Normalization Scripts

The main purpose of these scripts is to read in column data from multiple files (in formats such as .txt, .csv, and .dat) and normalize each trajectory to occupy values between 0 and 1. In this way, fluorescence level data may be compared at a common "scale" or input to programs such as ebFRET which require normalized data.

The main function is `normalizeTrajectorySet.m`, which is located in the main folder of the repository. It contains calls to functions inside the "functions" directory, so the working directly should be set to the sourceCode folder before running (otherwise the user will be prompted to switch directories). The user is first prompted for the number of channels. As described in the KERA documentation, channels are independent signals. In a dual-illumination experiment observing Cy3 and Cy5 simultaneously, and colocalizing the signals, there would be two channels. In a study with only one kind of fluorophore, there is likely only one channel. This was the case for the RPA study (1). Then, the user is prompted to select a directory which contains data files. The program assumes that all files of a certain type (specified by the user as .txt, .csv, or .dat) in that directory contain columns of data which should be imported. One column from each file is imported, according to the column the user specifies. For example, many .dat files contain two columns: a time series and an intensity trajectory. In that case, the user would specify that column 2 contains the desired data. If there is more than one channel, the user is prompted to go through these steps for each channel. The goal is to keep associated (colocalized) trajectories sorted in the same order; assuming the filenames are consistent, this is done automatically by the `dir` function upon input, especially if the colocalized signals are represented by different columns in the same files. If your data is in some other form, rearranging it into the format found in the `intensity` and `fileNames` variables from the sample data folder (normalizationSampleData.zip). By loading it in and setting "importOldSession" to 1, the default importation steps may be skipped (you may have to manually assign some other variables, like `channels`).

Once the data is imported, the user is prompted to view the traces. This is a good idea, unless the viewing/trimming process has already been completed for this data set and you are working from old imported data. Upon viewing, the user is guided through each trajectory in the set sequentially, viewing all channels (if there is more than one) at once, but only editing them one at a time. So, trajectory 1 channel 1 is seen first, followed by trajectory 1 channel 2, and so on. For each trajectory, the user can perform four actions:

- Discard the trace: this removes it from analysis, and also removes the associated trajectories in all other channels (if it is a multi-channel data set).
- Trim trace: clicking this button allows the user to select two points (the start and end) of the portion of the trajectory they would like to keep. In a multi-channel dataset, the corresponding trajectories from all channels will be trimmed to this length. The appearance of the data does not change to reflect the trimming. Clicking this button again on the same trace overwrites the previous trim region.
- Select low: select a region by clicking twice (start and end points) which represents a continuous region of the "low" state of the trajectory. This is the state which will be set

at the baseline when the trajectories are aligned and normalized. Clicking this button again on the same trace overwrites the previous low region selection.

- Select high: select a region in the same way, but which should be normalized as a "high" value (aligned to 1 or near 1). Clicking this button again on the same trace overwrites the previous high region selection.
- Go back: go to the previous trajectory

If no trimming is done, the full length of the trajectory will be included in the analysis. A feature reminds the user to trim, select low, and select high on every single trace, but this can be suppressed by clicking the appropriate option on the dialogue box. It is still recommended that at least every trace be assigned a low and a high region; this is because automatic normalization is highly data-dependent and the assumptions made by that algorithm may not be appropriate for your data.

When all traces have been viewed, the user is taken back to the main code, and a full export package is saved; an example of this is in the zipped file normalizationSampleData.zip in the sample data folder. It holds the original imported data as well as the edits made during the viewing session. Saving at this point means the user can always come back and make additional edits (running the code with importOldSession set to 1 after opening the .mat saved variables), either in the view window or in the normalization.

If all traces have had a low or high region selected, the normalization proceeds entirely automatically from that point, and the last step is to select the export formats and export locations. The original filenames are appended with flags specifying the filetype and saved in the locations indicated by the user. hFRET requires that all data within a set be the same length; if the trajectories have been trimmed to non-uniform lengths, the normalization script will ask the user for a padding value which will be appended to all trajectories until they are the same length. By setting this to a value like -.5, the user can assure that hFRET will assign the padded values to their own state (and only the padded values to that state) though this does mean that the hFRET parameters must be set to fit an additional state than would normally be part of the model. However, these details are only relevant to post-processing and dependent on the discretization program the user is planning to import the normalized traces to.

For questions about the code, or to contribute to the repository, email Joseph Tibbs at jtibbs2@illinois.edu.

1. Pokhrel, N., Caldwell, C.C., Corless, E.I., Tillison, E.A., Tibbs, J., Jocic, N., Tabei, S.M.A., Wold, M.S., Spies, M. and Antony, E. (2019) Dynamics and selective remodeling of the DNA-binding domains of RPA. *Nature Structural & Molecular Biology*, **26**.