Twine User Guide

version 5/17/2013

http://labs.bio.unc.edu/crews/twine/

Joseph Pearson, Ph.D.

Stephen Crews Lab

http://www.unc.edu/~crews/

Overview: Twine is designed to help the user efficiently visualize the most common information used in detailed analyses of cis-regulatory modules (enhancers): clusters of binding sites for putative regulators and conservation of those sites. Twine takes one or more FASTA alignments as input and generates several representations of those alignments. DNA sequence motifs (IUPAC consensus sequences and Position Count/Frequency Matrices) can be added, using thresholds set to user specification. Patterns of clustering and conservation can often be easily visually identified using the Aligned Species View, so putative "minimal enhancers" can be identified from larger elements by virtue of conserved sub-regions enriched for motifs; FASTA alignments of these sub-regions can then easily be exported for further analysis. Each graphical representation ("View") can be exported to a file that can then be manipulated in Adobe Illustrator or equivalent vector graphics programs (*e.g.* Inkscape), to make figure panels for publication. A text output of all motif matches can also be generated. Alignments can be sent to user-written Java plugins that analyze or manipulate the alignments and then read the output back into Twine; in principle, command-line programs such as multiple alignment software could use Twine as a front-end.

How to start:

1) Download and unzip Twine: Twine is distributed as a .zip file containing a .jar file (an executable Java class package) containing the Twine classes and source code, as well as several libraries (Batik and JSPF) that are important for functions such as exporting Views as SVG files and plugin functionality. Unzip the package into a folder, then run the Twine .jar file.

a) shortcuts: If double-clicking the Twine .jar file doesn't run the program, first update to Java SE 7 (http://www.oracle.com/technetwork/java/javase/downloads/index.html, download the Java Runtime Environment (JRE)). If it still won't run, you may need to create a shortcut that emulates the command-line (it's easy, don't worry). Java Jar files can be run on the command line by "java -jar ***Twine.jar"*** (the name of the .jar file may be slightly different), and this can be emulated in the shortcut properties.

b) The Twine .zip file also contains several default folders for libraries and output files, including several example files. If Twine.jar (and the lib folder) are moved, these default folders will be recreated, but will not have the sample files.

2) Download an aligned sequence: The FASTA format is the most common format for representing short regions of DNA aligned to orthologs. Orthologous sequences can be found using BLAST, then aligned using software optimized for non-coding and diverged DNA (T-COFFEE and Dialign-TX work well, and web servers exist hosting these programs). Alternatively, pre-computed alignments can be retrieved from web pages such as UCSC Genome Browser, then converted from MAF (Genome-scale alignment format that accommodates inversions) to FASTA using Galaxy. Obviously, the better the alignment, the better the analysis. See appendix for detailed instructions about getting FASTA alignments from the UCSC Genome Browser.

3) Open alignment(s) in Twine: Start Twine, then click File > Open FASTA alignment. Find one or more FASTA files (usually .fas or .fasta, but any FASTA format text file will open), and click "Open".

Representations of each alignment should now appear in the Comparison View, and the selected alignment will also appear in the other two Views.

4) Add Motifs: Click Analyze > Add Motif to bring up the motif dialog. Two types of motifs are recognized:

> **a) IUPAC consensus sequences.** These use the letters ACGT (unique nucleotides) and RYSWKMBDHVN (degenerate nucleotides) for a consensus sequence that represents the set of sites that a transcription factor of interest can bind. For example, "YMATTA" is commonly used as the consensus Hox binding site, which represents (C/T)(A/C)ATTA. More than one sequence (or consensus) can be entered. You can increase the number of nucleotide mismatches allowed to decrease stringency, and change the color of the motif representation.
> **See: http://labs.bio.unc.edu/crews/twine/Twine_IUPAC_motifs.html for details.**
>
> **b) Position Frequency (Weight) Matrix.** Each position of the binding site is represented on the X-axis (left-to-right), and the rows indicate the frequency of A,C,G,T (top-to-bottom) at each position. Vertical matrices can be rotated by clicking "Rotate Matrix". Count matrices (integers in each position) will be converted to frequency matrices by Twine, so each column will add up to 1. The threshold calculates the minimum similarity of the matrix to a given window of the sequence, as calculated by multiplying each position's frequency in the matrix. It's represented in -ln, so a lower number (closer to 0) is a stricter threshold; higher thresholds will match more sites.
> **See:http://labs.bio.unc.edu/crews/twine/Twine_position_frequency_matrices.html for details.**
>
> **c) Motif from library (see step 10d).** Previously saved motifs can quickly be reloaded from libraries. If the library is large, entries can be filtered using regular expression strings, which in its simplest form is a case-sensitive text filter.

Motif matches will be displayed on the Comparison View and Conservation Views as blocks. Opacity indicates the strength of the match, relative to the upper and lower thresholds set by the user; i.e. a strong match will be opaque, while a weak match will be more transparent. Conserved matches (as dictated by the threshold slider) will have a black border around the blocks. In the Sequence View, matching sites will be boxed by colored rectangles.

5) Adjust Motif Parameters. Motif display parameters can be adjusted in the Motifs table on the right-hand side. Motifs can be displayed or hidden by toggling the checkbox under "Display". Motif names can be changed by double-clicking on their "Names". Motif colors can be changed by clicking on the color box next to each motif in the Motif table on the right. More detailed settings can be changed by clicking on the blue gear icon.

> a) IUPAC motifs. The list of consensus sequences and the number of mismatches allowed can be adjusted in the same manner as the new motif dialog. In addition, the "Drift" allowed for a motif match to be conserved can be changed (default 0 bp). For example, if the alignment isn't optimal (or compensatory loss and gain of sites removed orthologous sites in some species), motif matches may not be completely aligned, and would not be considered "conserved" in Twine. Increasing the "Drift" increases the sliding window for considering matches in orthologous

sequences to be considered conserved, even if they aren't perfectly aligned. Finally, matches can be filtered to show only conserved matches. This is especially helpful for low-stringency searches with consensus motifs that find many sites, and where phylogenetic conservation might be an indication of the important subset of all matches that are functional.

b) PWM motifs. Conserved Drift and conservation filters can be applied, and the Threshold can be adjusted "on-the-fly" to see how different thresholds change the density of motif matches. In the color chooser, select the "RGB" tab to view sliders for red, green, blue and alpha components of the motif color. The "alpha" component of RGB color controls opacity, so reducing alpha will display low-scoring matches with reduced opacity. Sliders controlling the maximum threshold (the worst score that will be totally opaque) and the minimum threshold (the worst score that will be displayed, and will be displayed with the minimum opacity/alpha) can be adjusted to change the number and opacity of matches.

Motifs are drawn on each view in the same order as listed in the Motifs panel, top to bottom; motifs can be re-arranged to adjust the order in which matches to each motif are drawn to minimize hidden matches, using the “Up” and “Down” buttons, and can be deleted by clicking the "Delete" button, or saved to a library by clicking the "Save" button (see below).

6) View motif scores. For each motif, statistics based on the observed vs. expected number of matches can be viewed by clicking "Analyze > Motif Statistics".

a) Observed matches can be altered to include or exclude completely overlapping matches (i.e. palindromes), which can inflate statistics.

b) Expected matches are calculated by a zero- to third-order Markov background chain (using a file format that can be generated from the downloadable version of the MEME package, http://meme.nbcr.net/meme/downloads.html).

c) Statistical significance of the observed number of matches (i.e. the probability of at least *n* matches being observed in random DNA) is calculated by binomial probability (Papatsenko, 2007) and Poisson probability (which approximates binomial at reasonable sequence sizes). Additional statistical techniques will eventually be incorporated (e.g. Monte Carlo).

7) Adjust Comparison View values. By default, only perfect (100%) conservation is considered "conserved". But incomplete assemblies or poorly conserved enhancers might benefit by a less stringent standard, which can be adjusted by the Threshold slider. Reducing the slider value will increase the density of conserved blocks for all sequences in the Comparison View, and will indicate more motif matches to be conserved. Scaling can also be adjusted by sliding the Zoom slider (default 1x=1bp/pixel).

8) Adjust Conservation View values. The Conservation View has three sub-views that represent the selected sequence (selected in the Comparison View by clicking on it). A zoom slider for the Conservation Views re-scales each sub-view for the selected sequence.

a) Aligned Species View. This is a scaled representation of the selected alignment, where each nucleotide is indicated by a grey block, and gaps are indicated by lines. Positions that are conserved in each sequence within conservation blocks, as displayed in the comparison view, will

be displayed as black blocks; these can be hidden by toggling the "Plot Conservation Blocks" checkbox. Motif matches to each aligned species are indicated. In this View, potential mistakes in alignment can be identified by slightly offset motif matches in individual species sequences.

b) Conservation Plot. This is a plot of the conservation level (0-100%) along the selected alignment, along with matches to the reference sequence. The line can be smoothed by increasing the "Blur", which increases the number of nucleotides of the alignment used to calculate the conservation level. The Blur will also affect the Comparison View conservation blocks for that alignment only.

c) Unaligned Species View. Similar to the Aligned Species View, but with all gaps removed. This is the "raw" View of the sequence data, which can be useful to identify possible incomplete sequences for a species, for example if one species sequence is significantly shorter than all others.

9) Analyze, hypothesize, annotate. Inspect the sequence(s) in Comparison View and Conservation view and click on a region, which will shift the Sequence View to give you a close-up look at the alignment at that location. Adjust the thresholds to try to find motif clusters in conserved regions.

10) Save/Export data. All of the Views can be exported to allow further analysis and manipulation in other programs. The graphical Views can be saved as SVG format, which is readable by Adobe Illustrator, Inkscape, and web browsers. Sub-sequences can be selected in the Aligned Species View or Conservation Plot View (click-drag a region), and saved as a FASTA file. All matches to current motifs can be exported in a tab-delimited file (Excel compatible).

a) **SVG files.** Right-click the graphical representation of interest (e.g. Aligned Species View). Select "Save SVG", and choose a file name (suggested extension: .svg). Then open this .svg file in a vector graphics program (e.g. Illustrator). **The different layers are grouped, so you need to enter isolation mode to adjust an individual component.** Alternately, you can release all groups and the clipping mask, then work with each element or navigate through the layers tab. SVG files can be used to make high-quality panels for figures.

b) **FASTA sub-sequences.** If a sub-fragment of an enhancer seems interesting (e.g. a highly-conserved region with multiple motif matches), this sub-region can be saved as a separate alignment. In the Aligned Species View, click and drag across the region of interest, right-click, and select "Save selection as FASTA." You can then open this new FASTA file in Twine, and all motifs will automatically be loaded.

c) **Export matches.** If a list of all locations of the current set of matches would be useful, click File > Export matches, then select a file name (suggested extension: .tab). The file will contain a list of all matches. Each line will contain the alignment name, species name, sequence match, position, orientation, motif match length, motif name, type, and conservation status of the match (as a True or False value).

d) **Motif Libraries.** A new motif library can be made by selecting Analyze > Motif Libraries > New Library. Motifs that are open in the current analysis can be saved by selecting a motif, then clicking "save" in the Motifs panel. Library motifs from can be added to the current analysis by

selecting Analyze > Motif Libraries > Motif From Library, then choosing one or more motifs from a library using the checkbox. Settings for each motif can be viewed (but not edited) by clicking the Settings icon. Library motifs can be edited by adding a motif from a library, changing parameters, then saving it back to the library. Motifs can be deleted from a library by clicking the red 'X' icon. Changes to motif libraries are not saved until Twine is closed, at which point you will be prompted to save changes. Backup versions of motif libraries are created when you save, so you can recover motifs that you may have inadvertently deleted.

e) **Save Twine Analyses.** A file containing the set of open motifs and alignments can be saved by clicking File > Save Twine Analysis, and re-opened by clicking File > Open Twine Analysis.

11) Develop plugins. Using the Java Simple Plugin Framework, Twine will recognize plugins that can extend the functionality of Twine. See the appendix for a more detailed description. Plugins should accept an array of the currently opened alignments (as AlignedSequence objects), then return zero or more AlignedSequence objects to Twine, which will be added to the alignments. In theory, this can be used to run command-line programs. Examples of Java-based plugins ("RevComp" and several motif library importers) and a front-end for a Windows command-line program ("WinDotter") are included, as well as the source for a "skeletal" plugin that can serve as the template for user plugins. As I generate more plugins, I will make them available at the Twine web site. **The WinDotter example will not work unless you have WinDotter installed in the right location (or change the source and re-package as a .jar file).**

Appendix

FASTA alignments from UCSC Genome Browser

Generating your own alignment by identifying orthologous sequences from the latest genome/contig assemblies (or even individual traces), then aligning them using various multiple alignment programs (e.g. Dialign) is best. But, it's much easier to download pre-computed alignments. The UCSC Genome Browser (genome.ucsc.edu) contains whole-genome alignments using MultiZ for many popular model organisms. They're in MAF format, so they need to be converted into FASTA in order for Twine to recognize them.

1. Go to genome.ucsc.edu
2. Click "Genome Browser".
3. Select species of interest, and region of interest (e.g. gene, or specific genome coordinates).
4. Zoom in/out to encompass the entire region of interest in the browser window.
5. At the top, click "Tools > Table Browser".
6. Under "group", select "Comparative Genomics".
7. Under "table", select "multiz*n*way" (*n* is the number of species used for this organism).
8. Under "region", select "position" (not "genome").
9. Select "Send output to Galaxy."
10. Click "get output". Your sequence will be forwarded to the Galaxy web site, which contains a suite of applications for analyzing and manipulating data.
11. Click "Convert Formats" (on the left), then "MAF to FASTA". The MultiZ alignments were in MAF format, so they need to be converted to FASTA.
12. Under "MAF file to convert:", select your sequence.
13. Under "Type of FASTA Output", select "One Sequence per Species."
14. Select the species to extract (checkboxes), then "Execute."
15. When the conversion is complete, it will appear on the right, ready to download!

Creating/Analyzing Mutation Series Figures

Twine is designed to help identify important regulatory sequences, which are commonly tested by introducing combinations of mutations to the enhancer, commonly called "Enhancer Bashing". Twine can also be used both to double-check mutation constructs and to generate schematics of mutation series experiments. For an example alignment, open "link-5'-sitemutsAlignment.fas", which contains the mutation series from Pearson et al., Dev. Biol. 366, 420-432 (2012).

A FASTA alignment of the wild-type enhancer to all site variants (site mutations or deletions) can be loaded into Twine. The Comparison View will show gaps in conservation blocks at any place where a mutation is present in at least one variant. This is especially valuable when checking sequenced clones for unintended changes. Consensus motifs can then be added, revealing the presence of each site in the set of site variants. The Aligned Species View can then be saved as an SVG file, to be used in schematics and figure panels.

Creating Plugins for Twine

Hopefully, Twine is useful enough for analysis of enhancers that you will want to have an easy way to export alignments to perform a modification (alignment, motif analysis, find primers), then import it immediately back into Twine to see the results. Plugins are generated using the Java Simple Plugin Framework (JSPF), which uses a simple annotation-based system for loading and keeping track of plugins. Once you've written your custom plugin, package it in a jar and place it in the plugins folder, and it should be seen the next time you run Twine.

Right now, all plugin implementations implement AlignedSequencePlugin, which is part of the Twine package. You can see the source in the Twine JAR file, but there are only two methods that you need to override.

- getPluginName, which returns a string, the unique identifier for your plugin, which will be displayed in the menu under "Resources."
- manipAlignedSequence, which receives an array of all currently opened AlignedSequence objects, and returns any new (or modified) AlignedSequence objects. This is where your plugin implementation gets the AlignedSequence (and motifs).

You will need to import the following, which require the JSPF core jar and Twine as libraries. You do not need to include these libraries in your actual build, because they will already be included in the Twine distribution.

- AlignedSequencePlugin.AlignedSequencePlugin;
- net.xeoh.plugins.base.annotations.Capabilities;
- net.xeoh.plugins.base.annotations.PluginImplementation;
- twinerebuild.AlignedSequence;
- twinerebuild.Motif;
- Your implementation should have the annotation @PluginImplementation just above the class declaration, and your class needs to implement AlignedSequencePlugin.

You will need to specify a String variable "pluginName", and have it returned to AlignedSequencePlugin via getPluginName(). You can use the code in the examples. Same with capabilities(), which is used to find the particular plugin implementation chosen among AlignedSequencePlugins.

You will need to override manipAlignedSequence to get an array of AlignedSequence objects, do something to one or more of those AlignedSequence objects, then return modified AlignedSequence objects; a **null** object can be returned if you don't want to add an AlignedSequence to the set of opened alignmentss.

You can either perform the manipulation on all AlignedSequences, generate a popup window to have a user select a subset of open AlignedSequences, or find the selected AlignedSequence by calling the static AlignedSequence method "getSelectedAlignedSequence (AlignedSequence[] allAlignedSequences)", passing an array of AlignedSequence objects, getting the selected AlignedSequence in return.