CS505 NLP: Milestone 1 - Project Proposal

Chris Krenz
Boston University
ckrenz@bu.edu

Abstract

In biological research and manufacturing, biologists often submit orders to wet labs for various functions, such as DNA synthesis or sequencing. This interface between the client and service provider can be fraught with various inefficiencies and barriers to entry. To simplify this process and facilitate the client's ability to submit complex biodesign and biomanufacturing orders, I propose a machine learning model to take in natural language prompts to predict the user's desired workflow (i.e. sequence of lab services) and (potentially) the associated parameters. The DAMP Lab at Boston University offers a manageable number of services (and service bundles) from which the model could choose, and the training corpus could be acquired through targeted internet searches of these keywords to find the terminology and phrases that tend to accompany them (for example, references to 'gene editing' or 'modifying sequences' may tend to accompany gene editing services services, such as Gibson Assembly and Modular Cloning).

1 Background

In the context of both biological research and biomanufacturing pipelines, biologists will often submit orders to companies for a variety of wet lab functions, such as DNA synthesis, DNA sequencing, or any one of numerous other services. Companies like Azenta/GENEWIZ, IDT, and Twist Biosciences—as well as academic entities, such as the DAMP Lab (Design Automation Manufacturing Processes) here at Boston University—provide such services to their customers and colleagues to fulfill their experimentation and manufacturing needs. However, the process of submitting these orders can often be convoluted and difficult to process—both for the client and the provider. The order submission, processing, approval, and execution steps can all be slowed and complicated by mistakes in the order details or other miscommunications between the parties involved. In addition to such inefficiencies, the process can often be opaque and counter-intuitive to non-experts, limiting the accessibility and cross-disciplinary work of biological research.

2 Concept

Given the constraints above, I propose the development of a machine learning model that can take in a natural language prompt from a user and produce a recommended workflow (or sequence of services/protocols). The development of such a model would facilitate the process of submitting orders to a wet lab by allowing users to simply type in an English description of what they want to do and have the model produce a suggested order—at least as a starting point—for the customer to submit. This could both reduce errors and miscommunications by guiding the client's decision making, as well as democratize access to these biological services[1] by reducing the expertise required on the part of the client. This, in turn, could increase the potential for interdisciplinary research by allowing researchers in a wider variety of fields to benefit from the revolutionary potential of genomic and other biological research.

3 Example

To provide a (hopefully) clarifying example, we could imagine the following scenario. A biology student is interested in utilizing campus (or other) wet lab resources for a class project or thesis. Perhaps they want to conduct a study to identify the micro-biome most conducive to the growth and survival of certain plants. To do this, they will be tweaking the DNA of certain bacteria and experimentally testing the effects on plant growth when the bacteria are mixed with the soil.

Being a student, they are relatively inexperi-

enced and uncertain how to specify some of the parameters. At this point, the user might need to message a lab technician to clarify the lab's offerings and how to properly construct the order. This would cause delays and potentially even deter the student from pursuing the project further. With the proposed model in place, we could instead imagine a window that pops up asking the student, in natural language, what they are trying to do. The student then enters something like the following:

> **Prompt**: *I'd like to modify some genetic sequences in bacteria. I also need it to be completed quickly.*

The model could detect the reference to modifying genetic sequences, suggesting a Gibson Assembly or Modular Cloning workflow might be appropriate. Given that, in general, Modular Cloning has a faster turnaround time than Gibson Assembly, the model could recommend a Modular Cloning workflow. So it might return something like:

> **Response**: *You may want to consider the following Modular Cloning workflow: Send Sample to Sequencing > Design and Order Primers > Rehydrate Primers > PCR*

To further improve the model's robustness and utility, its implementation could act like a walkthrough that prompts the user with a sequence of specific questions to simplify the context and narrow the search space for each question.

4 Data

The DAMP Lab offers a variety of wet lab services,[2] on the order of several dozen individual services and a dozen or so bundles (preset sequences of services). This provides a manageable set of labels the model would be tasked with generating. These labels could be cross-referenced with other data sources to further build the training corpus. By searching for articles, conferences, web pages on each of these services or bundles, one could identify the terminology and phrases that tend to accompany them. In addition, I may be able to access additional sources of data from my existing relationship with the BU DAMP Lab, though I still need to establish the feasibility of this approach.

5 Conclusion

Part of the advantage of this project is that it is highly scalable. In its simplest form, the model could essentially just identify keywords in a prompt and suggest a preset workflow tagged with the most such keywords. In a more complicated form, the model could consider nuances in the request and adjust various workflow parameters accordingly. As such, I can start with a relatively simple implementation and then increase the model's sophistication as time permits.

References

1 P.-E. Li et al., "Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform," Nucleic Acids Res, vol. 45, no. 1, pp. 67–80, Jan. 2017, doi: 10.1093/nar/gkw1027.

2 "SERVICES," damp-lab. Accessed: Sep. 23, 2024. [Online]. Available: https://www.damplab.org/services