Pre-requisites for Sequedex

Sequedex is written in Java. Thus, to run Sequedex requires that you have Java installed. For the binary distribution, Sequedex has been compiled and tested using Java 8. Thus you will need at least Java 8 JRE or JDK installed to run Sequedex. Later versions of Java should also work. For example, very brief test runs with Java 10 and Java 11 were successful. Since Sequedex is written in Java it should run on Mac, Linux, or Windows. However, the authors themselves use Mac or Linux and therefore testing has been primarily on those platforms. Hardware requirements will depend to a large extent on which data modules are run. The current binary distribution comes with two data modules: Life2550-8GB.1 and virus10k.1. The larger, Life2550-8GB.1, requires 8 gb of free memory to run. In the default mode, Sequedex will warn the user if there is not enough free memory. A reasonably new laptop or desktop computer with a minimum of 16 gb of RAM should be sufficient for Life2550-8GB.1 while still leaving memory for every day tasks on your computer (reading email, browsing the web, etc). It may even run with only 8 gb of RAM if Sequedex is the only thing running on the computer. For larger data modules, more RAM is required. In addition, If a user needs to analyze multiple sample files using a single data module, Sequedex can run multiple threads (queueing up files to be processed one sample file per thread at a time) without requiring significantly more RAM. In this case, the more cores the better.

To find what version of Java is installed on your computer, at the operating system command line execute: java -version. The response should be something like the following:

```
java version "1.8.0_172"
Java(TM) SE Runtime Environment (build 1.8.0_172-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.172-b11, mixed mode)
```

A version starting with 1.8 indicates Java 8.

Installing Sequedex

A Sequedex release is provided as a zipped or compressed tar file (sequedex.zip or sequedex.tgz) which should be unzipped or uncompressed using the appropriate unzip or tar utility.

On the same github page where you download the release, you should find test data in files sequedex_v2.testdata.tgz or sequedex_v2.testdata.zip.

When sequedex.zip or sequedex.tgz are unzipped or uncompressed, you should have a directory containing:

data/
doc/
etc/
lib/
LICENSE.txt
licenses/
README.txt

where / indicates a directory.

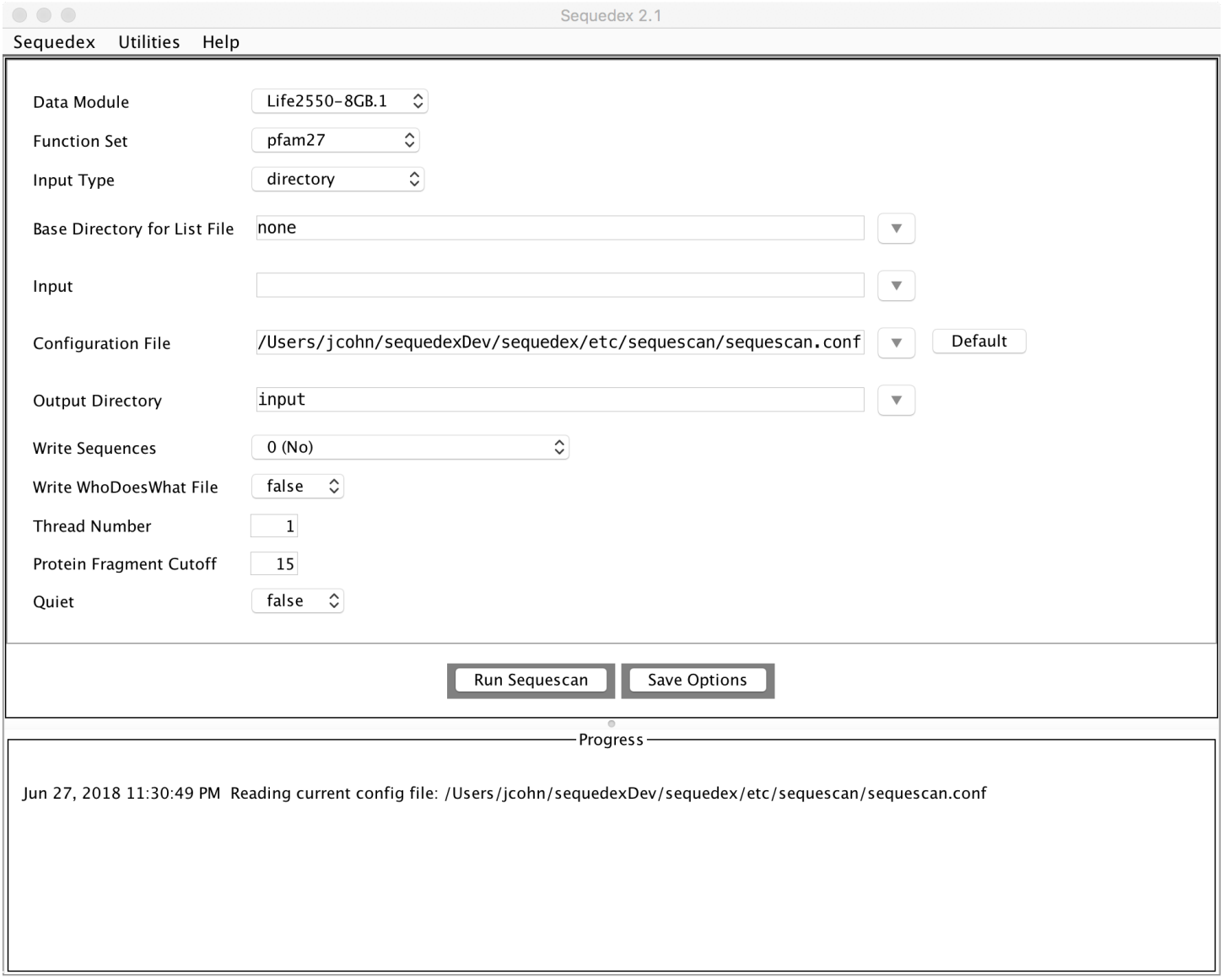
Unzip or untar either sequedex.zip or sequedex.tgz, either of which will uncompress into a directory called sequedex. This directory is the executable directory for Sequedex. Where you put the sequedex directory will depend on your usage case. If you are an individual user who does not plan to share Sequedex with other users, you may wish to put it in your home directory. Otherwise, if it will be used by multiple users, it can be moved to a shared directory (e.g. on the Mac to the Applications folder) with appropriate permissions.

Running Sequedex

Sequedex runs either from a GUI or the command line. For purposes of this Quick Start, the GUI interface will be described. The easiest way to open the GUI is to click on the icon for the application jar file sequescan.jar. This file is to be found in the lib subdirectory of sequedex. For convenience, one may create an alias or symbolic link to this file and put it on the Desktop. On the Mac, this is easily accomplished by dragging the icon for sequescan.jar to the Dock. One can then click on the alias or symbolic link. From the command line, the GUI can be started by executing:

java -jar <path-to-sequedex-dir>/lib/sequescan.jar.

Once you have either clicked on the icon for sequescan.jar or called it from the command line, you will get the application window, which is displayed below.



The application window is comprised of a menu bar at the top with 3 menus (Sequedex, Utilities, and Help). Below this are the options and buttons for running Sequescan, the algorithm in Sequedex which assigns phylogeny and function to DNA or protein sequences. And finally at the bottom of the window is the Progress Panel.

The rest of this document describes running Sequescan on samples using Version 2.1.1 of Sequedex. Many of the features of 2.1.1 were not present in earlier versions. The version is displayed at the top of the application window. The specific build version may also be important when reporting problems or asking for help on running Sequescan. To get the build version, click on the Help menu and select “Sequescan

Build Version". Current build is 20181211. The Help menu also has an option for help using the command line.

To run Sequescan, you will need to select the appropriate options from the fields in the form above the Run Sequescan button. Once selected, you can save the options you have chosen using the Save Options button. When you next open the GUI, the last set of saved options will be displayed.

A brief description of the options (note: these options parallel the command line options and thus the option for help using the command line may also be useful.

Data Module: Select a data module from the drop down menu listing available options (by default the distribution comes with Life2550-8GB.1 and virus1252.1).

Function Set: Select a function set from the drop down menu listing available function sets (if any) in the selected data module. Choose "none" if you do not want to assign functions to sequences.

Input Type: Select from drop down menu with options as follows

directory (process all files in the chosen directory which have a supported extension)
sequencing file (process a single file with a supported extension)
list file (process all files, with supported extension, included one per line in the list file)

Default supported extensions are:
fasta fst fna fas ffn fa fastq fq faa and their gzipped (.gz) versions
Content of input files with extension .fna, .ffn, .fq, or .fastq will be treated as DNA sequences. Content of input files with extension .faa will be treated as protein sequences. All other input files with a valid extension will be tested to determine if content is DNA or protein.

Base Directory for List File: Enter directory path or select directory (using button on the right) to be used with list file; enter none (the default) if list file includes full paths. Option will be ignored for **directory** or **list file** options.

Configuration File: type or select (using button on the right) configuration file. Default is the configuration fie included in the distribution. Do not change config file unless you understand in some detail how Sequedex works. Making changes can break Sequedex. Therefore if you do create an alternative configuration file, make sure you keep a copy of the original.

Output Directory: type or select (with button on the right) directory for sequedex output including log files. Default is "input", meaning the directory where input files reside. If the user does not have write permission to the output directory, this will result in an error.

Write Sequences: Select option from drop down menu for writing out sequences which match kmers in the data module (for DNA, sequences are written for the appropriate reading frame). There are 5 options to choose from: 0 (no), 1 (yes), 2 (Yes, with Kmers), 3 (Yes, Translate DNA), 4 (Yes with Kmers, Translate DNA). Default is 0 (no).

Write WhoDoesWhat File: select true or false (default is false)

Thread Number: Enter number of threads and thus number of files which can be analyzed in parallel. This choice should be guided by the number of cores on the computer and the number of files to be analyzed. Adding extra threads for a single file will not help as the program only uses stupidly parallel operation (i.e. one thread per file).

Protein Fragment Cutoff: Enter minimum length of amino acid sequence fragment which will be matched to kmers. Default is 15, in which case a read of less than 45 base pairs will automatically be ignored. Note: a fragment is a contiguous length of amino acid sequence which does not contain a stop codon.

Quiet: select true or false. If true, fewer messages will be written to log files or Progress Panel. Default is false.

Once you have selected/entered all the appropriate options, clicking on the **Run Sequescan** button will start the job. Only a single job (which, of course, may include an entire directory or list of files) can be run at one time using the GUI. After clicking on the **Run Sequescan** button, the program will check if the computer has enough available memory to run the selected data module. If not, a warning message will appear and offer the user a chance to stop the run. When starting a job, you should see something similar to the following appear in the Progress Panel:

Jun 20, 2018 3:01:44 PM Start Sequescan Program (mode=run)
Jun 20, 2018 3:01:45 PM Checking memory requirements
Jun 20, 2018 3:01:45 PM Sequescan will run with AutoMaxHeap, externally executing the following command:
Jun 20, 2018 3:01:45 PM java -Xmx7000m -jar /home/jcohn/sequedex/lib/sequescan.jar run -m -n -g /home/jcohn/test1/log/sequescan_run_20180620_150144.html -d Life2550-8GB.1 -s pfam27 -c /home/jcohn/sequedex/etc/sequescan/sequescan.conf -o /home/jcohn/test1 -t 4 -f 1 -w 0 -a 15 /home/mcmahon/V2.0/spike2
Jun 20, 2018 3:01:45 PM AutoMaxHeap execution starting now. Process output will go to /home/jcohn/test1/log/sequescan_run_20180620_150144.html

Messages will appear as the run progresses. At the end of the run, you should see something similar to:

Jun 20, 2018 3:12:44 PM AutoMaxHeap run exited with status: 0

Status of 0 is good - that means the run completed successfully.

In addition to messages in the Progress Panel, an html log file is written (in some cases with more detailed messages). The path to the html log file is displayed in the Progress Panel (see above). Two additional log files are generated along with the html log file. The names of these files start with the name of the html log file and end with .processErr.txt and .processOut.txt and include additional error and informational messages respectively generated by the running job.

Quit Sequedex

To exit Sequedex, one can either choose “Quit Sequedex” from the Sequedex menu or simply close the application window. In either case, you will be prompted by a dialog box warning:

This will not stop any runs executing in an external process.
They will continue to run until completion.
Continue to Exit?

All jobs executed in the GUI are currently executed in an external process (which allows Sequedex to run with just the right amount of memory). So if you have not seen the “AutoMaxHeap run exited with status” message in the Progress Panel, the job is still running in an external process.

Output from Sequedex

In addition to the log files described above, which are for the entire job, Sequedex creates a subdirectory in the output directory for each file analyzed. The subdirectory by default is named <sequencefilename>.sqdx. In this directory, you will find up to five output files (depending on what you choose for **Write Sequences** and **Write WhoDoesWhat File**).

These files are as follows.

Matching sequences: name starts with db and ends with extension of .fa, .fq or .faa depending upon original sequencing file and options chosen for **Write Sequences**. Example name: db-Life2550-4GB.0xseed_0911.m1.faa.

Job stats: name ending in stats.tsv – e.g. Life2550-4GB.0-stats.tsv.

Who file (phylogeny assignments): starts with who, e.g. who-Life2550-4GB.0.tsv.

What file (function assignments): starts with what, e.g. what-Life2550-4GB.0xseed_0911.m1.tsv

WhoDoesWhat (matrix of counts by who and what): starts with wdw, e.g. wdw-Life2550-8GB.1xpfam27.tsv.

Running Sequedex from the Command Line

Instructions for running sequedex from the command line can be found in two ways:

1) In the GUI, from the Help menu, select option “Sequescan Command-Line Help”.

2) From the command line:

java -jar <path-to-sequedex-dir>/lib/sequescan.jar run -h

Note: Currently when you run Sequedex from the command line, you will see on the console (and in the processOut.txt file) the following at the beginning of the run:

LOGBACK: No context given for
ch.qos.logback.core.spi.ContextAwareBase@60215eee

This should be ignored and will be fixed in due course.

Adding or Removing Data Modules

To add a data module to Sequedex, you merely have to put the data module jar file into path-to-sequedex-dir/data directory. Data modules have a version number and will not use a data_module with the wrong version number. To remove a data module, simply remove the data module jar file from the data directory. Note: All data module files end in .jar. However, when they appear in the GUI selection box or when running Sequedex from the command line, the .jar is not included. Thus, the data module file virus10k.1.jar will appear in the selection box as virus.10k.

J. Cohn
November 28, 2018