ortho_seqs

Miles Woollacott

What is ortho_seqs? What does ortho_seqs do?

Results

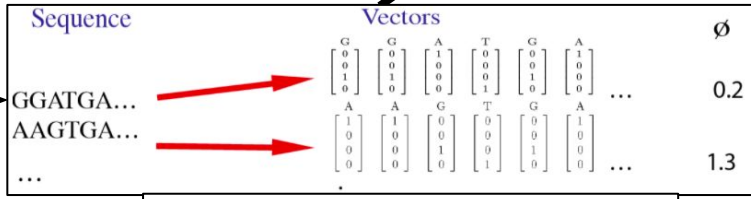
ortho_seqs is a command line interface (CLI) written in Python that applies the **tensor-based orthogonal polynomial method*** to sequence & phenotype data.

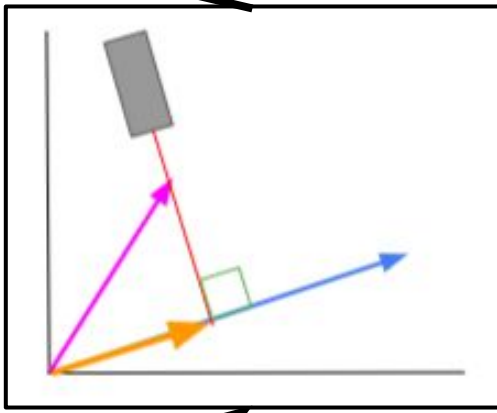
ic50_seq.txt

ARDLDVYGLDV	42.0
ARDYGDLYFDY	140.0
ARDFGDFYFDY	25.0
ATESWVYGSGSYSSGAFDI	1300.0
ARGPRYSGTYFDY	480.0
ARGPRYSGTHFDY	160.0
ARDPYGYSSIWDGQGGH	58.0
ARGGYCSDGSCYVQDRLIYYYSGLDV	39.0
ARDRRRRYCTNGVCYRPEEIDY	4900.0
ARDPFPGAVAGTGYLQY	590.0
AKSSGSYYYYYYGMDV	44.0
ARLHCGGDCYLDY	21.0
AKGSGSGSYPNYYYYYGMDV	19.0
AKANKYSSSEFDF	66.0
AVYYYDSGSPGWFDP	15.0
ARDFRYCSSTRCYFWFDP	12.0
ARWYDSTGSIDY	25.0
AKDQGYSYGNYFDY	13.0
ALRNQWDLLVY	1500.0
ARDLVVYGMDV	24.0
ARDPIRNGMDVW	18100.0
ARDAMSYGMDV	18.0
ARDAAVYGIDV	16.0
ARDLISRGMDV	33.0
ARDRVVYGMDV	12.0
ARDLVSYGMDV	553.0
ARDAQNYGMDV	15.0
ARDRGLVSDYW	22.0
ARPQGGSSWYRDYYYGMDV	4800.0
ARDLSVRGGMDV	1900.0

ic50_phi.txt

```
ortho_seq orthogonal-polynomial ../data/cov2/big_mabs_ic50s/ic50_seq.txt
--pheno_file ../data/cov2/big_mabs_ic50s/ic50_phi.txt --molecule protein
--poly_order first --out_dir /data_sm/home/miles/results/CDRH3/cov2_unique_ic50
```



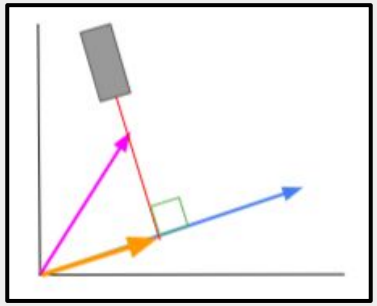


*See Nafees et al., 2020 for method application and Rice, 2020 for method background

(Image from https://github.com/snafees/ortho_seqs)

Why is ortho_seqs important?

ortho_seqs converts each site in a sequence into a vector, which allows for mathematical computations and analyses.



Matrix of Covariances

Covariance measures the relationship between two items.

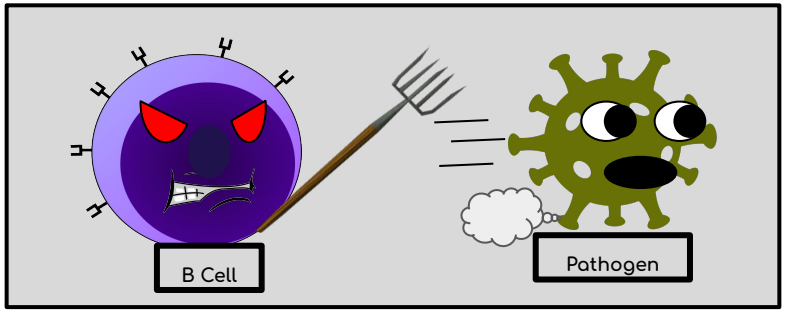
Having a positive covariance for a pair of site → values of the first item increase with the second (and vice versa).

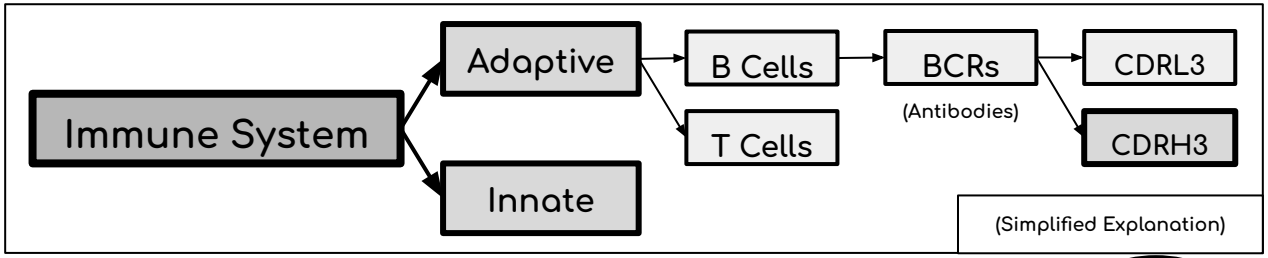
Regressions of Phenotype Onto the X Order Conditional Polynomial (rFonXD)

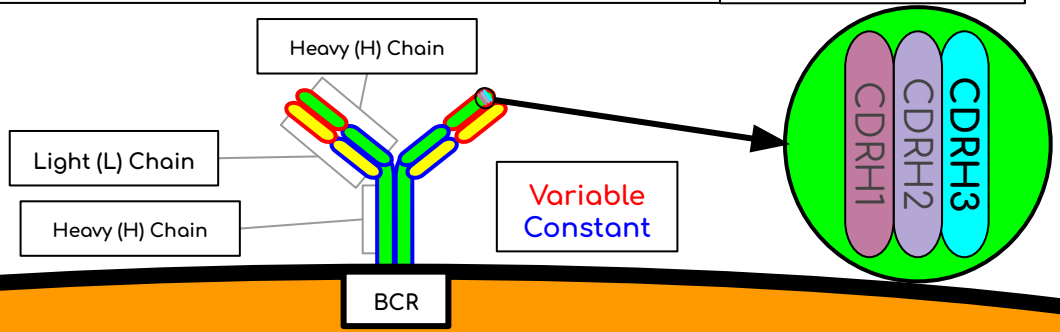
rFon1D: measures the impact of having a given amino acid/nucleotide at that site. This impact can be positive or negative.

Regression onto second order conditional polynomial (rFon2D): measures the impact of having a given nucleotide *pair* at pairs of sites (not supported for proteins yet).

Application Time!







(Image Adapted From "AIRR-seq data analysis and processing" by Victor Greiff, Slide 9)

cov2 Dataset

Application Time!

The cov2 dataset is a conglomeration of antibodies that neutralize SARS-CoV-2, along with the sequences, IC50, and EC50 values (where provided).

IC50 = Half Maximal Inhibitory Concentration
EC50 = Half Maximal Effective Concentration

SHARE

RESEARCH ARTICLE

Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model

Thomas F. Rogers[1,2,*], Fangzhu Zhao[1,3,4,*], Deli Huang[1,*], Nathan Beutler[1,*], Alison Burns[1,3,4], Wan-ting He[1,3]...
+ See all authors and affiliations

Science 21 Aug 2020:
Vol. 369, Issue 6506, pp. 956-963
DOI: 10.1126/science.abc7520

mAb ID	VH Gene	% SHM (nt)	CDRH3 (AA)
CC12.1	IGHV3-53*01	1.05	CARDLDVYGLDVW
CC12.2	IGHV3-53*01	0.70	CARDYGDLYFDYW
CC12.3	IGHV3-53*01	1.40	CARDFGDFYFDYW
CC12.4	IGHV1-2*04	1.74	CATESWVYGSGSYSSGAFDIW
CC12.5	IGHV1-2*02	4.86	CARGPRYSGTYFDYW
CC12.6	IGHV1-2*02	4.17	CARGPRYSGTYFDYW
CC12.7	IGHV1-2*02	7.99	CARGPRYSGTYFDFW
CC12.8	IGHV1-2*02	3.12	CARGPRYSGTYFDYW
CC12.9	IGHV1-2*02	3.47	CARGPRYSGTYFDYW
CC12.10	IGHV1-2*02	4.51	CARGPRYSGTHFDYW
CC12.11	IGHV1-2*02	3.47	CARGPRYSGTYFDYW
CC12.12	IGHV1-2*02	3.47	CARGPRYSGTYFDYW
CC12.13	IGHV3-53*01	0.35	CARDPYGYSSIWDGGGGHW
CC12.14	IGHV3-21*01	3.47	CARGGYCSDGSCYVQDRLIYYYSGLDVW
CC12.15	IGHV3-48*03	1.04	CARDRRRRYCTNGVCYRPEEIDYW
CC12.16	IGHV3-33*01	1.74	CARDPFPGAVAGTGYLQYW
CC12.17	IGHV3-30*03	1.39	CAKSSGSYYYYYYGMDVW
CC12.18	IGHV1-46*01	0.00	CARLHCGGDCYLDYW
CC12.19	IGHV3-23*04	1.04	CAKGSGSGSYPNYYYYYGMDVW
CC12.20	IGHV3-30*03	1.04	CAKDQAYYDILTGYLNPPKNYYYYGMDVW
CC12.21	IGHV1-24*01	1.39	CATAFSIFGVVPPDYW
CC12.22	IGHV1-24*01	1.04	CATGFAQNMVLLTPYW
CC12.23	IGHV4-39*01	0.69	CARQGDCSTTSCAYDYW
CC12.24	IGHV3-30*03	1.04	CAKDRTGNYYYGMDVW

mAb name	Predicted Unique Epitope	ELISA IC_{50} (pg/mL) SARS-CoV-2 S	SARS-CoV-2 RBD	SARS-CoV-1 S
CC12.1	RBD-A	0.017	0.042	>50
CC12.2	RBD-A	0.021	0.14	>50
CC12.3	RBD-A	0.018	0.025	>50
CC12.4	RBD-A	0.062	1.3	>50
CC12.5	RBD-A	0.33	1.4	>50
CC12.6	RBD-A	0.36	3.1	>50
CC12.7	RBD-A	0.091	1.2	>50
CC12.8	RBD-A	0.10	0.48	>50
CC12.9	RBD-A	20	>50	>50
CC12.10	RBD-A	0.013	0.16	>50
CC12.11	RBD-A	0.36	1.2	>50
CC12.12	RBD-A	0.15	0.67	>50
CC12.13	RBD-A	0.044	0.058	>50
CC12.14	RBD-A	0.014	0.039	>50
CC12.15	RBD-A	0.87	4.9	>50
CC12.16	RBD-B	0.073	0.59	>50
CC12.17	RBD-B	0.020	0.044	>50
CC12.18	RBD-B	0.017	0.021	0.018
CC12.19	RBD-B	0.012	0.019	>50
CC12.20	SPIKE-A	0.036	>50	>50
CC12.21	SPIKE-A	0.0050	>50	>50
CC12.22	SPIKE-A	>50	>50	>50
CC12.23	SPIKE-A	0.014	>50	>50
CC12.24	ND	0.013	>50	>50

Source: https://science.sciencemag.org/content/suppl/2020/06/15/science.abc7520.DC1

	A	B	C	D	E	F	G
1	Status	name	author	phenotype_t	ic50_ng	ec50_ng	cdr-h3
10	Verified	CC12.1	Rogers	IC50	42		ARDLDVYGLDV
11	Verified	CC12.2	Rogers	IC50	140		ARDYGDLYFDY
12	Verified	CC12.3	Rogers	IC50	25		ARDFGDFYFDY
13	Verified	CC12.4	Rogers	IC50	1300		ATESWVYGSGSYSSGAFDI
14	Verified	CC12.5	Rogers	IC50	1400		ARGPRYSGTYFDY
15	Verified	CC12.6	Rogers	IC50	3100		ARGPRYSGTYFDY
16	Verified	CC12.7	Rogers	IC50	1200		ARGPRYSGTYFDY
17	Verified	CC12.8	Rogers	IC50	480		ARGPRYSGTYFDY
18	Verified	CC12.10	Rogers	IC50	160		ARGPRYSGTHFDY
19	Verified	CC12.11	Rogers	IC50	1200		ARGPRYSGTYFDY
20	Verified	CC12.12	Rogers	IC50	670		ARGPRYSGTYFDY

ic50_seq.txt

```
ARDLDVYGLDV
ARDYGDLYFDY
ARDFGDFYFDY
ATESWVYGSGSYSSGAFDI
ARGPRYSGTYFDY
ARGPRYSGTHFDY
ARDPYGYSSIWDGQGGH
ARGGYCSDGSCYVQDRLIYYYSGLDV
ARDRRRRYCTNGVCYRPEEIDY
ARDPFPGAVAGTGYLQY
AKSSGSYYYYYGMDV
ARLHCGGDCYLDY
AKGSGSGSYPNYYYYYGMDV
AKANKYSSSEFDF
AVYYYDSGSPGWFDP
ARDFRYCSSTRCYFWFDP
ARWYDSTGSIDY
AKDQGYSYGNYFDY
```

ic50_phi.txt

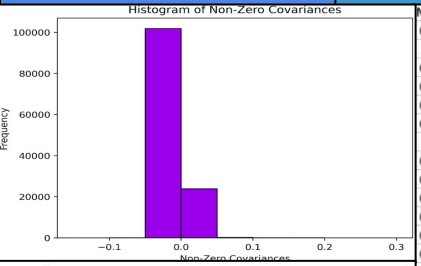
```
42.0
140.0
25.0
1300.0
480.0
160.0
58.0
39.0
4900.0
590.0
44.0
21.0
19.0
66.0
15.0
12.0
25.0
13.0
```

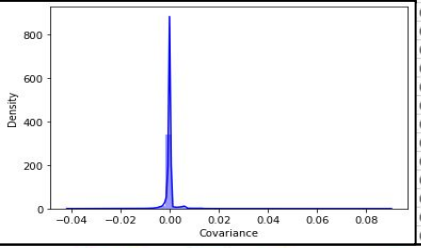
CLI Input:

```
ortho_seq orthogonal-polynomial ../data/cov2/big_mabs_ic50s/ic50_seq.txt --pheno_file
../data/cov2/big_mabs_ic50s/ic50_phi.txt --molecule protein --poly_order first --out_dir
/data_sm/home/miles/results/CDRH3/cov2_unique_ic50
```

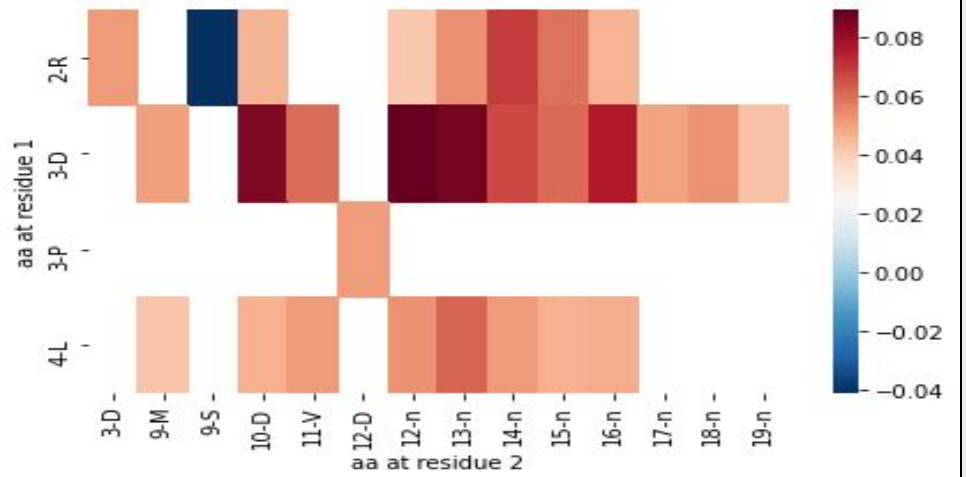
cov2 Dataset

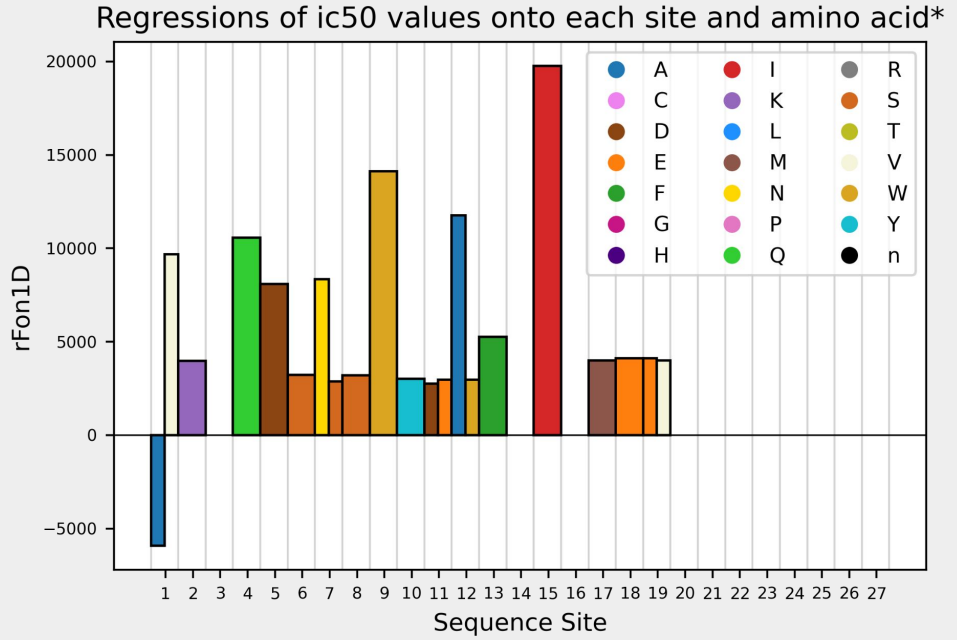
Application Time!





Magnitude	Covariance	First Site	First Group	Second Site	Second Grou	Percentile
0.08920745	0.08920745	3	2	12	20	99
0.0861943	0.0861943	3	2	13	20	99
0.08459642	0.08459642	3	2	10	2	99
0.07619613	0.07619613	3	2	16	20	99
0.06902849	0.06902849	2	14	14	20	99
0.06720234	0.06720234	3	2	14	20	99
0.0619065	0.0619065	4	9	13	20	99
0.06108473	0.06108473	3	2	15	20	99
0.06058254	0.06058254	3	2	11	17	99
0.05971512	0.05971512	2	14	15	20	99
0.05387144	0.05387144	2	14	13	20	99
0.05355186	0.05355186	3	2	18	20	99
0.05314098	0.05314098	4	9	12	20	99
0.05204529	0.05204529	2	14	3	2	99
0.05204529	0.05204529	4	9	14	20	99
0.05195398	0.05195398	4	9	11	17	99
0.05136048	0.05136048	3	12	12	2	99
0.05104091	0.05104091	3	2	9	10	99
0.05044741	0.05044741	3	2	17	20	99
0.04830168	0.04830168	4	9	16	20	99
0.04720599	0.04720599	4	9	10	2	99
0.04711468	0.04711468	4	9	15	20	99
0.04674945	0.04674945	2	14	10	2	99
0.04656684	0.04656684	2	14	16	20	99
0.04346238	0.04346238	3	2	19	20	99
0.04254931	0.04254931	4	9	9	10	99



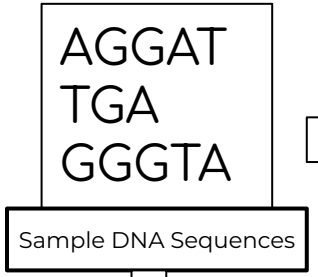


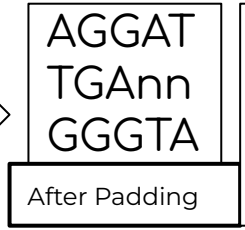
*rFon1D values between -2500 and 2500 were omitted for graph legibility

Updates to ortho_seqs

What I've Done:

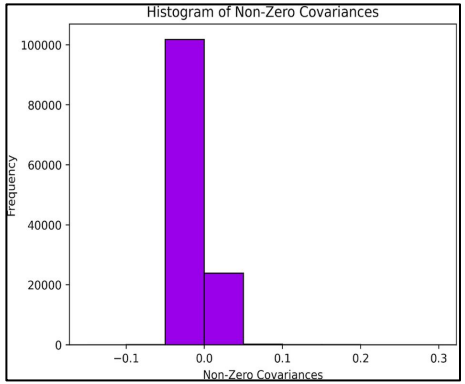
- User-Friendly Updates and Efficiency
  - Automatic Padding
  - --dm, --sites, --pop_size flag Removal
  - Dimension Reducer
  - No Overwriting Data from --out_dir
- Features
  - Covariance Histogram
  - Covariance .csv File
    - --min_pct flag
  - rFon1D Bar Plot
  - --alphbt_input flag



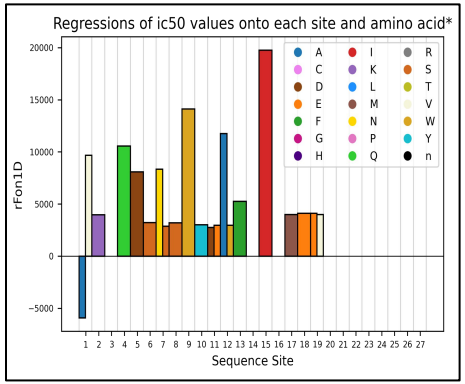


Dimensions: 4
(A, G, T, n)
Sites: 5
(5 columns)
Pop Size: 3
(3 Sequences)

"C" is never used in any sequence, and is thus removed to reduce runtime.



Magnitude	Covariance	First Site	First Group	Second Site	Second Grou	Percentile
0.08920745	0.08920745	3	2	12	20	99
0.0861943	0.0861943	3	2	13	20	99
0.08459642	0.08459642	3	2	10	2	99
0.07619613	0.07619613	3	2	16	20	99
0.06902849	0.06902849	2	14	14	20	99
0.06720234	0.06720234	3	2	14	20	99
0.0619065	0.0619065	4	9	13	20	99
0.06108473	0.06108473	3	2	15	20	99
0.06058254	0.06058254	3	2	11	17	99
0.05971512	0.05971512	2	14	15	20	99
0.05387144	0.05387144	2	14	13	20	99
0.05355186	0.05355186	3	2	18	20	99
0.05314098	0.05314098	4	9	12	20	99
0.05204529	0.05204529	2	14	3	2	99
0.05204529	0.05204529	4	9	14	20	99
0.05195398	0.05195398	4	9	11	17	99
0.05136048	0.05136048	3	12	12	2	99
0.05104091	0.05104091	3	2	9	10	99
0.05044741	0.05044741	3	2	17	20	99
0.04830168	0.04830168	4	9	16	20	99
0.04720599	0.04720599	4	9	10	2	99
0.04711468	0.04711468	4	9	15	20	99
0.04674945	0.04674945	2	14	10	2	99
0.04656684	0.04656684	2	14	16	20	99
0.04346238	0.04346238	3	2	19	20	99
0.04254931	0.04254931	4	9	9	10	99



Updates to ortho_seqs

What Will Be Done:

- User-Friendly Updates and Efficiency
  - GUI
  - Histogram Improvements
  - Runtime Efficiency
  - Only one file for both seq and phi
- Features
  - Third-Order Calculations for DNA
  - Second-Order Calculations for Proteins

...and more!

Special Thanks:

- Saba Nafees: Mentor
- Advisors: Eric Waltari, Joan Wong
- Code review: Pranathi Vemuri
- Server help: Saransh Kaul
- CZ Biohub

Contact Me:

- miles.woollacott@gmail.com