# Starcode tutorial

Guillaume Filion

June 17, 2015

## 1 Build instructions

Starcode runs on Mac and Linux as a command line application. You can download the source code of Starcode from Github with the following command on a standard terminal.

```
git clone git@github.com:gui11aume/starcode
```

Note that this requires that you already have a Github account and that the computer you are working on has an SSH key registered on Github. If this is not the case, follow the instructions from `https://help.github.com/articles/generating-ssh-keys/`.

This should download a directory named `starcode`. To build Starcode, execute the following.

```
cd starcode
make
```

This should succeed on most Linux systems because `make` is available by default. If this is not the case, you can obtain it by typing `sudo apt-get install make` on Ubuntu. On Mac, you need to install XCode, which may take some time. First, you will need an Apple ID, then you will need to download it from the developer website of Apple `https://developer.apple.com/xcode/downloads/`. Then, you may need to follow the instructions shown on the following link to install the command line version of `make` `http://stackoverflow.com/q/10265742/1248687`.

Calling `make` should create an executable called `starcode`. To check that the building is successful, execute the following command in the same directory.

```
./starcode test/test_file.txt
```

If you obtain the output shown below, then everything went fine and you are done with the build. If not, then something went wrong. In this case, you can explain how to reproduce the problem on `https://github.com/gui11aume/starcode/issues`.

```
running starcode with 1 thread
reading input files
raw format detected
sorting
setting dist to 2
progress: 100.00%
AGGGCTTACAAGTATAGGCC  7
CCTCATTATTTGTCGCAATG  7
GGGAGCCCACAGTAAGCGAA  7
TAGCCTGGTGCGACTGTCAT  7
TGCGCCAAGTACGATTTCCG  7
```

# 2 Starcode basics

The previous example is a standard invocation of Starcode on sequence data. The input file `test/test_file.txt` is in raw format (*i.e.* plain text), it contains 35 lines, each with one sequence of 20 nucleotides. By default, Starcode is verbose and prints some information about the clustering process. In the example above, Starcode used a single CPU, and it clustered the sequences within edit distance 2 of each other.

The output starts after the line `progress: 100.00%`. It consists of five lines (one per cluster) where the sequence is the centroid of the cluster (the most representative sequence) and the number is its size (the number of sequences in the cluster).

We can make Starcode quiet (non verbose) with the `-q` option, and we can set the clustering distance to 0 with `-d0`.

```
./starcode -q -d0 test/test_file.txt
```

The output is the following.

```
AGGGCTTACAAGTATAGGCC  6
CCTCATTATTTGTCGCAATG  6
GGGAGCCCACAGTAAGCGAA  6
TAGCCTGGTGCGACTGTCAT  6
TGCGCCAAGTACGATTTCCG  6
AGGGGTTACAAGTCTAGGCC  1
CCTCATTATTTACCGCAATG  1
GGAAGCCCACAGCAAGCGAA  1
TAACCTGGTGCGACTGTTAT  1
TGCGCCAAGTAAGAATTCCG  1
```

This time we obtain five clusters of size 6 and five clusters of size 1. Each centroid of the clusters of size 1 has 2 mismatches with one of the centroids of the clusters of size 6, which is why they form separate clusters when the distance is less than 2.

As a side note, this example output illustrates that the clusters are sorted first by size and then by alphabetical order of the centroid.