Identifying Cyclic Genes in Time-Course Gene Expression Studies Using Proc Traj

Gen İfade Süreç Çalışmalarında Proc Traj Kullanarak Döngüsel Genlerin Belirlenmesi

Mehmet KOCAK[a]

[a]Department of Preventive Medicine,
University of Tennessee
Health Sciences Center,
Memphis, TN

Geliş Tarihi/*Received:* 28.07.2015
Kabul Tarihi/*Accepted:* 19.08.2015

Yazışma Adresi/*Correspondence:*
Mehmet KOCAK
University of Tennessee
Health Sciences Center,
Department of Preventive Medicine,
Memphis, TN,
USA/ABD
mkocak1@uhtsc.edu

ABSTRACT Jones and Nagin developed PROC TRAJ to study developmental trajectories for the areas of psychology, sociology, and criminology. We tested the utility of the TRAJ procedure in identifying cyclic (i.e., periodic) genes based on their time-course expression measurements. We have shown through extensive simulations that PROC TRAJ offers practical solutions in identifying gene-sets with different profiles including cyclic patterns over cell division cycles. We then applied the procedure to an S. Pombe gene-expression data and showed that through the use of TRAJ procedure, different sets of periodic genes can be obtained, where truly periodic genes are ranked much higher overall compared to not-periodic genes, thus making their identification easier. As a conclusion, we present convincing preliminary evidence for the utility of PROC TRAJ in cell-cycle gene expression data analysis although its utility must be tested more stringently in future simulation studies as well as by its application to other real-experimental data where identifying cyclic patterns is the primary goal.

Key Words: PROC TRAJ; periodicity; trajectory; time-course experiment

ÖZET Jones ve Nagin psikoloji, sosyoloji ve kriminoloji alanlarında gelişimsel yörüngeleri çalışmak için PROC TRAJ yöntemini geliştirmişlerdir. Süreç ifade ölçümlerine dayalı döngüsel (örneğin; periyodik) genleri belirlemede TRAJ yönteminin faydalarını test ettik. Hücre bölünme döngüleri üzerinde döngüsel modelleri içeren farklı profillerle gen setlerini belirlemede pratik çözümler öneren PROC TRAJ yöntemini kapsamlı simülasyonlarla aracılığıyla gösterdik. Daha sonra bu yöntemi S. Pombe gen-ifade verisine uyguladık ve TRAJ yönteminin kullanımıyla tamamen periyodik genlerin periyodik olmayan genler ile kıyaslandığında daha yüksekte sıralandığında periyodik genlerin farklı setlerinin elde edilebildiğini dolayısıyla belirlenmelerinin daha kolay olduğunu gösterdik. Sonuç olarak, yöntemin faydasının gelecekteki simülasyon çalışmalarıyla birlikte ilk amacın döngü yapılarının belirlenmesi olan diğer gerçek deneysel çalışmalara uygulanması ve daha kesin bir şekilde test edilmesi gerekli olmasına rağmen hücre döngüsü gen ifade veri analizinde PROC TRAJ yönteminin faydası için ilk ikna edici kanıtı sunduk.

Anahtar Kelimeler: PROC TRAJ; dönemsellik; yörünge; süreç deneyi

Turkiye Klinikleri J Biostat 2015;7(2):47-54

Investigating the cyclic behavior of genes during cell cycle led to a substantial body of works in the biology and statistical methodology literature, among which we can include the independent cell cycle experiments conducted by Oliva et al., Peng et al., and Rustici et al.[1-3] The main aim of these research is not only to identify cyclic genes but also describe that cyclic behavior. In an era of the rise of targeted therapies, it is critical to identify genes that have cyclic behavior during the cell division

doi: 10.5336/biostatic.2015-47362

cycle as such findings may help researchers develop and fine-tune therapies that target cells with genes having certain cyclic behaviors of interest.

To search for genes with cyclic patterns, which may be in any shape and form from sinusoidal to spiky expression at a certain phase of cell division, say during the G2 phase or M phase, the researcher needs time-course measurements of expression for every gene of interest. This is made possible with the microarray technology where the expression of thousands of genes can be measured simultaneously. Once expression of genes are obtained through such technology at selected pre-determined time points to cover multiple cell cycles, the desired time-course data would be obtained.

Several methods were proposed to identify genes with cyclic behavior, among which we can count Fisher's G-test by Fisher, the permutation test described by Lichtenberg et al., and an empirical Bayesian approach by Kocak et al.[4-6] Kocak et al. showed through extensive simulation that their Bayesian approach was superior to Fisher's G-test and the Permutation method as a test for periodicity against not only noise but also some other 'non-periodic' patterns.[6] However, all these tests operate on each gene independently and do not aim at describing what type of cyclic behavior a given gene has. To address this issue, in this study, we put to trial a new SAS procedure by Jones and Nagin named PROC TRAJ, whose base was established by, to identify and describe genes with cyclic behavior through extensive simulations.[7-9]

In Section 2, we briefly describe SAS TRAJ procedure, followed by the simulation design in Section 3. We provide the simulation results in Section-4 and provide a real-data application in Section-5. We then end with some discussions.

METHOD

PROC TRAJ is a SAS procedure developed by Jones and Nagin to study developmental trajectories, utilizing a semiparametric, group-based modelling approach where the underlying model is a mixture probability distributions with a multinomial modeling strategy.[7,8] The researcher decides how many trajectories may describe the data at hand, and the authors also provided a strategy to choose the 'best' number trajectories based on the log-likelihood values of the resulting models.

TABLE 1: Call Syntax of PROC TRAJ.

```
PROC TRAJ DATA=SAMPLEDATA
OUT=<Output data with all variables in the model including final GROUP assignments>
OUTPLOT=<Output data with final trajectories by time variable>
OUTSTAT=<Output data with parameter estimates for all trajectories>
OUTEST=<Output data with all covariance estimates as well as log-likelihood values>;
VAR D1-Dn; /* Dependent Variable laid horizontally over n-time points */
INDEP T1-Tn; /* Time variable laid out horizontally over n-time points */
MODEL CNORM; /* Censored Normal Model */
MIN 0; /* User defined Lower Censoring Point */
MAX 10; /* User defined Upper Censoring Point */
NGROUPS K; /* To specify K trajectories for the data */
ORDER d1 d2 d3 … dK; /* Polynomial degree for each of the K-trajectories */
RUN;
```

The procedure is available as a free download and its installation is also straightforward and an a step-by-step approach is provided by the authors (https://www.andrew.cmu.edu/user/bjones/index.htm). The call syntax of the procedure is as follows (Table 1):

The authors also developed a SAS macro program (%TRAJPLOT) to plot the resulted trajectories to visually inspect what types of patterns the data at hand displays based on the K trajectories selected, which may guide the user to increase or decrease the number of trajectories to be studied.

We present a simple run for illustration purposes below (Table 2):

We have captured the results as they appear on the SAS output window and the figure from the %TRAJPLOT macro (Figure 1A, Figure 1B).

In this sample run with two trajectories, we see that one of two trajectories is periodic, and the other seems to be just noise. We then ran the same example with three trajectories (Figure1C) and see that the noise part can actually be split into two trajectories of noise, one with higher, one with lower noise level.

TABLE 2: A sample run of PROC TRAJ.

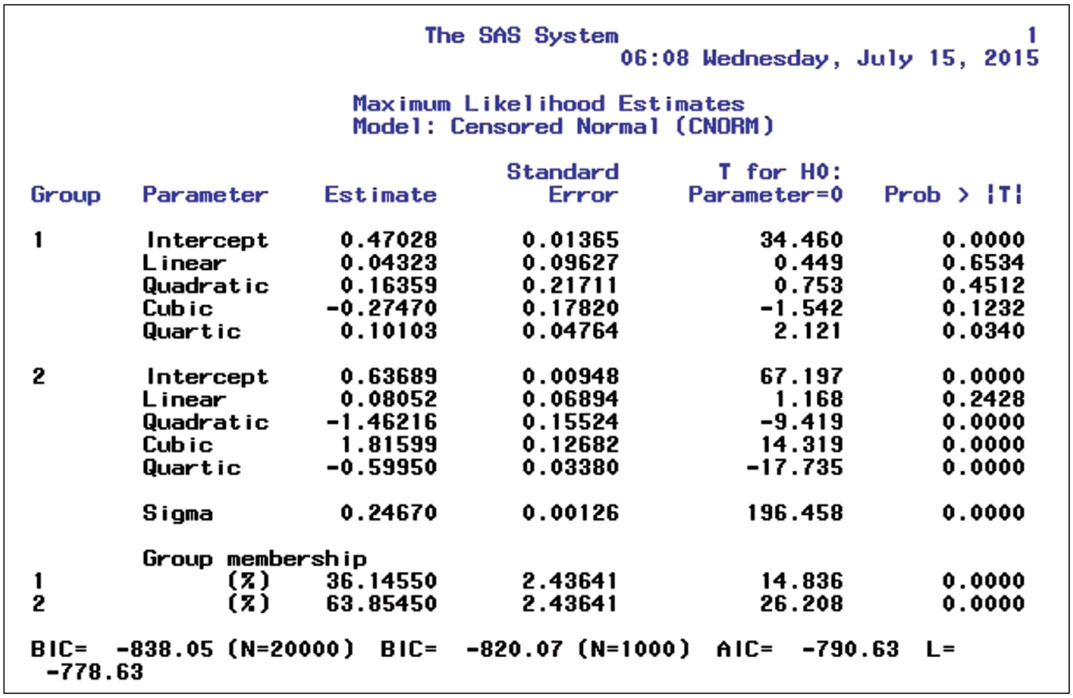
```
PROC TRAJ DATA=alltraj_sinusoidal
OUT=period.OF OUTPLOT=period.OP OUTSTAT=period.OS;
VAR e1-e20; *Expression Variables;
INDEP T1-T20; *Time Variables;
MODEL CNORM; *Censored Normal Model;
MIN -10; **Lower Censoring Point;
MAX 10; *Upper Censoring Point;
NGROUPS 2; *No. of Groups fitted;
ORDER 4 4;*4-degree Polynomial for each trajectory;
RUN;
%TRAJPLOT (period.OP, period.OS,"Expression Trajectories",, "Expression","Time");
```

SIMULATION STUDY

To test for the efficiency and feasibility of PROC TRAJ in identifying cyclic profiles (e.g., periodic genes), we have used the combination of cyclic and non-cyclic patterns that Kocak et al. used, which is listed below:[6]

- Cyclic Patterns (n=5): Sinusoidal, Spike, Double Spike, Beta, Bi-Model.
- Non-Cyclic Patterns (n=5): White Noise, Linear, Low-Plato, High-Plato, Random Spikes.
- Each pattern was represented by 500 randomly generated profiles that cover two complete cycles of data, where each cycle had 10 equidistant measurements.
- Number of trajectories (n=6): 3, 4, 5, 6, 7, 8
- Degree of Polynomials (n=5): 1, 2, 3, 4, 5

The total number of combinations of the experimental parameters can go up exponentially to more than 750 scenarios, in fact, even higher, if we start combining subsets of cyclic and non-cyclic patterns together, which also has relevance as one may want to assess whether or not the proposed approach can different a given cyclic pattern, say Sinusoidal, not only against Noise Pattern but Noise and Linear pattern combined. Also, one may what to use different combination of polynomial degrees for a given list of trajectories rather than using the same degree across the board. After each simulation run, the relevant output datasets were saved for further data processing.



```
                              The SAS System                              1
                                        06:08 Wednesday, July 15, 2015

                         Maximum Likelihood Estimates
                         Model: Censored Normal (CNORM)

                                          Standard       T for H0:
Group    Parameter     Estimate            Error     Parameter=0     Prob > |T|

1         Intercept      0.47028          0.01365          34.460         0.0000
          Linear         0.04323          0.09627           0.449         0.6534
          Quadratic      0.16359          0.21711           0.753         0.4512
          Cubic         -0.27470          0.17820          -1.542         0.1232
          Quartic        0.10103          0.04764           2.121         0.0340

2         Intercept      0.63689          0.00948          67.197         0.0000
          Linear         0.08052          0.06894           1.168         0.2428
          Quadratic     -1.46216          0.15524          -9.419         0.0000
          Cubic          1.81599          0.12682          14.319         0.0000
          Quartic       -0.59950          0.03380         -17.735         0.0000

          Sigma          0.24670          0.00126         196.458         0.0000

          Group membership
1              (%)    36.14550          2.43641          14.836         0.0000
2              (%)    63.85450          2.43641          26.208         0.0000

BIC=  -838.05 (N=20000)  BIC=  -820.07 (N=1000)  AIC=  -790.63  L=
 -778.63
```

FIGURE 1A: Model results of the simple PROC TRAJ run.

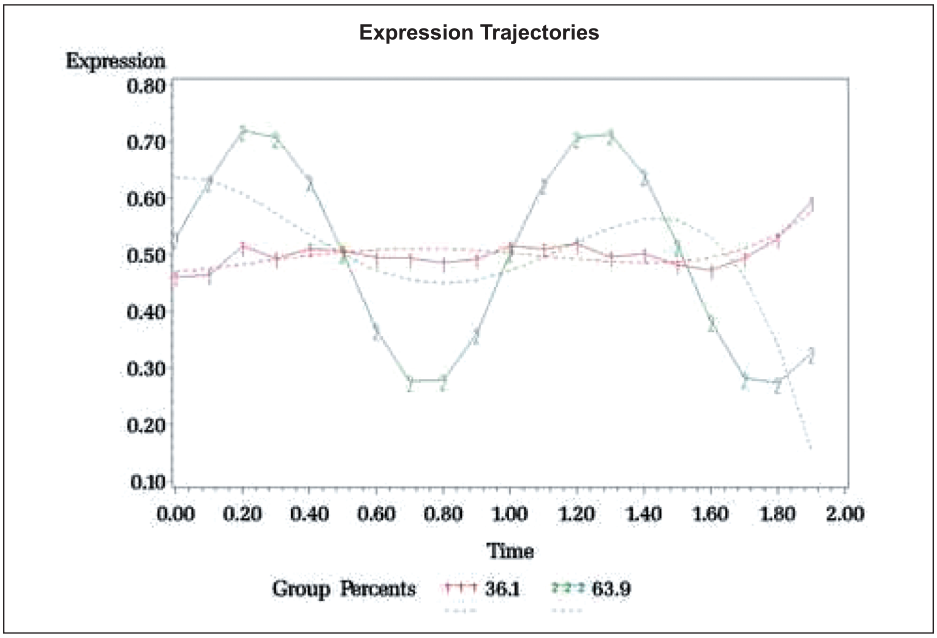


FIGURE 1B: A sample run of PROC TRAJ and %TRAJPLOT macro.

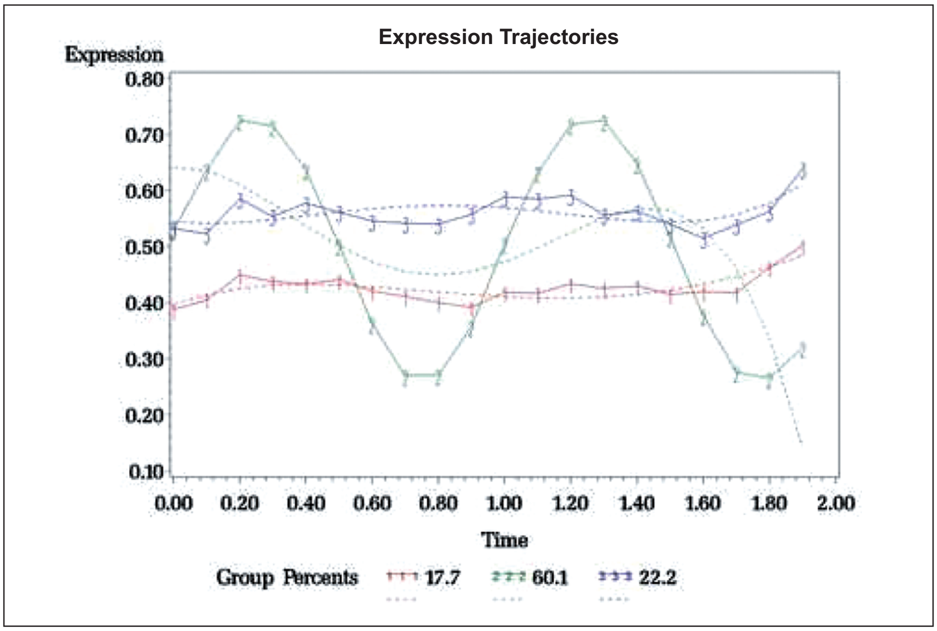


FIGURE 1C: A sample run of PROC TRAJ and %TRAJPLOT macro with three trajectories.

RESULTS

We summarized the simulation results as sensitivity, specificity, and total accuracy. The entire simulation results were provided in Table 3.

We present the total accuracy summaries in a graphical form in Figure 2.

Overall, we conclude that using polynomial degree of 1 as well as polynomial degree of 2 for each trajectory does not provide a compatible accuracy measure for any given pattern against all non-cyclic patterns including noise. Considering the above comparisons include each periodic pattern against all five non-periodic patterns includ-

TABLE 3: Sensitivity and specificity results.

No. of		1st Degree		2nd Degree		3rd Degree		4th Degree		5th Degree	
Trajectories		Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Sinusoidal	3	0.98	0.80	1.00	0.43	0.99	0.79	0.99	0.78	0.99	0.80
	4	0.98	0.80	0.99	0.80	0.99	0.81	0.99	0.81	0.99	0.84
	5	0.54	0.89	0.99	0.80	0.99	0.81	0.99	0.81	0.99	0.84
	6	0.51	0.89	0.98	0.81	0.98	0.82	0.97	0.95	0.99	0.99
	7	0.74	0.94	0.60	0.90	0.97	0.94	0.97	0.95	0.99	0.99
	8	0.74	0.94	0.60	0.90	0.97	0.94	0.88	0.97	0.99	0.99
Spike	3	1.00	0.61	1.00	0.62	1.00	0.63	1.00	0.64	1.00	0.64
	4	1.00	0.64	1.00	0.65	1.00	0.67	1.00	0.68	1.00	0.68
	5	1.00	0.66	1.00	0.85	1.00	0.85	1.00	0.68	1.00	0.87
	6	1.00	0.66	1.00	0.85	1.00	0.85	1.00	0.85	1.00	0.68
	7	1.00	0.69	1.00	0.87	1.00	0.87	1.00	0.88	1.00	0.90
	8	0.99	0.70	1.00	0.85	1.00	0.87	1.00	0.90	1.00	0.93
Double Spike	3	1.00	0.60	1.00	0.61	1.00	0.62	1.00	0.62	1.00	0.62
	4	1.00	0.63	1.00	0.64	1.00	0.66	1.00	0.66	1.00	0.66
	5	0.97	0.66	1.00	0.83	1.00	0.82	1.00	0.83	1.00	0.83
	6	0.62	0.89	1.00	0.83	1.00	0.82	1.00	0.84	0.99	0.86
	7	0.63	0.90	1.00	0.83	0.99	0.85	0.99	0.85	0.99	0.86
	8	0.63	0.89	0.98	0.86	0.99	0.86	0.98	0.87	0.99	0.89
Beta	3	1.00	0.60	0.99	0.61	1.00	0.57	1.00	0.72	1.00	0.74
	4	1.00	0.64	0.99	0.67	1.00	0.74	1.00	0.75	1.00	0.75
	5	0.97	0.92	1.00	0.86	1.00	0.96	1.00	0.97	1.00	0.99
	6	0.97	0.92	1.00	0.86	1.00	0.77	1.00	0.98	1.00	0.78
	7	0.97	0.94	1.00	0.92	1.00	0.98	1.00	0.98	1.00	0.99
	8	0.97	0.94	1.00	0.94	1.00	0.98	1.00	0.98	1.00	0.99
Bi-Modal	3	0.94	0.59	1.00	0.44	1.00	0.65	1.00	0.65	1.00	0.65
	4	0.92	0.62	1.00	0.73	1.00	0.72	1.00	0.72	1.00	0.75
	5	0.99	0.86	0.99	0.80	0.99	0.80	1.00	0.81	1.00	0.86
	6	0.99	0.86	0.99	0.80	0.99	0.80	1.00	0.81	1.00	0.94
	7	0.99	0.86	0.99	0.80	1.00	0.89	1.00	0.89	1.00	0.95
	8	0.97	0.88	1.00	0.86	1.00	0.89	1.00	0.89	1.00	0.95

ing noise, ideally we expect that the number of trajectories must be at least 6. The above results confirm this expectation in the sense that all the simulation scenarios where we had less than six trajectories specified resulted in lower total accuracy overall than the scenarios with ≥6 trajectories. These findings are also highly encouraging in that as we increase the number of trajectories to higher number than six as we would not usually know the number of true trajectories in reality, the total accuracy still increases in differentiating the periodic trajectory from those for non-periodic patterns.

In this particular simulation study, we achieved the highest total accuracy for Sinusoidal Pattern and Beta Patterns. It was a bit more difficult to achieve high total accuracy (say, >90%) for Spike and Double Spike patterns, which is expected as the indication of periodicity is evident only by a single or two spikes in every cycle, rest of which is just noise.

APPLICATION TO S.POMBE CELL CYCLE GENE EXPRESSION EXPERIMENTS

As an illustration, we have applied PROC TRAJ to one of the time-course gene expression experi-

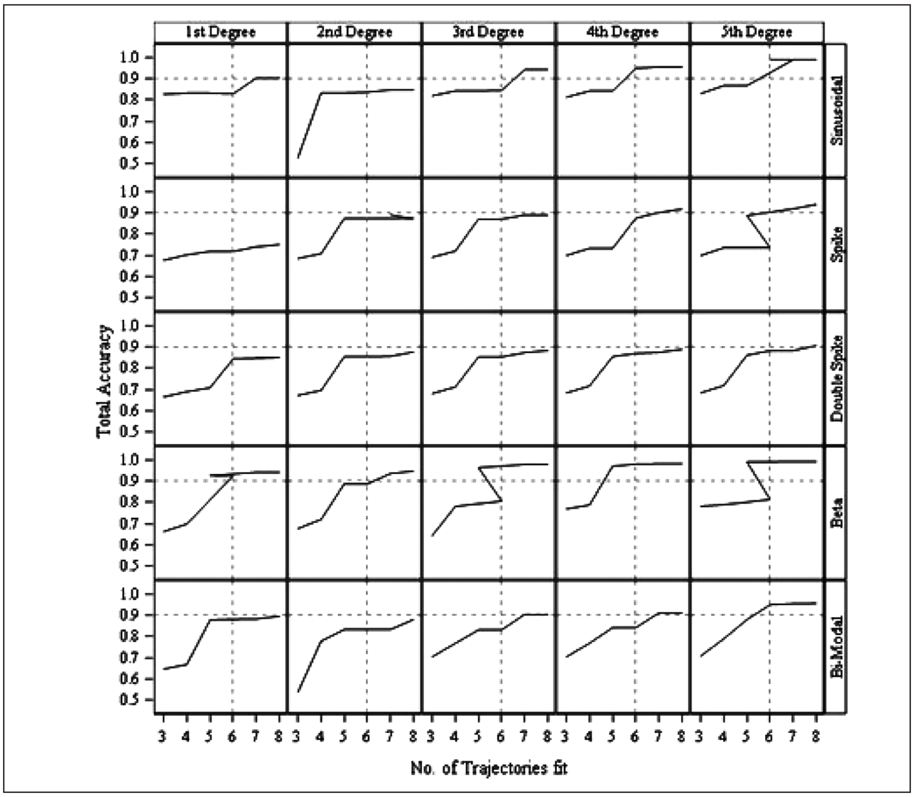


FIGURE 2: Total accuracy measure by Pattern and Polynomial Degree.

ments by Peng et al[2]. In this experiment, expression data was available for 4,929 genes. We fit 10 trajectories to the data and obtained the following trajectories (Figure 3).

Of these 10 gene expression trajectories, it is clear that Trajectories 4, 6, 8 and 10 show a strong periodic pattern. These four trajectories successfully captured 34 out of the benchmark set of 38

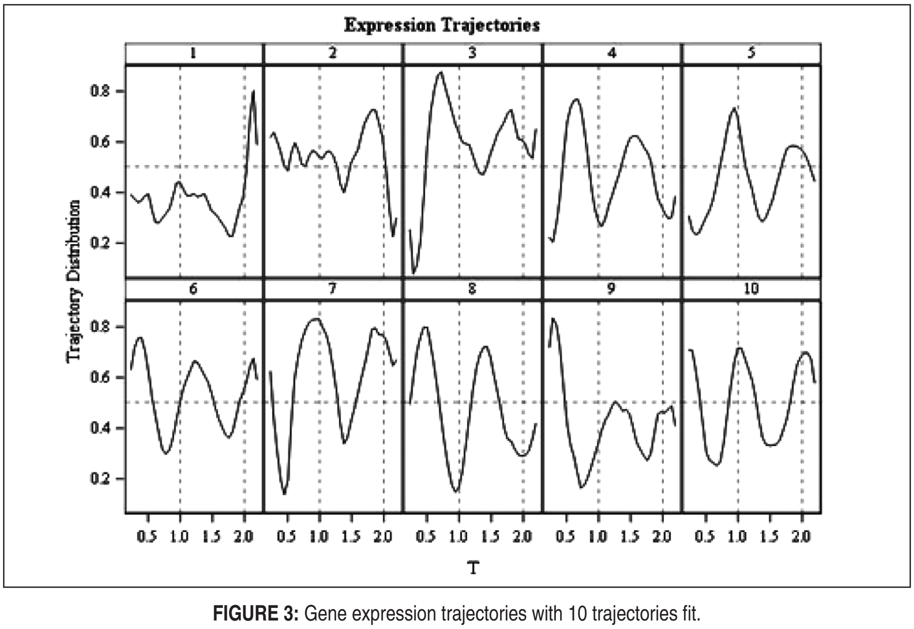


FIGURE 3: Gene expression trajectories with 10 trajectories fit.

periodic genes reported by Marguerat et al.[10] When we ranked the genes based on their group membership probabilities overall and within each trajectory, then the median rank for these 34 truly periodic genes was 34 within their respected trajectories and 304 overall, while the median rank for the remaining genes called as 'periodic' by the above procedure was 276 within their respected trajectories and 2392 overall.

DISCUSSION

Jones and Nagin developed PROC TRAJ to study developmental trajectories for the areas of psychology, sociology, and criminology, where the number of measurements over time is not expected to be very high.[7,8] We tested their procedure in terms of feasibility (i.e., computing time) as well as efficiency in successfully identifying cyclic genes and it proved to provide a rich and efficient platform for such aims.

We have shown by extensive simulations the utility of PROC TRAJ in identifying different cyclic patterns in time-course data, specifically targeting its use in time-course gene-expression data. We then applied the procedure to an S. Pombe gene-expression data and showed that TRAJ procedure offers practical solutions for the identification of cyclic patterns by clustering genes that behave similarly following a specific average pattern, which can be easily visually inspected. The procedure is vulnerable to this visual inspection in that it may bring in subjectivity as to what can be considered as 'cyclic'. For example, in the example we used for illustration (Figure 3), another pair of eyes may consider Trajectory-3 or even Trajectory-7 as potentially periodic based on the location of the peaks.

The choice of the degree of polynomials for each trajectory can be considered as another area of subjectivity. Although our simulations showed that Polynomial Degree-5 was the best among all scenarios considered in our simulations, lower degree polynomials may work equally as well. The procedure only supports up to 5-degree polynomials and we suspect that higher degree polynomials could perform better by definition in identifying other potential trajectories, which, however, may increase the subjectivity issue discussed above much stronger as the resulting trajectories get finer and finer with higher degree polynomials.

As a general clustering issue, choosing the right number of trajectories may pose another practical challenge for the researchers as the true number of underlying trajectories may not be necessarily known in a given experiment. For small number of trajectories, the true underlying trajectories start blending with each other and the true signal start getting blurry; on the other hand, for large number of trajectories, another type of blending potentially occurs with finer trajectories making the final decision process more difficult as well as their potentially becoming 'similar' to other fine trajectories as an off-shoot of a different underlying trajectory. Jones and Nagin describe a procedure based on the change in BIC values obtained as the number of trajectories increase; however, their approach may not complete address the issue when the sample size and the number of measurements increase for each subject (e.g, genes).[7,8] This issue must be studied perhaps through simulations in the setting of periodicity testing with 2 or more cycles of data.

Finally, especially in the gene expression setting, the position of the peak of the cycle for each gene does not necessarily need to be synchronized, which can only be resolved by increasing the number of trajectories to capture cyclic patterns whose peaks occur at different times during a complete cycles. The researcher should keep this in mind and potentially synchronize the expression data, if possible, before applying PROC TRAJ for the best performance.

As a conclusion, we have shown both based on simulation as well as on a true cell-cycle gene expression data that PROC TRAJ offers practical solutions to identify gene-sets that have similar expression profiles within their sets, yet behave totally differently when the gene-sets are compared with each other. PROC TRAJ can be downloaded freely (https://www.andrew.cmu.edu/user/bjones/

index.htm), and its installation and use are straightforward. We plan to further our investigation on the utility of this procedure in the analysis of long time-course profiles where cell-cycle gene expression data, or circadian rhythm data, can be considered as immediate application areas.

REFERENCES

1. Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, Skiena S, et al. The cell-cycle regulated genes of Schizosaccharomyces pombe. PLoS Biol 2005;3(7):e225.
2. Peng X, Karuturi RK, Miller LD, Lin K, Jia Y, Kondu P, et al. Identification of cell-cycle-regulated genes in fission yeast. Mol Biol Cell 2005;16(3):1026-42.
3. Rustici G, Mata J, Kivinen K, Lió P, Penkett CJ, Burns G, et al. Periodic gene expression program of the fission yeast cell-cycle. Nat Genet 2004;36(8):809-17.
4. Fisher RA. Tests of significance in harmonic analysis. Royal Society 1929;125(796):54-9.
5. de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S. Comparison of computational methods for the identification of cell-cycle-regulated genes. Bioinformatics 2005;21(7):1164-71.
6. Kocak M, George EO, Pyne S, Pounds S. An empirical bayes approach for analysis of diverse periodic trends in time-course gene expression data. Bioinformatics 2013;29(2): 182-8.
7. Jones BL, Nagin DS, Roeder K. A SAS procedure based on mixture models for estimating developmental trajectories. Sociol Method Res 2001;29(3):373-93.
8. Jones BL, Nagin DS. Advances in group-based trajectory modeling and a sas procedure for estimating them. Sociol Method Res 2007;35(4):542-71.
9. Nagin DS. Analyzing developmental trajectories: a semiparametric, group-based approach. Psychol Methods 1999;4(2):139-57.
10. Marguerat S, Jensen TS, de Lichtenberg U, Wilhelm BT, Jensen LJ, Bähler J. The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. Yeast 2006;23(4):261-77.