**Readme**

CYCLOPS (**Cycl**ic **O**rdering by **P**eriodic **S**tructure) is a tool designed to reconstruct the temporal ordering of high dimensional that is generated by a common periodic process

Descriptions of the CYCLOPS approach can be found online at PNAS

* This readme is NOT a comprehensive guide to using CYCLOPS. Please see the example scripts for more direction. If you have further troubes/questons –please email.

CYCLOPS is based on a circular node autoencoder (see work of Kirby and Miranda).

Most backpropogation scripts are written under the assumption that all derivatives can be computed from the inputs to a given node. In a circular node device this is not the case. It is for this reason that generic NN programs may not be suited for this problem

The infrastructure of CYLCOPS was designed so that it could eventually handle various network structures and activation functions. For now, however, CYCLOPS is running with linear neurons projecting onto a singular circular node.

Non Linear PCA tools in matlab have implemented circular node autoencoders (see the work of Matthias Scholz and also the work of William W. Hsieh). I have not had a Matlab license for a while! Earlier I was getting errors in the Scholz NLPCA output (when using a circular node) and could not track down the bug. That being said, for those with matlab, these are alternatives to consider

To run the CYCLOPS program you will need:

A gene expression data set (each column is a different sample, each row is a different gene)

A list of “seed genes” believed to be enriched for rhythmic expression and used to base orderings

There are several parameter used in the CYCLOPS ordering and should be set in the main program calling the various CYCLOPS routine

**Frac\_Var**=0.85 # The Number of Dimensions of SVD retained is set to maintain this fraction of variance

**DFrac\_Var**=0.03 # The Number of Dimensions of SVD retained is set so that the incremetal fraction of variance increases by at leas this much with each added dimenstion

**N\_trials** =40 # Number of random initial conditions to try for each optimization

**MaxSeeds** = 10000 # Maximum number of seed genes accepted

**total\_background\_num**=200; # Number of background runs for global background refrence statistics (note that computing the background distribution and assessing significance is by far the most time consuming aspect. You may decided to hold on running background statistics until after you have explored CYCLOPS and the results)

**n\_cores**=5; # Number of machine cores (not really used right now –program is just –might need to adjust multicore stats routine)

* At present CYCLOPS is written to optimally run on a computer with 6 physical cores.
* While there is a parameter describing the # of cores – this is presently not used
* If your computer has a different number of cores – you will (unfortunately) need to go through the code. We aim to correct this

Cleaning, normalization, and scaling of data performed before temporal reconstruction, as is a dimensional reduction scheme. All of these can have a marked effect on results. Much of data cleaning takes place within the getseed function (an example function call is bellow)

getseed(**data**::Array{Any,2},**symbol\_list**,**maxcv**=.75,**mincv**=.07,**minmean**=500,**blunt**=.99)

The first argument is the full data matrix.

The second is the list of seed genes to be extracted to use in reconstruction

However it is likely that some of these seed genes are likely poor candidates upon which to base reconstruction. The next arguments further constrain the list.

For example, it makes no sense to base an ordering on a gene with no variation in expression.

The mincv and maxcv and minimum and maximum coefficient of variation (in expression) allowed for genes in the seed list. Genes with expression variation outside this bound are not used for the basis of CYCLOPS ordering.

Similarly genes must have a mean expression above the threshold minmean

Finally, extreme values are capped above/bellow the blunt percentile.

Data are then normalized as a percent dispersion from the mean.

Finally dimensional reduction is implemented by SVD.

Depending on your problem different strategies may be more effective or appropriate.

**CYCLOPS quality metrics**:

Two metrics are used to assess the quality of the CYCLOPS ordering.

compares the total sum of squares error (SSE) of the circular autoencoder in reconstructing characteristic expression patterns with the residual variance/error unexplained by the first principle component. The first principle component analytically reproduces the results of a fully linear autoencoder with a single bottleneck node. Thus compares the improvement in model fit when a circular rather than linear bottleneck node is used. mirrors the definition of the F statistic in nested regression models used to evaluate the inclusion of additional parameters with increasing model complexity. Defining as the autoencoder reconstruction error and as the variance remaining unexplained by the first principal component.

is a measure of the smoothness of the ordered expression trajectory. It compares the mean distance, in expression space, between sequential samples when using a circular ordering to the distance between sequential samples assuming a linear ordering. Given the samples j and associated CYCLOPS phases ( we create circular ordering *c* which is a permutation of the indexes *j* such that for all c. Similarly we define a linear ordering *l* based on the magnitude of the first eigengene so that for all *l.* Denoting as the Eulcidean norm and the as the eigengenes expression profile of sample *j*

)

**Evaluating significance of error metric**: The significance of was assessed by Bootstrap. To create a background distribution for each row *i* of was independently permuted. This removed any fixed phase relationship between the probes while maintaining their marginal distributions. These data were then expressed eigengenes and the CYCLOPS autoencoder was re-trained and the error metric recomputed. This process is repeated to create a background distribution used to assess the significance of .

**Modified Cosinor Regression:** Cosinor regression has long been used to identify circadian processes. The datasets typically analyzed in this way include more than a single cycle so that monotonic expression patterns are not misidentified as circadian. CYCLOPS, however, assigns each sample a phase along a single reconstructed cycle (

To better exclude monotonic trends we compared the fit of a best-fit line with the fit of nested model including both that line and sinusoidal functions of CYCLOPS phase.

For each probe *(i)* we fit the data to two models:

(M1)

(M2)

The are first optimized in M1 by a brute-force search. The F-Test is used to evaluate the null hypothesis 0

The amplitude and acrophase of each probe are given by and using the standard regression model