Using Principal Component Analysis to Better Understand G-quadruplexes

R. Sgallová,[1,2] M. Volek,[1,3] P. Srb,[1] V. Veverka,[1,4] and E. A. Curtis[1]

[1] Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czech Republic.
[2] Department of Low-Temperature Physics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.
[3] Department of Genetics and Microbiology, Faculty of Science, Charles University, Prague, Czech Republic.
[4] Department of Cell Biology, Faculty of Science, Charles University, Prague, Czech Republic.

Abstract. G-quadruplexes are noncanonical nucleic acid structures. Their high functional and structural diversity shows the importance of better understanding the relationship among G-quadruplex primary sequence and biochemical function. We are exploring this question in the context of a DNA library with 496 sequences, which we screened for five biochemical properties. We analyzed results using a principal component analysis. It revealed a correlation between multimeric state and biochemical activities.

Introduction

The best-known DNA structure is the double helix, but many other folds have also been observed. One of them is the G-quadruplex [*Davis*, 2004]. This structure is formed from stacked guanine tetrads connected by loops. G-quadruplexes have many interesting properties: for example, they bind to many biologically important small molecules [*Li et al.*, 2013] and proteins [*Mishra et al.,* 2016], have various biochemical functions, including intrinsic fluorescence [*Mendez et al.,* 2009, *Kwok et al.,* 2013, *Majerová et al.,* 2018], and some can catalyze peroxidase reactions [*Sen et al.,* 2011, *Travascio et al.,* 1998], or act as an obstacle to replication forks and polymerases [*Paeschke et al.,* 2013]. This long list of activities and properties raises an important question: what determines the biochemical specificity of a G-quadruplex? We hypothesize that at least part of this specificity can be rationalized based on analysis of its primary sequence.

To test this hypothesis, we designed a G-quadruplex library consisting of 496 variants of a monomeric reference G-quadruplex. The entire library was then screened for the ability to bind GTP, promote a model peroxidase reaction, generate fluorescence, form dimers, and form tetramers. [*Majerová et al.,* 2018, *Kolesnikova et al.,* 2017, *Kolesnikova et al.,* 2019, *Švehlová et al.,* 2016, *Volek et al.,* 2021]. This leaves us with a dataset too large for reliable manual evaluation, and necessitates the use of statistical techniques to process the data. We decided to use principal component analysis (PCA) [*Jaumot et al.,* 2010, *Jolliffe,* 2002]. It is a technique which facilitates identification of statistically significant patterns in complex datasets, such is the one in this study.

Use of PCA revealed that the functional properties of the G-quadruplexes in this library can indeed be rationalized based on primary sequence. Furthermore, the structural basis for this relationship appears to be related to the multimeric state of the G-quadruplex. Our analysis also showed that translation of DNA sequences into PCA leads to highly symmetrical correlation matrices in which symmetry can be destroyed if the numerical accuracy of calculations is not handled correctly. This can strongly affect the results of PCA. Our study also highlights importance of use of negative control datasets to distinguish artifacts caused by library design from meaningful patterns in the experimental data.

Methods

Dataset analyzed in this study

The DNA library used in this study consists of 496 mutational variants of a reference monomeric G-quadruplex. It comprises four parts (Figure 1). The first and the largest one is the tetrad library, which contains all 256 possible variants of the central tetrad of the reference G-quadruplex (referred to as "the central tetrad" in the rest of the manuscript). The second one is the 17.3 loop library which contains all 81 variants of loops (A, C, or T, but not G) of the reference G-quadruplex. The third one is the 17.4 loop library which contains all 81 variants of loops (A, C, or T, but not G) of a representative dimeric G-quadruplex. The fourth one is the 17.10 loop library which contains all 81 variants of loops (A, C, or T, but not G) of a representative tetrameric G-quadruplex. All 496 sequences were scored according to five properties: ability to produce intrinsic fluorescence, promote model peroxidase reaction, bind GTP, form dimers, and form tetramers. Each sequence can be uniquely described using eight letters (four for tetrad positions and four for loop positions).

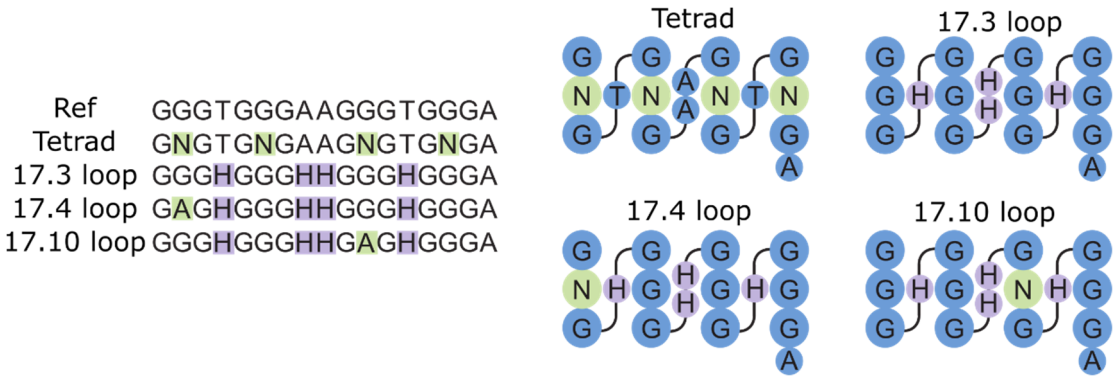

Figure 1. Library design, left: Ref = the reference G-quadruplex, Tetrad = tetrad library containing all 256 possible variants of the central tetrad of the reference G-quadruplex, 17.3 loop = 17.3 loop library containing all 81 variants of nucleotides at loop positions in the reference G-quadruplex (A, C, or T, but not G), 17.4 loop = 17.4 loop library containing all 81 variants of loops with a G to A mutation at position 2 in the central tetrad of the reference G-quadruplex, 17.10 loop = 17.10 loop library containing all 81 variants of loops with a G to A mutation at position 11 in the central tetrad of the reference G-quadruplex. Right: Depiction of mutated positions in all four libraries with the respect to a secondary structure of the reference monomeric G-quadruplex.

Principal component analysis

Principal component analysis (PCA) is a statistical technique useful for the analysis of complex datasets such as the one in this study. It facilitates identification of patterns and makes it possible to reduce dimensionality while losing only a small amount of information. In order to apply PCA to our dataset, it was necessary to convert sequence variables into numeric variables without losing any information. For each tetrad position, we created four variables. For example, tetrad position 2 was represented by the variables 2A, 2C, 2G, and 2T. The value of variable 2A was 1 for sequences which have adenine (A) at position 2 and 0 for sequences which have cytosine (C), thymine (T), or guanine (G) at position 2. Other variables were constructed in the same way. Each loop position was translated into three variables. For example, position 4 was represented by the variables 4A, 4C, and 4T. Variable 4G was omitted since it would contain 0 for all sequences due to the design of the library. Each sequence was therefore represented by 28 sequence variables (16 for tetrad positions and 12 for loop positions) and five variables representing the five experimentally determined functions, resulting in a final dataset with 33 variables.

The first step of PCA is to calculate a covariation or correlation matrix. We decided to calculate a correlation matrix, since it is not dependent on the scaling of variables [*Jolliffe,* 2002]. This is a big advantage for our dataset, which contains many binary variables and variables obtained by multiple experimental approaches. Elements of the correlation matrix $\boldsymbol{R}$ can be calculated from the variables:

$$\boldsymbol{R_{ij}} = \frac{E(x_i x_j) - \mu_i \mu_j}{\sigma_i \sigma_j},$$

where $E(x_i x_j)$ is the expected value of the product of the variables x_i and x_j, σ_i is the standard deviation of variable x_i, σ_j is the standard deviation of variable x_j, μ_i is the mean value of the variable x_i, and μ_j is the mean of the variable x_j.

The next step is to calculate the eigenvalues and eigenvectors of the correlation matrix. We did this using Wolfram Mathematica 12.1. Eigenvectors represent each principal component (PC) and indicate the extent to which each variable contributes to each PC, while eigenvalues indicate how much of the total variability of the dataset is described by a given PC. The first n PCs, which together describe between 70 % and 90 % of the variability of the dataset, are typically analyzed [*Jolliffe,* 2002]. n is usually considerably smaller than the number of variables in the original dataset, which allows reduction of dimensionality with minimal loss of information.

To visualize data using PCs, it is necessary to renormalize them. This can be accomplished using the formula $x' = \frac{x-\mu}{\sigma}$, where μ is the mean of a given variable and σ is the standard deviation of given variable. After doing this it is possible to use information obtained from the eigenvectors to transform data from variables into PCs.

Negative control

The goal of PCA is to reveal patterns in data, and it is critical to distinguish interesting ones created by experimental results from uninteresting ones created by experimental design. This is particularly true in the case

of the G-quadruplex library analyzed here because the majority of variables are sequence variables. Sequence variables also exhibit many uninteresting patterns which will likely be recognized by PCA. For example, when 2T is 1, then 2A, 2C, and 2G are 0. To identify such patterns, we decided to create a negative control dataset containing only sequence variables and analyze it by PCA. By comparing results obtained from analysis of the complete dataset with those of a negative control dataset, it was possible to distinguish results which are artifacts of the library design from those which reflect meaningful patterns in the experimental data.

Effect of limited accuracy of computer calculations

Computer calculations are usually done with some fixed number of digits of accuracy, which can cause problems during the calculation of eigenvectors of a correlation matrix of sequence variables. Correlation matrices of sequence variables of homogenous datasets (such as a dataset only containing data from the tetrad library) contain many zero values and exhibit a high level of symmetry (Figure *2*). However, this symmetry is disrupted if exact zeros are replaced by small random numbers (for example, numbers of the order 10^{-17}) due to the limited accuracy of computer calculations (specifically finite floating-point precision). This can significantly change the eigenvectors of a matrix (i.e., PCs) and make the interpretation of results more difficult (compare Figure *3* with Figure *4*). To avoid this issue, we made sure during each calculation that fields which should contain zero do in fact contain zero, and that the symmetry of the matrix was maintained.

Results

Here we used PCA to study two datasets. One contained all data and the second only contained data from the tetrad library (see [*Volek et al.,* 2021] for analysis of three additional datasets, each of which only contained data from one of the three loop libraries). The reason to study smaller datasets is that the complete dataset is not homogeneous due to the design of the library. For example, the complete dataset contains 81 sequences with a GGGG tetrad, but only one with an ATTA tetrad. With analysis of the smaller dataset, this problem disappears.

	2A	2C	2G	2T	6A	6C	6G	6T	11A	11C	11G	11T	15A	15C	15G	15T
2A	1	−1/3	−1/3	−1/3	0	0	0	0	0	0	0	0	0	0	0	0
2C	−1/3	1	−1/3	−1/3	0	0	0	0	0	0	0	0	0	0	0	0
2G	−1/3	−1/3	1	−1/3	0	0	0	0	0	0	0	0	0	0	0	0
2T	−1/3	−1/3	−1/3	1	0	0	0	0	0	0	0	0	0	0	0	0
6A	0	0	0	0	1	−1/3	−1/3	−1/3	0	0	0	0	0	0	0	0
6C	0	0	0	0	−1/3	1	−1/3	−1/3	0	0	0	0	0	0	0	0
6G	0	0	0	0	−1/3	−1/3	1	−1/3	0	0	0	0	0	0	0	0
6T	0	0	0	0	−1/3	−1/3	−1/3	1	0	0	0	0	0	0	0	0
11A	0	0	0	0	0	0	0	0	1	−1/3	−1/3	−1/3	0	0	0	0
11C	0	0	0	0	0	0	0	0	−1/3	1	−1/3	−1/3	0	0	0	0
11G	0	0	0	0	0	0	0	0	−1/3	−1/3	1	−1/3	0	0	0	0
11T	0	0	0	0	0	0	0	0	−1/3	−1/3	−1/3	1	0	0	0	0
15A	0	0	0	0	0	0	0	0	0	0	0	0	1	−1/3	−1/3	−1/3
15C	0	0	0	0	0	0	0	0	0	0	0	0	−1/3	1	−1/3	−1/3
15G	0	0	0	0	0	0	0	0	0	0	0	0	−1/3	−1/3	1	−1/3
15T	0	0	0	0	0	0	0	0	0	0	0	0	−1/3	−1/3	−1/3	1

Figure 2. Correlation matrix of a negative control dataset containing only sequences from the tetrad library.

	2A	2C	2G	2T	6A	6C	6G	6T	11A	11C	11G	11T	15A	15C	15G	15T
1	0	0	0	0	0	0	0	0	0.71	–0.7	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	–0.4	–0.4	0.82	0	0	0	0	0
3	0.71	–0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	–0.4	–0.4	0.82	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0.71	–0.7	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	–0.4	–0.4	0.82	0
7	–0.4	–0.4	0.82	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0.71	–0.7	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	–0.3	–0.3	–0.3	0.87	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	–0.3	–0.3	–0.3	0.87	0	0	0	0
11	–0.3	–0.3	–0.3	0.87	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	–0.3	–0.3	–0.3	0.87
13	–0.5	–0.5	–0.5	–0.5	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	–0.5	–0.5	–0.5	–0.5	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	–0.5	–0.5	–0.5	–0.5	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	–0.5	–0.5	–0.5	–0.5

Figure 3. PCs (i.e., eigenvectors) calculated for the correct correlation matrix of a negative control dataset containing only sequences from the tetrad library. Rows: PCs. Columns: contributions from individual variables.

	2A	2C	2G	2T	6A	6C	6G	6T	11A	11C	11G	11T	15A	15C	15G	15T
1	–0	0.05	–0	–0	0	–0	–0	0.03	–0	–0	0.09	–0	–0.1	–0.6	0.76	0
2	–0	–0.1	0.31	–0.1	–0.1	0.65	0.02	–0.6	–0	0.03	0.15	–0.2	–0.2	0.08	0.07	0
3	–0.1	0.06	0.09	–0	–0.2	0.09	–0.2	0.31	0.28	–0.6	0.54	–0.3	0.14	–0	–0.1	0
4	0.29	0.11	–0.2	–0.2	0.2	0.29	–0.5	–0	–0.4	–0.2	0.11	0.47	0.19	–0.1	–0.1	0
5	0.61	–0.6	–0.2	0.21	0.11	–0.1	–0	0.02	–0	–0.1	0.22	–0.2	–0.2	0.12	0.08	0
6	0.48	0.32	–0.1	–0.7	–0.1	–0.1	0.18	0.03	0.06	0.2	0.04	–0.3	0.01	0.02	–0	0
7	0.03	–0.3	0.38	–0.1	–0.6	–0	0.27	0.31	–0.4	0.06	0.06	0.3	–0	0.01	–0	0
8	–0	–0.2	0.59	–0.4	0.49	–0.4	0.01	–0.1	0.1	–0.2	–0	0.12	0.03	–0	–0	0
9	0.21	–0.1	–0	–0.1	–0.2	0.2	–0	0.06	0.55	–0.3	–0.5	0.32	–0.1	0.01	0.09	0
10	0.11	–0.3	0.09	0.08	–0.1	0.09	0.02	–0.1	0.12	0.2	–0.2	–0.2	0.7	–0.5	–0.2	0
11	0	0	0	0	0	0	0	0	0	0	0	0	–0.3	–0.3	–0.3	0.87
12	0.06	0.11	–0.3	0.09	–0	–0.1	0.62	–0.4	–0.1	–0.4	0.19	0.26	0.16	–0.1	–0.1	0
13	0	0	0	0	0	0	0	0	0	0	0	0	–0.5	–0.5	–0.5	–0.5
14	0.17	0.17	0.17	0.17	0.26	0.26	0.26	0.26	–0.4	–0.4	–0.4	–0.4	–0	0	0	0
15	0.42	0.42	0.42	0.42	–0.3	–0.3	–0.3	–0.3	–0	–0	–0	–0	–0	–0	0	0
16	–0.2	–0.2	–0.2	–0.2	–0.3	–0.3	–0.3	–0.3	–0.3	–0.3	–0.3	–0.3	0	–0	–0	0

Figure 4. PCs (i.e., eigenvectors) calculated for the incorrect correlation matrix of a negative control dataset containing only sequences from the tetrad library. Rows: PCs. Columns: contributions from individual variables.

Complete library

If we compare eigenvalues of the correlation matrix of the complete dataset with those of a negative control (Figure 5), we can see that they are quite similar. They both contain eight zero values — one for each mutated position in the library. This was expected, since we created sequence variables in such a way that one linearly dependent variable will exist at each sequence position. For example, $6G = 1 - 6A - 6C - 6T$ (in other words, exactly one nucleotide can occur at position six and this can be either A, C, G, or T). We decided to create these unnecessary varibles to make interpretation of results more straightforward, since it is easier to determine if there is (for example) some connection between fluorescence and 6G, then between fluorescence and $1 - 6A - 6C - 6T$.

If we would want to use the typical cutoff of between 70% and 90% of variability of the original dataset to determine how many PCs should be analyzed further, we would have to analyze between 12 and 18 PCs. However, if we look at Figure 5, we can see that only the first two or three PCs (corresponding to 34.8% and 39.7% of the variability of the dataset) contain significantly more variability than the rest of PCs and are therefore important. The rest of the variability is hidden in the sequence part of the correlation matrix, which is not interesting for the interpretation of results.

The first PC (Figure 6) highlights the necessity of a negative control — contributions from sequence variables are in all cases similar for the original dataset and the negative control and contributions from all experimentally determined properties are positive. This indicates that the first PC is an artifact of library design.

For the second PC (Figure 7), contributions from sequence variables are also similar for both the original dataset and the negative control. For example, position 2 contains a positive contribution from 2G and a negative contribution from 2A. This is due to the design of the 17.4 loop library, in which all 81 sequences contain a G to A mutation at position 2. Similarly, position 11 contains a positive contribution from 11A and negative contribution from 11G. This is due to the design of the 17.10 loop library, in which all 81 sequences contain a G to A mutation at position 11. If we look at contributions from experimentally determined properties, we can see that there is a positive contribution from tetramerization and GTP binding and a negative contribution from dimerization and peroxidase activity. This highlights two main trends in the data: tetramers tend to bind GTP well and usually contain G at position 2 and A at position 11, while dimers tend to promote peroxidase reaction efficiently and contain A at position 2 and G at position 11.

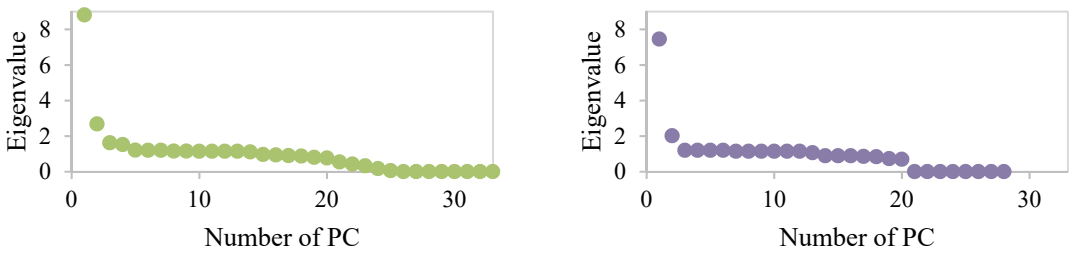

Figure 5. Eigenvalues of the correlation matrix of the dataset containing all data (left) and of the negative control matrix (right).

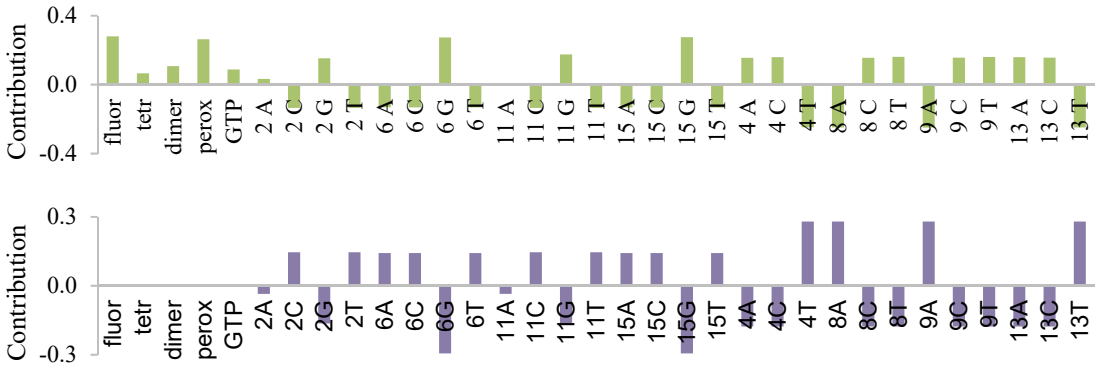

Figure 6. Contributions to the first PC from each variable in the original dataset (above) and the negative control (below).

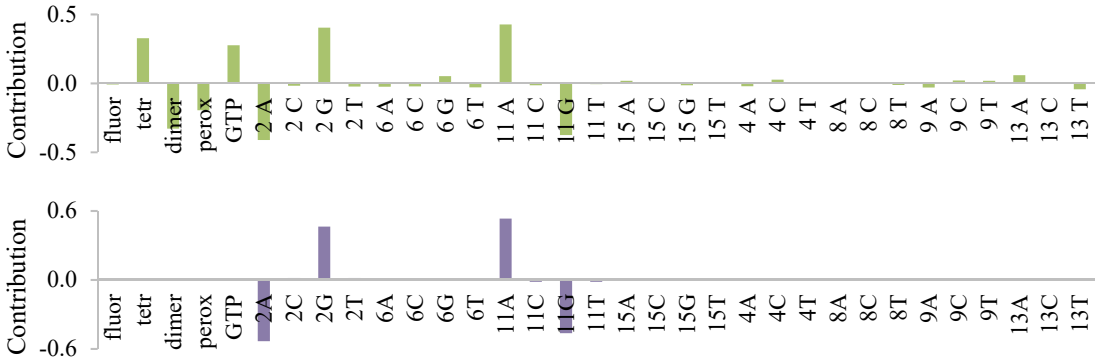

Figure 7. Contributions to the second PC from each variable in the original dataset (above) and the negative control (below).

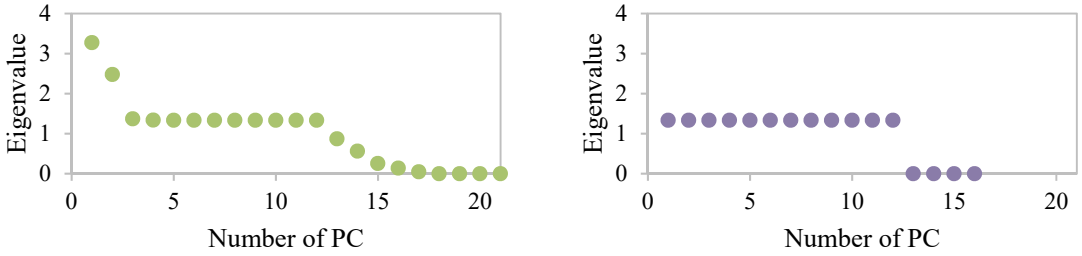

Figure 8. Eigenvalues of the correlation matrix of the dataset containing data from the tetrad library (left) and the negative control matrix (right).

Tetrad library

In the case of the tetrad library, the first two PCs are important (Figure 8). Together they describe 27.4% of the variability of the dataset. The rest of the variability is again contained in the sequence part of the correlation matrix.

Contributions to the first PC (Figure 9) from sequence variables differ significantly in the original dataset and the negative control. The original dataset contains positive contributions from all experimentally determined properties and Gs at all tetrad positions. This represents the most significant pattern in the tetrad library — the more Gs a sequence contains, the better it is at forming a G-quadruplex, and all five of the functions we studied are associated with G-quadruplex formation.

The second PC (Figure 10) also differs significantly from the negative control. It contains positive contributions from fluorescence, tetramerization, GTP binding, 2G, 6G, 11A, and 15A, and negative contributions from dimerization, peroxidase activity, 11G, and 15G. It highlights two interesting trends in the tetrad library: sequences with 2G, 6G, and mutations at positions 11 and 15 (preferably to A) tend to be tetramers, bind GTP well, and have high fluorescence values, while sequences with 11G, 15G, and mutations at positions 2 and 6 (preferably to A) tend to be dimers and catalyze peroxidase reaction efficiently.

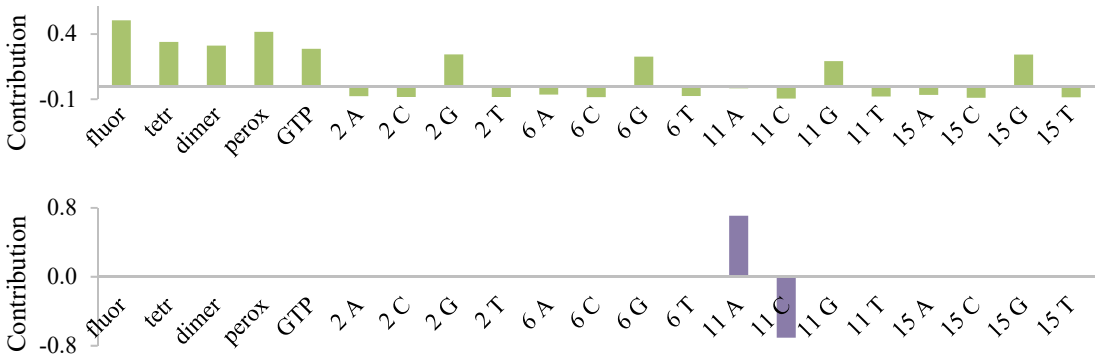

Figure 9. Contributions to the first PC from each variable in the dataset containing data from the tetrad library (above) and the negative control (below).

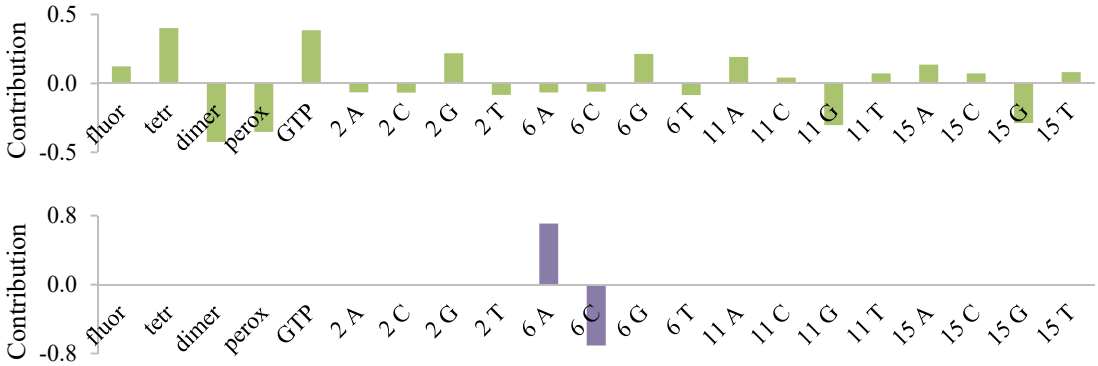

Figure 10. Contributions to the second PC from each variable in the dataset containing data from the tetrad library (above) and the negative control (below).

Discussion

In this study, we used PCA to analyze a dataset made up of 496 DNA sequences in G-quadruplex library, each of which was experimentally tested for five different biochemical functions. Our goal was to try to better understand the relationship between G-quadruplex primary sequence and biochemical function.

During our analysis, we observed that it is important to understand the mathematical details of PCA in order to correctly interpret results. Without this understanding, it would be possible to misinterpret or completely miss results due to the absence of negative control and/or the disruption of symmetry in correlation matrices. Figure 6 highlights importance of a negative control — the first PC describes significantly more variability in the dataset then the rest of PCs (Figure 5). If this was interpreted without a deeper understanding of PCA in the context of this dataset, a logical conclusion would be that the first PC describes the most important pattern in the dataset. However, when we compare the first PC with the negative control, it is clear that this PC is an artifact of library design and does not provide insights into the meaning of experimental results.

One way to reduce artifacts caused by library design is to analyze smaller homogenous datasets within a larger dataset. In the context of this study, we did this by analyzing the tetrad library by itself. This library contains each of the 256 possible nucleotide patterns in the central tetrad of the reference G-quadruplex. In contrast, the complete library contains 81 sequences with the tetrad sequence GGGG and only one with the sequence AACC. This eliminated an artifact described by the first PC in our analysis of the complete dataset. However, there are still several other ways to misinterpret these results. First, because a significant part of the variability of the dataset is still contained in the sequence variables, nine to twelve PCs would have been analyzed if the usual cutoff was used to determine which PCs are significant. However, only two of these are actually significant for interpretation of experimental results (Figure 8). Second, the negative control would have been more difficult to interpret if symmetry had not been maintained in correlation matrices.

One advantage of using PCA is a reduction of dimensionality. A dataset as complex as the one analyzed here cannot be easily represented in graphs with respect to the original variables. If we instead use PCs, however, trends in the dataset can be visualized in two dimensions while preserving most of the information significant for interpretation of experimental results. Figure 11 shows all sequences in the tetrad library plotted with the respect to the first two PCs of the dataset containing only sequences from the tetrad library. In this graphical

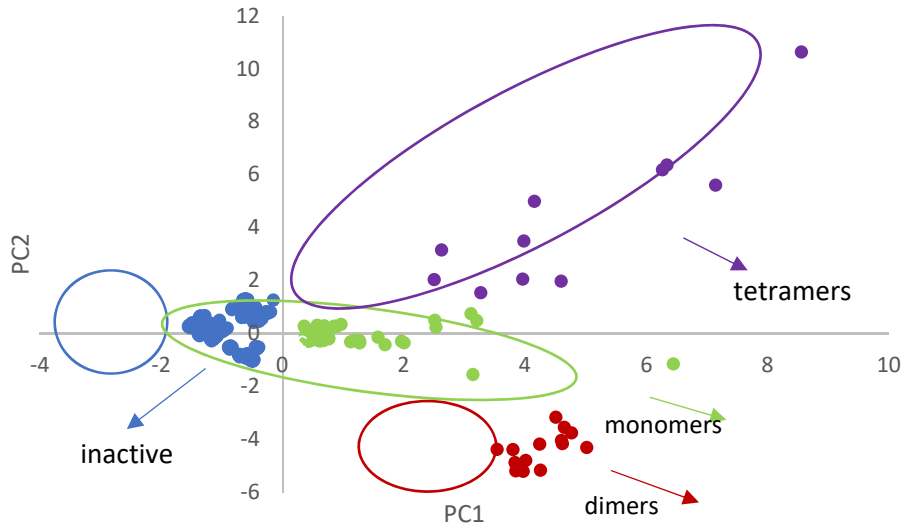

Figure 11. All sequences in the tetrad library plotted with respect to the first and second PC (calculated for a dataset containing only sequences from tetrad library).

representation, sequences form four clusters: tetrameric G-quadruplexes, dimeric G-quadruplexes, monomeric G-quadruplexes, and inactive sequences. These correspond to the four major groups of sequences identified in previous studies of this library.

PCA analysis of our dataset highlighted two of the main trends in the data: tetramers tend to bind GTP well and dimers tend to promote the peroxidase reaction efficiently. It also revealed sequence patterns connected with multimer formation. For example, sequences with mutation(s) in the second half of the tetrad (especially G to A mutations; such sequences include those in the 17.10 loop library) tend to form tetramers, while those with mutation(s) in the first half of tetrad (especially G to A mutations; such sequences include those in the 17.4 loop library) tend to form dimers. These results indicate the functions of the G-quadruplexes in this library can be rationalized based on primary sequence. They also suggest that the structural basis for this connection is related to multimeric state.

Conclusion

We used PCA to analyze a dataset made up of 496 sequences from a G-quadruplex library, each of which was experimentally tested for five different biochemical functions. Our goal was to determine whether the functions and biochemical specificities of these G-quadruplexes could be rationalized based on primary sequence. We have shown that PCA can be used to study this type of dataset. However, it is necessary to understand the mathematical details in order to interpret results correctly. We found that sequences with mutation(s) in the first half of the central tetrad of the reference G-quadruplex form dimers and catalyze peroxidase reaction efficiently and that sequences with mutation(s) in the second half of the central tetrad form tetramers and bind GTP well. This indicates that the functions of these G-quadruplexes can be rationalized based on sequence. Our results also suggest that the structural basis for these patterns is related to multimeric state.

Acknowledgments. This study was supported by the Charles University, project GAUK No. 152120.

References

Davis, A. T. “G-quartets 40 years later: from 5′-GMP to molecular biology and supramolecular chemistry”, *Angewandte Chemie International Edition, 43.6: 668–698*, *2004.*

Jaumot, J. and Gargallo, R. “Using principal component analysis to find correlations between loop-related and thermodynamic variables for G-quadruplex-forming sequences,” *Biochimie,* vol. 92, p. 1016–1023, 2010.

Jolliffe, I. T. Principal component analysis for special types of data, Springer, 2002.

Kolesnikova, S., Srb, P., Vrzal, L., Lawrence, M. S., Veverka V., and Curtis, E. A. “GTP-dependent formation of multimeric G-quadruplexes,” *ACS Chemical Biology,* vol. 14, p. 1951–1963, 2019.

Kolesnikova, S., Hubálek, M.,Bednárová, L., Cvačka, J., and Curtis, E. A. “Multimerization rules for G-quadruplexes,” *Nucleic acids research,* vol. 45, p. 8684–8696, 2017.

Kwok, C. K., Sherlock, M. E., and Bevilacqua, P. C. “Effect of loop sequence and loop length on the intrinsic fluorescence of G-quadruplexes,” *Biochemistry,* vol. 52, p. 3019–3021, 2013.

Li, Q., Xiang, J.-F., Yang, Q.-F., Sun, H.-X., Guan, A.-J., and Tang, Y.-L. “G4LDB: a database for discovering and studying G-quadruplex ligands,” *Nucleic acids research,* vol. 41, p. D1115–D1123, 2013.

Majerová, T., Streckerová, T., Bednárová, L., and Curtis, E. A. “Sequence requirements of intrinsically fluorescent G-quadruplexes,” *Biochemistry,* vol. 57, p. 4052–4062, 2018.

Mendez, M. A. and Szalai V. A., “Fluorescence of unmodified oligonucleotides: A tool to probe G-quadruplex DNA structure,” *Biopolymers: Original Research on Biomolecules,* vol. 91, p. 841–850, 2009.

Mishra, S. K, Tawani, A., Mishra, A., and Kumar, A. “G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins,” *Scientific reports,* vol. 6, p. 1–9, 2016.

Paeschke, K., Bochman, M. L., Garcia, P. D., Cejka, P., Friedman, K. L., Kowalczykowski, S. C., and Zakian, V. A. “Pif1 family helicases suppress genome instability at G-quadruplex motifs,” *Nature,* vol. 497, p. 458–462, 2013.

Sen, D. and Poon, L. C. H. “RNA and DNA complexes with hemin [Fe (III) heme] are efficient peroxidases and peroxygenases: how do they do it and what does it mean?” *Critical reviews in biochemistry and molecular biology,* vol. 46, p. 478–492, 2011.

Švehlová, K., Lawrence, M. S., Bednárová, L., and Curtis, E. A. “Altered biochemical specificity of G-quadruplexes with mutated tetrads,” *Nucleic acids research,* p. gkw987, 2016.

Travascio, P., Li, Y., and Sen, D.”DNA-enhanced peroxidase activity of a DNA aptamer-hemin complex,” *Chemistry & biology,* vol. 5, p. 505–517, 1998.

Volek, M., Kolesnikova, S., Svehlova, K., Srb, P., Sgallová, R., Streckerová, T., Redondo, J. A., Veverka, V., and Curtis, E. A. “Overlapping but distinct: a new model for G-quadruplex biochemical specificity,” *Nucleic acids research,* vol. 49, p. 1816–1827, 2021.