor quality and luckily we have quality control measures in place. This assessment can be performed by any research assistant and backed up, however more importantly, those who will potentially be assigned this analysis work need to be aware of it so that they do not waste their

time. Rest assure, our database can capture this information through the relationship between raw data and CLiC Dashboard. Some of the hardline quality control parameters are saved directly as attributes, so that in most cases, with a quick reference of those values, a dataset can be trashed immediately if not qualified.

2.2.4 Data analyzed!! Great, where is the one I want?

The greatest pain often comes from actually finding the particular subsets of data out of the pools of dataset. For example, find PATHs of all DNA unwinding analysis result in after Sept, 2018 that are decent counts. This is such a tedious task that every researcher suffers from in the lab. Nevertheless, with this database, all the PATHs can be output immediately with a line of query and potentially we can implement this as a user-friendly function (the magic button).

2.2.5 Who is responsible??

This is a general question asked for everything in the lab. Conveniently, our database captures the person responsible for every task, or in other way around, what a person is responsible for in his work. This even includes the supervision responsibility between researcher and research assistant.

3 Relations

Entities

labMember (Name, contactInformation, Status)
Researcher (Name, contactInformation, Status, Level)
researchAssistant (Name, contactInformation, Status)

CLiCDashboard (dID, Date, PATH, Focus, SNR, Contaminants, crID, Name) (crID ref CLiC Raw Data(crID), Name ref researchAssistant (Name))

nanoparticleAnalysisResults (narID, Date, PATH, MSD, Name, crID) (crID ref nanoparticleData (crID), Name ref researchAssistant (Name))

bindingAnalysisResults (barID, Date, PATH, nnalysisType, Name, crID) (crID ref DNAUnwindingData (crID), Name ref researchAssistant (Name))

CLiCRawData (crID, Date, PATH)
nanoparticleData (crID)
DNAUnwindingData (crID)
Protocol (ProtocalID, PATH, Name) (Name ref Researcher (Name))
Microscope (MicroscopeID, Status)

Weak Entity

Experiment(protocolID, tryNumber, microscopeID, crdID, Name) (microscopeID ref Microscope(microscopeID), protocolID ref Protocol(protocolID), crdID ref CliCRawData(crdID), Name ref Researcher(Name))

Relationship

Supervise (Period, rName, raName) (rName ref Researcher (Name), raName ref researchAssistant (Name))

We did not feel there was a way to combine any relations. For example, if we did not use a relation table for the ER relationship **Supervise**, but save this as an attribute in the **researchAssistant** table, we would not be able store past records. This sitation could occur if a research assistant had worked with multiple researcher.