## Readme introduction

This Readme describes in detail how to use the bash scripts *barcoder.sh* and *reconstructer.sh* locally as well as the data analysing excel spreadsheet *loopproperties.xlsx*. For information on how the bash scripts actually work, see the annotations inside each script.

Using Linux

The scripts are in the BASH programming language and can are run in linux. A tutorial of the necessary basics to run bash scripts is here:

<http://ryanstutorials.net/linuxtutorial>

The scripts can be opened by any text editor to read and edit. They are annotated to make them easier to use and edit.

User inputs

Both scripts need the user to define some basic inputs. All user-defined inputs required are indicated in the top section of each script. Inputs are entered simply by editing the text after the '=' sign.

For barcoder.sh:

1. Input fasta file name (default = example.fa)
2. Columns to be barcoded
3. Barcode sequences

For reconstructer.sh:

1. Input fasta file name (default = newalign.fa)



Figure 1 | User inputs section for *barcoder.sh* (left) and *reconstructer.sh* (right)

Using *barcoder.sh*

*Barcoder.sh* replaces user-defined columns of a fasta alignment with barcode sequences 12 amino acids long. To use it follow these steps:

1. Open *barcoder.sh* in a text editor and input the required parameters in the ‘User inputs’ section
2. Save the newly edited *barcoder.sh*
3. Place the input fasta file and *barcoder.sh* in a folder
4. Open the command terminal
5. Navigate to the folder containing *barcoder.sh* with the command: **mv C:/folderlocation**
6. Run *barcoder.sh* with the command: **bash barcoder.sh**
7. The scripts should output the following:
   1. A BARCODED.fa file that contains the new barcoded alignment
   2. A set of column.fa files that contain the columns replaced with barcodes
8. The BARCODED.fa file can be re-aligned by any standard alignment program (e.g. ClustalΩ, Probcons, Muscle etc.)

Notes:

The *barcoder.sh* script can barcode up to 8 columns. If you need to barcode fewer than that, simply leave the later column definitions empty.

When several alignments need to be combined into a single alignment, the user has two options. Firstly, aligning homologous cysteines across all the sequences can be done manually (with no attempt to align the non-cysteine loops) before running *barcoder.sh* on the single ‘alignment’ file. Secondly, the individual alignments can be barcoded independently with *barcoder.sh* (as in the example figures used in the manuscript). When these barcoded alignments are concatenated into a single file they can be re-aligned by any standard alignment program to align the barcoded columns and allow the inter-barcode sequence to align with homologous regions. The column output files from *barcoder.sh* will also need to be concatenated in the same order as the concatenated barcoded alignments in order to run *restorer.sh* on the concatenated alignment.

Using reconstructer.sh

Once the barcoded alignment has been re-aligned by an appropriate program, *Resconstructer.sh* replaces the barcoded columns of a fasta file with the original columns that were removed and saved by *barcoder.sh*. Additionally, it provides summary file of the length, hydrophobicity and charge of each inter-cysteine region. To use it follow these steps:

1. Open *reconstruter.sh* in a text editor and input the required parameters in the ‘User inputs’ section
2. Save the newly edited *reconstruter.sh*
3. Place *reconstruter.sh* and its input fasta file in the folder containing the files produced by barcoder.sh
4. Open the command terminal
5. Navigate to the folder containing *reconstruter.sh* with the command: **mv C:/folderlocation**
6. Run *reconstructer.sh* with the command: **bash reconstructer.sh**
7. The scripts should output the following:
   1. A FINAL.fa file that contains the new alignment with barcodes removed
   2. A loop\_statisitics.csv file containing the properties for each sequence
   3. A folder containing the column.fa files
   4. A folder containing the loop.fa files
8. The FINAL.fa file is the final alignment with the barcodes removed and the true sequences restored
9. The loop\_statisitics.csv file can be copy-pasted into the Excel spreadsheet *loopproperties.xlsx* which will graphically summarise the data

Notes:

If using custom barcode sequences, ensure that they match in both the *barcoder.sh* and *reconstructer.sh* inputs and are 12 letters long. For calculating loop properties, all cysteines are assumed to be oxidised (i.e. involved in disulphide bridges).

If several alignments were generated with *barcoder.sh* that were later combined, the column files (column1.fa etc) for each alignment will need to be concatenated in the same order that the sequence alignments were combined before running *reconstructer.sh*. This allows *reconstructer.sh* to replace the correct amino acids in the barcoded columns for all sequences in the combined alignment. This cannot be performed by the web tool, and requires running the bash scripts locally.

Using *loopproperties.xlsx*

Once *reconstructer.sh* has generated the summary file loop\_statistics.csv, the data from this can be copy-pasted into *loopproperties.xlsx* which will provide summary graphs. To use it follow these steps:

1. In Excel, open both loop\_statistics.csvand *loopproperties.xlsx*
2. Copy the whole spreadsheet of loop\_statistics.csv
3. Paste the copied contents into the first worksheet page of *loopproperties.xlsx*
4. The subsequent worksheets of *loopproperties.xlsx* give the following information:
   1. Length – amino acid length
   2. Hydropathy – Doolittle index of hydrophobicity
   3. Charge – net coulomb charge
   4. Figures – a summary of bar charts and histograms from the other sheets

Notes: parameters with no data (e.g. the charge of a loop with length zero) are marked with 1E-100 to distinguish from true zero values (e.g. the charge of a neutral loop). These 1E-100 values are ignored when generating the subsequent charts.



Figure 2 | *loopproperties.xls* spreadsheet for worksheet ‘Raw data’ with pasted data (above) and worksheet ‘Figures’ with summary charts (below)