

Circular_FSA.sas

Suppose A and B are two species of organisms where $\{A_1, A_2, \dots, A_G\}$ is a set of cell cycle genes of species A and $\{B_1, B_2, \dots, B_G\}$ is the set of corresponding orthologs of species B . Suppose the phase angles of the set $\{A_1, A_2, \dots, A_G\}$ satisfy the following relative order around the unit circle: $A_1 \prec A_2 \prec A_3 \prec \dots \prec A_G \prec A_1$, where the notation $A_1 \prec A_2$ refers to “gene A_2 follows gene A_1 in its time to peak expression”, then the program Circular_FSA.sas tests whether the orthologs $\{B_1, B_2, \dots, B_G\}$ satisfy the same relative order as the genes $\{A_1, A_2, \dots, A_G\}$, i.e., it tests the null hypothesis $B_1 \prec B_2 \prec B_3 \prec \dots \prec B_G \prec B_1$ against the alternative that the order is not satisfied. If the null hypothesis is rejected then Circular_FSA.sas identifies the subset of g genes $\{B_{s_1}, B_{s_2}, \dots, B_{s_g}\}$ which have the same relative order of time to peak expression as the corresponding subset $\{A_{s_1}, A_{s_2}, \dots, A_{s_g}\}$ by eliminating genes which are out of order.

1. **Open FSA.sas:** Open FSA.sas using notepad or using SAS.
2. **Inputs:** Provide the following inputs in the order they appear in the SAS code:
 - a. **NAMES:** Specify the names of genes of species B (e.g. *S. pombe*) according to the relative order of their orthologs in species A (e.g. *S. cerevisiae*) in the format given in the following example. Thus, we are testing the null hypothesis that the order of cell cycle genes in species A genes is conserved in species B . Each gene is separated by at least one empty space. If a gene’s name has a “period” in it, then replace the period by “_”. For example, replace SPAC1705.03C by SPAC1705_03C.

Example: In the following we provide the names of fourteen *S. pombe* genes arranged in the same order as their *S. cerevisiae* orthologs. You don’t specify the *S. cerevisiae* genes.

NAMES={ SPAC1705_03C eng1 cdc18 ssb1 cdc22 msh6 mrc1 pol1 psm3 rad21
rhp51 cig2 pol2 mik1};

- b. **DATA0:** This is an input matrix of phase angles of species *B* genes. Rows correspond to the experiment and the columns correspond to the phase angles of species *B* genes arranged according to the order specified in the variable NAMES. Set a missing value to 500. Each row should end with “,” except the last row which ends with “;”.

Example: In the following we provide the phase angles of the above fourteen *S. pombe* cell cycle genes from 10 experiments (rows). The phase angles for each gene may be obtained by using the random periods model (RPM) (Liu et al., 2005) or any other method.

```
DATA0={
4.88 6.03 5.79 0.20 0.22 6.26 6.02 5.62 5.76 0.89 4.93 5.61 5.34 6.26,
2.81 2.91 2.18 2.94 3.26 2.81 2.68 2.00 2.85 1.60 2.25 2.38 1.94 1.71,
0.45 0.12 5.43 0.44 0.45 5.26 5.52 5.93 6.21 4.38 0.66 5.46 0.58 6.04,
3.19 3.03 2.53 3.33 3.57 3.39 3.03 3.01 2.81 3.19 2.47 3.26 2.85 3.03,
2.74 3.88 3.67 3.33 3.91 3.89 3.14 3.57 3.44 3.65 2.42 3.97 3.38 4.30,
1.81 1.77 1.41 1.97 2.15 2.03 1.78 2.06 2.03 1.74 1.50 2.07 500 1.73,
1.33 1.41 1.70 1.81 2.21 1.41 2.30 2.19 1.35 1.96 1.98 1.94 1.37 1.98,
0.96 1.43 1.33 500 1.46 1.29 1.49 1.31 500 1.53 1.66 1.37 1.36 1.38,
1.30 1.71 1.73 2.21 1.79 1.73 1.78 1.82 1.99 1.88 1.86 1.88 1.39 3.07,
2.20 2.60 2.10 2.34 2.70 2.70 2.22 2.69 2.53 2.98 2.28 2.32 500 2.28};
```

- c. **eq:** This input contains information on the equalities in the null hypotheses. That is, some of the cell cycle genes in species *A* (and hence in species *B*) are hypothesized to have same phase angle. Genes which are hypothesized to have same phase angle are entered in groups by providing the gene number of the first and the last member of the group separated by space. Each group is separated by a comma. If none of the genes are hypothesized to have same phase angle then we enter {0 0} for this variable **eq**.

Example: In the above example, we hypothesize that genes (cdc22, msh6, mrc1 pol1 and psm3) have the same phase angle. Similarly (rad21 and rhp1) and (cig2

and pol2) have the same phase angle. Note that cdc22 is the 5th gene in our list and psm3 is the 9th gene in our list. Similarly rad21 is the 10th gene, rhp51 is the 11th gene, cig2 is 12th and pol2 is the 13th gene in our list. Hence in this example we enter:

eq={5 9, 10 11, 12 13};

- d. **orden:** Orden is a vector specifying the order in which the genes are introduced into the FSA algorithm. This vector sets the order in which the genes are introduced in the model by the FSA algorithm according with their periodicity rank. For details refer to Fernandez et al. (2011) Supplementary file.
- e. **Initnum:** Initnum is the number of genes whose periodicity rank in both species A and B is less than 100. The FSA algorithm begins with this set of genes. For details refer to Fernandez et al. (2011) Supplementary file.
- f. **FSAp1, FSAP2, FSA n1 and FSA n2:** These are the input parameters used to decide if a gene is included in a given step of the FSA or not. A gene is rejected if, among all the experimental data used, at least FS n1 of the p-values is lower than FSAP1 or at least FSA n2 of the p-values are lower than FSAP2.

Example: Suppose we chose FSAP1 = 0.2, FSAP2 = 0.3, FSA n1 = 2, FSA n2 = 3. In the above example, we have 10 experiments. For given set of genes, the FSA tests the null hypotheses for data in each of the 10 experiments. This results in 10 p-values. Suppose the 10 p-values are; 0.0001, 0.25, 0.35, 0.40, 0.50, 0.65, 0.8, 0.9, 0.9 and 0.95. We see that X = 1 p-value, namely, 0.0001, is less than FSAP1 = 0.2 and Y = 2 p-values, namely, 0.0001 and 0.25, are less than 0.30. Thus since X < FSA n1 and Y < FSA n2, therefore the null hypothesis regarding the relative order among the set of genes is not rejected.

3. **Outputs:** The outputs consist of the following:

- a. For each step of FSA the program returns the step number, the serial number and name of the gene processed in that step, the sum of circular error (SCE) and the p-values corresponding to each experiment. It prints a value of 1 if the gene is selected or 0 (or a negative number) if the gene is rejected in that step using the FSAP1, FSAP2, FSA n1, and FSA n2 criteria.

- b. Upon completion of the program, FSA prints out the names of all genes that conserve the hypothesized relative order among the two species (labeled as finalgen). It also provides the serial number of genes in the conserved set (labeled as solu).

For illustration purposes, the FSA.sas file contains data and all above input parameters corresponding two data sets discussed in the paper entitled “A core set of signature cell cycle genes with relative order of time to peak expression conserved across species”, namely, the FB and FH core sets.

In the case of FB data, the above input parameters are:

- **eq**={3 7, 8 9, 10 11, 21 23, 25 27, 28 29};
- **orden**={33 3 14 17 13 15 16 20 10 4 12 2 27 7 31 8 28 1 21 35 34 22 5 23 30 26 6 19 29 24 18 11 32 9 25};
- **initnum**=16;
- FSAp1=0.2; FSAp2=0.3; FSAn1=1; FSAn2=2;

In the case of FH data, the above input parameters are:

- **eq**={16 17, 19 21, 22 23};
- **orden**={22 4 3 9 19 20 18 13 24 21 6 14 5 10 16 11 15 12 1 17 8 2 23 7};
- **initnum**=10;
- FSAp1=0.2; FSAp2=0.3; FSAn1=1; FSAn2=2;

References:

1. Fernandez, M., Rueda, C. and Peddada SD (2011). A core set of signature cell cycle genes with relative order of time to peak expression conserved across species. *Under review*.
2. Rueda,C., Fern´andez,M. and Peddada,S.D. (2009) Estimation of parameters subject to order restriction on a circle with application to estimation of phase angles of cell-cycle genes. *J. Amer. Statist. Assoc.*,**104**, 338-347.
3. Liu,D., Umbach,D.M., Peddada,S.D., Li,L., Crockett,P.W. and Weinberg,C.R. (2004). A Random-Periods Model for Expression of Cell-Cycle Genes. *Proc Natl Acad Sci USA*, **101(19)**, 7240-7245.