A Problem from Genetics and the Singular Value Decomposition

Stephen L. Campbell* and Matthew A Campbell†

December 7, 2018

Abstract

The singular value decomposition is widely used for applications such as least squares. The question is whether a period three peak pattern existed in certain species specific genetic data. The pattern was expected to be nonmonotonic and not trigonometric. To illustrate the technique data is used from human, mice, and rice genetic data. The biological significance is also discussed. This example both provides a different type of application of the singular value decomposition which will be of interest to educators and also introduces an approach for searching for certain types of repeating patterns in biological data.

1 Introduction

The singular value decomposition, or SVD, is widely used in a number of ways in applied mathematics from least squares data fitting to discussing the condition numbers of algebraic equations [7]. This paper discusses its use in a different setting that arose in examining genetic data. The idea could also find use in solving related biological problems. We will first describe the basic mathematical problem and its importance in very general terms. Then there is a quick review of the singular value decomposition, Then it will be applied to the specific problem at hand. While we have looked at eight different species, we will consider three species here. Finally there is a conclusion with some comments.

The basic questions are as follows. We have a finite sequence of non-negative integers $x = \{x_1, \ldots, x_M\}$ where M is often greater than 100, sometimes much greater than 100. There is also great variability in the magnitude of a given x_i. Is there a pattern of period 3 peaks that plays a dominant role in the sequence of numbers? Also, how long does this pattern persist?

For a given species there are two types of data which we are interested in, AA and AD. In referring to the data sets we will use terms like Mouse AA or Mouse AD for example. One typical set of data is plotted in Figure 1. This particular data is from a mouse. The left side of Figure

*Department of Mathematics, North Carolina State University, Raleigh, NC 27695. slc@ncsu.edu.
†Genus, Madison, WI.

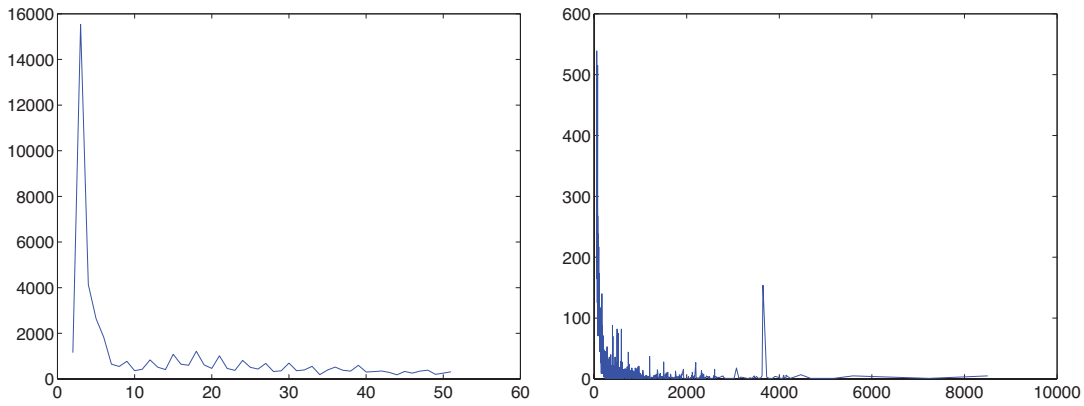

Figure 1: Mouse AD data. First 50 points (left) rest of data (right).

1 shows the first 50 data points, the right side shows the remaining values. Several things are immediate. There is a large variance in size with the largest values at the beginning. There is a long tail to the right with very small and sometimes zero values. The pattern of 3, if it exists, clearly is not a simple trignometric function nor is there a constant rate of decay.

Prior research to determine periodicity in biological sets used Maximum Entropy Spectral Analysis, or MESA [3]. However, MESA was not appropriate in our case because the AA and AD data sets had a strong initial maximum followed by significant dampening along the data series. The Fast Fourier Transform (FFT) was also discounted from this analysis due to the dampening within the data set as well as the following factors: 1) the data set exhibits a bias (i.e. exclusively non-negative) whereas the FFT is centered on zero, 2) the expected period is three and this leads to a frequency of one-third which approaches the end of the frequency interval for FFT (and yields only two interior data points per period), and 3) the FFT is based upon cosines which will lead to an excess of terms in the expansion in this situation. For those reasons, the Singular Value Decomposition (SVD) was chosen to evaluate periodicity in the AA and AD data sets.

2 The Singular Value Decomposition

Suppose that A is an $m \times n$ matrix of real numbers. The singular value decomposition of A is available in essentially all software packages that deal with matrices, we note only commercial software MatLab and the open source software SciLab and OML to name a few [1, 2, 6]. The SVD is defined for any sized matrix. It is U, Σ, V where

$$A = U \begin{bmatrix} \Sigma & 0_1 \\ 0_2 & 0_3 \end{bmatrix} V^T. \tag{1}$$

Here U, V are both orthogonal matrices. They are square and their rows (or columns) form an orthonormal basis. 0_i are appropriate sized matrices of zeros. For a particular matrix either $[0_2\ 0_3]$

or $\begin{bmatrix} 0_1 \\ 0_3 \end{bmatrix}$ can be missing. Matrix Σ is a block diagonal matrix with positive entries σ_i and $\sigma_i \geq \sigma_{i+1}$,

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & & & \sigma_r \end{bmatrix}.$$

The σ_i are called the singular values of A.

The two key facts we will use are that the first r columns of U, which we denote $\{U_1, \ldots, U_r\}$ form an orthonormal basis of the column space of A. That is, the set of all linear combinations of columns of A. There are many ways to generate an orthonormal basis for a column space but the ones from the SVD have a special property. The singular values tells us the relative importance of each U_i in accounting for the columns of A. The larger σ_1 is in relation to the other singular values the more important U_1 is in accounting for the columns of A.

Suppose that a is a vector in the span of the columns of A. Then

$$a = \sum_{I=1}^{r} a_i U_i,$$

where $a_i = a^T U_i$ where a^T is the transpose of a. Letting $\|a\| = \sqrt{a^T a}$, we have $c_i = a_i/\|a\|$ can be used to measure how much of a comes from U_i since $1 = \sum_{i=1}^{r} c_i^2$. We shall be interested in c_1.

3 The Mathematical Problem and Its Solution

The sequences of interest here come from alternative acceptor (AA) and alternative donor (AD) isoforms. These data sequences were constructed from publicly available transcript data. We considered eight eukaryotic species in our prelminary work but in this paper we look at Homo sapiens (human), Mus musculus (mouse), and Oryza sativa (rice).

For a given data set X, the first few values always are very large and as will be explained later are not important for our purposes here. Accordingly we delete the first few entries. The number of deleted entries is called Chop. The results in this paper use Chop=6. Then the remaining data was cut into segments of length SqL where SqL was a multiple of 3. SqL needed to be large enough to show the period three behavior if it existed but small enough to produce useful results. Experimentation showed that the results were not very sensitive to this number so we shall use 15 here. The first N_1 segments are used to compute the dominant mode using the SVD. Here we use $N_1 = 4$. The singular vector that goes with the largest singular value is denoted U_1. Then we examine how much each of the first N_2 segments are accounted for by this dominant mode. Here we use $N_2 = 12$ since some of the data sets were not much longer than this. Since the size of segments decreased as we moved down the data, and we were looking for a geometric pattern, the segments were all normalized to have norm one before applying the SVD. For a given data set N_S is the number of initial segments of length 15 that do not have a zero entry. This is one measure of how much of that data is available to be analyzed by this approach.

In order to more clearly show when a pattern was significant we decided to modify the c_1 value. All of our segments have nonzero entries which creates positive inner products. We ran a large number or inner products of randomly generated vectors using a uniform distribution. We found that while theoretically 0 was the lower bound over enough trials, in practice 0.6 was a good lower bound. Accordingly in what follows we plot $c_1 - 0.6$ for all segments. Values close to 0.4 show a high correlation to U_1 and values close to zero or negative show a very weak correlation. To simplify the graphs and make comparison easier we set $c_1 - 0.6$ to zero if it is negative.

To serve as a comparison we randomly generated a sequence of nonzero numbers and performed the analysis. The results are in Figure 2. Two things should be noted about Figure 2. First the

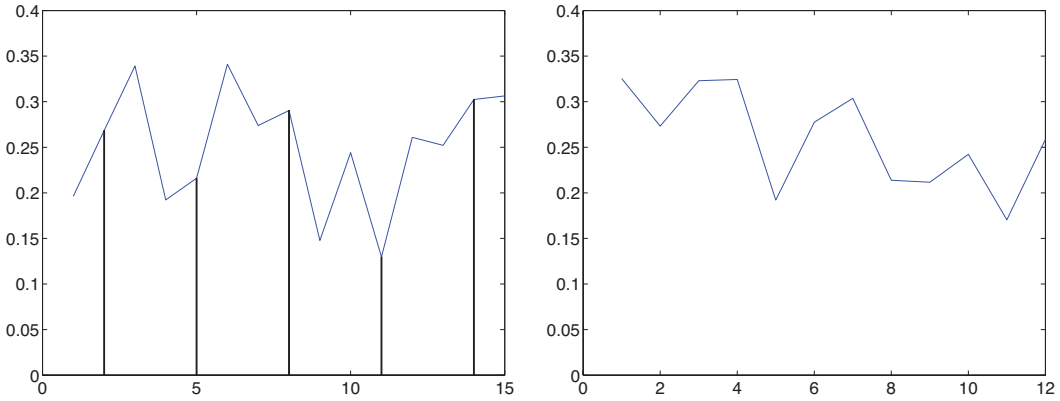

Figure 2: Analysis of randomly generated sequence. U_1 on the left and coefficients on the right.

graph on the left does not have a dominant period 3 pattern. Secondly, the graph on the right shows that U_1 does not consistently give a much larger than expected per cent of the coefficient.

By comparison consider the same analysis applied to the mouse data. Figure 3 is for the mouse AA data and Figure 5 is the Mouse AD data.

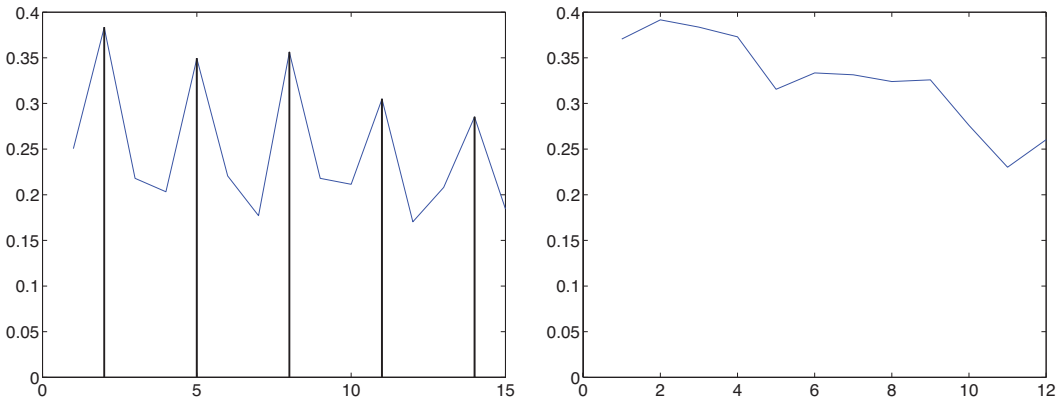

Figure 3: Analysis of mouse AA data. U_1 on the left and coefficients on the right. U_1 peaks clearly exhibits a period 3 pattern. $N_S = 15$.

The right hand side of Figure 3 shows that the period 3 peak pattern still plays an important role in what is observed. Out around the 5th segment the normalized inner product is above 0.9. However, if we just look at the raw data, this is not obvious. Figure 4 shows what the data in the 5th segment looks like when normalized. It is not a nice period three pattern like in Figure 3. The the SVD has found a period three pattern that is a large part of the fifth segment but this is not obvious just from looking at the fifth segment.

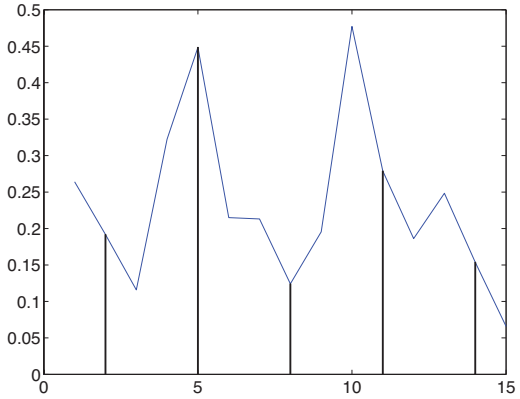

Figure 4: Mouse AA data on the 5th segment.

Looking at the mouse AD results in Figure 5 we see similar behavior. Note that the dominant mode we have found in not monotonically decreasing.

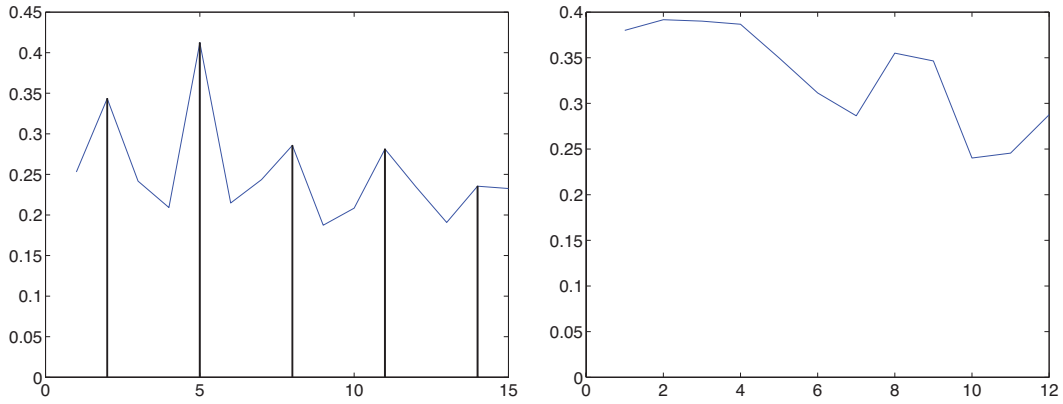

Figure 5: Analysis of mouse AD data. U_1 on the left and coefficients on the right. U_1 peaks clearly exhibits a period 3 pattern. $N_S = 18$.

On the other hand suppose we consider the rice AA data. Rice is known to have a much more flexible genome.

Two things are immediate in Figure 7. We see that the dominant mode on the left does not have a period 3 pattern. However, as the right figure shows, the pattern in U_1 is repeated for rice

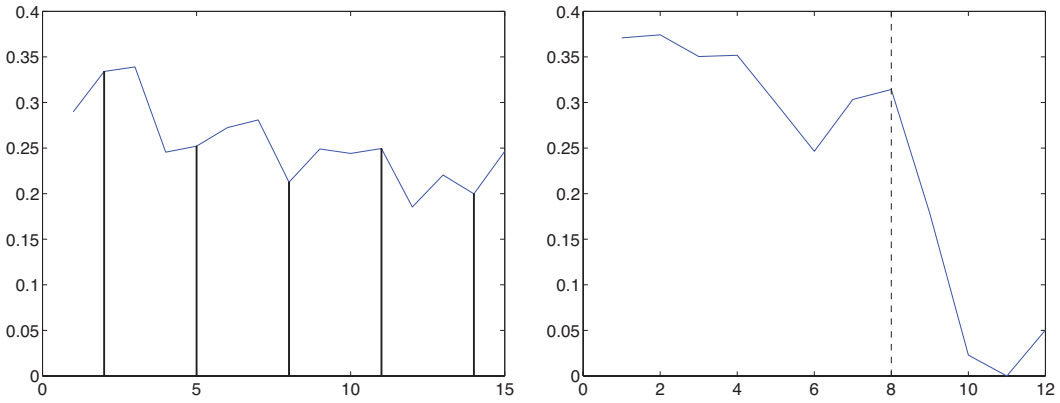

Figure 6: Analysis of rice AA data. U_1 on the left and coefficients on the right. $N_S = 8$

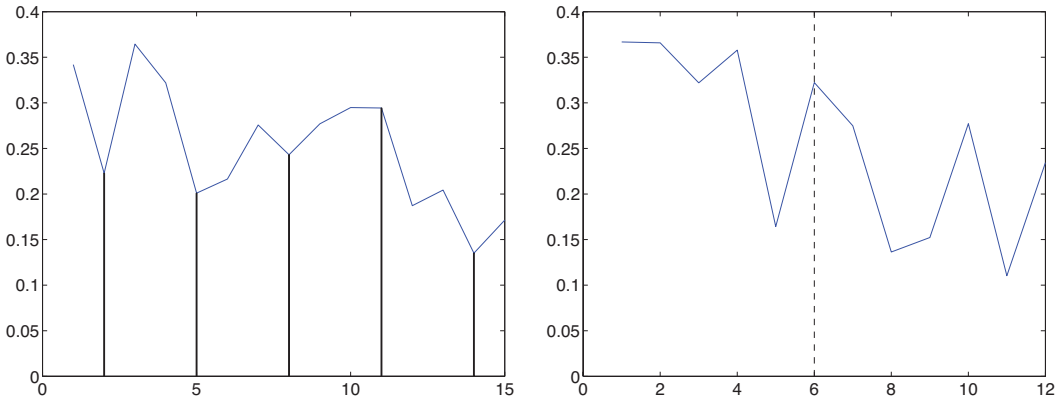

Figure 7: Analysis of rice AD data. U_1 on the left and coefficients on the right. $N_S = 6$

AA as shown by the large coefficient up until N_S. The reason for this persistence is an interesting question that lies outside the scope of this paper.

As a second example we will consider human AA and AD data. The results are in Figures 8 and 9. The human AA has a strong period 3 pattern that persists at a very high level out to 12 segments. The human AD data is almost period 3 with the peak slightly displaced at 10. This pattern also strongly persists out to 12. As noted the first few values of human AA are very large. If we rerun the human AA with Chop=0 we get Figure 10. The left side shows a strong period 3 pattern. The right side is interesting. Recall it is looking at normalized coefficients. Thus the pattern on the left is only a small part of the first segment of data, but from then on it is a large part of the remaining data segments.

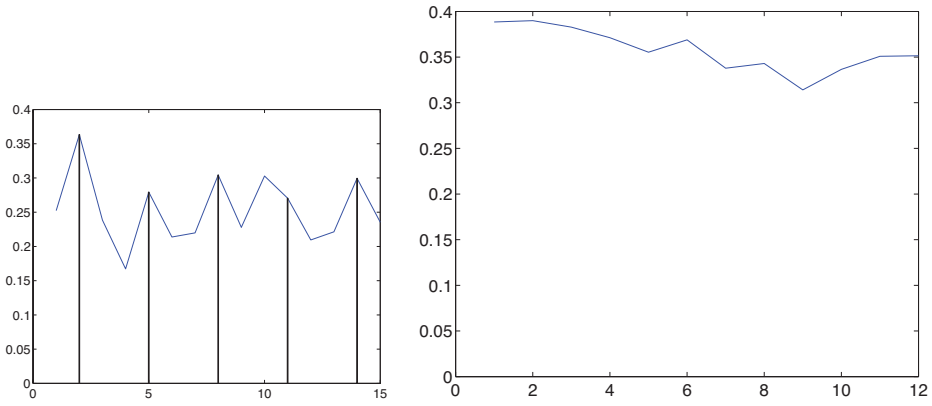

Figure 8: Analysis of human AA data. U_1 on the left and coefficients on the right. $N_S = 31$

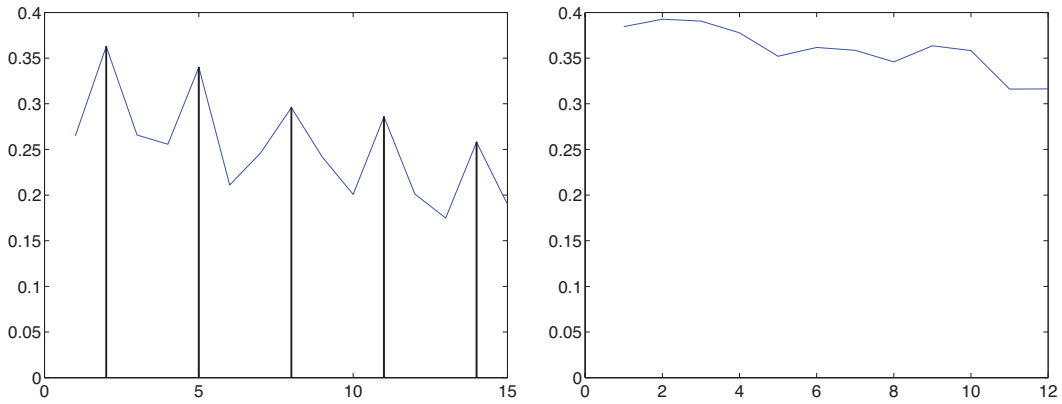

Figure 9: Analysis of human AD data. U_1 on the left and coefficients on the right. $N_S = 31$

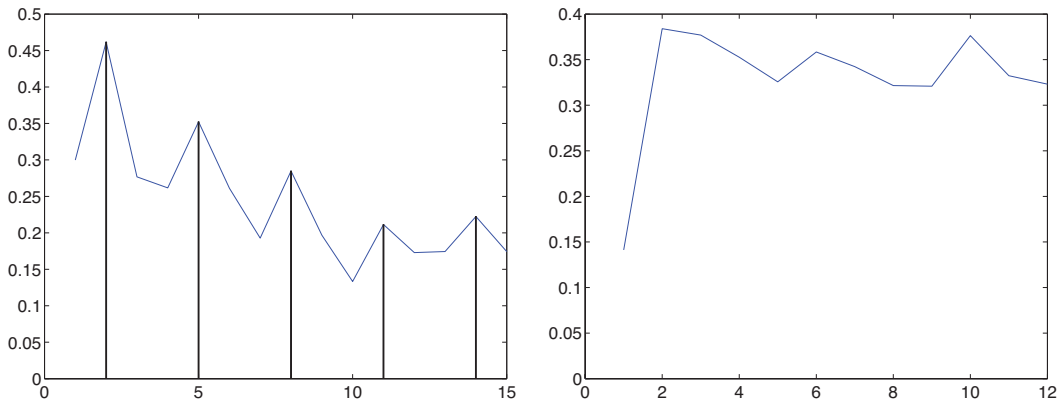

Figure 10: Analysis of human AA data with Chop=0. U_1 on the left and coefficients on the right. $N_S = 31$

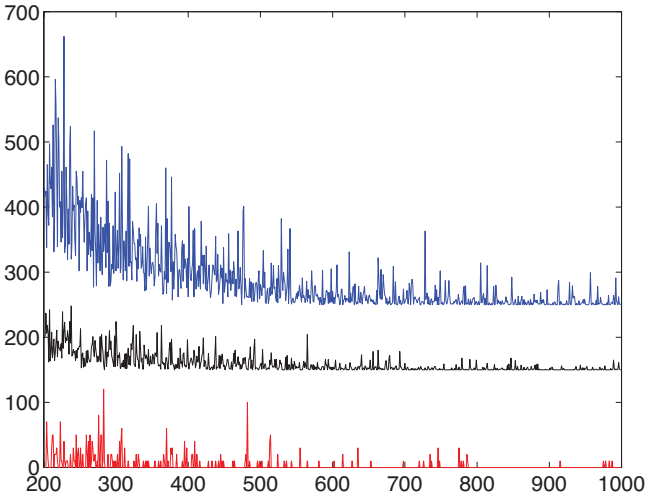

Figure 11: The three data sequences after first 200 chopped. Mouse AD + 250 (top), human AD + 150 (middle), 5 times rice AD (bottom).

4 Conclusion

The alternative acceptor (AA) and alternative donor (AD) isoforms are frequently observed alternative splicing isoforms for a range of eukaryotic species. In this analysis, transcripts from three diverse eukaryotic species were aligned to their respective genomic sequences using the Program to Assembly Spliced Alignments (PASA) to identify AA and AD isoforms in each genome. The AA isoforms have a higher frequency than the AD isoforms for all of the genomes indicating a fundamental difference in splice site selection for these two isoforms. Notably, a periodicity of three basepairs was observed in both the AA and AD data sets except for rice. Singular value decomposition (SVD) revealed that a significant and strong periodicity of three was observed in all species save rice. This period is of particular interest as this is the length of a codon and would preserve the conceptual translational frame of the transcript. Therefore this periodicity of three, which would preserve the translational frame, are retained across a diverse set of eukaryotes. The exceptional behavior of rice is hypothesized to be the result of a relatively dynamic genome prone to segmental duplications and extensive paralog formation. SVD, in conjunction with PASA, represents a novel application for signal analysis to reveal periodicity in complex biological data sets.

Similar results were observed for several species not reported here. The broad conservation of this periodicity suggests that evolutionary pressures have favored the maintenance of the translational frame for a subset of the AA and AD class in a non-random fashion. The outlying species in this analysis, rice, displays no significant periodicity for either the AA or AD data sets. Given that the rice genome has undergone extensive segmental duplication and tandem duplications in recent geologic time, evolutionary adaptation to maintain the translational frame may have yet to occur among these diversifying paralogous sequences [5, 8, 9, 10, 4, 11]. The adaptation and validation of SVD to identify periodicity in alternative splicing adds a new tool to biological researchers who are searching for non-apparent patterns in complex biological data sets.

Acknowledgements

The authors thank Brian J. Haas of the Broad Institute for assistance with the original data.

References

[1] S. L. Campbell and R. Nikoukhah, Modeling and Simulation with Compose and Activate, Springer, 2019.

[2] S. L. Campbell and R. Nikoukhah, Modeling and Simulation in Scilab/Scicos with ScicosLab 4.4, Springer, 2009

[3] Dowse HB, Ringo JM (1989) The search for hidden periodicities in biological time series revisited. Journal Theor Biol 139: 487-515.

[4] Guyot R, Keller B (2004) Ancestral genome duplication in rice. Genome 47: 610-614.

[5] Lin H, Zhu W, Silva JC, Gu X, Buell CR (2006) Intron gain and loss in segmentally duplicated genes in rice. Genome Biol 7: R41.

[6] The MathWorks Inc., Natick, MA, 2000.

[7] C. D. Meyer Jr., *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics, Philadelphia, 2000.

[8] Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A 101: 9903-9908.

[9] Paterson AH, Bowers JE, Peterson DG, Estill JC, Chapman BA (2003) Structure and evolution of cereal genomes. Curr Opin Genet Dev 13: 644-650.

[10] Yu J, Wang J, Lin W, Li S, Li H, et al. (2005) The genomes of Oryza sativa: A history of duplications. PLoS Biology 3: e38.

[11] Vandepoele K, Simillion C, Van de Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell 15: 2192-2202.