**Project Report for MSA Viewer**

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 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, and MView creates a new web view using HTML and CSS from an input MSA. Additionally, of the tools performing an MSA, only Clustal Omega, MAFFT, MUSCLE, and T-Coffee can output alignments in Clustal/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 other 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)

As I started this project, 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.

At that point 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 Clustal/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 (just like how EMBL-EBI has their input and output on separate pages).

\*There is a JavaScript file included with the project which is not used in the final products.

Project Discussion

Project File Structure

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

**/css/** contains the css used for the entire project

**/files/**

**list.txt** a file containing the job IDs from running the program

**\*** Multiple files containing the Clustal results from running the program

**/img/** contains images used by the project

**/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

**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”))

**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)

**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)

**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**)

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](http://europepmc.org/articles/PMC5570243). DOI: [10.1093/nar/gkx273](https://doi.org/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](http://europepmc.org/abstract/MED/20439314). DOI: [10.1093/nar/gkq313](http://dx.doi.org/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](http://www.ebi.ac.uk/ebisearch/search.ebi?db=literature&query=15215339). DOI:  [10.1093/nar/gkh405](http://dx.doi.org/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](http://www.ebi.ac.uk/ebisearch/search.ebi?db=literature&query=17576686)   [Abstract](http://europepmc.org/abstract/MED/17576686). DOI:  [10.1093/nar/gkm291](http://dx.doi.org/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](http://journal.embnet.org/index.php/embnetnews/article/view/95/111)

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](http://europepmc.org/abstract/MED/25845596). DOI: [10.1093/nar/gkv279](https://doi.org/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](https://doi.org/10.1093/nar/gkac240). EuropePMC: [35412617](https://europepmc.org/article/MED/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](https://doi.org/10.1093/nar/gkz268). EuropePMC:  [30976793](http://europepmc.org/abstract/MED/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](https://doi.org/10.1002/cpbi.74). EuropePMC: [31039604](https://europepmc.org/abstract/MED/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](http://europepmc.org/abstract/MED/23671338). DOI: [10.1093/nar/gkt376](https://doi.org/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](http://www.ebi.ac.uk/ebisearch/search.ebi?db=literature&query=19435877). DOI:  [10.1093/nar/gkp302](http://dx.doi.org/10.1093/nar/gkp302)

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](http://www.ebi.ac.uk/ebisearch/search.ebi?db=literature&query=15980463). DOI:  [10.1093/nar/gki106](http://dx.doi.org/10.1093/nar/gki106)