Author: Lui, Nyalia
PI: Rekepalli, Bhanu PhD.
University of Tennessee, Knoxville
Oak Ridge National Laboratory
August 7, 2013

# Bioinformatics Applications & Analysis using PoPLAR Gateway

## Abstract

Over the years, scientific data in the life sciences has been generated at exponential rates and continues to grow with the rise of new technologies. We are focusing on the parallel analysis tools and resources used to generate the data. Our focus has led us to using the PoPLAR research project, a scientific gate designed to provide biologists a means to access High Performance Computing (HPC) resources. Currently, the gateway is under development as a web application with a user friendly interface, eliminating the need for biologists to have basic, and in some cases advanced, computer science skills [1].

## Background

Sciences similar to Biology, Chemistry, and Physics require researchers who are willing to sacrifice many hours for the development of their respective fields. The amount of time and effort they work on their projects is always represented by the vast amount of data the researchers generate. The abundance of data is so large that researchers cannot access, analyze, or utilize the data. To compensate for these issues, computers and other forms of technology are blending with the life sciences. One of the sciences that generate the most data is Biology. In this science, computers are primarily used to access and upload data to and from remote locations. Data is uploaded to various online resources where other researchers can access that data. This is the basis for the bioinformatics research field, the creation of applications that organize and provide researchers easy access to a growing pool of data.

Currently, Bioinformatics is centered on genomic sequencing, the analysis of proteins, DNA, and RNA sequences. A Genome is the entirety of an organism's hereditary information [4]. This information is encoded in long strings of acids called the Nucleotides. The different Nucleotides are Adenine, Thymine, Guanine, Cytosine, and Uranine. Nucleotides combine to create pairings called base pairs. For DNA strands, Adenine pairs with Thymine and Guanine pairs with Cytosine. For RNA strands, Adenine pairs with Uranine and Guanine still pairs with Cytosine. The nucleotides pair together to help form the double-helix shape the DNA strands form. Aside from the pairings, the nucleotides form different types of acids, amino acids, in their individual strands. Amino acids are formed from a specific three character string of nucleotides. Each string is called a codon. The nucleotides have 64 different codons to create amino acids; however, there are only 20 unique amino acids [5]. Three of the codons are called stop codons because they signal where the end of a strand is.

DNA strands are the base to the genome of an organism because they are used to create proteins. Proteins are the basic molecular structure used to perform various functions in an organism. Proteins are created from Nucleotides in a process called DNA transcription. In this process, the double-helix structure is unwound by other proteins called transcription factors. Then an enzyme called RNA Polymerase transcribes strands of DNA into an RNA polymer called messenger RNA (mRNA). The nucleotides are paired with their appropriate partners during this step. These two steps occur within a cell's nucleus. Once the polymerase is finished creating the mRNA, the mRNA moves outside of the nucleus and into the cytoplasm where ribosomes (rRNA) work together with transfer RNA (tRNA) to translate the mRNA into amino acid chains [5]. The appropriate amino acid is appended to a growing polypeptide chain which when finished, is called a protein.

## Bioinformatics

Previously discussed was the biological side to bioinformatics, this section introduces how computers and technology blend with biology. In bioinformatics there are four explored fields, in order they go as follows: Genome Assembly, Sequence Analysis, Structure, and Drug Discovery. Genome Assembly is where researchers focus on assembling the hereditary sequences of a specific organism i.e. the genome. Sequence Analysis is where researchers or developers use and create tools to analyze hereditary sequences. Within Sequence Analysis is a sub-category called phylogenetic. Phylogenetic is about the development of tools to help the analysis of genetic relationships through phylogenetic trees. The Structure Creation field is the use of data from sequence analysis to develop 3 dimensional structures. The last, Drug Discovery, is about utilizing and analyzing the 3D structures so life scientists may create new medicine.

Sequence Analysis is the focus of this research project. Teams of computational biologists use various tools and applications to help analyze sequences. These tools help analyze by aligning sequences together, this is called sequence alignment. Sequence alignment is the process of taking two different sequences or strands and finding if they have similar regions anywhere within their string of nucleotides. There are two types of sequence alignment, global and local alignment. Global alignment is an attempt to align every nucleotide from one sequence to a nucleotide in another sequence. Global alignment is useful when comparing similar sequences. Local alignment is an attempt to align a region from one sequence to a region in another and is useful for dissimilar sequences [6].

Four of the many sequence analysis tools are the National Center for Biotechnology Information's Basic Local Alignment Search Tool (BLAST), HMMER, MUSCLE, and MUMMER. These four are local alignment tools however MUSCLE is different from the rest. MUSCLE can conduct multiple sequence alignment (MSA) whereas BLAST, HMMER, and MUMMER do not. Multiple sequence alignment is the alignment of three or more sequences instead of two. There are two methods that MSA tools use. The first utilizes a heuristic search, a progressive technique, which will align pairs of sequences together starting at the most similar pair and progressively moving to the most dissimilar pair [7]. The second method is an iterative method. The tool will align two sequences first and then realign the last two to the next sequence. Previously aligned sequences will be realigned to the next sequence until the tool has aligned all sequences together. These tools may be found online for free download.

Author: Lui, Nyalia
PI: Rekepalli, Bhanu PhD.
University of Tennessee, Knoxville
Oak Ridge National Laboratory
August 7, 2013

## Methods

To understand more about bioinformatics sequence analysis applications; BLAST, HMMER, MUSCLE, and MUMMER were run on local computers. However, BLAST and MUMMER were compared against each other. Two files were submitted to each program, a sequence file from the researcher and a sequence file from a database. In these files the sequence(s) are broken up into IDs with specific lengths. For simplicity, the IDs from the researcher's sequence file are called Nodes and the IDs from the database sequence file are called Contigs. Recalling that alignment finds regions of similarity between two sequences, the alignment started and ended at any spot in the sequence. This information is used when comparing BLAST and MUMMER against each other.

BLAST and MUMMER were compared by calculating the coverage of each ID in a sequence. The coverage of an ID is calculated in three steps. First, every times a unique ID was aligned, the starting and ending points of the alignment had to have been recorded. Second, take one of the IDs and all of its starting and ending points from the alignment and calculate the total length covered by the alignment. One must make sure to accommodate for overlap as adding the same length multiple times ruins data. Third, divide the total length covered by the alignment by the total length of the ID. The total length of the ID can be found in either the database file or the researchers file depending on what sequence the ID came from. The quotient will be the coverage of that ID. Use the coverage of each ID as the measuring factor when comparing BLAST and MUMMER. Their coverage data was compared in two different ways. The first way was by analyzing each individual ID coverage. The second was by analyzing the coverage and how many IDs covered that same amount. The results of the Contig IDs are shown in figure 1 and 2. The results show that the alignment algorithm used in MUMMER cover more of and IDs original length than BLAST.

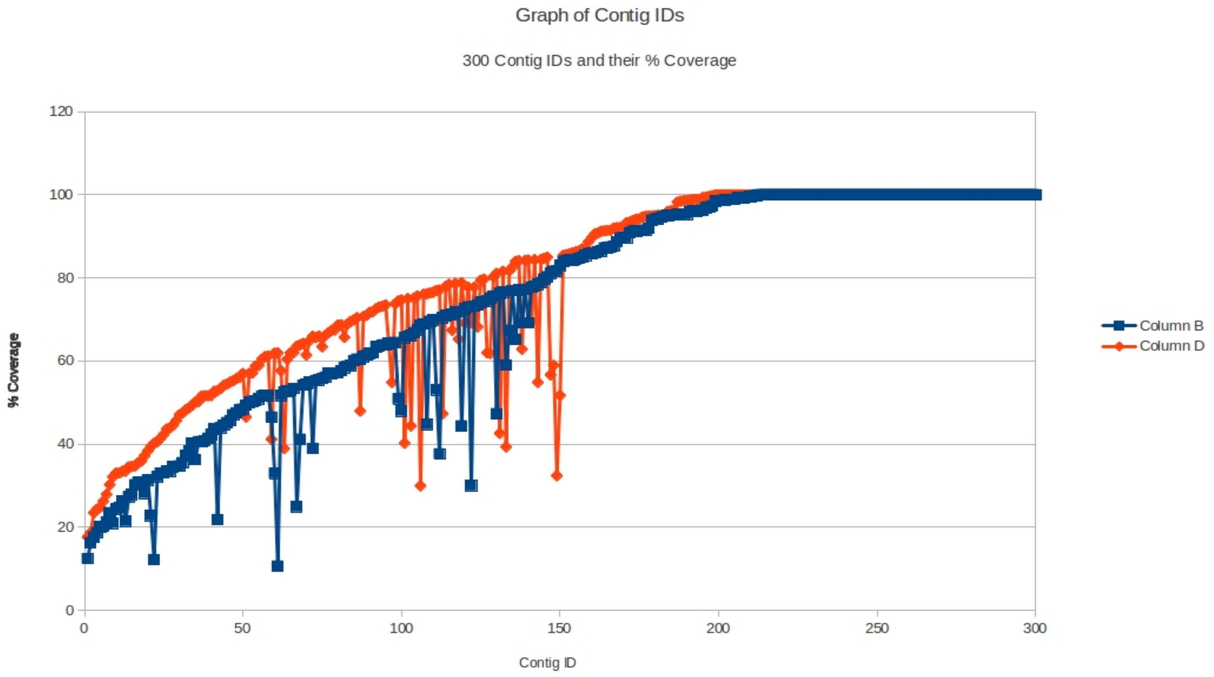


**Figure 1:** A graph of  300 Contig IDs and each IDs coverage. The orange indicates the MUMMER results and the blue indicates the BLAST results.

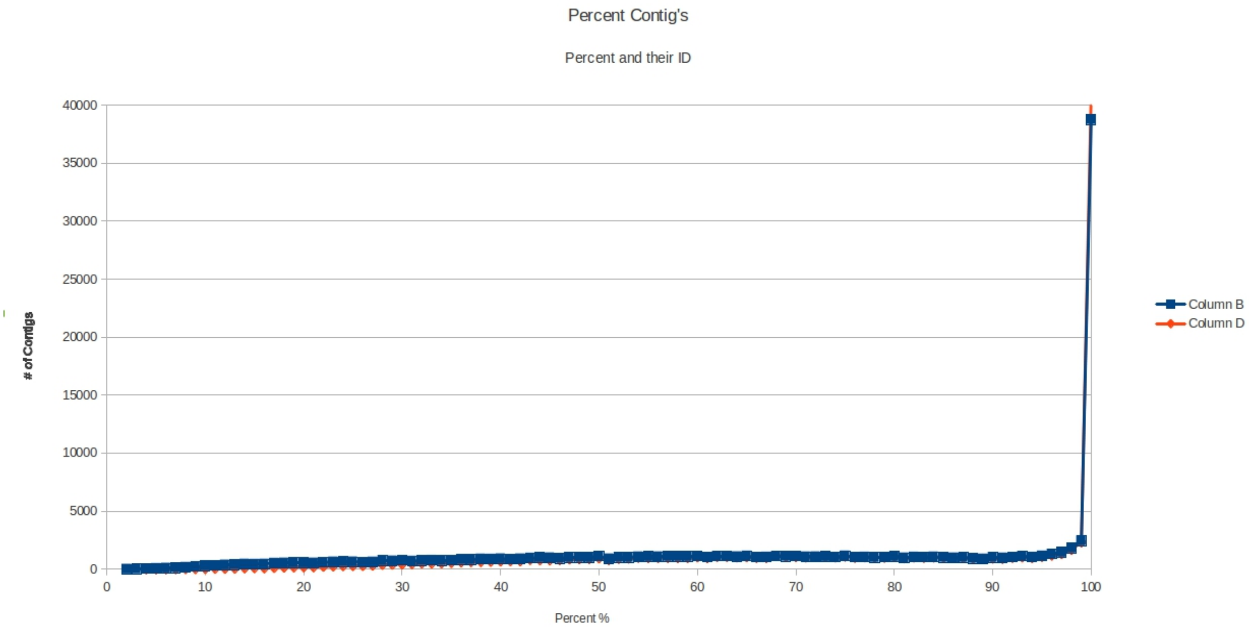
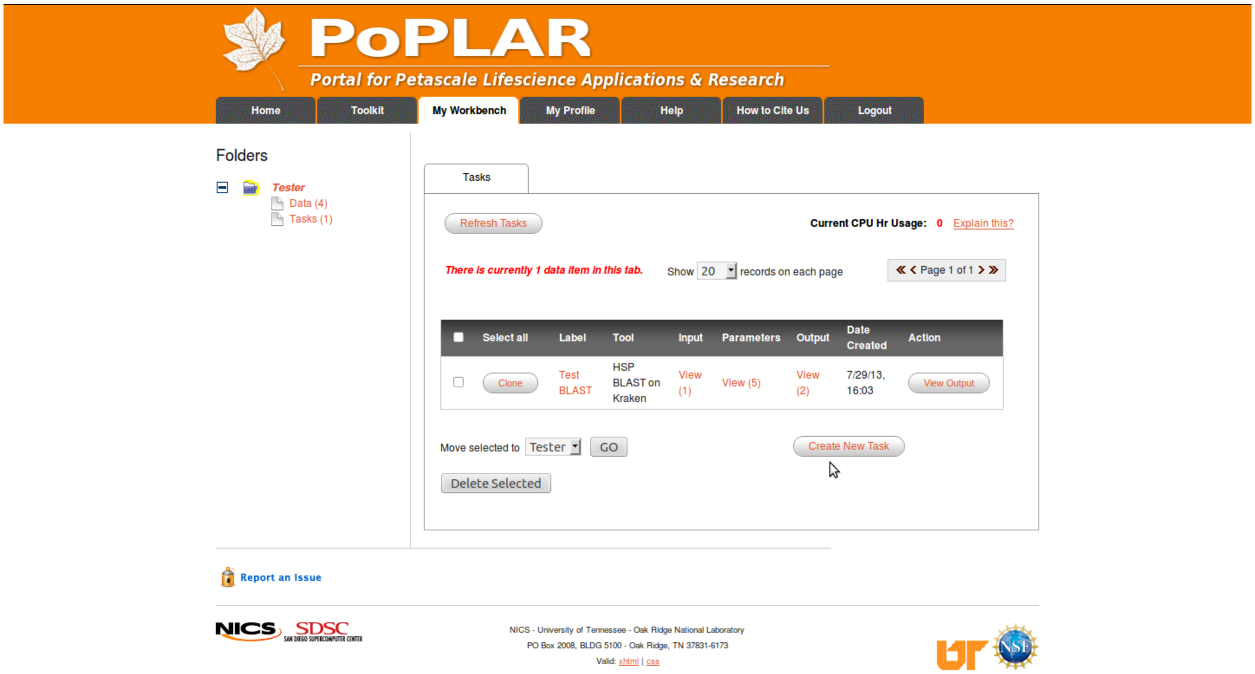


**Figure 2:** A graph of each coverage and the number of Contig IDs that covered the same amount.

## The PoPLAR Science Gateway

BLAST, HMMER, MUSCLE and MUMMER are programs that are run within the command line (Windows) or the terminal (Linux). This means that they require a basic understanding of computer science and navigation. Gaining these skills can be time consuming, tedious, and stressful. The focus is to create analytical applications similar to current tools as well as develop resources that provide biologists easy access to analytical apps. Ultimately, the objective is to eliminate the need for biologists to learn computer science skills. This led to the idea of utilizing a science gateway[1]. A science gateway is an application that provides researchers easier access to Highly Scalable Parallel (HSP) tools on High Performance Computing (HPC) resources. HSP tools are command line applications that in the past sequentially performed computational calculations but have been rewritten for parallel computing [1]. The optimal gateway for this research project is the Portal for Petascale Lifescience Applications & Research (PoPLAR) scientific gateway. The PoPLAR gateway is based off of the CIPRES scientific gateway for phylogenetic research and is specifically devoted to providing biologists easy access to data and HSP tools. PoPLAR is under development as a web application that gives researchers a customizable interface for running HSP tools. Instead of typing various commands in the terminal a user may instead login to PoPLAR. Once logged in, they will be shown a page where they may create a new directory. Every directory has two sub directories, a data folder and a task folder. The data folder is where the results from various HSP tasks will be sent. A researcher may also upload data from their personal machine to their data folder. The second sub directory, tasks folder, is where a researcher may run a HSP task on a HPC resource. Currently, PoPLAR supports BLAST, HMMER, and MUSCLE as HSP tasks and all tasks will be ran on the Kraken supercomputer located at Oak Ridge National Laboratory in Knoxville, Tennessee. See figure 3 for an image of the directory tree.

Author: Lui, Nyalia
PI: Rekepalli, Bhanu PhD.
University of Tennessee, Knoxville
Oak Ridge National Laboratory
August 7, 2013



Current CPU Hr Usage: 0

There is currently 1 data item in this tab.

Show 20 records on each page

Page 1 of 1

| □ Select all | Label | Tool | Input | Parameters | Output | Date Created | Action |
|---|---|---|---|---|---|---|---|
| □ Clone | Test BLAST | HSP BLAST on Kraken | View (1) | View (5) | View (2) | 7/29/13, 16:03 | View Output |

NICS - University of Tennessee - Oak Ridge National Laboratory
PO Box 2008, BLDG 5100 - Oak Ridge, TN 37831-6173
Valid: xhtml | css

**Figure 3:** An image of the directory tree and the screen where the researcher may create or edit a task. The directory tree is in the top left right below the “home” button.

After a task is submitted through the gateway a number of actions happen. The task will be ran on an HPC resource but after the task finishes either the HPC will return the raw output data back to the gateway and consequently the data folder, or the output will go through a back end process. This process is used to parse the output into a format that is easier to understand. There are three choices for parsed data output, which are indicated by numerical options 7, 8, or 9. The choices are, flat query anchored without identities (7), tab-delimited (8), or XML (9). The parsed output is then returned to the gateway where the file will be stored in the researcher’s online data folder. See figure 4 for the flow chart of an initiated task.

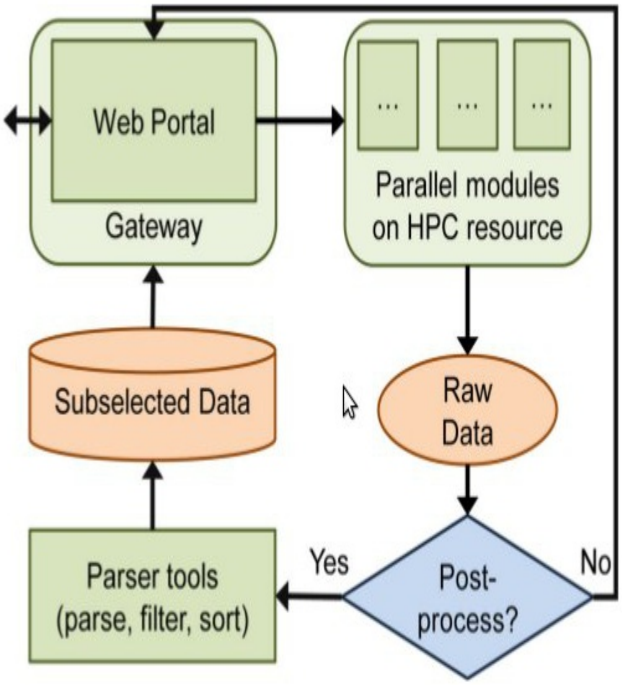


**Figure 4:** A flow chart of what happens to an initiated task in the back-end of the PoPLAR gateway.

Author: Lui, Nyalia
PI: Rekepalli, Bhanu PhD.
University of Tennessee, Knoxville
Oak Ridge National Laboratory
August 7, 2013

## Conclusions

This research project will contribute to the bioinformatics community by providing easy access to command line tools such as BLAST and MUMMER through the Portal for Petascale Lifescience Applications & Research. Likewise, the project will educate biologists on accessing High Performance Computing resources for large scale data analysis. The goal for the future is to finish the PoPLAR gateway, but also further develop the gateway to not only include the Sequence Analysis field of bioinformatics but also Genome Assembly, Structures, and Drug Discovery. Recalling that PoPLAR currently supports BLAST, HMMER, and MUSCLE, there is also an aspiration to include more Highly Scalable Parallel tools in the gateway. Including more HSP applications will contribute to the next generation of biologists and researchers by providing them an easy to use interface to further the communities' knowledge.

## References

1. Rekapalli et al. "PoPLAR: Portal for Petascale Lifescience Applications and Research." BMC Bioinformatics. 2013, 14(Suppl 9):S3. Web. http://www.biomedcentral.com/1471-2105/14/S9/S3
2. M. Gerstein, et al. "What is bioinformatics? An introduction and overview." Yearbood of Medical Informatics 2001. June 2013, 18. Web. http://www.ebi.ac.uk/luscombe/docs/imia_review.pdf
3. Milller, Mark A. et al. "The CIPRES Science Gateway: A Community Resource for Phylogenetic Analyses." Gateway Computer Enviornment 2010. July 2013, 25. Web. http://www.phylo.org/sub_sections/portal/sc2010_paper.pdf
4. "A Basic Introduction to the Science Underlying NCBI Resources." National Center for Biotechnology Information. March 2004, 31. June 2013, 18. Web. http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html
5. Bailey, Regina. "Genetic Code." About.com. June 2013, 18. Web. http://biology.about.com/od/genetics/ss/genetic-code.htm
6. Altschul, Stephen F. "Global and Local Alignment." National Center for Biotechnology Information. June 20, 2013. Web. http://www.cs.umd.edu/class/fall2011/cmsc858s/Alignment.pdf
7. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity BMC Bioinformatics, (**5**) 113. Web. http://www.biomedcentral.com/1471-2105/5/113