ortho_seqs

Release 1.0.1

Saba Nafees

Mar 01, 2023

CONTENTS

ortho_seqs

CHAPTER

ONE

INTRODUCTION:

This documentation accompanies the ortho_seqs python command line tool that computes multivariate tensor-based orthogonal polynomials based on DNA, RNA or protein sequence data and maps corresponding phenotypic information onto the sequence space.

CHAPTER

TWO

GUIDE

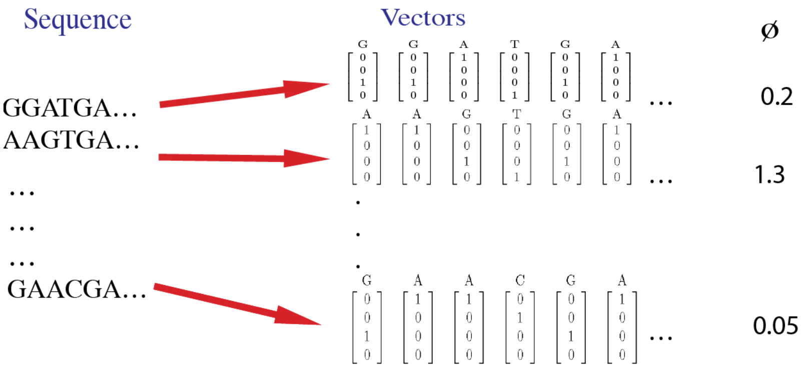
2.1 Background & Quickstart

2.1.1 ortho_seqs

Ortho_seqs is a command line tool that implements a mathematical approach to convert sequence data (DNA/protein) to multivariate tensor-valued orthogonal polynomials and project phenotypes onto the polynomial space. We are currently working to update this mathematical approach and will post these updates soon.

We do this by first converting the sequence information into 4-dimensional (for DNA) or 20-dimensional (for amino acids) vectors. The method can also be used for padded sequences to deal with unequal sequence lengths. Find out more about the original approach in this paper Analyzing genomic data using tensor-based orthogonal polynomials with application to synthetic RNAs. The paper gives an example of this method as applied to a case of synthetic RNA from a previously published dataset.

For example, the sample data inputs for this tool are shown in this image. Here, each site in a sequence is first converted to a 4-dimensional vector. The input data includes phenotype values for each sequence.



2.1.2 Documentation

To view documentation and detailed tutorials for *ortho_seqs*, click here.

2.1.3 Usage

First, install an environment with dependencies for this package:

```
conda create -n ortho_seqs pip
conda activate ortho_seqs
pip install -r requirements.txt
```

or

```
conda env create -f conda_environment.yml
conda activate ortho_seq
```

Then, install the package:

```
python setup.py install
```

Gather the input file(s) needed.

There are three main ways to submit your sequence and phenotype files to *ortho_seqs*. The first method is to submit them separately, in their own .txt files. Recently, however, an update was added that allows you to submit them both in the same file. For this to apply: 1) The file must be either a .xlsx or a .csv file. 2) The sequences must be in the first column, and the phenotypes must be in the second column. 3) The columns must not have header names.

If you use a single file for the sequence and phenotype, you would submit the file path where you would submit the sequence file path, and do not include the *–pheno_file* flag. **note: the GUI does not support single-file uploads yet.**

The phenotypes must be real numbers.

Then, to run the commandline tool:

To start with a test example, you can run the sample command below:

```
ortho_seq orthogonal-polynomial ./ortho_seq_code/tests/data/nucleotide/first_order/test_
↪seqs_2sites_dna.txt --molecule DNA --pheno_file ./ortho_seq_code/tests/data/nucleotide/
↪first_order/trait_test_seqs_2sites_dna.txt --poly_order second --out_dir ../results_
↪ortho_seq_testing/DNA_2sites_test_run/
```

The above sample command line is building the tensor-valued orthogonal polynomial space based on the sequence data which consists of 12 sequences, each with two sites. Since these are DNA sequences, the vectors are 4-dimensional. These used to be flags for sites, dimensions, and population size, but new functionality will automatically calculate these. Corresponding to each sequence is a phenotype value (a real number) as given in the phenotype file. For DNA, the tool can run first and second order analyses currently. We'll implement third order in a future version. For amino acids, the current version supports first order analysis and we hope to expand this in the future.

Amino acids/nucleotides that do not appear in any sequence will be removed from the alphabet when the letters are being converted to first order vectors. For example, if the residue 'R' (Arginine) never occurs in the sequence dataset, the first order vectors will now have 19 dimensions (instead of 20) and 20 dimensions (instead of 21) if the sequences are padded with 'n'. This is done to greatly reduce runtime for larger sequence datasets and for longer sequences. When the program will run, it will return this sentence:

```
Will be computing p sequences with s sites, and each vector will be d-dimensional.
```

Where p represents the population size (number of rows in sequence file), s represents the number of sites, and d represents the number of amino acids/nucleotides detected in the sequence file (adds on 1 for lowercase n's). For the above example, the program will return

```
Will be computing 12 sequences with 2 sites, and each vector will be 4-dimensional.
```

Along with regressions on each site independent of one another and onto two sites at a time, the above command also computes *Fest* which is the phenotype estimated by the regressions. This shows that the mathematical calculations are done correctly as we now have an equation that accurately captures our initial data points. This only works here for sequences with 2 sites. If we had more sites, we'd need to do higher order calculations in order to capture all our combinations. Therefore, when running the tool with more sites, as will probably be the case for most users, even just going up to second order gives us useful information about our system. First order tells us the importance of each site (independent of any correlations it might have with another site) and second order tells the importance of pairs of nucleotides independent of other pairs. Please take a look at the paper linked above to learn more about this method.

Flags & Functionality

```
--pheno_file
```

Input a file with phenotype values corresponding to each sequence in the sequence file. If you have a .xlsx or .csv file, do NOT use this flag (more details above in the ***Gather the Input Files Needed*** section).

```
   --molecule

Currently, you can provide DNA or protein sequences. Here, you can also provide␣
↪sequences of unequal lengths, where sequences will be padded with lowercase 'n's until␣
↪it has reached the length of the longest sequence.
```

```
--poly_order
```

The order of the polynomials that will be constructed. Currently, one can do first and second order for DNA and first order for protein.

```
--out_dir
```

Directory where results can be stored.

```
--precomputed
```

Let's say you have a case where you have the same set of sequences but two different corresponding sets of phenotypes. You can build your sequence space and then project the first set of phenotypes onto this space. Then, if you wish to see how the other set of phenotypes maps onto the same sequence space, you can use this flag so that you're not wasting time and memory to recompute the space. When doing this, be sure to add your results from the first run to the **out_dir** when rerunning the command with the **precomputed** flag.

```
--alphbt_input
```

Used to group amino acids/nucleotides together, or specify certain amino acids/nucleotides. If you don't want to group anything, don't include this flag when running *ortho_seqs*. For example, putting *ASGR* for a protein molecule will tell the program to have 6 dimensions: one for each amino acid specified, and one for *z*, where every unspecified amino acid or nucleotide will be converted to *z*, and one for *n* (whenever sequences have unequal lengths, *ortho_seqs* will pad the shorter sequences with *n* at the end). You can also comma-separate amino acids/nucleotides to group them. For example, putting *AS,GR* will make the vectors 4-dimensional, one for *AS*, one for *GR*, one for every other amino acid (*z*), and one for *n*.

There are also built-in groups:

protein_pnp will group by polar and non-polar amino acids, every other amino acid, and *n*.

essential groups by essential and non-essential amino acids, every other amino acid, and *n*. Group 1: Essential - ILVFWHKTM Group 2: Non-Essential - Everything else Group 3: n (Source: https://www.ncbi.nlm.nih.gov/books/NBK557845/)

alberts groups by categories set by Alberts. Group 1: Basic - KRH Group 2: Acidic - DE Group 3: AVLIPFMWGC Group 4: Everything else Group 5: n (Source: https://www.ncbi.nlm.nih.gov/books/NBK21054/)

sigma groups by categories set by Sigma. Group 1: Aliphatic - AILMV Group 2: Aromatic - FYV Group 3: Polar Neutral - NQCST Group 4: Acidic - KRH Group 5: Basic - DE Group 6: Other - G Group 7: Other - P Group 8: n (Source: https://www.sigmaaldrich.com/US/en/technical-documents/technical-article/protein-biology/protein-structural-analysis/amino-acid-reference-chart)

hbond groups by strength of hydrogen bond attractions. Group 1: Can Make Hydrogen Bonds - NQSTDERKYHW Group 2: Can Not Make Hydrogen Bonds - Everything else Group 3: n The first group is able to make hydrogen bonds, whereas the second group is not.

hydrophobicity groups by hydrophobicity. Group 1: Very Hydrophobic - LIFWVM Group 2: Hydrophobic - CYA Group 3: Neutral - TEGSQD Group 4: Hydrophilic - Everything else Group 4: n The first group is very hydrophobic, the second group is slightly hydrophobic, the third group is neutral, and the last group is hydrophilic.

```
--min_pct
```

When ortho_seqs is run, a .csv file of covariances will be saved in the specified path. This matrix of covariances is one of the main results of the program (as shown in {sequence_file_name}.npz output below). The csv file will contain the covariance of each nucleotide at each site with another nucleotide at another site (or amino acids at each site). Suppose there are 5 covariance values of 2, 1, 0, 0, -1. For the percentiles, all unique *magnitudes* will be considered when assigning covariances, which will be 2, 1, and 0. 0 will be the 0th percentile (therefore, assigning 0 to the *–min_pct* flag will return every covariance), 1 (and -1) will be 33.33…, and 2 will be 66.66… Specifying 50 as *–min_pct* will only return the row with the covariance of 2, since only 66.6…>50. The min_pct flag is short for minimum percentile, which will remove any covariances from the .csv file that are below the given percentile. The default value is 75.

```
--pheno_name
```

Let's say you know that your phenotype values represent IC50 values. You could then add *–pheno_name IC50* as a flag, and on the rFon1D plot that is automatically generated, the y-axis label will include IC50. Default is **None**.

2.1.4 Results & Outputs

Generating logo plot

Before even running the tool, we can generate a logo plot to visualize the different nucleic/amino acids in the sequence dataset. This is implemented as a command line function.

Refer to the *logo-plot* tutorial on the ReadTheDocs for more information on how to generate this.

Running ortho_seqs

The tool will provide updates as the run is progressing regarding which parts of the calculations are done being computed. For example, when the mean is computed, it'll say "computed mean". All the different elements that it is computing are different parts of building the multivariate tensor-valued orthogonal polynomial space based on the sequence information. To get a general idea of what the calculations mean, please refer to the supplementary methods in the paper linked above. The program will save outputs in npz format. See below for what is stored.

```
{sequence_file_name}.npz
```

This will store the calculations that went into constructing the polynomial space. This also includes information about the statics of our sequence space, such as mean, variance and a matrix of covariances. See figures 4 and for ideas on how mean and the matrix of covariances can be visualized. All of these calculations go into building the orthogonal polynomial space based on sequence information and at this point of the program, we have not connected the phenotype (the functional variable) with the sequence information.

```
{sequence_file_name}_covs_with_F.npz
```

This will store the covariance of the phenotype (or trait) with the polynomials. This is when we start connecting the phenotype with the sequence space.

```
{trait_file_name}_Fm.npz
```

This contains the mean trait value. This is a scalar.

```
{trait_file_name}_regressions.npz
```

This set of files contains the main results which includes the following:

1. **rFon1D**: This is the regression of the trait onto the first order conditional polynomial orthogonalized within. This tells us the regression of the phenotype onto each site and onto each nucleotide (or amino acid) at that site independent of any correlations that site might have with other sites. For the case of nucleotides, this can be visualized as bar plots as shown in Figure 6 in the paper linked above.
2. **rFon2D**: This gives 4 matrices which give the regression of the pheonotype onto (site1)x(site1), (site 1)x(site 2), (site 2)x(site 1) and (site 2)x(site 2), in that order. The second matrix here is the important one and it is the same as rFon12. See description of rFon12.
3. **rFon12**: This is the regression of the trait onto *pairs* of sites for given nucleotides at each site. These are regressions on (site 1)x(site 2) independent of first order associations. Since we're looking at 2 sites at a time and there's a possibility of having 4 nucleotides at each site (for the case of DNA), we can visualize this via a 4x4 matrix as shown in Figure 8 in the paper linked above.

```
cli_output.txt
```

Everything that prints out on the CLI, when running *orthogonal_polynomial*, will be saved to this document, in the defined out_dir.

2.1.5 The rf1d class

The newest update to *ortho_seqs* involves adding a new class of objects, called *rf1d* (short for rFon1D). To run *rf1d*, use the CLI, and type in *rf1d-viz* like you would *orthogonal_polynomial* when running *ortho_seqs*.

Note: *rf1d-viz* requires you to have run *orthogonal_polynomial* beforehand.

```
--filename
```

This is the same as the *{trait_file_name}_regressions.npz* file that is returned from *ortho_seqs*, as it contains the rFon1D values that are used.

```
--alphbt_input
```

Similarly to *orthogonal_polynomial*, this flag takes in a comma-separated list of the groupings (**Note:** this list must be comma-separated for the code to work). *orthogonal_polynomial* will print out the *rf1d form of alphabet input* in the CLI before any mathematical calculations are made, which will work if you choose to copy/paste it.

```
--molecule
```

Identical to how it is in *orthogonal_polynomial*. It doesn't matter much what you put here, as this is purely for visual purposes only.

```
--phenotype
```

Identical to how it is in *orthogonal_polynomial*. This is used as the y-axis labeling for the barplot.

```
--out_dir
```

The path where you want the visualizations saved, if applicable.

```
--action
```

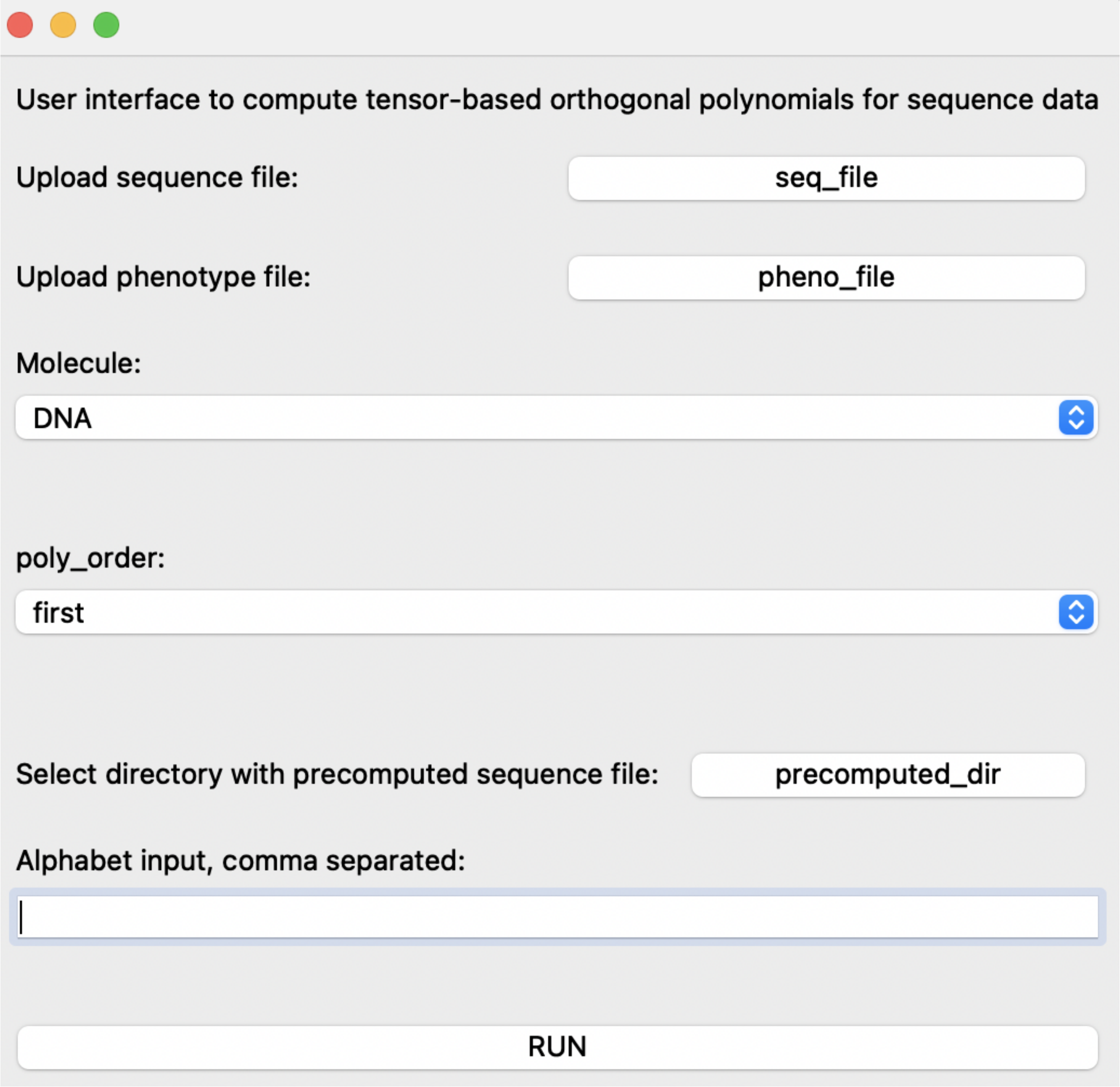
This flag is where you specify which visualization you want.

Options:

1. *barplot* - Prints and saves a barplot of the rFon1D values, and saves it, if an *out_dir* is specified. This is what is called in *orthogonal_polynomial* at the end.
2. *density* - Prints a histogram plot of the rFon1D values, and saves it, if an *out_dir* is specified.
3. *summary* - Prints out the number of sites and dimensions, the alphabet input, the molecule, and calls *sort* (explained in further detail below). This is called in *orthogonal_polynomial* automatically, at the very end of the program. This will be saved to the *out_dir* as *summary.txt* **Note:** there won't be a separate *sort.txt* file created.
4. *heatmap* - Prints a heatmap of the rFon1D values, and saves it, if an *out_dir* is specified.
5. *boxplot* - Prints a boxplot of the rFon1D values, and saves it, if an *out_dir* is specified.
6. *sort* - This will print out the top 10 rFon1D values *by magnitude*, including the rFon1D value, the site, and the group it belongs to. This will be saved to the *out_dir* as *sort.txt*.

2.1.6 To run the GUI

A GUI version of the CLI is also available to make it easier for users to utilize the tool. The GUI allows the user to upload the sequence and phenotype information via an upload button, specify the molecule, the polynomial order they wish to run, and provide the path to the directory which contains precomputed sequence space if the user wishes to project a different phenotype onto the same space (i.e., given same sequence data but different corresponding phenotypes). The GUI is in its early form and will include further updates resembling the cli in future versions.



User interface to compute tensor-based orthogonal polynomials for sequence data

Upload sequence file: seq_file

Upload phenotype file: pheno_file

Molecule:

DNA

poly_order:

first

Select directory with precomputed sequence file: precomputed_dir

Alphabet input, comma separated:

To run the gui, open a terminal and make sure you're in the ortho_seqs environment just as you would do if you were running the cli (see above). Then type in the following:

```
ortho_seq gui
```

This will pull up the gui window and allow you to input the relevant information.

```
cov_hist_{trait_file_name}.png
```

This is a histogram of all non-zero covariances. Its bin width is 0.5.

```
cov_data_frame_{trait_file_name}.csv
```

This file is a csv file of covariances between every item at every site. This includes the item ID and site for both items in the pair used to calculate the covariance, the covariance value, the covariance magnitude, and an ID for the pair (s1-g2,s3-g4 represents the pairing of an element from the first group in the alphabet at the second site, and an element from the third group at the fourth site). The sites are one-indexed, meaning a value of 3 for the First Site or Second Site column corresponds to site number three along the sequence.

```
rFon1D_graph_{trait_file_name}.png
```

This is a bar plot of all nonzero rFon1D values of every item at every site.

2.1.7 Support

If you have specific or general questions, feel free to open an issue and we'll do our best to address them. If you have any comments, suggestions or would like to chat about this method or similar ideas, feel free to reach out via email at saba.nafees314@gmail.com.

2.1.8 Roadmap

We hope to implement third order analysis for DNA in the near future. For amino acids, we hope to implement second order analysis. We'll add visualization ideas soon but if you have any thoughts on this, please feel free to reach out.

2.1.9 Contribution

We hope to make the tool run faster as with higher dimenions and higher order analysis of longer sequence data, we can run into memory and time issues. Any thoughts on this or visualization are welcome.

2.1.10 Authors and acknowledgements

The derivation of the method and the construction of an initial version of the program was done by Dr. Sean Rice who served as Saba's Ph.D. advisor. Thank you to Isaac Griswold-Steiner for helping write the function to compute generalized inner and outer products. Thank you to Pranathi Vemuri for helping with the very initial draft of the CLI, adding CI integration testing, and to Phoenix Logan for helping write unit-tests. Thank you to AhmetCan for helping initiatie the first GUI version. Thank you, Aaron, for always being ready to review PRs and for your insights/help in the development process. Thank you to Vijayanta Jain and Saugato Rahman Dhruba for being the guinea pigs and running lots of sample commands, discussing the mathematics with me, and for their ideas on visualizations. Their efforts are deeply appreciated!

2.1.11 License

MIT

2.2 Tutorial: Running a sample dataset (protein sequences)

This document will walk you through the steps of how to run a dataset on ortho_seqs, and what the various outputs are. This tutorial uses protein sequence data from *The Intrinsic Contributions of Tyrosine, Serine, Glycine, and Arginine to the Affinity and Specificity of Antibodies* by Birtalan & Sidhu et al., 2008.

2.2.1 1. Setting Up Your Computer to Run ortho_seqs

The first thing you have to do (aside from gathering data!) is set up your computer to run ortho_seqs.

> You first need to have Miniconda installed on your computer, in order to do the shell commands. To do so, follow the link here, and choose the appropriate version, with regards to your computer.

After you have installed Minoconda, open up Terminal, or an equivalent Command-Line Interface (CLI). Run either this:

```
conda create -n ortho_seqs pip
conda activate ortho_seqs
pip install -r requirements.txt
```

Or, alternatively:

```
conda env create -f conda_environment.yml
conda activate ortho_seqs
```

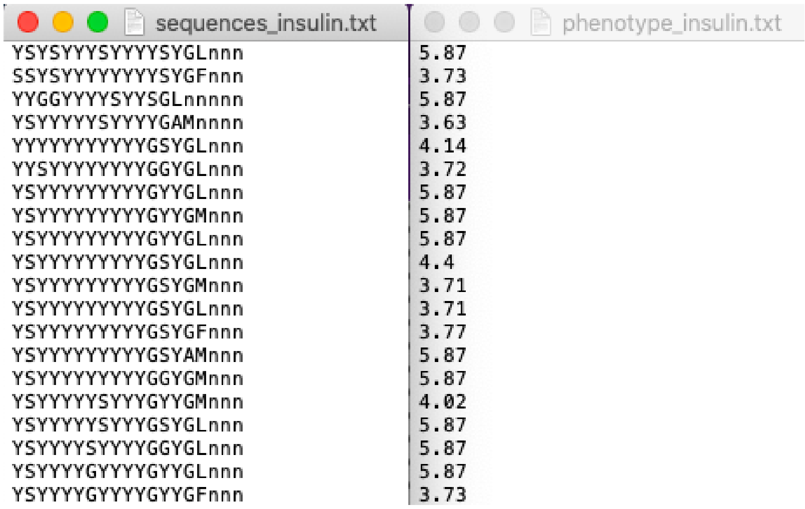
To activate ortho_seqs on your device. You will also need to run:

```
conda install openpyxl
python setup.py install
```

This line must be run every time ortho_seqs is updated, so you are using the most recent version. If the above steps have worked, congrats! You now have ortho_seqs on your computer. It's time to input some data.

2.2.2 2. Your dataset

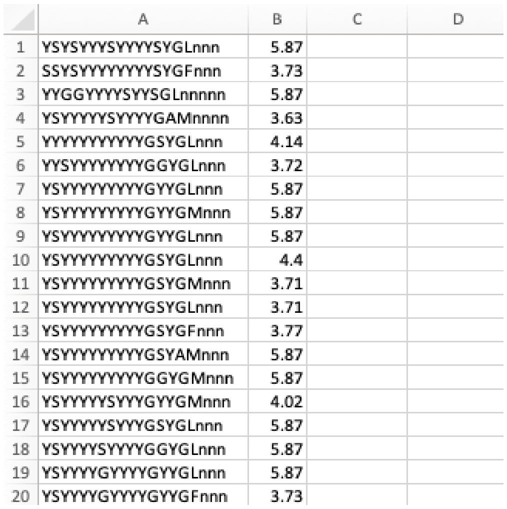
The data that is input to ortho_seqs must include a column of sequences, and a column of their corresponding phenotype values. These two columns can either be separate .txt files, or a single .xlsx or .csv file. Take, for instance, our toy example, which is a dataset originating from a paper titled The Intrinsic Contributions of Tyrosine, Serine, Glycine, and Arginine to the Affinity and Specificity of Antibodies by Sidhu et al. In this work, the authors constructed synthetic antibody Fab libraries to measure the impact of four different amino acids, Tyr, Ser, Gly and Arg on antigen recognition. Affinity and specificity data for these antigen-binding Fabs is provided for 3 different antigens (insulin, VEGF, and HER2). For this tutorial, we look at CDRH3 sequences of Fabs binding to insulin, along with corresponding phenotypes which is given by Specificity ELISA Signal Optical Density (see Figure 4a in the paper). This will be referred to as the "Sidhu dataset" for this tutorial. The dataset, when input into ortho_seqs, should look like



```
YSYSYYYSYYYYSYGLnnn
SSYSYYYYYYYYSYGFnnn
YYGGYYYYSYYSGLnnnnn
YSYYYYYSYYYYGAMnnnn
YYYYYYYYYYYGSYGLnnn
YYSYYYYYYYYGGYGLnnn
YSYYYYYYYYYGYYGLnnn
YSYYYYYYYYYGYYGMnnn
YSYYYYYYYYYGYYGLnnn
YSYYYYYYYYYGSYGLnnn
YSYYYYYYYYYGSYGMnnn
YSYYYYYYYYYGSYGLnnn
YSYYYYYYYYYGSYGFnnn
YSYYYYYYYYYGSYAMnnn
YSYYYYYYYYYGGYGMnnn
YSYYYYYSYYYGYYGMnnn
YSYYYYYSYYYGSYGLnnn
YSYYYYSYYYYGGYGLnnn
YSYYYYGYYYYGYYGLnnn
YSYYYYGYYYYGYYGFnnn
```

```
5.87
3.73
5.87
3.63
4.14
3.72
5.87
5.87
5.87
4.4
3.71
3.71
3.77
5.87
5.87
4.02
5.87
5.87
5.87
3.73
```

or



	A	B	C	D
1	YSYSYYYSYYYYSYGLnnn	5.87		
2	SSYSYYYYYYYYSYGFnnn	3.73		
3	YYGGYYYYSYYSGLnnnnn	5.87		
4	YSYYYYYSYYYYGAMnnnn	3.63		
5	YYYYYYYYYYYGSYGLnnn	4.14		
6	YYSYYYYYYYYGGYGLnnn	3.72		
7	YSYYYYYYYYYGYYGLnnn	5.87		
8	YSYYYYYYYYYGYYGMnnn	5.87		
9	YSYYYYYYYYYGYYGLnnn	5.87		
10	YSYYYYYYYYYGSYGLnnn	4.4		
11	YSYYYYYYYYYGSYGMnnn	3.71		
12	YSYYYYYYYYYGSYGLnnn	3.71		
13	YSYYYYYYYYYGSYGFnnn	3.77		
14	YSYYYYYYYYYGSYAMnnn	5.87		
15	YSYYYYYYYYYGGYGMnnn	5.87		
16	YSYYYYYSYYYGYYGMnnn	4.02		
17	YSYYYYYSYYYGSYGLnnn	5.87		
18	YSYYYYSYYYYGGYGLnnn	5.87		
19	YSYYYYGYYYYGYYGLnnn	5.87		
20	YSYYYYGYYYYGYYGFnnn	3.73		

Note that for .xlsx (and .csv) files, the first column must be the sequences, and the second column must be the phenotypes. In addition, there must not be any header names for any files.

2.2.3 3. Executing Ortho_Seqs

We now turn towards our CLI to execute ortho_seqs. Using the Sidhu dataset, our input would look like:

```
ortho_seq orthogonal-polynomial ortho_seq_code/tests/data/protein/Sidhu.xlsx --molecule␣
↪protein --poly_order first --out_dir docs/source/tutorial_outputs --alphbt_input SYG,R␣
↪--min_pct 40 --pheno_name ELISA
```

Let’s explore what these flags are, and how you can use them.

The file input (ortho_seq_code/. . . /sidhu.xlsx) is our sequence AND phenotype data.

```
--molecule
```

This flag is where you indicate what kind of molecule this is. This can be DNA, RNA, or protein. For the Sidhu dataset, the molecules are protein molecules.

```
--poly_order
```

This flag is to indicate the highest degree of polynomial order you want. Currently, DNA and RNA can go up to 2, and protein can only be 1. For the Sidhu dataset, we will look at first-order interactions.

```
--pheno_file
```

This flag is not in the example, because we don't need it. If you were to present your data as two separate .txt files, then this would be where you put the file path for the phenotype data, and the first file path is for your sequence data.

```
--out_dir
```

This flag indicates where you want the output files to go (more on what exactly is saved there later). If the folder path already exists, ortho_seqs will create a new directory with a very similar name, and it will tell you what the new path's name is.

```
--alphbt_input
```

(Note: "Characters" in the following section refer to the nucleotides for DNA, the bases for RNA, and all 21 amino acids for proteins, plus one additional character, "n", which indicates nothing is at that spot to deal with protein sequences of unequal lengths.) This flag indicates the groupings of characters you want. The default will be no groupings, or every character gets counted on its own. If you include (uppercase) letters here, then only those characters will be used (every other character, except "n", gets converted to a "z" and treated as one group). If you comma-separate somewhere in that group, then characters will be grouped based on what comma(s) they are in between. For the Sidhu dataset, to show proof of concept, the groupings will be:

1. SYG
2. R
3. Everything else (z)
4. n

If we were to leave out the commas, the groups would be:

1. S
2. Y
3. R
4. G
5. Everything else (z)
6. n

```
--min_pct
```

One output will be an .xlsx file containing all of the first-order covariances between each amino acid at each side with another amino acid at another site. However, this file can get pretty big pretty quick. Therefore, this flag will only print out covariance values whose magnitudes are at or above the PERCENTILE value specified. The default is 75, meaning it will only save the covariances which range from the 75th to the 100th percentiles in magnitude. To keep it at the default, leave out this flag when inputting what you want. For the Sidhu dataset, we want all magnitudes at or above the 40th percentile (as proof of concept).

```
--pheno_name
```

The pheno_name will label the y axis of the rFon1D graph with whatever the phenotype value represents, if desired.

2.2.4 4. Obtained Outputs

CLI Outputs

The CLI will first print out whether or not it expects one file for the sequence and phenotype, or two separate files. If it is one file, it will then identify whether or not it is a .xlsx or a .csv file. The example outputs:

```
Pheno file is not separate from sequence file, assuming seq_file is either a .csv or a .
↪xlsx file.
Reading .xlsx file.
```

The CLI will then print out the groupings used for the alphabet. If you specified groups with a comma (such as in the example), it will print a map of what every numerical group corresponds to, and a list of the groupings, which is useful for creating rf1d objects. The example outputs:

```
Groupings according to --alphbt_input:
{0: SYG | 1: R | 2: z | 3: n}
rf1d form of alphabet input:
SYG,R,z,n
```

After that, it will output the following text for all first-order calculations as it calculates what it's on:

```
computed mean
computed variance
computed covariance
saved covariance histogram as {out_dir}/cov_hist_{seq_filename}.png
Saved covariance data frame as {out_dir}/cov_data_frame_{seq_filename}.csv
computed reg11
computed Pa: first order orthogonalized within each vector
computed P1i1
computed varP1i1
computed cov11i1
computed reg11i1
computed Pa1i1
computed P1D
computed varP1D
Saving to {out_dir}/{seq_filename}.npz
Saving to {out_dir}/{seq_filename}_covs_with_F.npz
```

where {out_dir} is replaced with the out_dir that was supplied to it, and {seq_filename} is the name of the sequence file.

Next, the CLI outputs the first order regression outputs. The example outputs:

```
Regression of trait on site 1
[[-1.2409090909  1.2409090909  0.            0.          ]
 [ 0.5716578947 -0.5716578947  0.            0.          ]
 [ 0.1729480519 -0.1729480519  0.            0.          ]
 [-0.1468831169  0.1468831169  0.            0.          ]
 [-0.1652922078  0.1652922078  0.            0.          ]
 [-0.2701158301  0.2701158301  0.            0.          ]
 [-0.1246911197  0.1246911197  0.            0.          ]
 [ 0.8338       -0.8445       -0.67875       0.          ]
 [-0.7543831169  1.2094936709  1.194375     -0.67875     ]
```

(continues on next page)

(continued from previous page)

```
 [-0.4118        0.2073076923  1.194375      0.2610759494]
 [-0.3639382239  1.2094936709 -0.8821518987  0.5846153846]
 [-0.2904929577  0.5495526316  0.           -0.0067894737]
 [-0.3020833333  0.4392307692  1.194375     -0.0067894737]
 [-0.1501690141  1.194375      0.0443181818 -0.0067894737]
 [-0.1723809524  0.0611392405  0.1521575342  0.1660273973]
 [-0.1069047619  0.            0.0424064516  0.0627828054]
 [-0.3776712329 -1.073625      0.2858333333  0.1069047619]
 [-0.4165068493  0.           -0.7238461538  0.5358701299]
 [ 0.            0.           -0.4165068493  0.4165068493]]
Regression on 1st order polynomial - orthogonalized within - rFon1D
[[-1.3013918017  1.3013918017  0.            0.          ]
 [ 0.656        -0.656         0.            0.          ]
 [ 0.211342155  -0.211342155   0.            0.          ]
 [-0.0709090909  0.0709090909  0.            0.          ]
 [-0.2096854147  0.2096854147  0.            0.          ]
 [-0.3441860465  0.3441860465  0.            0.          ]
 [-0.1927906977  0.1927906977  0.            0.          ]
 [ 0.9240104167 -0.9316923077 -0.7539130435  0.          ]
 [-0.6834090909  1.1394117647  1.1228985507 -0.7539130435]
 [-0.5211924119  0.5136781609  1.1228985507  0.1872058824]
 [-0.2885714286  1.1394117647 -0.9605882353  0.5121393035]
 [-0.2600081934  0.5730053805  0.           -0.0852307692]
 [-0.2246265938  0.3658706468  1.1228985507 -0.0852307692]
 [-0.0686666667  1.1228985507 -0.0325       -0.0852307692]
 [-0.1128708756  0.2267178503  0.0726612903  0.0867741935]
 [ 0.4901587302  0.           -0.1979       -0.0246423752]
 [-0.1385915493 -0.7828571429  1.0344444444  0.          ]
 [ 0.            0.            0.            0.          ]
 [ 0.            0.            0.            0.          ]]
Regression of trait on site 2 independent of 1
[0. 0. 0. 0.]
computed rFon1
computed rFon1D
Saving regression results to to {out_dir}/{seq_filename}_regressions.npz
Trait values estimated from regressions
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,␣
↪0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,␣
↪0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

After that, the CLI will create an rf1d object, and present the rf1d.summary() and rf1d.barplot() function outputs. The example returns:

```
rf1d Object:

Number of sites: 19
Number of dimensions: 4
Alphabet input: ['SYG', 'R', 'ACDEFHIKLMNPQTVW', 'n']
Molecule: protein

Phenotype represents ELISA values
Highest rFon1D magnitudes:
```

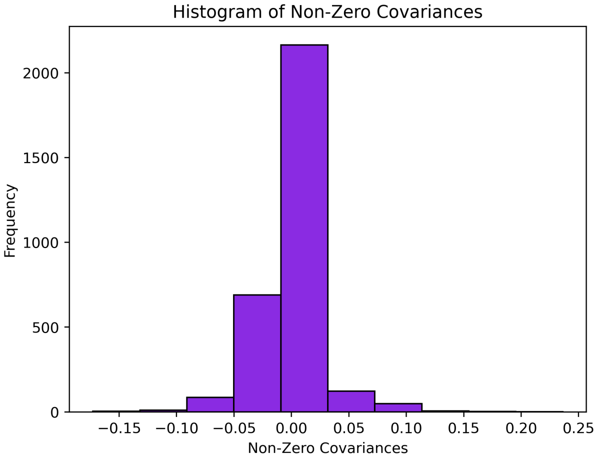
(continues on next page)

(continued from previous page)

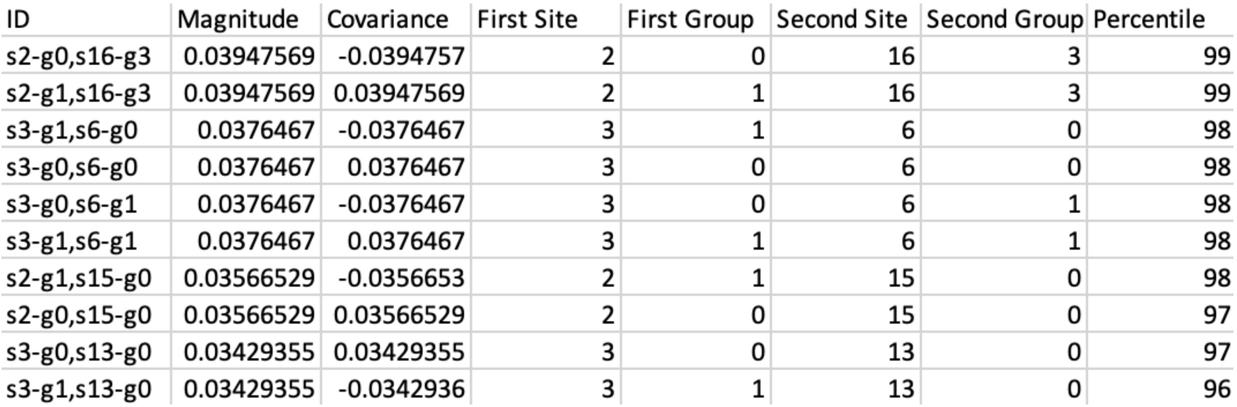
```
-1.3014       Site: 0         Key: SYG
1.3014        Site: 0         Key: R
1.1394        Site: 8         Key: R
1.1394        Site: 10                Key: R
1.1229        Site: 9         Key: ACDEFHIKLMNPQTVW
1.1229        Site: 12                Key: ACDEFHIKLMNPQTVW
1.1229        Site: 8         Key: ACDEFHIKLMNPQTVW
1.1229        Site: 13                Key: R
1.0344        Site: 16                Key: ACDEFHIKLMNPQTVW
-0.9606       Site: 10                Key: ACDEFHIKLMNPQTVW
saved regression graph as {out_dir}/rFon1D_Regressions_of_{phenotype}_values.png
```

Histogram and Spreadsheet of Covariances

The covariances between every character at every site with every other character at another site is recorded in a .csv file, and includes everything at or above the minimum percentile you specified in the input (or defaults to 75th percentile). In addition, the program outputs a histogram of the non-zero covariances, with the bin widths always being 0.5. For the Sidhu dataset, it looks like



And will have the file name cov_hist_{name}.png The .csv file has 8 columns, and looks like:



ID	Magnitude	Covariance	First Site	First Group	Second Site	Second Group	Percentile
s2-g0,s16-g3	0.03947569	-0.0394757	2	0	16	3	99
s2-g1,s16-g3	0.03947569	0.03947569	2	1	16	3	99
s3-g1,s6-g0	0.0376467	-0.0376467	3	1	6	0	98
s3-g0,s6-g0	0.0376467	0.0376467	3	0	6	0	98
s3-g0,s6-g1	0.0376467	-0.0376467	3	0	6	1	98
s3-g1,s6-g1	0.0376467	0.0376467	3	1	6	1	98
s2-g1,s15-g0	0.03566529	-0.0356653	2	1	15	0	98
s2-g0,s15-g0	0.03566529	0.03566529	2	0	15	0	97
s3-g0,s13-g0	0.03429355	0.03429355	3	0	13	0	97
s3-g1,s13-g0	0.03429355	-0.0342936	3	1	13	0	96

They are:

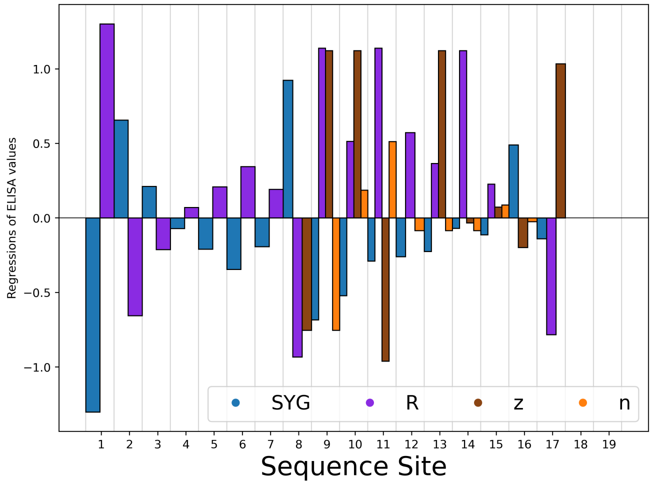
1. ID: Useful for searching for a specific pairing. Ordering will be s{Site 1)-g{Group 1}, s{Site 2)-g{Group 2}. For example, s1-g2,s10-g8 refers to the pairing between Group 2 at Site 1, and Group 8 at Site 10.
2. Magnitude: Absolute value of the covariance value, used to assign percentile values and plot histogram.
3. Covariance: The obtained covariance value.
4. First Site: The site of one of the groups of the covariance pairing. Site 1 for ID column.
5. First Group: The group the character belongs to, identifiable through the –alphbt_input dictionary. Group 1 for ID Column.
6. Second Site: The site of the other group of the covariance pairing. Site 2 for ID column.
7. First Group: The group the character belongs to, identifiable through the –alphbt_input dictionary. Group 2 for ID Column.
8. Percentile: The percentile the respective magnitude is, relative to the entire dataset (including magnitudes that were omitted from the .csv file).

E.g., in our case, say we have a value of -.051 for the covariance corresponding to s1-g1,s5-g2. This means that the group SYG at site 1 covaries negatively with an Arg at site 5. This covariance analysis tells us things about the statics of the sequence space. At this point, we have not projected our phenotype onto the sequence space. Here, we can discover patterns of covariation between amino acids at a given site with amino acids at other sites. When looking at the distribution of the magnitude of covariances, we can identify the ones at the tail ends of this distribution. This information is denoted in the output csv and the allows for the identification of highly covarying (negative or positive) sites.

rFon1D Graph

The main result for the first order analysis is the regression of the phenotype (in this case, ELISA values) onto the first order conditional polynomial (denoted as rFon1D). This tells us the effect of having a given amino acid at one site independent of its correlations with other amino acids at other sites. Here, we can use this result to understand the independent effects of a given amino acid at a given site on the phenotype.

One output is a graph of the nonzero rFon1D values. For the Sidhu dataset, it looks like



At the bottom, it lists the dictionary of the groups and their corresponding number, which then can be used to determine which color bar belongs to which group. The rFon1D graph will always have the name rFon1D_graph_{name}.png. The rFon1D values can also be found in the _regressions.npz file which can be opened up by the user in a jupyter notebook for further visualization.

2.2.5 4. The *rf1d* Class

To obtain more visualizations of your rFon1D results, there is another CLI tutorial here that you can follow.

2.3 Tutorial: Visualizing your rFon1D results

This document will walk you through the steps of how to visualize your rFon1D outputs from *ortho_seqs* using the *rf1d-viz* CLI command.

> **Note:** *rf1d-viz* assumes that you have already run *orthogonal_polynomial* on the dataset. For a tutorial on how to run *orthogonal_polynomial*, view the tutorial here.

2.3.1 1. Requirements for *rf1d-viz*

- The *{trait_file_name}_regressions.npz* file that is returned from *orthogonal-polynomial*.
- The *rf1d* form of the alphabet input.

When you run *orthogonal-polynomial*, the CLI will output the following text towards the beginning:

```
rf1d form of alphabet input:
```

The line **beneath** that line is the *rf1d* form of the alphabet input.

- The molecule type of the sequence (mostly *DNA* or *protein*).
- What the phenotype values are representing.

2.3.2 2. *rf1d-viz* flags:

rf1d-viz will require you to input the following flags, many of which have counterparts in *orthogonal-polynomial*:

```
--filename
```

This will be the *{trait_file_name}_regressions.npz* file that is returned from *orthogonal-polynomial*.

```
--alphbt_input
```

This will be the *rf1d* form of the alphabet input.

```
--molecule
```

This is the molecule type.

```
--phenotype
```

This is the phenotype type. It will be used for labelling the graphs.

```
--out_dir
```

This is where you want the graphs stored. **Note:** the path must exist prior to running *rf1d-viz*.

```
--action
```

This is where you specify what kind of visualization you want. The current options are:

1. *barplot* - This will create a barplot of the rFon1D values, grouped by site and alphabet input. This is called automatically when you run *orthogonal-polynomial*.
2. *density* - This will create a density plot of the rFon1D values.
3. *summary* - Prints out the number of sites and dimensions, the alphabet input, the molecule, and calls *sort* (another *rf1d-viz* action that is explained in further detail below). This is called in *orthogonal-polynomial* automatically, and will be saved to the *out_dir* as *summary.txt*.
4. *heatmap* - This will create a heatmap of the rFon1D values, grouped by site and alphabet input.
5. *boxplot* - This will create a boxplot of the rFon1D values, grouped by .
6. *sort* - This will print out the top 10 rFon1D values by magnitude, including the rFon1D value, the site, and the group it belongs to. This will be saved to the *out_dir* as *sort.txt*.
7. *ALL* - This will produce a barplot, histogram, heatmap, and boxplot simultaneously.

> **Note:** For now, you will need to close the first graph once it displays on your computer for the rest of the graphs to run.

2.3.3 3. Running *rf1d-viz*

Similarly to *orthogonal-polyomial*, you will run *rf1d-viz* in your CLI, first starting with the keyword *ortho_seq*, but now followed by *rf1d-viz*, instead of *orthogonal-polynomial*. The general format is

```
ortho_seq rf1d-viz filename --alphbt_input --molecule --phenotype --out_dir --action
```

where *filename* represents the *–filename* flag.

2.3.4 Guided example with the Sidhu dataset

The example uses the *Sidhu* dataset, which is the same as was used for the *orthogonal-polynomial* tutorial. Recall that the input for *orthogonal-polynomial* was:

```
ortho_seq orthogonal-polynomial ortho_seq_code/Sidhu/Sidhu.xlsx --molecule protein --
↪poly_order first --out_dir docs/source/tutorial_outputs --alphbt_input SYG,R --min_pct␣
↪40 --pheno_name ELISA
```

The regression file that will be used for *rf1d-viz* will thus be called

```
Sidhu_regressions.npz
```

Using the CLI output, we obtain

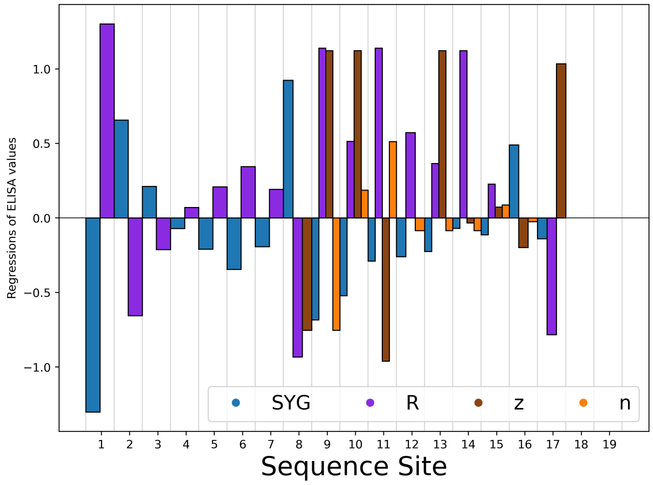
```
rf1d form of alphabet input:
SYG,R,z,n
```

which reveals that the *rf1d* form of the alphabet input is **SYG,R,z,n**.

With these in mind, the CLI input for *rf1d-viz* for a **barplot** will be

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action barplot
```

This line of code will reproduce the graph that is automatically run, and looks like

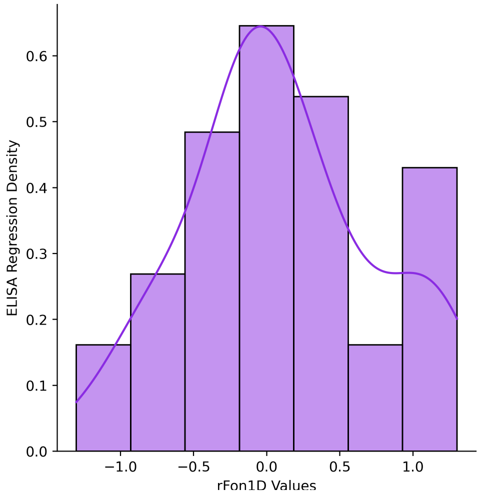


Notice how the y axis is labelled with the phenotype name specified

The CLI input for *rf1d-viz* for a **density plot** will be

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action density
```

The graph looks like



Run **summary** with

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action summary
```

The output will be

```
rf1d Object:

```

(continues on next page)

(continued from previous page)

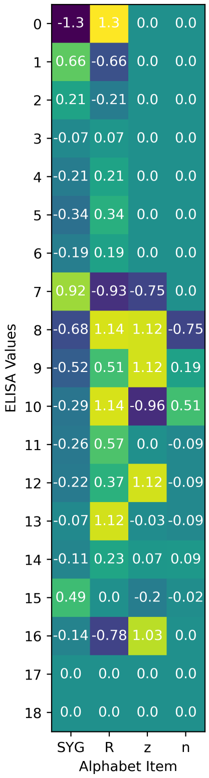
```
Number of sites: 19
Number of dimensions: 4
Alphabet input: ['SYG', 'R', 'z', 'n']
Molecule: protein

Phenotype represents ELISA values
Image output directory: docs/source/tutorial_outputs
Highest rFon1D magnitudes:
-1.3014       Site: 0         Key: SYG
1.3014        Site: 0         Key: R
1.1394        Site: 8         Key: R
1.1394        Site: 10                Key: R
1.1229        Site: 9         Key: z
1.1229        Site: 12                Key: z
1.1229        Site: 8         Key: z
1.1229        Site: 13                Key: R
1.0344        Site: 16                Key: z
-0.9606       Site: 10                Key: z
```

The CLI input for *rf1d-viz* for a **heatmap** will be

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action heatmap
```

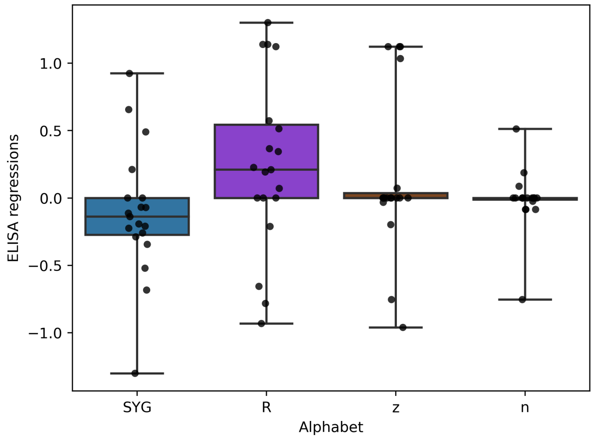
The graph looks like



The CLI input for *rf1d-viz* for a **boxplot** will be

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action boxplot
```

The graph looks like



Lastly, this is the input for **sort**:

```
ortho_seq rf1d-viz docs/source/tutorial_outputs/Sidhu_regressions.npz --alphbt_input SYG,
↪R,z,n --molecule protein --phenotype ELISA --out_dir docs/source/tutorial_outputs --
↪action sort
```

The output will be

```
-1.3014       Site: 0         Key: SYG
1.3014        Site: 0         Key: R
1.1394        Site: 8         Key: R
1.1394        Site: 10                Key: R
1.1229        Site: 9         Key: z
1.1229        Site: 12                Key: z
1.1229        Site: 8         Key: z
1.1229        Site: 13                Key: R
1.0344        Site: 16                Key: z
-0.9606       Site: 10                Key: z
```

As you can see, this prints out the second half of the *summary* output, since *summary* calls *sort*.

2.4 Tutorial: Generating a sequence logo plot

This document will walk through how to generate a frequency/probability-based logo plot from sequence data in the format of an input to *ortho_seqs*, using the *logo-plot* CLI command. The logo plot can provide information on the frequencies of nucleotides/amino acids present in your sequence dataset before running *orthogonal-polynomial*.

Logo plots are generated using the logomaker package in Python. More customization options exist (font, color schemes, etc) that are not (yet) implemented here.

2.4.1 1. Requirements for *logo-plot*

The sequence data, formatted as input to *ortho_seqs*. This can take the form of either:

- .txt file: single file containing sequences, separated by line breaks.
- .csv or .xlsx file: single file containing sequence data in the first column. Can (but doesn't have to) contain phenotype data in the second column.

2.4.2 2. *logo-plot* flags:

logo-plot will require the following flags.

```
--filename
```

This will be the sequence data file, formatted as described in (1).

```
--molecule
```

This is the molecule type. Should either be DNA, RNA, or protein. Default is DNA.

```
--out_dir
```

This is where you want the logo plot to be stored.

2.4.3 3. Running *logo-plot*:

You will run *logo-plot* in the CLI the same way you would run *orthogonal-polynomial* or *rf1d-viz*. The general format is:

```
ortho_seq logo-plot filename --molecule --out_dir
```

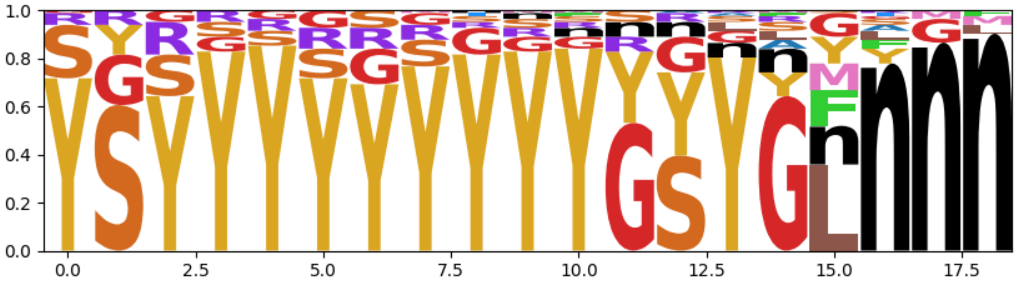
2.4.4 4. Guided example with test dataset:

The example uses the Sidhu dataset that has also been used in the other tutorials.

The sequence data that will be used for this example is called *sidhu_insulin_cdrh3_seqs.xlsx*. Given that this dataset is about proteins, our CLI input will be

```
ortho_seq logo-plot docs/source/sidhu_insulin_cdrh3_seqs.xlsx --molecule protein --out_dir docs/source/tutorial_outputs/
```

The saved figure will look like:



2.5 License

2.6 Get Further Help

If you have more questions, please refer to the github repo at https://github.com/snafees/ortho_seqs for existing issues or start an issue there. You can also refer to the PyPI page at https://pypi.org/project/ortho-seq-code/

2.6.1 Contact

If you have more questions or comments, please feel free to reach out to me at saba.nafees314@gmail.com. We look forward to hearing from you!

CHAPTER

THREE

INDICES AND TABLES

- genindex
- modindex
- search