Katharina Gees
AS.410.712.81.SP22

Project Report for MSA Viewer (aka apcc-bfx)

Project Background

Multiple sequence alignment (MSA) is a process of aligning three or more biological sequences (DNA, RNA, or protein) in order to infer sequence homology and conduct phylogenetic analysis. Results of these analyses can be used to predict structures, functions, and evolutionary relationships between biological sequences. There are multiple tools which perform MSA, such as Clustal Omega, MAFFT, MUSCLE, and T-Coffee. These and others are available as web interfaces from the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) at https://www.ebi.ac.uk/Tools/msa/. They are also available as web services for programmatic access using representation state transfer (REST) and simple object access protocol (SOAP) sample clients, and as Open API Interfaces.

All of the MSA tools available from EMBL-EBI are related to the process of MSA, but they are unique in their function and requirements. For example, EMBOSS Cons creates a consensus sequence from an input MSA. Additionally, of the tools performing an MSA, only Clustal Omega, MAFFT, MUSCLE, and T-Coffee can output alignments in ClustalW format which is simple but informative output (detailed at https://mccb.umassmed.edu/meme/doc/clustalw-format.html).

Project Proposal (Original)

MUSCLE (Multiple Sequence Comparison by Log-Expectation) is one of the MSA tools available from EMBL-EBI. It is mostly used for protein alignments (though it can be used for nucleotide sequences) due to its high accuracy and high speed for medium-sized protein sequence alignments. The web interface tool is available at https://www.ebi.ac.uk/Tools/msa/muscle/, and documentation on its REST and SOAP web service clients and open API interface is available at https://www.ebi.ac.uk/seqdb/confluence/display/JDSAT/MUSCLE+Help+and+Documentation.

My original project proposal was to build the MUSCLE web interface tool using the REST API for only ClustalW formatted output and representing this output in a new way. One of the limitations of the EMBL-EBI views of this output is that the annotations (such as asterisks (*)) are never quantified and are not described in the other results files either. I wanted to make this information more accessible by providing counts and descriptions in an output view.

Project Proposal (Revised)

I found that working with the web services APIs (specifically the REST API) was very difficult. What was challenging was submitting the job to the service, because the parameters required were difficult to use and encode (the form inputs needed to be translated to “x-www-form-urlencoded” for sending), and at the same time getting a job ID so that the results could be retrieved.

I decided to not include submitting a job as part of the project and chose to use previously submitted job’s job IDs as input. When I switched to using this input, I found that I could augment my project differently by making it possible to view the results of multiple MSA tools, as long as their results

were ClustalW formatted (this format is important because it is the only format which uses the asterisk, colon, and period annotations which I wanted to highlight in my results view).

Project Technologies and Methods

The following technologies and methods covered in the class which this project will use are the UNIX OS and filesystem, the placement and organization of files within a web application, python CGI programming, CGI and HTML templating, HTML5 document markup, and page styling with CSS. The technologies which the project will not use are the relational database schemas and design, MySQL, the python module to connect to MySQL, and JavaScript and jQuery client-side interaction.

The project will use the python requests module as an "alternative" to the MySQL connector module in order to perform HTTP requests which are an "alternative" to retrieving data from a relational database, and data from these requests is saved (the job IDs and alignment results are saved).

The inclusion of JavaScript and/or jQuery was considered* for this project. However, I did not use it because the requests module was being used to send HTTP requests and Ajax was not necessary, the desired response from the API endpoint was Unicode text and not a JSON, and Jinja2 provided sufficient scripting in the HTML. Also, I preferred the format of a separate input and output page for this project (like how EMBL-EBI has their input and output on separate pages). (*There is a JavaScript file included with the project which I did not end up finishing/using.)

Project Discussion

The main finished project is a web interface which accepts an MSA job ID through a form, and then uses this job ID to retrieve the results and generate a view of these results. The view shows the input job ID, the alignment, a table quantifying the annotations, and a description of the annotations' meanings. This project can be used by someone who is performing MSAs for DNA, RNA, or protein sequence analysis, and wants to quickly retrieve some summary statistics about the MSA. This is particularly helpful in cases when the MSA input is set of many (hundreds) of sequences and/or long (hundreds of nucleotides or amino acids) sequences. I find this project and the tools it includes to be very useful because it is something that I could have used on previous projects and assignment in my other bioinformatics classes.

Due to the fact that I mostly ran the same MSA input (the example sequences) for my project testing, the annotation results were not particularly "meaningful" for me. What I did notice was that this alignment text had to handled carefully so that the spacing was maintained between the annotations and that there is a potential for error in the annotation counts if these characters are used within the name of a sequence (which is possible to do). This reiterates the importance of following specific formats for data in bioinformatics.

While working on the project I also noticed some of its potential limitations and potential extensions. I think that the view could show more information, such as information on the nucleotides and amino acids, like frequencies and properties, and sequence identify information (accessed by another request call to another API endpoint). Also, if the project did not save information, it would be easier to use because issues such as file permission errors are unlikely to exist.

Katharina Gees
AS.410.712.81.SP22

Project File Structure

/var/www/html/kgees1/apcc-bfx

- **/css/** contains the css used for the entire project
- **/files/**
  - **alignment.txt** a file containing the "current" alignment from running the program
  - **alignment_list.txt** a file containing the alignments from running the program
  - **job_id_list.txt** a file containing the job IDs from running the program
- **/img/** contains images used by the project
- **/js/** contains the js file I did not end up using for the project (unfinished)
- **/proposal_and_report/** contains final and draft versions of the proposal and report
- **/scripts/** contains scripts which were used within the cgi
- **/templates/** contains template files (used for MSA views)
- **demo.cgi** creates an MSA view without input (has hardcoded job ID)
- **input.html** template used for job ID input
- **main.cgi** creates and MSA view with input
- **README.md** provides documentation on the project

Project Links

GitHub https://github.com/katgees/apcc-bfx

input.html http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/input.html
(Real input form html file)

output.html http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/templates/output.html
(Real output html file, this shows the "empty" html template)

main.cgi http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/main.cgi
(cgi file for **input.html**, this should be showing results using the **output.html**)

demo_input.html http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/demo_input.html
(**demo_input.html** is a fake version of **input.html**: the form input is fake and the submit button is a button link to **demo_view.html**)

demo_output.html http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/templates/demo_output.html
(**demo_output.html** is identical to **output.html** besides the heading ("MSA Viewer – Demo"), this shows the "empty" html template)

demo.cgi http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/demo.cgi
(cgi file for **demo_output.html**, this should be showing **demo_view.html**)

demo_view.html http://bfx3.aap.jhu.edu/kgees1/apcc-bfx/templates/demo_view.html
(**demo_view.html** was generated by running **demo.cgi** using the **demo_output.html** template)

Katharina Gees
AS.410.712.81.SP22

References

Chojnacki S, Cowley A, Lee J, Foix A, Lopez R. (2017). **Programmatic access to bioinformatics tools from EMBL-EBI update: 2017.** Nucleic Acids Res Volume (2017) p. PMID: 28431173. DOI: 10.1093/nar/gkx273

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. (2010). **A new bioinformatics analysis tools framework at EMBL-EBI.** *Nucleic Acids Res* 38 (Web Server issue) : W695-9. PMID: 20439314. DOI: 10.1093/nar/gkq313

Harte N., Silventoinen V., Quevillon E., Robinson S., Kallio K., Fustero X., Patel P., Jokinen P. and Lopez R. (2004). **Public web-based services from the European Bioinformatics Institute**. Nucleic Acids Research 32: W3-W9. PMID: 15215339. DOI: 10.1093/nar/gkh405

Labarga A., Valentin F., Andersson M. and Lopez R. (2007). **Web Services at the European Bioinformatics Institute**. Nucleic Acids Research 35: W6-W11. PMID: 17576686 Abstract. DOI: 10.1093/nar/gkm291

Labarga A., Pilai S., Valentin F., Anderson M. and Lopez R. (2005). **Web services at the European Bioinformatics Institute**. EMBnet.news 11(4):18-23. full-text PDF

Li W., Cowley A., Uludag M., Gur T., McWilliam T., Squizzato S., Park YM., Buso N., Lopez R. (2015). **The EMBL-EBI bioinformatics web and programmatic tools framework.** Nucleic acids research 43 (W1) : W580-4. PMID: 25845596 . DOI: 10.1093/nar/gkv279

Madeira, F., Pearce, M., Tivey, ARN., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., Lopez, R. (2022). **Search and sequence analysis tools services from EMBL-EBI in 2022.** DOI: 10.1093/nar/gkac240. EuropePMC: 35412617

Madeira, F., Park, YM., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, ARN., Potter, SC., Finn, RD. and Lopez, R. (2019). **The EMBL-EBI search and sequence analysis tools APIs in 2019**. *Nucleic Acids Res*, 47(W1), W636–W641. DOI: 10.1093/nar/gkz268. EuropePMC: 30976793

Madeira, F., Madhusoodanan, N., Lee, J., Tivey, A. R. N., and Lopez, R. (2019). **Using EMBL-EBI services via web interface and programmatically via web services.** *Curr Protoc in Bioinformatics*, 66(1):e74. DOI: 10.1002/cpbi.74. EuropePMC: 31039604

McWilliam H., Li W., Uludag M., Squizzato S., Park YM., Buso N., Cowley A., Lopez R. (2013). **Analysis Tool Web Services from the EMBL-EBI.** Nucleic acids research 41 (Web Server issue) : W597-600. PMID: 23671338. DOI: 10.1093/nar/gkt376

McWilliam H., Valentin F., Goujon M., Li W., Narayanasamy M., Martin J., Miyar T. and Lopez R. (2009). **Web Services at the European Bioinformatics Institute** . Nucleic Acids Research 37: W6-W10. PMID: 19435877. DOI: 10.1093/nar/gkp302

Katharina Gees
AS.410.712.81.SP22

Pillai S., Silventoinen V., Kallio K., Senger M., Sobhany S., Tate J., Velankar S., Golovin A., Henrick K., Rice P., Stoehr P. and Lopez R. (2005). **SOAP-based services provided by the European Bioinformatics Institute**. Nucleic Acids Research 33: W25-W28.PMID: 15980463.
DOI: 10.1093/nar/gki106