Functional Dynamics Recognition by Machine Learning with Applications to Drug Discovery

Tyler Grear[1], Chris Avery[1,2], Donald J. Jacobs[1]
[1] Department of Physics and Optical Science, University of North Carolina at Charlotte, NC, USA
[2] Department of Bioinformatics and Genomics, University of North Carolina Charlotte, NC, USA

INTRO

- The grand challenge for Machine Learning (ML) in Bioinformatics is to recognize the causal connections between the microscopic molecular properties of a system, to macroscopic outcomes and consequences.
- Once critical features that make a system functional are identified, computational modeling and classification can be employed in a bottom up approach for rational experimental design strategies.
- The primary challenges for ML that we address are: data reduction without losing critical information from features hidden by many irrelevant degrees of freedom with respect to the function of interest, and how to identify critical features that do not separate into well-defined categories.

METHODS

- Small synthetic molecules were generated using a Monte Carlo simulation approach.
- Each synthetic molecule consists of 29 atoms and the oscillations on the xy-plane were generated for both 500 samples and 20000 samples.
- The small molecules were categorized by the geometric signature for the three sections of the synthetic molecules. A designation of functional was assigned to molecules with a middle section that contained a geometric signature of linear, this was considered the xLy training set.
- Methods of Discriminant Analysis (DA) were applied to the synthetic trajectory data to identify the dynamical motions that were similar with the xLy training set and could be assigned classification.
- Supervised Projective Learning for Orthogonal Congruences (SPLOC) utilizes a generalized Jacobi eigenvalue algorithm. The discriminant modes obtained by SPLOC are able to contain more higher frequency modes than the use of principal components which are widely used for Normal Mode Analysis (NMA).

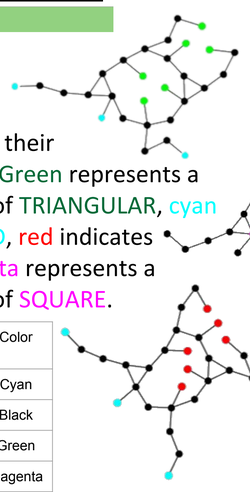
Results

- It is important to note that the small variance contained in the atomic trajectories played a significant role in the feature selection approach for the applied methods. The geometric signatures of the synthetic molecules at 500 samples are 'confused' by the standard methods
- A highlighted example of the classification 'confusion' avoided by SPLOC is presented when compared to the standard method predictions of the synthetic molecule ELF.

Discussion

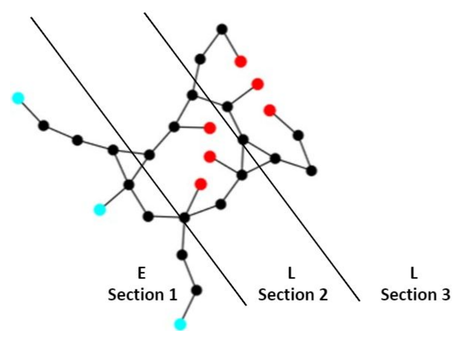
- Many of the standard DA methods have a tendency to overfit when presented with unknown atomic trajectories. SPLOC did not suffer from this issue.
- As shown with the discriminant modes found by SPLOC, the information lost in PCA from protein dynamics with little variance is conserved.
- SPLOC is able to discern functional motions that are not detected by the standard methods when applied to mutations of Beta Lactamase.

Synthetic Molecule Generation

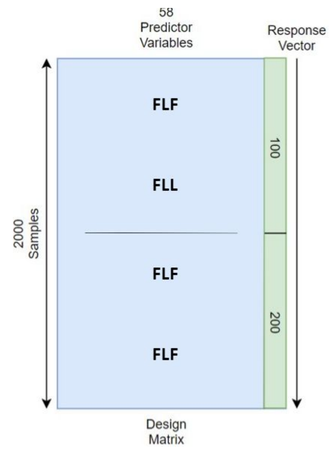


The synthetic molecules were categorized based on their geometric signature. Green represents a geometric signature of TRIANGULAR, cyan represents EXTENDED, red indicates a LINEAR, and magenta represents a geometric signature of SQUARE.

Geometric Signature	Designation	Color
Extended	E	Cyan
Free	F	Black
Triangular	T	Green
Square	S	Magenta
Linear	L	Red



Each synthetic molecule consists of three sections. The example depicted above shows the molecule ELL.



The data generated was configured with the rows as the samples and the columns being the independent variables. The synthetic molecules were generated in such a way as to contain little variance within the motions on the xy-plane. Mathematically this indicates that when the dimensions of the data are reduced in PCA the information is spread across many principal components.

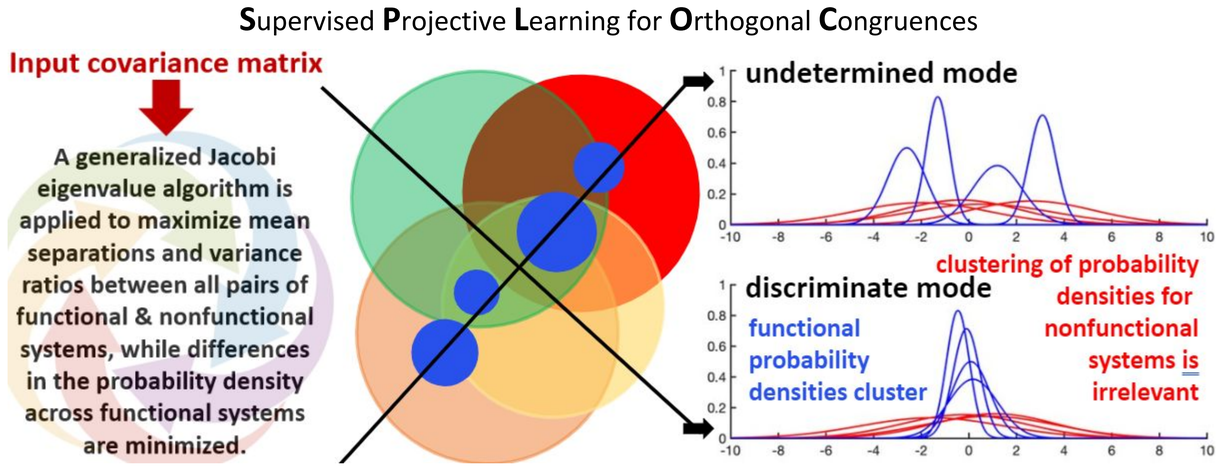
Discriminant Analysis

With the aim of identifying the functional dynamics of the synthetic molecules, different methods of discriminant analysis were employed to be compared with Supervised Projective Learning for Orthogonal Congruences (SPLOC). A training set xLy was created which contained a central section of Linear.

Standard DA Methods:

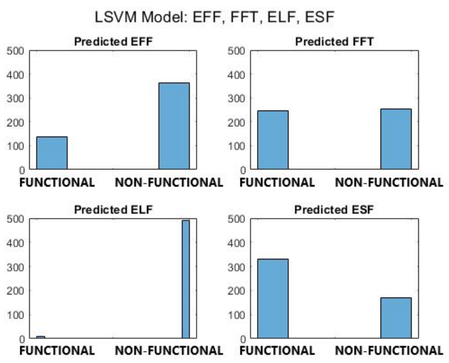
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Support Vector Machines (Linear Hyperplane)
- Support Vector Machine (Quadratic Hyperplane)
- Logistic Regression

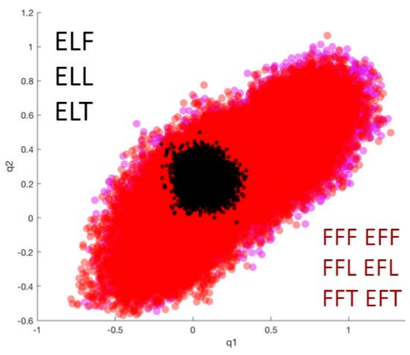
Supervised **P**rojective **L**earning for **O**rthogonal **C**ongruences



Results

Model Type	Training Time	Model Accuracy	Sample Size
LDA	.73 s	99.15%	500
LDA w/o PCs	.41 s	99.90%	500
QDA	.67 s	99.95%	500
QDA w/o PCs	.40 s	100.0%	500
LSVM	.73 s	99.45%	500
LSVM w/o PCs	.54 s	99.95%	500
QSVM	.81 s	100.0%	500
QSVM w/o PCs	.47 s	100.0%	500
Logistic Regression	1.19 s	100.0%	500
Logistic Regression w/o PCs	2.18 s	100.0%	500
LDA	4.83 s	82.21%	20000
LDA w/o PCs	3.33 s	90.38%	20000
QDA	4.64 s	86.33%	20000
QDA w/o PCs	3.29 s	95.77%	20000
LSVM	1105.8 s	82.94%	20000
LSVM w/o PCs	903.98 s	91.85%	20000
QSVM	3900 s	89.16%	20000
QSVM w/o PCs	989.35 s	98.53%	20000
Logistic Regression	7.07 s	81.98%	20000
Logistic Regression w/o PCs	10.923 s	90.79%	20000

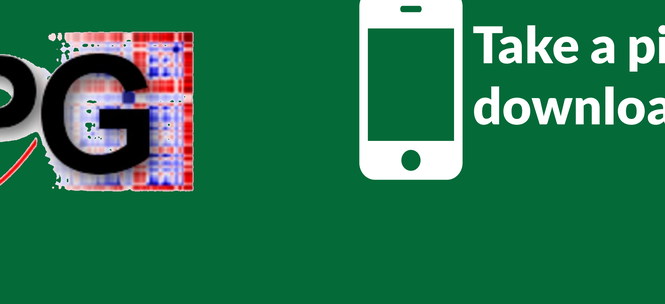


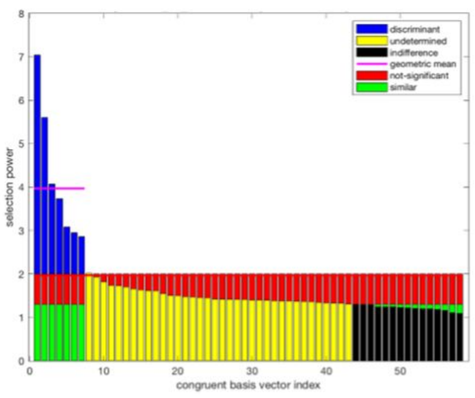


Molecule	Functional Probabilities	Functional Probabilities	Molecule	Functional Probabilities	Functional Probabilities
ELL	.79773	.89640	FSL	.092810	.12313
ELT	.74673	.87788	FST	.076147	.09980
FLF	.68065	.90480	FTF	.066015	.01829
FLT	.67391	.88533	ESF	.062109	.22311
ELF	.59607	.91410	EST	.060206	.12955
FFT	.54156	.00834	FTT	.043889	.00822
EFL	.48609	0.0	ETF	.031512	.05474
FTL	.37432	.01622	ESL	.017328	.12857
FLL	.31709	.90892	EFF	.001915	.00058
FSF	.22117	.22225	EFT	0.0	.00191
ETT	.14277	.02994	ETL	0.0	.02359
FFF	.10283	0.0	FFL	0.0	0.0
	500 samples	20000 samples		500 samples	20000 samples

Discussion

SPLOC applied to three dimensional trajectories of Beta-Lactamase. The red portion of the image represents the omega loop. The trajectories of the omega loop is classified differently than the active site Serine at 130 and the primary active Serine at 70. The blue portions of the image show many of the dynamical motions of the Beta-Lactamase. Similar to the two dimensional representation presented, the aim is to identify and classify the different atomic trajectories related to the functional motions of the protein.





The congruent basis vector index found by SPLOC using the generalized Jacobi eigenvalue method. These discriminant modes differ from the principal components normally used in Principal Component Analysis. SPLOC is identifying basis vectors that contain small amounts of variance within the trajectories of molecules/proteins.

UNC CHARLOTTE

BMPG

Take a picture to **download** the **full poster**