Biol 360 – Bioinformatics Lab
Lab Handout – Bioinformatics Workflows

So far this semester, you've been doing web-based bioinformatics analyses on small input, such as BLASTing a single sequence. However, researchers often need to analyze much larger amounts of data. Imagine if, instead of just BLASTing one sequence, you had to BLAST 1000 sequences. Running each BLAST individually, as you've done in this course, would take a very long time! Fortunately, we can automate bioinformatics analyses so that manually performing them on large amounts of data is not required. One way to do this automation is through computer programming - you got a glimpse at how this is done in last week's Python lab. Another option, and the one we'll focus on today, is bioinformatics workflows.

Workflows are series of connected steps to be performed on a set of input data. Each step in a workflow builds on the previous step, so that the entire series of steps can take place without stopping or needing further input. In this way, you can perform many different analyses on a dataset without needing to perform each operation manually (which can be extremely time-consuming when you are dealing with large amounts of data and several operations to be performed on those data). Additionally, the input for a workflow can be much larger in size than a single sequence. Workflows can be quite complicated, involving many layers of analysis, as shown in Figure 1.

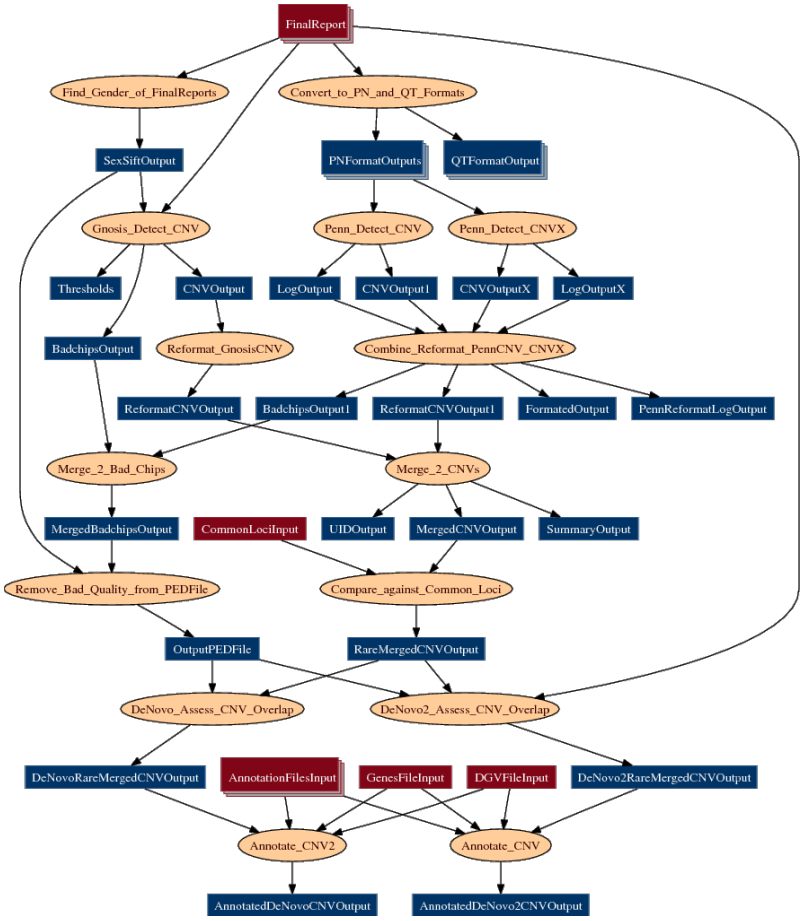

Figure 1: Bioinformatics workflow example. We won't be doing anything this complicated today, but this is an example of the complex analysis that can be performed in a workflow. From http://pegasus.isi.edu/static/pegasus/pegasus_bio.php.

Workflows are often designed using programming-heavy approaches, but recently several programs have been designed to be used by biologists with little to no computer programming experience. We'll be working with one of these programs today, Unipro UGENE. UGENE includes many tools for bioinformatics analysis, but we'll be focusing on the Workflow Designer, a tool that allows users to create multi-step workflows to analyze data (usually gene or protein sequence data). Figure 1 shows an example of a UGENE workflow and diagrams its components. In this example, the first step is to read in an input sequence from a file. The next step BLASTS that input sequence and outputs a set of annotations identifying the search hits. Finally, the annotations are given as input to the last step, which saves the annotations to a file. The direction of the arrows indicates the order in which the steps will take place, i.e. the order in which data will "flow" from one element to the next.

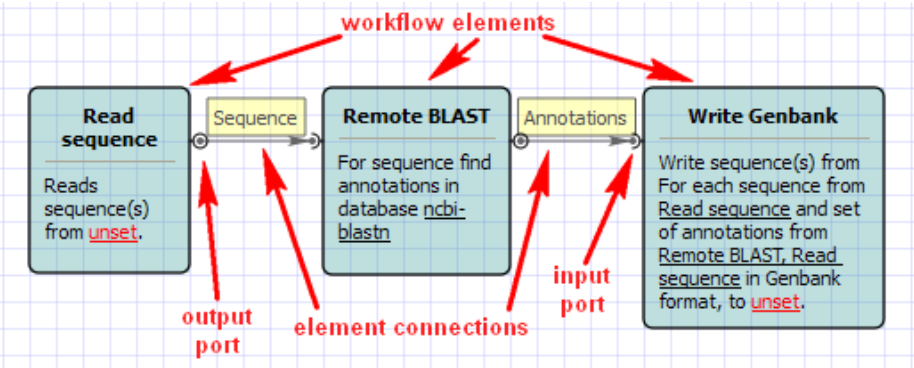

Figure 2: UGENE workflow diagram, showing the parts that make up a schema (visualization of a workflow). From http://ugene.unipro.ru/documentation/wd_manual/introduction/schema_terms.html.

In today's lab exercise, you'll use UGENE to create basic workflows as an introduction to the topic. In future lab sessions, you'll use the software to perform slightly more complicated analyses (although not as complicated as the workflow shown in Figure 1!). Since this is a likely a new concept for most of you, you may want to take a look at UGENE's Workflow Designer manual. UGENE also has a YouTube channel with instructional videos that you may find helpful.

Your lab writeup for this week will consist of the workflows you create during the lab session and your notes about the process.

UGENE Basics

In this section of today's exercise, you will create a very simple workflow to get a feel for UGENE. Your first workflow will convert a DNA sequence from one file format to another.

You have been given a file containing several sequences in GenBank format. If you open the file in a text editor, you'll be able to see the GenBank sequences. Your workflow will have two steps: reading in the sequences from the file, and converting the sequences to FASTA format and saving them in a new file.

Step 1:

Open UGENE and select Workflow Designer from the Tools menu. Add the Read Sequence element to the scene. (You'll find the Read Sequence element, and the other elements you can work with, on the left-hand side of the window. An example of adding an element to the scene is shown in Figure 3). Set the input for Read Sequence to the file containing your sequences.

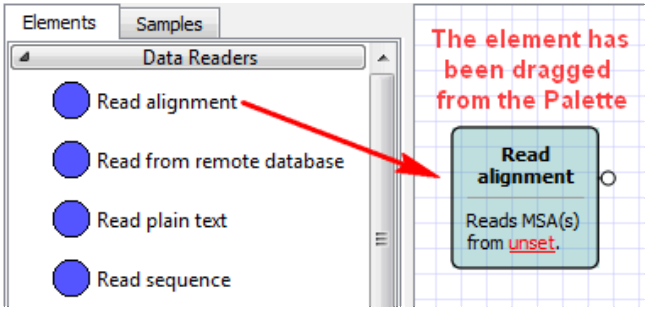

Figure 3: Adding workflow elements from the palette to the scene. From http://ugene.unipro.ru/documentation/wd_manual/introduction/create_and_run_schema.html.

Step 2:

You want to convert the sequences from GenBank to FASTA format, so select the element under the Data Writers heading that will convert a sequence to GenBank format and write it to a file. Add that element to the scene, and give the output file a name and save it in a place where you'll be able to find it easily. Finally, connect the two elements (Read Sequence and the data writing element). It's a good idea to save your workflow at this point to the desktop or a flash drive!

What happens if you change the value of the "Accumulate Objects" parameter to False?

Step 3:

Validate and run your workflow. If you open the output file you created in a text editor, you should see that the sequences are now in FASTA format. Congratulations, you just created your first workflow!

Workflows with Several Steps

Now that you've had a chance to create a simple workflow, you'll be able to build more complicated workflows involving multiple steps. While, as you've seen, it is possible to work on many sequences at once, we'll be working on one sequence at a time in this section to simplify the process. Each person in your group should choose one of the sequences from your sequence file to work with for the remainder of the exercise. Save that sequence in a text file to use as input for your workflows.

You may recall from your previous biology courses that when an mRNA sequence is translated into

amino acids, translation may begin from one of six possible positions. This variation means, of course, that there are several possibilities for the amino acid composition of the resulting protein. It can be useful to study all of the potential translations when trying to learn what the gene you're analyzing does. With that in mind, in this section of the exercise you'll design a workflow that takes a nucleotide sequence and translates it into amino acids using all of the possible reading frames.

The elements you'll need for this section include:
ORF Marker
Read Sequence
Get Sequences By Annotations
Write Sequence

Step 4:
Add the elements listed above to the palette. Select each element and note its input and output types (e.g. Sequence, Set of Annotations, etc.). The input of the current step you're working on needs to match the output of the previous step, so keeping track of the input and output of each element will help you order the steps of your workflow.

Step 5:
Create a new workflow schema. Using the elements listed above, design a workflow that will take your sequence, find all open reading frames, and write the resulting sequences to a file.

Another useful way to analyze your nucleotide sequence might be to determine its GC content. As we've discussed in previous labs, percent GC content often has effects on the properties of a sequence.

The elements you'll need for this step are:
Read from Remote Database
DNA Statistics
A data writer (look at the output type of the data to determine which one you need)

Step 6:
Design a workflow that will read in a sequence and determine its GC content. This time, instead of giving the sequence to UGENE in a file, add a step that will use the accession number to look up the sequence record at NCBI.

The DNA Statistics element has options for determining GC, GC1, GC2, and GC3 content. What are the differences between these options?

Finding repeat regions can be another useful way to analyze a nucleotide sequence. Repeat regions in DNA can help elucidate the function of a gene. Also, for human genes, short tandem repeats (repeats of usually 3-4 nucleotides that lie next to each other) have been used in forensics to identify the source of evidence at a crime scene, since these repeats are often unique to individuals (see Figure 4).

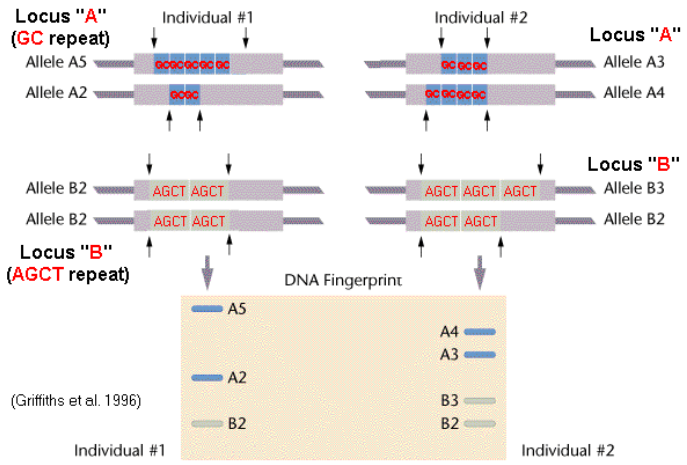

Figure 4: DNA fingerprinting example. From http://www.mun.ca/biology/scarr/VNTR_fingerprinting.html.

You'll need the following elements for this step:
Find Repeats
A data writer
A data reader

Step 7:
Design a workflow to find repeat regions in your sequence. (You can get the sequence input either from a file or by using the accession number to get it from NCBI.) Write the output to a file.

More Workflow Practice

Now you'll put all of the steps from the previous section together into one workflow with several steps. The ability to perform several forms of analysis on your sequence without having to perform each step manually is one of the main advantages of bioinformatics workflows, so it will be useful to practice setting up workflows that can perform several different operations on one set of input data.

Step 8:
First, design a workflow that will get an input sequence and determine its GC content. Add an element that will find repeat regions. Finally, add an element that will identify ORFs in your sequence. Write the output of each of these steps to a file.

Step 9:
Adjust the parameters of the ORF sequence element to translate the ORFs and output amino acid sequences. Write the amino acid sequences to a file.