Squiggle Documentation

Benjamin Lee

Nov 18, 2018

Contents

Squiggle is a two-dimensional DNA sequence visualization library that can turn FASTA sequence files like this:

```
>lcl|NC_000011.10_cds_NP_000509.1_1 [gene=HBB]
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAA
```

into gorgeous, interactive visualizations like this:

CHAPTER 1

Installation

If you don't have Python 3.4 or greater installed, be sure to get it. To get the current stable version of Squiggle, run:

```
$ pip install squiggle
```

Or, alternatively, if you want to get the latest development version:

```
$ pip install git+https://github.com/Lab41/squiggle.git
```

CHAPTER 2

Usage

Using Squiggle is as easy as:

```
$ squiggle your_sequence.fasta
```

Squiggle has tons of options available to make beautiful, interactive visualizations of DNA sequences. To get a full rundown of the various option, take a look at the *User Guide*.

CHAPTER 3

Citation

To be determined!

CHAPTER 4

Table of Contents

4.1 Visualization Methods

There are a variety of ways to visualize DNA sequences in two dimensions. Squiggle provides its own novel visualization method as well as implementations of various other methods. Each method captures a different aspect of a sequence, so it is highly recommended to try using multiple methods in order to get a feel for a sequence.

4.1.1 Squiggle

Squiggle's DNA visualization method is based on the UCSC .2bit format and the Qi et. al Huffman coding method. In essence, a DNA sequence is first converted into binary using the 2bit encoding scheme that maps T to 00, C to 01, A to 10, and G to 11. For example:

```
ATGC
```

becomes:

```
10001101
```

Then, starting at the origin, for each bit, the following vectors are layed end to end:

This mapping has the effect of giving each nucleotide a distinctive shape:

This encoding method has several handy features:

- Based on an open, common bioinformatics format.
- No degeneracy in the encoding (an encoding can only map to one sequence and vice versa).
- The overall GC-content can be inferred from at a glance based on whether the endpoint of the graph is above or below zero.

- Regions inside the gene with varying GC-content can be seen as peaks and valleys.
- Is limited to quadrants I and IV and is a function
- The x-axis corresponds directly with nucleotide position
- Supports ambiguous nucleotides (which are displayed as horizontal lines)

For an example, let's look at the human β-globin gene using the squiggle method:

```
$ squiggle example_seqs/human_HBB.fasta
```

4.1.2 Gates

In Gates's method, DNA sequences are converted into 2D walks in which Ts, As, Cs, and Gs are up, down, left, and right, respectively. This gives each sequence a "shape." However, there is degeneracy, meaning that a visualization is not necessarily unique. For example, `TGAC` is a square (up, right, down, and left), but so is `GTCA` (right, up, left, down).

To see an example of Gate's method, we'll again look at human β-globin:

```
$ squiggle example_seqs/human_HBB.fasta --method=gates
```

4.1.3 Yau

Yau et. al's method uses unit vectors with upward vectors indicating pyrimidine bases (C and T) and downward vectors indicating purine bases (A and G). Similar to Squiggle, this method has no degeneracy.

Specifically,

$A \to \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right), T \to \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right), G \to \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right), C \to \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right).$

Warning: The x-coordinate in Yau's method is not equivalent to base position.

It produces a visualization of β-globin like this:

```
$ squiggle example_seqs/human_HBB.fasta --method=yau
```

4.1.4 Yau-BP

Unique to Squiggle is the Yau-BP method, a slight modification of Yau's method that ensures that the x axis is equivalent to the base position. It preserves that salient feature of the method, which is the purine/pyrimidine split.

4.1.5 Randić and Qi

Randić et al. and Qi and Qi's methods are similar to tablature, with a different base (or 2-mer in the case of Qi's method) assigned to each y value. The best way visualize it is through an example.

Let's look at the Randić visualization of `GATC`:

Look's pretty good. However, this visualization method isn't well suited to long sequences, as we'll see when we look at β-globin:

```
$ squiggle example_seqs/human_HBB.fasta --method=randic
```

Qi's method produces very similar results, just with a much larger range of y values:

```
$ squiggle example_seqs/human_HBB.fasta --method=qi
```

4.2 User Guide

Squiggle is designed to be easy to use while still providing complete flexibility to the user. For the sake of demonstration, we'll be using four different species' β-globin genes (human, chimpanzee, rhesus macaque, and Norway rat).

For a full list of the command line options and their meanings, see the *CLI Reference*.

4.2.1 Basic Usage

The easiest way to visualize a sequence is by passing a FASTA file to Squiggle:

```
$ squiggle human_HBB.fasta
```

To use a different visualization method, provide `--method` with a setting (see *Visualization Methods* for a description of the supported methods):

```
$ squiggle human_HBB.fasta --method=gates
```

4.2.2 Plotting Multiple Sequences

If your FASTA file has multiple sequences, they will get plotted together automatically. If, however, your sequences are in separate files, you can still plot them together by passing multiple files to Squiggle:

```
$ squiggle human_HBB.fasta chimpanzee_HBB.fasta norway_rat_HBB.fasta rhesus_HBB.fasta
```

To put them on separate plots, use the `--separate` flag:

```
$ squiggle human_HBB.fasta chimpanzee_HBB.fasta --separate
```

By default, their x axes are linked. This can be disabled with `--no-link-x` (try if for yourself by panning around):

```
$ squiggle human_HBB.fasta chimpanzee_HBB.fasta --separate --no-link-x
```

Similarly, the y axes can be linked and unlinked with `--link-y`/`--no-link-y`.

Note that when plotting seperately, Squiggle will try to make the layout as square as possible. If you want to specify the number of columns, you can do so with the `-c` option.

If you want to compare FASTA files, you can use the `--mode=file` flag to treat each file as a separate entity, as opposed to each sequence. The `--mode=auto` flag (which is the default) will attempt to visualize each sequence independently unless there are too many, in which case it will switch to file mode.

As an example, let's compare the highly expressed genes of *E. coli* and *B. anthracis*:

```
$ squiggle ecol.heg.fasta banth1.heg.fasta
```

Also, be aware that the `--hide` flag will make it so that clicking on the name of a sequence or file in the legend will hide it for easier comparisons.

Note: Using the `--hide` flag may result in a significant slowdown when visualizing a large number of sequences.

Finally, the palette can be controlled by the `-p` option. Valid palettes can be seen here:

```
$ squiggle human_HBB.fasta chimpanzee_HBB.fasta -p Accent
```

4.2.3 Controlling Output

If you don't want to show your plot in a browser but would rather save it for later, you can do so with the `-o` option:

```
$ squiggle human_HBB.fasta -o output.html
```

If you don't have an internet connection, you can still use Squiggle by telling it to include the full Bokeh plotting library in its output with `--offline`:

```
$ squiggle human_HBB.fasta --offline
```

Warning: This will signifcantly increase the size of your output file.

To adjust the dimensions of your output, use the `-d` option, providing the width and height (in that order):

```
$ squiggle human_HBB.fasta -d 650 150
```

By default, Squiggle titles the plot with the name of the sequence, as determined by the FASTA file. If you want to override it, you can manually provide the title:

```
$ squiggle human_HBB.fasta -t β-globin
```

If applicable, you can also specify the location of the legend using the `--legend-loc` flag. The default setting is to put the legend in the top left.

4.3 API Reference

Squiggle is built around a single function, documented fully below:

`squiggle.`**`transform`**`(`*sequence*, *method='squiggle'*, *bar=False*`)`

Transforms a DNA sequence into a series of coordinates for 2D visualization.

Parameters

- **sequence** (*str*) – The DNA sequence to transform.
- **method** (*str*) – The method by which to transform the sequence. Defaults to "squiggle". Valid options are `squiggle`, `gates`, `yau`, `randic` and `qi`.
- **bar** (*bool*) – Whether to display a progress bar. Defaults to false.

Returns A tuple containing two lists: one for the x coordinates and one for the y coordinates.

Return type tuple

Example

```
>>> transform("ATGC")
([0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0], [0, 0.5, 0, -0.5, -1, -0.5, 0, -0.
↪5, 0])
>>> transform("ATGC", method="gates")
([0, 0, 0, 1, 0], [0, -1, 0, 0, 0])
>>> transform("ATGC", method="yau")
([0, 0.5, 1.0, 1.8660254037844386, 2.732050807568877], [0, -0.8660254037844386, 0.
↪0, -0.5, 0.0])
>>> transform("ATGC", method="yau-bp")
([0, 1, 2, 3, 4], [0, -1, 0, -0.5, 0.0])
>>> transform("ATGC", method="randic")
([0, 1, 2, 3], [3, 2, 1, 0])
>>> transform("ATGC", method="qi")
([0, 1, 2], [8, 7, 11])
```

Warning: The entire sequence must be able to fit in memory.

Raises `ValueError` – When an invalid character is in the sequence.

4.4 CLI Reference

Squiggle's CLI is documented fully below. Note that this reference is available at any time by invoking the `squiggle --help` command.

```
$ squiggle --help
Usage: squiggle [OPTIONS] [FASTA]...

Options:
  -w, --width FLOAT                The width of the line. Defaults to 1.
  -p, --palette TEXT               Which color palette to use. Choose from boke
                                   h.pydata.org/en/latest/docs/reference/palett
                                   es.html. Defaults to Category20.
  --color / --no-color             Whether to plot the visualizations in color.
  --hide / --no-hide               Whether to hide sequences when clicked in
                                   the legend. Defaults to false if plotting
                                   one sequence and true if plotting multiple.
  --bar / --no-bar                 Whether to show a progress bar when
                                   processing multiple sequences. Defaults to
                                   true.
  -t, --title TEXT                 A title to display when plotting sequences
                                   together.
  --separate                       Whether to plot the visualizations
                                   separately.
  -c, --cols INTEGER               The number of columns when plotting
```

(continues on next page)

(continued from previous page)

```
                                  separately. Defaults to a the closest to
                                  square layout as possible.
  --link-x / --no-link-x          Whether to link the x axes for separate
                                  plotting. Defaults to true.
  --link-y / --no-link-y          Whether to link the y axes for separate
                                  plotting. Defaults to false.
  -o, --output FILE               The output file for the visualization. If
                                  not provided, will open visualization in
                                  browser. The filetype must be .html
  --offline                       Whether to include the resources needed to
                                  plot offline when outputting to file.
                                  Defaults to false.
  --method [squiggle|gates|yau|yau-bp|randic|qi]
                                  The visualization method.
  -d, --dimensions WIDTH HEIGHT   The width and height of the plot,
                                  respectively. If not provided, will default
                                  to 750x500.
  --skip / --no-skip              Whether to skip any warnings. Defaults to
                                  false.
  --mode [seq|file|auto]          Whether to treat each sequence or file as
                                  the individual object. Defaults to automatic
                                  selection.
  --legend-loc [top_left|top_center|top_right|center_right|bottom_right|bottom_
↪center|bottom_left|center_left|center]
                                  Where to put the legend, if applicable.
                                  Defaults to top left.
  -h, --help                      Show this message and exit.
```

4.5 License

This software is licensed under the MIT License. If you're unfamiliar with software licenses, here is a handy summary of the license.

For reference, the license is reproduced below:

(continues on next page)

(continued from previous page)

```
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

CHAPTER 5

Indices and tables

- genindex
- modindex
- search

Python Module Index

Index