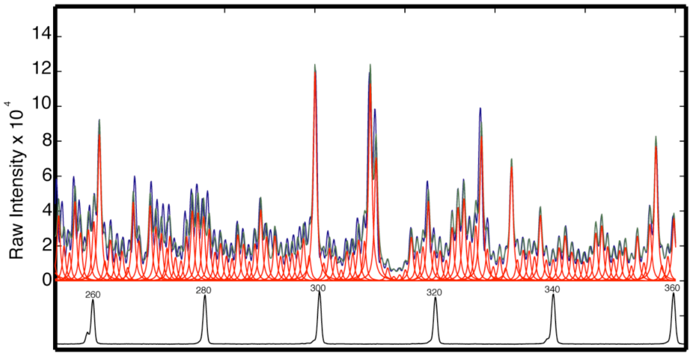
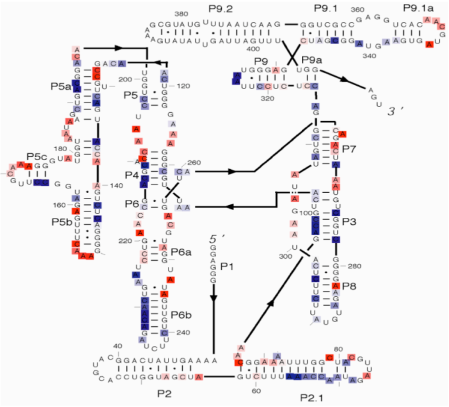
Capillary Automated Footprinting Analysis





Somdeb Mitra, Inna Shcherbakova, Russ B. Altman, Michael Brenowitz and Alain Laederach
http://simtk.org/cafa. E-mail: Alain Laederach (alain@helix.stanford.edu)

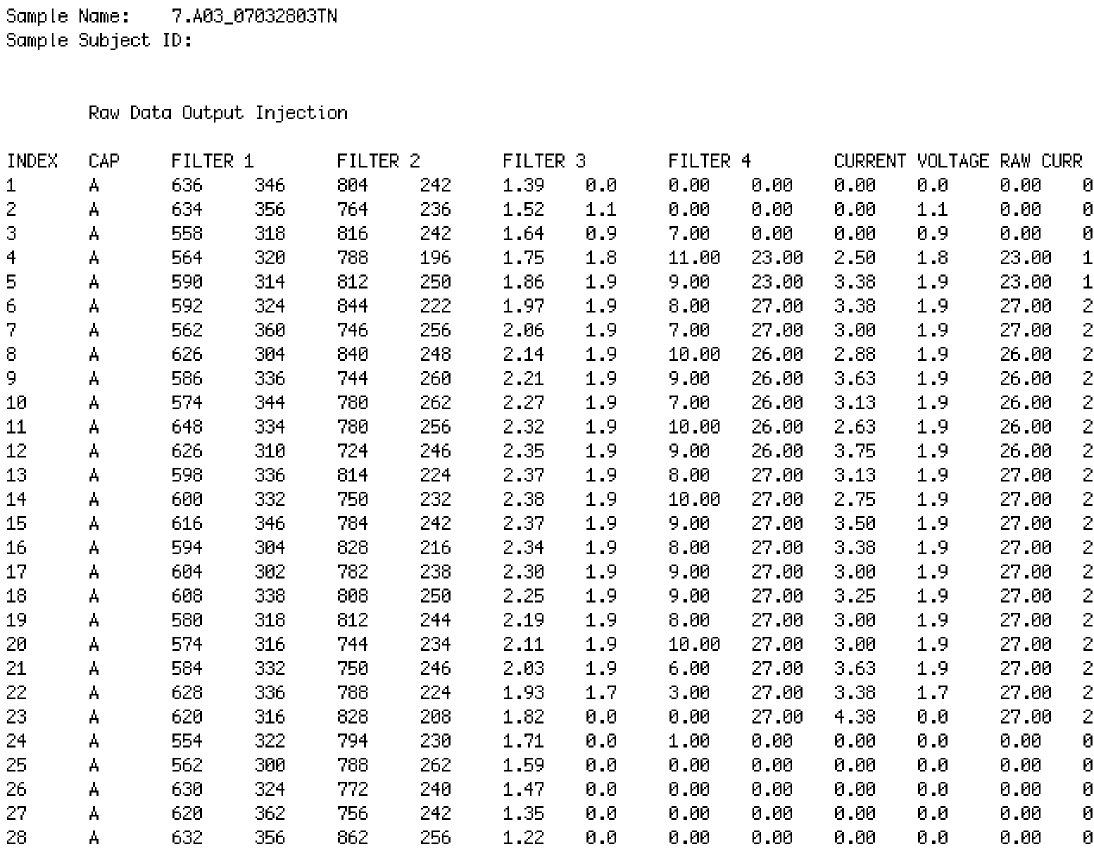
User’s Manual for the CAFA software.
March 2008, Stanford University
Alain Laederach

Installation

Platform specific installation instructions are provided in separate documents that can be downloaded from https://simtk.org/home/cafa. The figures in this manual are from the Mac Intel version of CAFA, but are equivalent to the Windows version.

General Data Philosophy

CAFA is currently designed to read in output from a Beckman CEQ-8000 sequencer. The data has to be exported from the Beckman software using the "Text Export" functionality in the database. The output files should look like this:



```
Sample Name:    7.A03_07032803TN
Sample Subject ID:


        Raw Data Output Injection

INDEX   CAP     FILTER 1        FILTER 2        FILTER 3        FILTER 4        CURRENT VOLTAGE RAW CURR
1       A       636     346     804     242     1.39    0.0     0.00    0.00    0.00    0.0     0.00    0
2       A       634     356     764     236     1.52    1.1     0.00    0.00    0.00    1.1     0.00    0
3       A       558     318     816     242     1.64    0.9     7.00    0.00    0.00    0.9     0.00    0
4       A       564     320     788     196     1.75    1.8     11.00   23.00   2.50    1.8     23.00   18
5       A       590     314     812     250     1.86    1.9     9.00    23.00   3.38    1.9     23.00   18
6       A       592     324     844     222     1.97    1.9     8.00    27.00   3.38    1.9     27.00   26
7       A       562     360     746     256     2.06    1.9     7.00    27.00   3.00    1.9     27.00   26
8       A       626     304     840     248     2.14    1.9     10.00   26.00   2.88    1.9     26.00   26
9       A       586     336     744     260     2.21    1.9     9.00    26.00   3.63    1.9     26.00   26
10      A       574     344     780     262     2.27    1.9     7.00    26.00   3.13    1.9     26.00   26
11      A       648     334     780     256     2.32    1.9     10.00   26.00   2.63    1.9     26.00   26
12      A       626     310     724     246     2.35    1.9     9.00    26.00   3.75    1.9     26.00   26
13      A       598     336     814     224     2.37    1.9     8.00    27.00   3.13    1.9     27.00   26
14      A       600     332     750     232     2.38    1.9     10.00   27.00   2.75    1.9     27.00   26
15      A       616     346     784     242     2.37    1.9     9.00    27.00   3.50    1.9     27.00   26
16      A       594     304     828     216     2.34    1.9     8.00    27.00   3.38    1.9     27.00   26
17      A       604     302     782     238     2.30    1.9     9.00    27.00   3.00    1.9     27.00   26
18      A       608     338     808     250     2.25    1.9     9.00    27.00   3.25    1.9     27.00   26
19      A       580     318     812     244     2.19    1.9     8.00    27.00   3.00    1.9     27.00   26
20      A       574     316     744     234     2.11    1.9     10.00   27.00   3.00    1.9     27.00   26
21      A       584     332     750     246     2.03    1.9     6.00    27.00   3.63    1.9     27.00   26
22      A       628     336     788     224     1.93    1.7     3.00    27.00   3.38    1.7     27.00   26
23      A       620     316     828     208     1.82    0.0     0.00    27.00   4.38    0.0     27.00   26
24      A       554     322     794     230     1.71    0.0     1.00    0.00    0.00    0.0     0.00    0
25      A       562     300     788     262     1.59    0.0     0.00    0.00    0.00    0.0     0.00    0
26      A       630     324     772     240     1.47    0.0     0.00    0.00    0.00    0.0     0.00    0
27      A       620     362     756     242     1.35    0.0     0.00    0.00    0.00    0.0     0.00    0
28      A       632     356     862     256     1.22    0.0     0.00    0.00    0.00    0.0     0.00    0
```

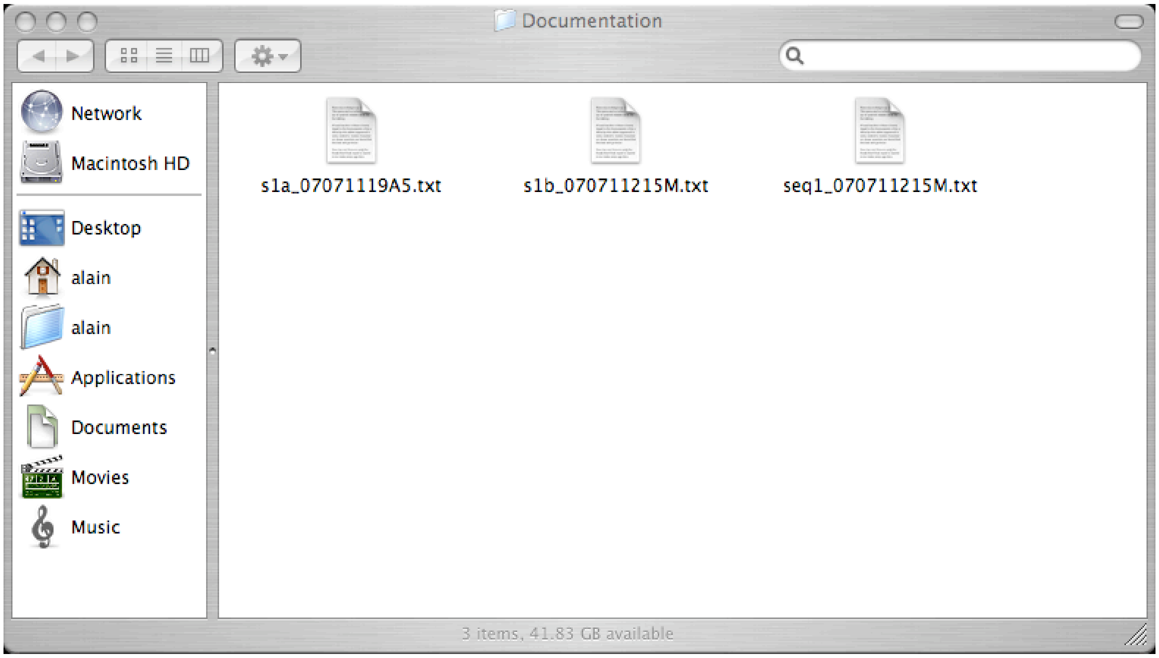
Currently, we do not directly support other types of input, but depending on need, please contact Alain Laederach (alain@helix.stanford.edu) if you would like to see other input formats supported.

The basic idea behind the CAFA data philosophy is that each well from the 96 well plate has one corresponding text file that contains the raw trace (Filter 1) and the corresponding reference ladder (Filter 2). It is important when setting up experiment to always remember to add a reference ladder to the sample. Without the reference ladder, CAFA will not be able to analyze your data and will crash.

As such, when setting up to analyze data, you will want to create a directory with all your data in it. You may want to setup a directory structure for each experiment to avoid an overwhelming amount of data being analyzed simultaneously. For example, if you are mapping the structure of an RNA under various solution conditions, or performing a titration, you may want to have a directory with all the data pertaining to a particular titration in /Titration1, and another Titration in /Titration2. CAFA will generate an output file with all the data in it in the end, and you can organize what data goes in that output file by organizing your data into different directories prior to analysis.

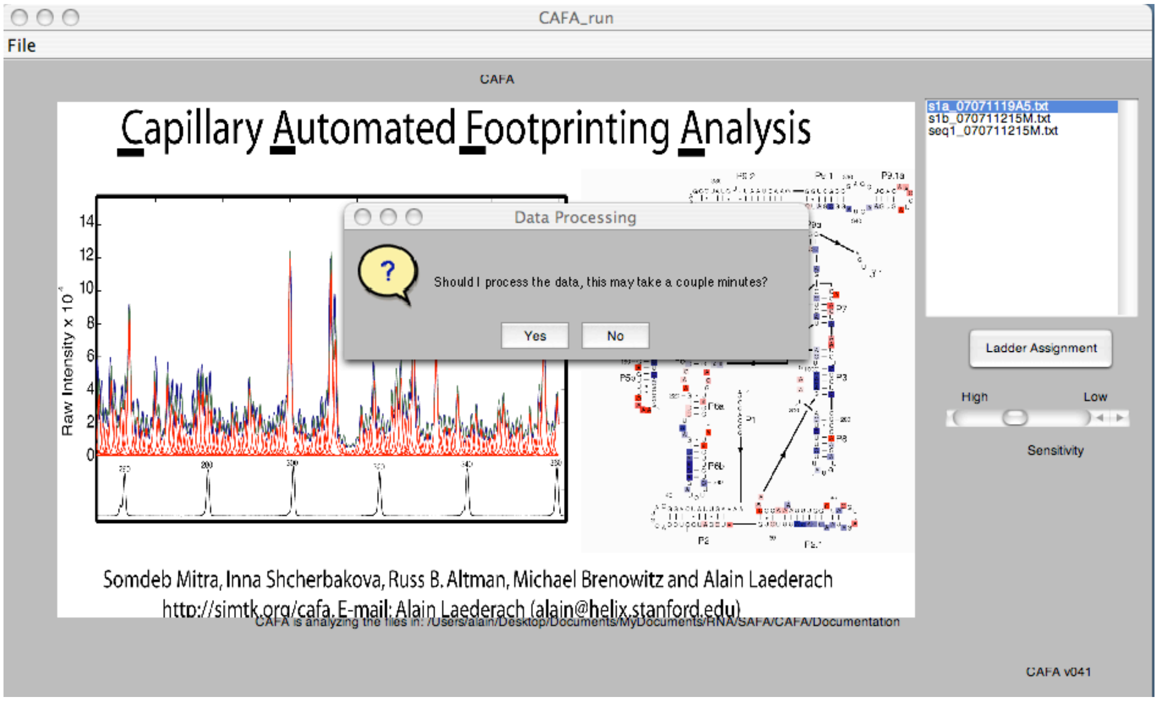
Reading in the Data

For this example we will consider a directory that has the following three traces (DMS maps of the L-21 group I intron), this data is also available for download from http://simtk.org/home/cafa.



The first two files are repeats of the experiment and the seq1_070711215M.txt file is control lane we will use as background for identifying RT stops.

Open CAFA (launching instructions differ depending on whether you are using a Mac or a PC) and select the working directory using the File-> Menu.



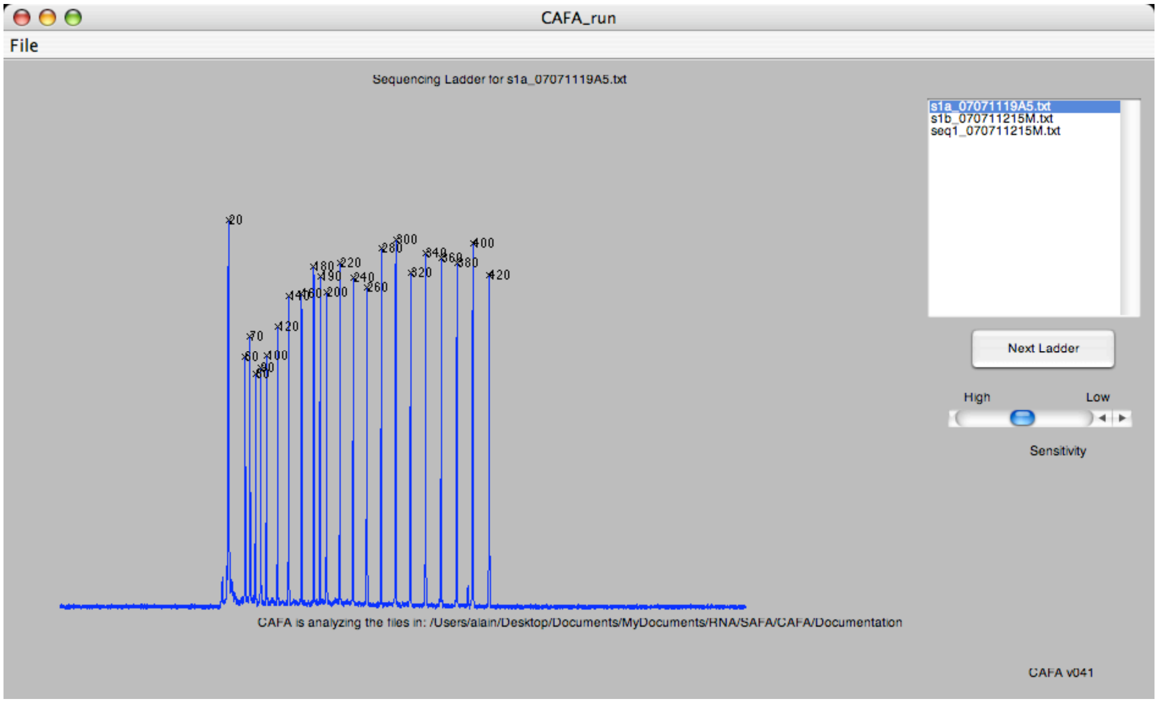
Somdeb Mitra, Inna Shcherbakova, Russ B. Altman, Michael Brenowitz and Alain Laederach
http://simtk.org/cafa, E-mail: Alain Laederach (alain@helix.stanford.edu)

Should I process the data, this may take a couple minutes?

CAFA is analyzing the files in: /Users/alain/Desktop/Documents/MyDocuments/RNA/SAFA/CAFA/Documentation

Go ahead and process the data by clicking on “Yes.” If you had previously worked with this directory and already processed it once, you can click “No.” This is useful when working with very large data sets you want to reanalyze.

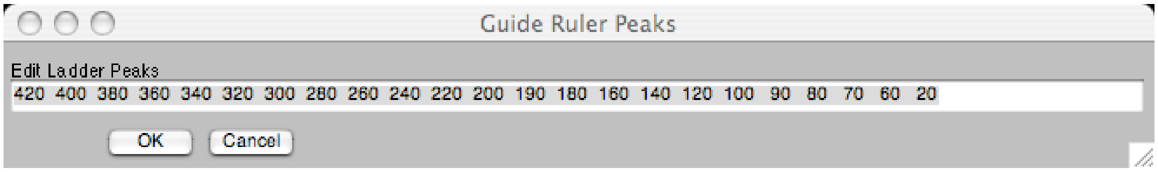
Once the data is processed, click on the Ladder Assignment button and CAFA will automatically begin assigning the DNA ladder in your data:



CAFA is analyzing the files in: /Users/alain/Desktop/Documents/MyDocuments/RNA/SAFA/CAFA/Documentation

Adjusting the sensitivity slider will adjust what CAFA considers to be a peak in the sequencing ladder. Normally, CAFA should be able to automatically guess all the peak positions correctly, but if the ladder trace is weak, going to a higher sensitivity can help.

CAFA defaults to the Beckman-400 ladder. If you are using a different ladder, you can edit the positions of the ladder peaks using the File-> Edit Ladder Positions dialog:



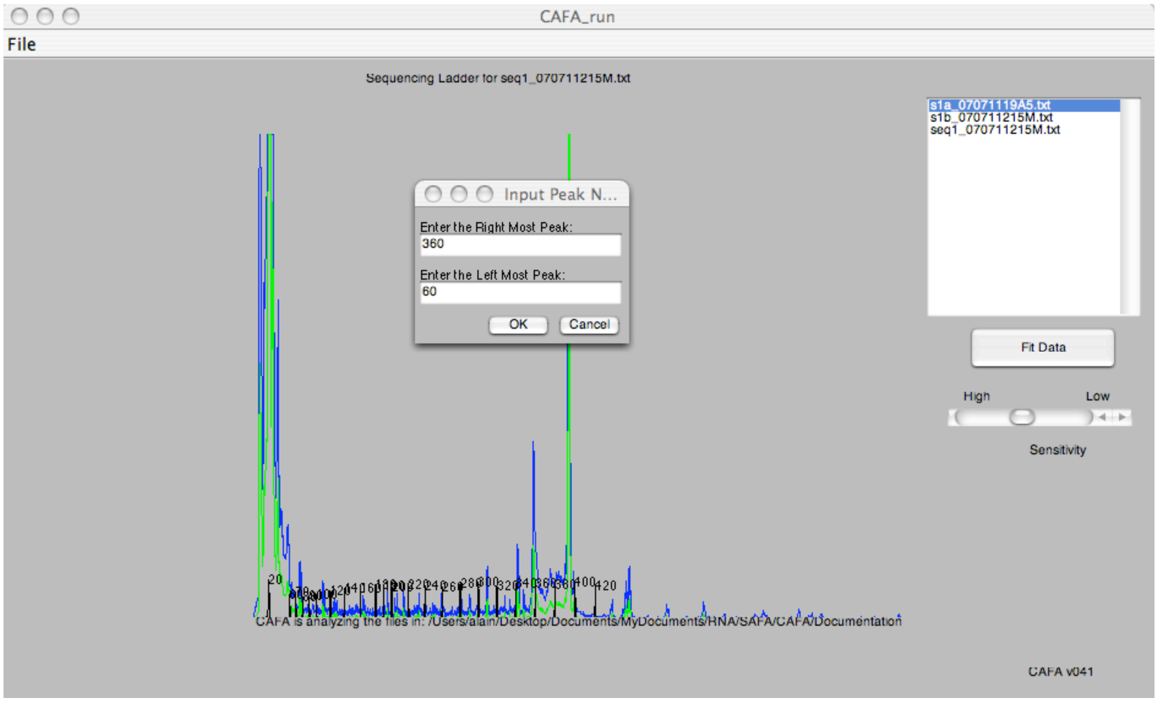
Guide Ruler Peaks

Edit Ladder Peaks

420 400 380 360 340 320 300 280 260 240 220 200 190 180 160 140 120 100 90 80 70 60 20

You can use different ladders for different traces, and each ladder is stored and associated with the particular trace you are looking at. If you do not have a ladder in you sample, CAFA may crash. The best is then to remove that data set from the directory, reprocess the data and continue the analysis.

Once you have assigned each ladder, CAFA will present you with the fitting parameters:



CAFA_run

Sequencing Ladder for seq1_070711215M.txt

s1a_07071119A5.txt
s1b_070711215M.txt
seq1_070711215M.txt

Input Peak N...

Enter the Right Most Peak: 360

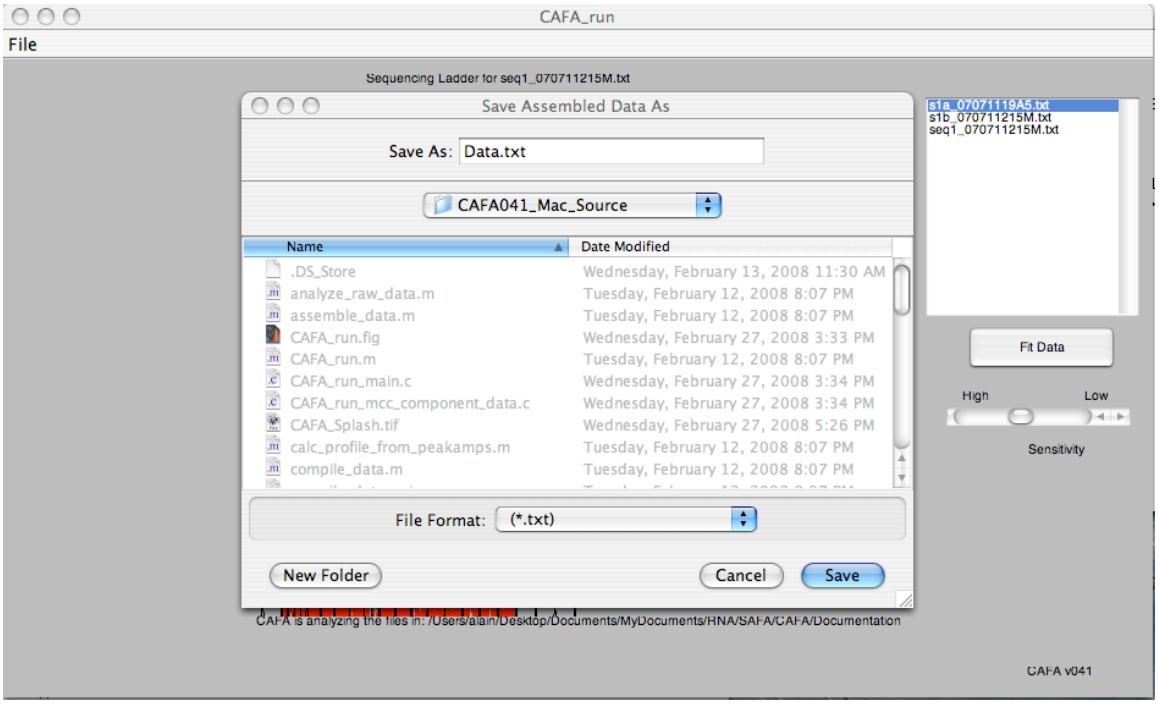
Enter the Left Most Peak: 60

CAFA is analyzing the files in: /Users/alain/Desktop/Documents/MyDocuments/RNA/SAFA/CAFA/Documentation

CAFA v041

Here you can enter the left and right most peaks you would like to fit. If your Nucleic Acid is shorter than your ladder, you can select to fit between the peaks where there is data. Simply click on “Ok” to begin fitting.

With the data fit, CAFA will ask you where to save the raw data:



Sequencing Ladder for seq1_070711215M.txt

Save Assembled Data As

Save As: Data.txt

CAFA041_Mac_Source

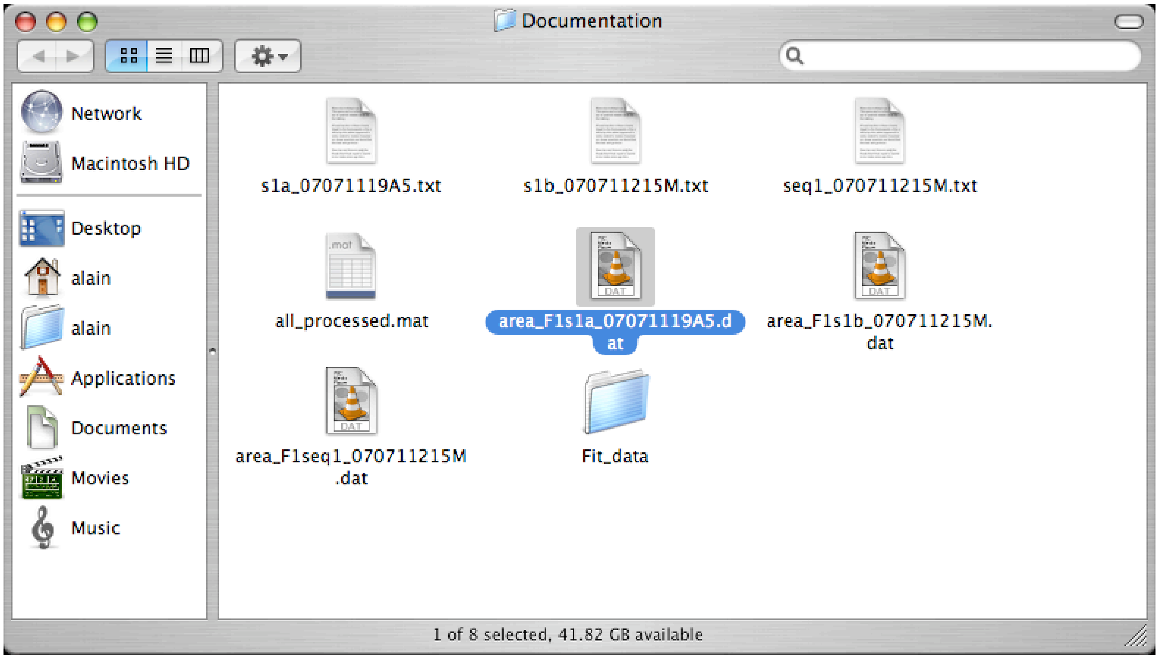
Name	Date Modified
.DS_Store	Wednesday, February 13, 2008 11:30 AM
analyze_raw_data.m	Tuesday, February 12, 2008 8:07 PM
assemble_data.m	Tuesday, February 12, 2008 8:07 PM
CAFA_run.fig	Wednesday, February 27, 2008 3:33 PM
CAFA_run.m	Tuesday, February 12, 2008 8:07 PM
CAFA_run_main.c	Wednesday, February 27, 2008 3:34 PM
CAFA_run_mcc_component_data.c	Wednesday, February 27, 2008 3:34 PM
CAFA_Splash.tif	Wednesday, February 27, 2008 5:26 PM
calc_profile_from_peakamps.m	Tuesday, February 12, 2008 8:07 PM
compile_data.m	Tuesday, February 12, 2008 8:07 PM

File Format: (*.txt)

s1a_07071119A5.txt
s1b_070711215M.txt
seq1_070711215M.txt

CAFA is analyzing the files in: /Users/alain/Desktop/Documents/MyDocuments/RNA/SAFA/CAFA/Documentation

At this stage you have finished the fitting of your data, congratulations! It is usually not a good idea to save this file in your original data directory, as it is a txt file like the raw data and may interfere with the analysis if you run the program again on the raw data. Instead, creating a directory called Fit_data separates the raw and fit data.



s1a_07071119A5.txt
s1b_070711215M.txt
seq1_070711215M.txt
all_processed.mat
area_F1s1a_07071119A5.dat
area_F1s1b_070711215M.dat
area_F1seq1_070711215M.dat
Fit_data

1 of 8 selected, 41.82 GB available

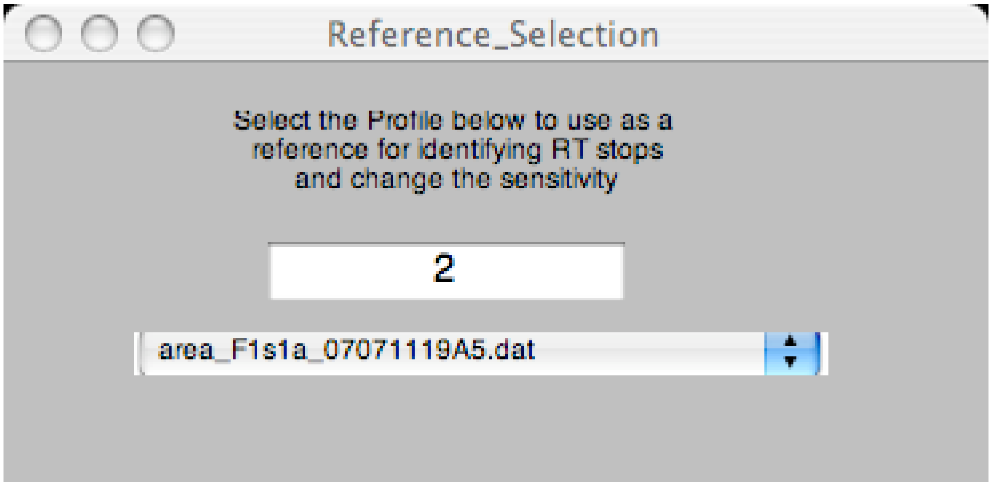
Above are the files that are now in your directory. The files starting with area_F1 correspond the individual peak areas and have this format:

```
F1s1a_07071119A5.txt
60      159922.574379
61      83176.303827
62      78713.837047
63      109612.246825
64      59842.400391
65      49383.548448
66      80645.881063
67      75769.886078
68      66772.737142
69      64594.606567
70      57230.873266
71      29542.321073
72      50833.046553
73      44031.947376
74      28988.165851
75      13190.880051
76      195933.832520
```

where the first column is the numbering corresponding to the DNA ladder and the second column is the raw peak area. The Data file saved after all the fitting has a similar format with all the data concatenated into columns.

One final step can be carried out to normalize and filter the data. This step is only if you are using a indirect labeling scheme for your nucleic acid (e.g. reverse transcription). To do this select File-> Normalize and Filter. This can be done anytime after the fitting even after you have quit and restarted CAFA.

Select the Data file you saved after fitting and a Reference_Selection window will popup:



Reference_Selection

Select the Profile below to use as a reference for identifying RT stops and change the sensitivity

2

area_F1s1a_07071119A5.dat

Select from this window the name of your experiment containing the background trace (i.e. a trace where the RT reaction was run on a non-modified nucleic acid). In this case it is the area_F1 seq1_070711215M.dat data.

CAFA will then ask you where you want to save the normalized data, which you can save in the Fit_data directory too. This data looks like this:

```
Residues area_F1s1a_07071119A5.dat area_F1s1b_070711215M.datarea_F1seq1_070711215M.dat
60       3.356953e+00       4.242148e+00       1.542471e+00
61       1.745963e+00       2.387782e+00       1.208065e+00
62       1.652291e+00       2.154473e+00       7.477427e-01
63       2.300882e+00       2.748971e+00       1.886696e+00
64       1.256158e+00       1.471044e+00       3.605530e-01
65       1.036616e+00       1.372553e+00       3.705500e-01
66       1.692847e+00       1.984043e+00       1.586979e+00
67       1.590494e+00       1.938537e+00       6.243484e-01
68       1.401634e+00       1.817204e+00       4.868283e-01
69       1.355913e+00       1.574855e+00       1.571141e+00
70       1.201340e+00       1.370064e+00       1.309624e+00
71       6.201261e-01       6.345996e-01       0
72       1.067042e+00       1.309437e+00       9.398702e-01
73       9.242796e-01       1.109162e+00       2.474074e-01
74       6.084939e-01       6.001110e-01       0
75       NaN       NaN       NaN
76       NaN       NaN       NaN
77       NaN       NaN       NaN
78       3.094050e+00       3.699825e+00       1.870406e+00
79       NaN       NaN       NaN
80       NaN       NaN       NaN
81       NaN       NaN       NaN
82       5.732181e-01       6.727585e-01       1.625434e-01
83       5.701988e-01       6.333145e-01       3.072298e-01
```

Where the data is now normalized to the mean intensity and data near strong RT stops is labeled as NaN. This is non-reliable data since it is a result of an RT stop.

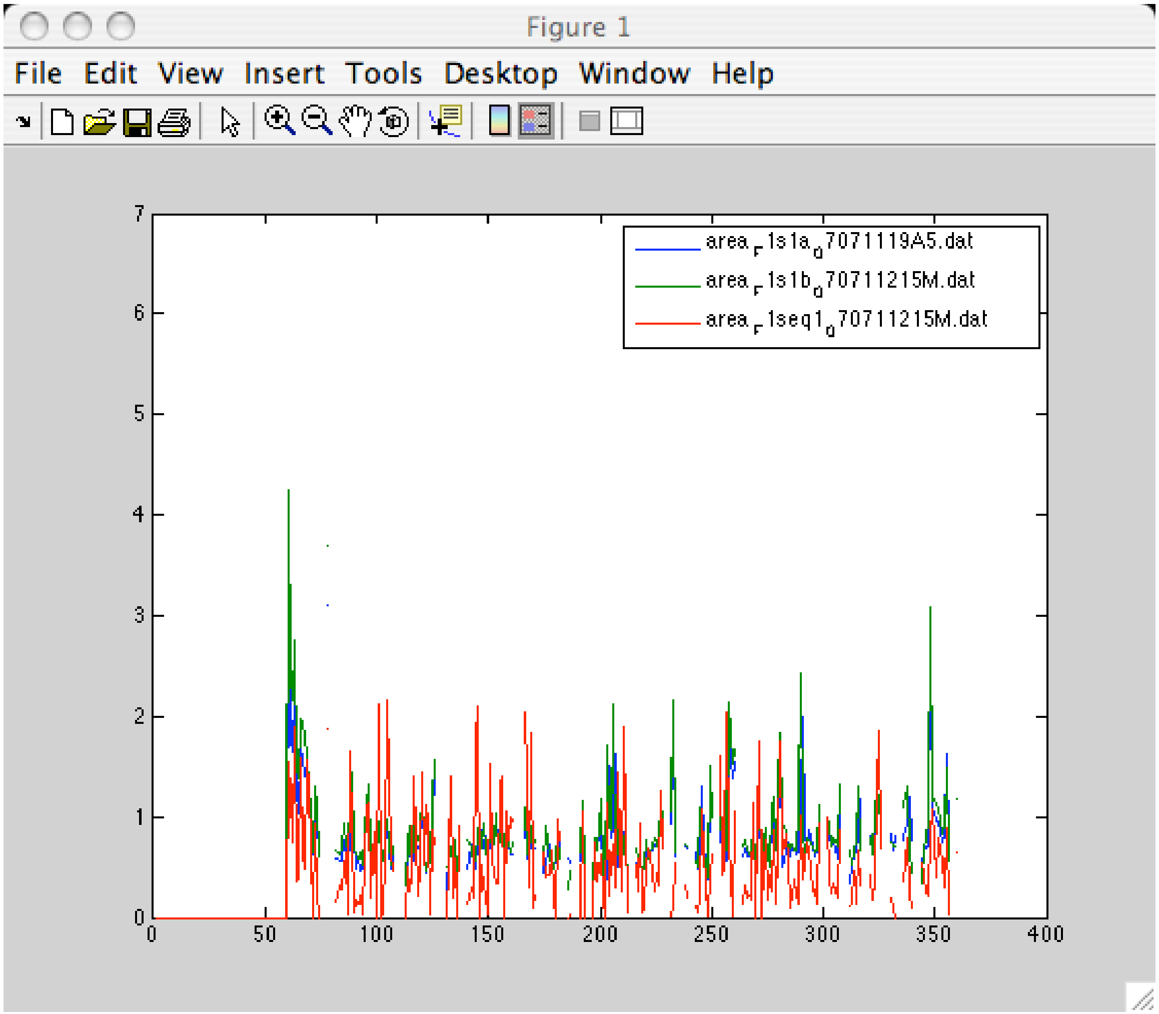


Figure 1
File Edit View Insert Tools Desktop Window Help
area$_F$1s1a$_d$7071119A5.dat
area$_F$1s1b$_d$70711215M.dat
area$_F$1seq1$_d$70711215M.dat
7
6
5
4
3
2
1
0
0
50
100
150
200
250
300
350
400

The above figure is plotted by CAFA to show where data was removed. CAFA also indicates the percentage of the data that was excluded (in this case 22.5%). You can adjust the sensitivity of the filtering by changing the Sensitivity (default 2) in the Reference_Selection window.

At this stage, the data can be imported into excel and further analyzed.