Analyzing Genomic Data Using Tensor-Based Orthogonal Polynomials

Saba Nafees[1], Sean Rice[1*], Catherine Wakeman[1]

1 Department of Biological Sciences, Texas Tech University, Lubbock, TX, USA

*sean.h.rice@ttu.edu

Abstract

Due to increasing computational power and experimental sophistication, extensive collection and analysis of genomic data is now possible. This has spurred the search for better algorithms and computational methods to investigate the underlying patterns that connect genotypic and phenotypic data. We propose a multivariate tensor-based orthogonal polynomial approach to characterize nucleotides or amino acids in a given DNA/RNA or protein sequence. Given quantifiable phenotype data that corresponds to a biological sequence, we can construct orthogonal polynomials using sequence information and subsequently map phenotypes on to the space of the polynomials. With enough computational power, this approach provides information about higher order interactions between different parts of a sequence in a dataset and ultimately illuminates the relationship between sequence structure and the resulting phenotype. We have applied this method to a previously published case of small transcription activating RNAs (STARs), quantifying higher order relationships between parts of the sequence and how these give rise to the distinct phenotypes.

Author summary

An important goal in molecular biology is to quantify both the patterns across a genomic sequence, such as DNA, RNA, or protein, and the relationship between phenotype and underlying sequence structure. With an increasing amount of genomic data being available, there exists an increasing need to identify patterns across sites in a sequence and ultimately connect sequence structure to output phenotype. In this work, we propose a mathematical method based on tensor-valued multivariate orthogonal polynomials that represents sequence data as vectors and subsequently builds polynomials onto which the corresponding phenotype is projected. This method captures higher order interactions between sequence states and resulting phenotype. We show proof of concept of this approach as applied to a case of regulatory RNA that were previously studied and demonstrate its potential applications to other biological systems.

Introduction

Advancements in sequencing technology and the rise in the availability of genomic data has led to the development of novel computational methods that seek to determine the relationship between underlying sequence and the resulting function or phenotype. These methods aim to predict protein function given sequence and structure information, identify novel and potential DNA binding motifs, including transcription factor binding sites, and determine RNA secondary structure based on the underlying sequence [1] - [4]. Though the development of these tools has expanded dramatically, there exists a continued need to improve upon existing methods and develop better tools with diverse functionality. For example, in the case of RNA secondary structure prediction, predicting pseudoknots without constraints is impossible and is an NP-complete problem [5]. Due to the computational complexity inherent in this and other RNA secondary structure prediction problems, this is an active area of algorithm development.

RNAs are extremely versatile molecules and play important regulatory roles in gene expression by affecting transcription and translation through various different mechanisms, including intermolecular (trans) and intramolecular (cis) interactions [6]. Of particular importance is the ability of RNA to control transcription initiation or termination through secondary structure formation of hairpins and loops [7]. In recent years, much work has been done to understand the role of sequence composition in the regulatory potential of RNAs by not only studying the ones that occur naturally in bacterial systems but also synthetically constructing these regulators de novo [7] - [9]. This presents a unique opportunity to employ novel computational tools to quantify the sequence-function relationship between RNAs and their resulting regulatory activity in both synthetic and natural systems.

In general, given sequence information and corresponding phenotype data, one important objective is to quantify exactly how the underlying sequence, whether DNA, RNA or protein, gives rise to variation in phenotype. In the case of RNA regulators, one measurable phenotype is the regulatory activity of an RNA as captured experimentally through the use of fluorescent reporters [10]. Here we describe a mathematical method using multivariate tensor-based orthogonal polynomials to convert sequence information into vectors and building orthogonal polynomials with the aim of quantifying the effect of the sequence states on the resulting phenotype [11].

We have applied this method to a case of regulatory RNAs called Small Transcription Activating RNAs (STARs) that were synthetically constructed and whose regulatory activity was quantified experimentally by monitoring levels of green fluorescent protein expression (Fig 1, [12]). Our methods showcase important characteristics about the sequence composition of these synthetic RNAs and the corresponding target RNAs that they bind to. In addition, projecting transcription termination activity of the target RNAs onto the sequence space reveals first and higher order effects of having a given nucleotide at a given site along the sequence. We demonstrate how this method's ability to capture nucleotide (or amino acids in the case of proteins) interactions across sequences and quantify sequence-phenotype interactions can be leveraged when applied to other biological systems.

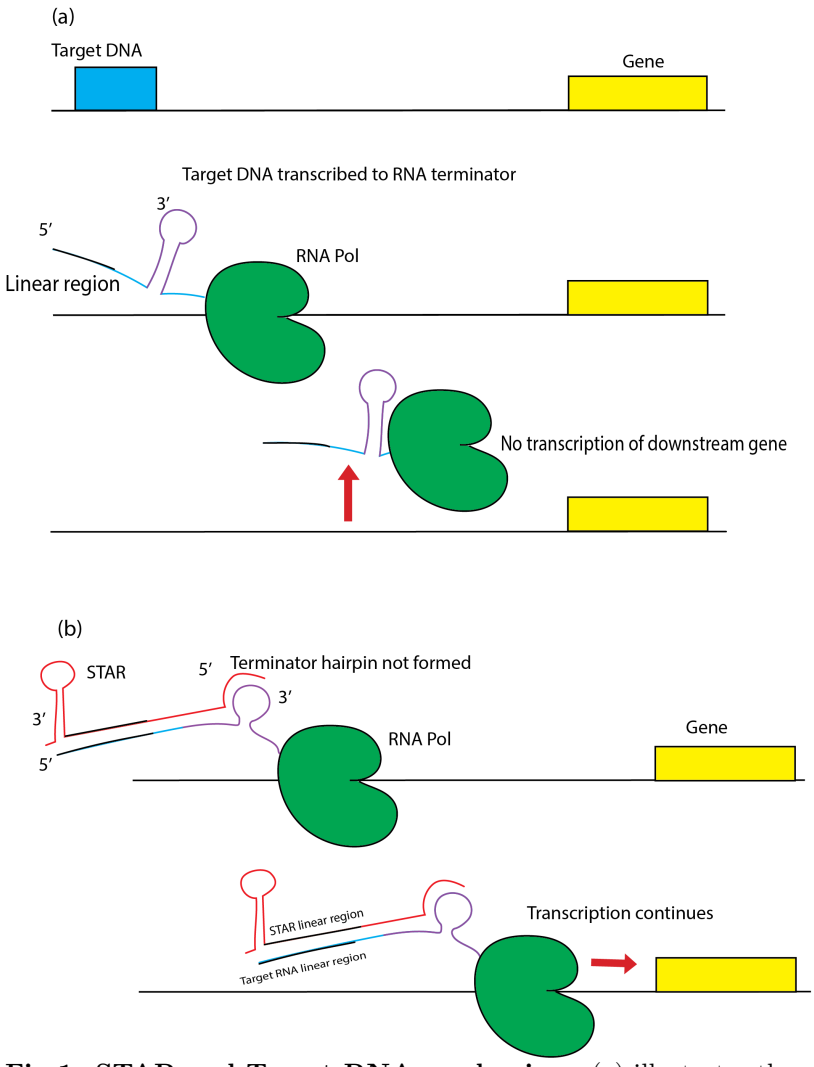

Fig 1. STAR and Target RNA mechanism. (a) illustrates the case in which a target DNA sequence is transcribed into an intrinsic terminator hairpin which displaces the polymerase, preventing transcription of the downstream gene. (b) shows that when STAR binds to the target RNA, the terminator hairpin is prevented from forming and transcription of the downstream gene continues. The purple region in the target RNA indicates the terminator helix which is disrupted upon binding of STAR. The 40 nucleotide linear region in both the target RNA and STAR is depicted with a black line. This was the sequence that was varied while the terminator hairpin sequence remained the same across all 99 variants. Figure adapted from [12].

The case of STAR RNAs is particularly interesting to study at the sequence level due to the impact that both intra- and intermolecular interactions may have on the function of this type of RNA. Our method described herein enables prediction of nucleotide sequence covariation both positively and negatively impacting the function of this RNA-based regulator. These findings have the potential to uncover critical intra- and intermolecular interactions within the STAR RNA regulatory system. The impact of nucleotide covariation on both intra- and intermolecular interactions is highlighted in the examples from Fig 2 depicting the ability of different hairpin sequences to potentially form an intermolecular interaction with an unchanging target sequence. In the scenario

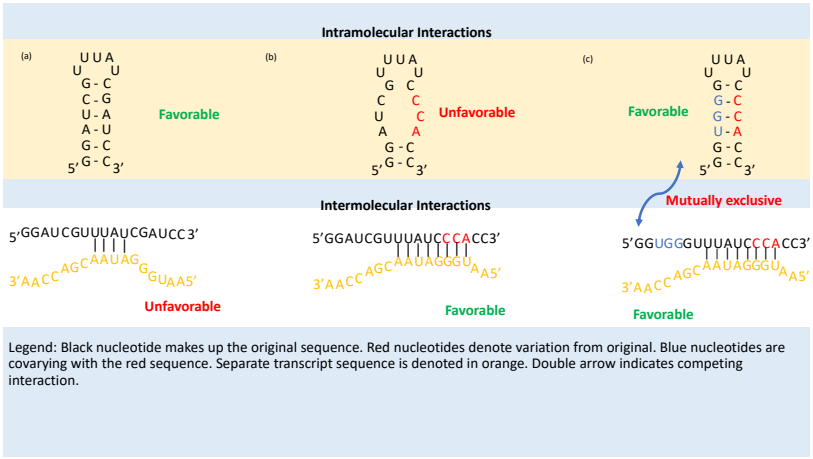

Fig 2. Importance of covariation for secondary structure. (a) shows a case where there exists favorable intramolecular interaction (top panel) and an unfavorable intermolecular interaction with another transcript (bottom panel). (b) Here, the same original sequence has three nucleotides mutated which gives rise to an unfavorable intramolecular interaction (top panel) but a favorable intermolecular interaction with the other transcript (bottom panel). (c) shows mutually exclusive structures with competing intramolecular and intermolecular interactions when there is a mutation (denoted in blue) that covaries with the nucleotides denoted in red.

shown in Fig 2(a), a favorable intramolecular interaction forming a strong hairpin is likely to prevent the formation of the less favorable intermolecular interaction. In the scenario shown in Fig 2(b), a slight sequence change in the hairpin RNA has shifted the balance to a far more favorable intermolecular interaction likely to outperform the relatively weak intramolecular interaction. Interestingly, a covariation on the other end of the potential hairpin restores a favorable intramolecular interaction without impacting the theoretical strength of the intermolecular interaction (Fig 2c). Despite the theoretical strength of the intermolecular interaction depicted in Fig 2(c), the competing intramolecular structure is likely to impair full intermolecular function. Thus, the study of both the positive and negative impacts of covariation on molecular function can provide insight into the nucleotides involved in important intra- and intermolecular interactions. Importantly, while the Watson-Crick base pairing interactions depicted in Fig 2 are rather easy to computationally predict, there are many other types of interactions occurring within structured RNAs that are less simple to predict. Studies providing insight into the functional importance of sequence covariation may assist in improving the predictions of these other types of structural elements in the future.

Materials and methods

Application of orthogonal polynomials

Given a set of DNA, RNA, or protein sequences, along with corresponding phenotypic data, our method consists of building tensor-valued first and higher order polynomials and projecting the phenotypic data into this polynomial space (see supplementary methods for detailed explanation and proofs). To apply our methods to the STAR system, we first converted each site in each sequence of the 99 sequences of STAR and target RNAs into a vector as depicted in Fig 3. For each type of RNA, each sequence had a corresponding experimentally derived OFF/ON fluorescence value which served as the numerical phenotype that we later project onto the polynomial space.

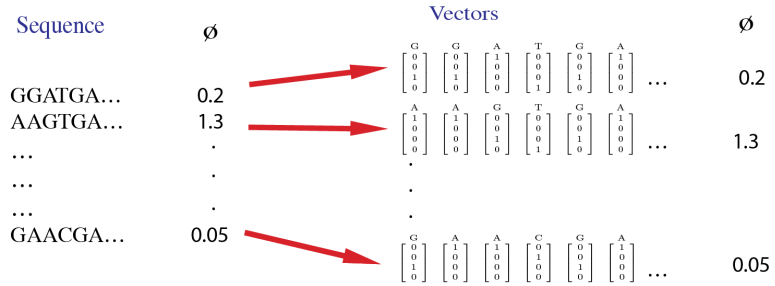

Fig 3. Example sequences with corresponding phenotypic values. Example sequences with corresponding phenotypic values. The first step in our methods is to convert each site in a DNA or RNA sequence to its respective 4-dimensional vector. In this example, a set of 6 sequences is shown, each corresponding to a phenotype (a real-valued number). For DNA and RNA, the vectors are 4-dimensional but this can be changed to a 20-dimensional vector in the case of proteins. The phenotype (ϕ) is the off/on fluorescence value associated with the sequence.

After subtracting out the means across all sites in the set of sequences, we get the first order M vectors (see example in described in supplementary methods). We use these to find the variances at each site and covariances between each pair of sites. The covariance analysis shows positive and negative relationships between a pair of two sites. It picks out not only the correlations between sites across a sequence but also the relationship between the *state* at one site (what nucleotide is present) and the state at another site (Fig 3).

Next, we constructed orthogonal polynomials based on our vectors and projected our variable of interest (OFF/ON values) onto the polynomial space as shown in Supplementary Methods. For the first order analysis, we are interested in the regression of the phenotype (OFF/ON values) onto the first order conditional polynomial. This is presented here as $[\![{}^{\phi}_{\mathbb{M}_\bullet}]\!]$ and distributions of these regressions onto target RNA and STAR linear regions are shown in Figures 6 and 7 respectively.

The covariance matrix for two sites is the mean, across all individuals, of the outer product of $\mathbb{M}^1$ and $\mathbb{M}^2$, where $\mathbb{M}^1$ and $\mathbb{M}^2$ are first order vectors for each individual in the population:

$$[\![\mu_1, \mu_2]\!] = \overline{\mathbb{M}^1 \otimes \mathbb{M}^2} \tag{1}$$

First order analysis is useful to get an overall idea of how a given trait is related to the underlying sequence. However, biological systems are complex and relationships between sequence and the corresponding phenotypic traits often exhibit higher order associations. The challenge with trying to capture higher order relationships, for example, quantifying the regression of the trait on the combination of two or more sites at once, is that this becomes computationally difficult for sequences with large numbers of sites. Thus, second and third order polynomials were constructed for a smaller set of sites. We identified these sites as likely candidates after doing first order analysis and noticing sites that potentially exhibited second and third order interactions. Polynomials up to third order were built for three interacting sites and the regressions of the phenotype on the third orthogonal polynomial were computed (Fig S3-S4). In addition, regressions of the phenotype on the second order orthogonal polynomial were built for six interacting sites in the 5 prime region of the STAR sequence (Figs S5-S8).

All code to construct up to third order orthogonal polynomials is written in the Python programming language. All computational analysis of results was also done in Python. A command line tool to compute these polynomials based on sequence data and corresponding phenotype data is currently under construction.

Application to STARs (small transcription activating RNAs)

We applied our methods to the case of a synthetic RNA regulator designed by [12] known as STARs or small transcription activating RNAs. In this system, a target DNA sequence, containing the necessary information for termination, is placed upstream of a gene. When transcribed, this sequence turns into an intrinsic terminator with a linear region and a hairpin structure. Upon formation of this structure, the polymerase gets knocked off, preventing the transcription of the downstream gene (Fig 1(a), [12] - [13]). This is known as the "OFF state". In the "ON state", a STAR is constructed such that upon binding to the target RNA, terminator hairpin formation is prevented and the polymerase continues transcription of the downstream gene (Fig 1(b)).

In this system, it was determined that the linear region of the STAR binding to the corresponding part on the target RNA was critical to the activation of the downstream gene. To establish this, linear regions of 100 STAR:target RNA variants were constructed *de novo* using a software package known as NUPACK [14]. In this construction, only the linear recognition region was varied while the terminator hairpin remained the same for all variants. This linear region was 40 nucleotides long and it was hypothesized that variation in this part of the sequence would give rise to distinct OFF and ON states for all STAR:target variants (see methods in [12] for details). For the application of our methods to this system, this set of 40 nucleotide long sequences in the STAR and target RNA were used to build first and higher order orthogonal polynomials. Fold and off values were then projected onto the space of the orthogonal polynomials.

Results

Covariances showcase nucleotide interactions across STAR sites

As part of first order analysis, means, variances and covariances were computed for all 40 sites across the population of STAR and target RNA sequences. As depicted by the

means of each nucleotide present at a site, there exists a greater frequency of pyrimidines than purines in the population of target RNA sequences (Fig 4). Pyrimidines consist of nucleotides cytosine (C), thymine (T), and uracil (U, in the case of RNA). Purines consist of adenine (A) and guanine (G).

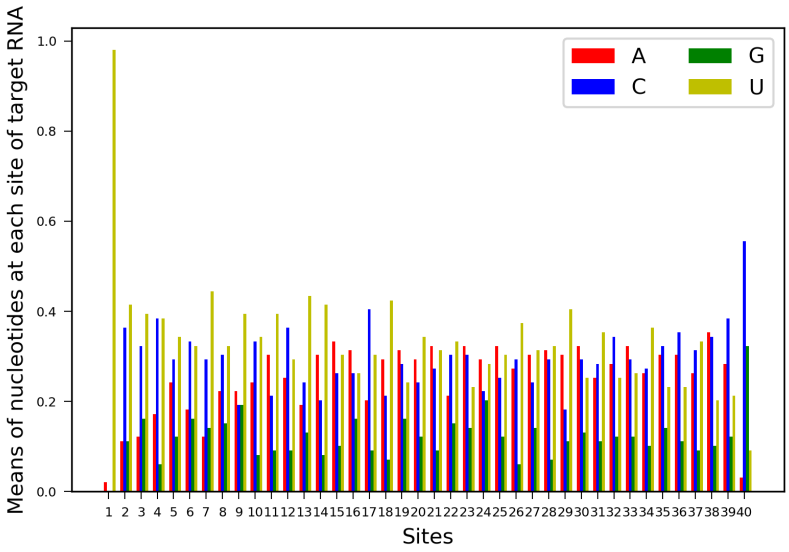

Fig 4. Means of nucleotides present at each site. Fig 3: Means of each nucleotide present at each site of the target RNA (from 3 prime to 5 prime). Across the population of all sequences, the initial pool of STAR-target RNAs designed using the NUPACK software demonstrated an enrichment for pyrimidines over purines.

The covariance analysis yields matrices for all pairs of sites, across the 40 sites, and quantifies the relation between having a particular nucleotide at one site and another nucleotide at another site. Since there are 40 sites in this system, there are 780 unique pairs of sites. The covariance between each pair of sites is a matrix and thus, there are 780 4x4 matrices. To understand the distribution of these covariances, a histogram was constructed (Fig S2). This plot depicted a small number of covariances (32 in total) that were greater than 0.05 and less than -0.05, indicating nucleotides at specific sites covarying highly and positively with each other and nucleotides at other sites covarying highly and negatively with each other (Supplementary Table 1).

These large positive and negative covariances are visualized as seen in Fig 5 and Fig S1. The start of the sequence, depicted as site 1, is the 5' end of the STAR while the end of the sequence is the 3' end. It can be immediately noted that sites near the 5' end of the sequence are correlated with sites throughout the sequence and there is a lack of notable correlation between sites in the middle portion of the sequence. One possible explanation for this result is that this interaction between sites at the ends of the sequence is preventing any potential binding between the linear region and the hairpin part of the RNA so that the hairpin region can stay intact and pursue its function of terminating transcription of the downstream gene. This type of function for transacting regulatory RNAs has been well described in other contexts ([7], [15] - [17]).

Another possible explanation for this result could be related to the way these sequences were designed by the NUPACK algorithm. The objective of this software program is to calculate equilibrium distributions of the given nucleic acid strands [14]. However, as has been noted by the authors of the STAR paper, STAR "regulation is

governed by kinetic, out-of-equilibrium folding regimes" and thus, the NUPACK design of STAR sequences may not resemble natural sequences [12]. Therefore, the apparent importance of sites at the 5' end of the sequence could be an artifact of the software used to design the sequences. This warrants further investigation.

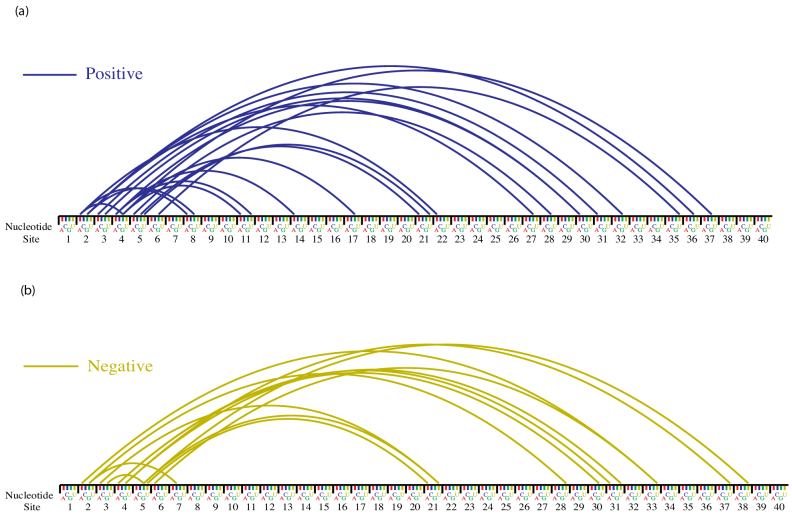

Fig 5. Strong correlations between nucleotides at sites across STAR. Covariances with absolute magnitude greater than 0.05 across the 40 site long linear region of STAR (shown here from 5' to 3'). (a) Large positive covariances. (b) Large negative covariances. See supplementary Table 1 for actual values.

Regressions of fold and off values onto linear binding regions

In the STAR system, the authors determined that the OFF state provides the best measure of the efficiency of the STAR:target complex as a whole. This motivates the analysis of the relationship between sequence and function of the target RNA (as measured by termination efficiency). To establish this, after building the first order orthogonal polynomials, we computed regressions of OFF values onto each nucleotide at each site of the target RNA. These regressions show how each nucleotide at each site contributes to the overall function of the STAR sequence.

Distribution of regressions

The magnitude of regressions is highly variable but large values seem to be uniformly distributed across the sequence. Regressions here refer to the projection of the phenotype (OFF fluorescence values) onto the first order conditional polynomial (see Methods). This indicates that this pattern is "unstructured" as noted by the authors of the STAR paper. This can be inferred from the shape of the regressions across the linear region of the target RNA (Fig 6a). Given the covariance structure in the STAR linear sequence sequence (which is complementary to the target linear region) (Fig 5), it would not be surprising if the regressions also showed a similar pattern of connection between sites at the 3' end of the sequence and sites at the 5' end. However, this does not appear to be the case.

Overrepresentation of purines

While there is no spatial pattern in the magnitudes of regression coefficients, there is a strong relationship between the kind of nucleotide at a site and the regression of the off phenotype values on it. Almost all the purines have positive regressions while almost all the pyrimidines have negative values. This means that there is a preference for having purines along the target RNA sequence while pyrimidines are disfavored. This could be due to a few reasons. The hairpin of the target RNA, the intrinsic terminator, includes a string of guanines and cytosines [12]. This hairpin structure must remain intact in order for it to efficiently terminate the downstream gene. Thus, if there are more cytosines present in the linear region of the structure, this could lead to binding between these cytosines and the GC rich region in the hairpin, decreasing the termination efficiency of the hairpin.

We note that beyond the purine/pyrimidine distinction, there is a positive relationship between the magnitude of regression and the molecular weight of the nucleotide. Guanine is the heaviest molecule, followed by adenine, uracil and cytosine (Fig 6b). The regressions were the largest for heaviest molecules and decreased as the molecular weights decreased.

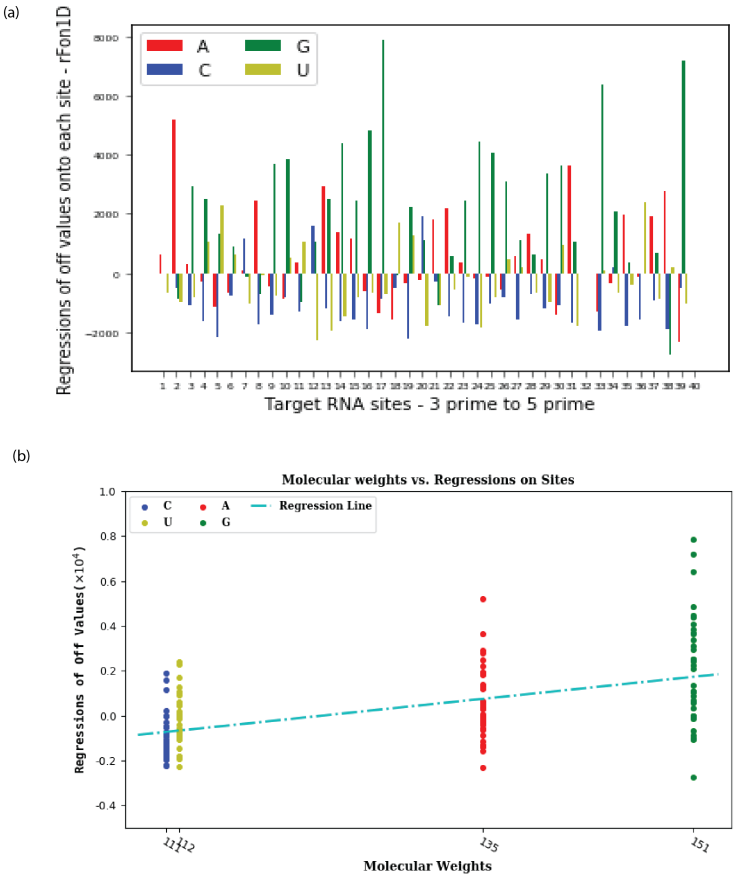

Fig 6. Regressions onto target RNA sites. (a) Regressions of off values onto each site of the target RNA orthogonalized within each vector, 3' to 5' ($([[\phi_{\mathbb{M}\bullet}]])$). (b) Regressions of off values onto target RNA sites increase with increasing molecular weights of the nucleotides.

Regressions of ON values onto STAR sites

In addition to projecting OFF values onto target RNA sites, ON values were projected onto the space of the sequences comprising the STAR linear region. The ON values are a measure of how good a STAR is at binding to the target RNA and activating the downstream gene. Since the STAR sequence is complementary to the target RNA, as expected, the regressions of having pyrimidines in the STAR linear region are positive while the regressions of having purines is negative (Fig 7).

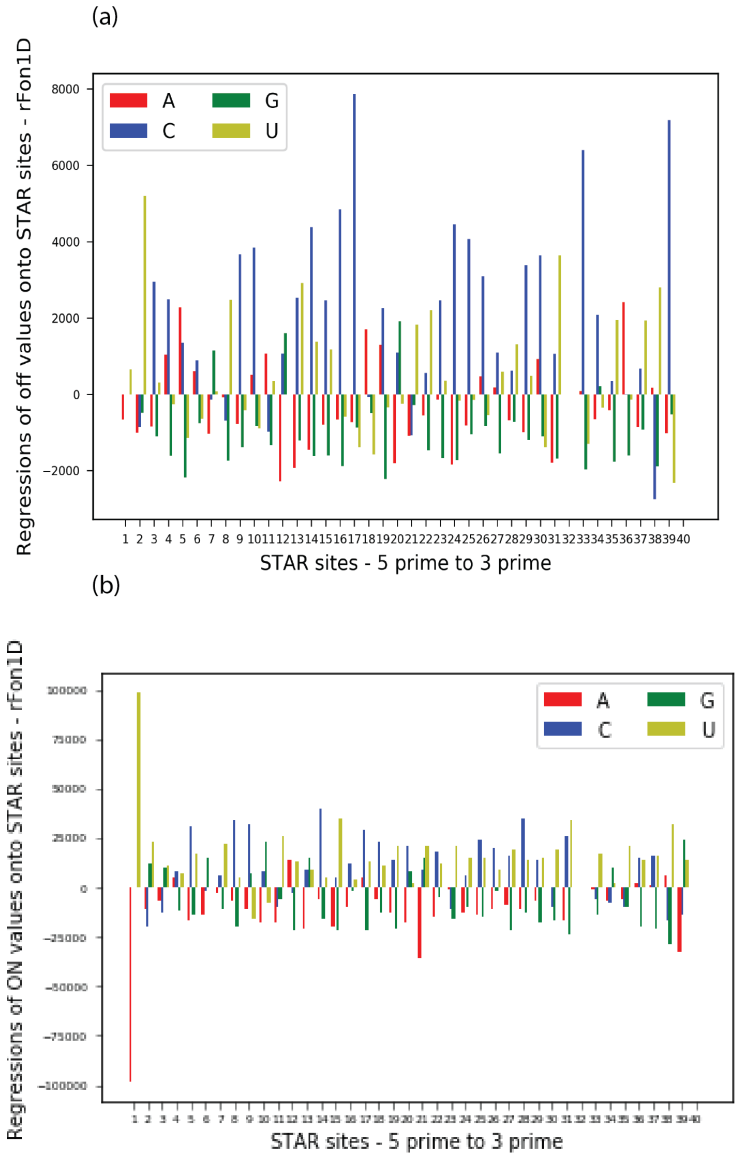

Fig 7. Regressions onto STAR sites. (a) Regressions of off values onto each STAR site orthogonalized within each vector (5' to 3'). (b) Regressions of ON values onto STAR sites (orthogonalized within each vector).

2nd order analysis on all pairs of sites across the STAR sequence

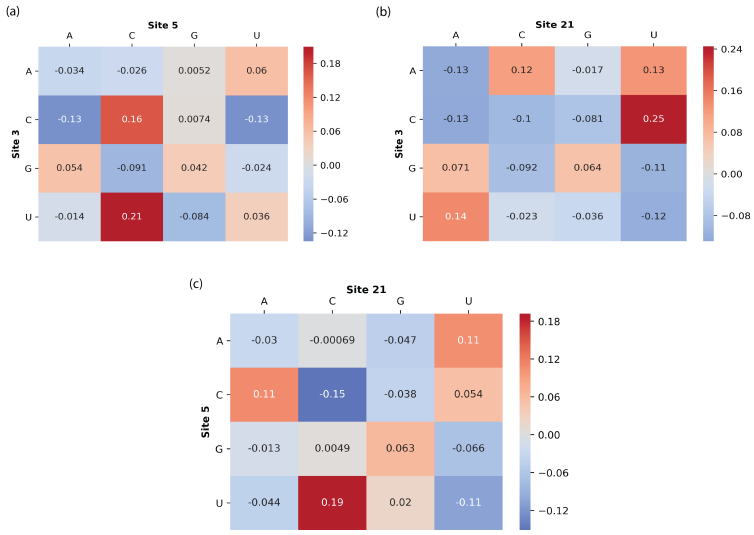

Fig 8. Regressions of off values onto the second order orthogonal polynomial when including the entire sequence. (a) shows regressions of OFF values on pairs of nucleotides, one at site 3 and one at site 5, independent of the first order contributions of each site. (b) shows regressions of OFF values on pairs of nucleotides at site 3 and at site 21. (c) shows regressions of OFF values on pairs of nucleotides at site 5 and at site 21. All values are scaled by the absolute value of the largest regression in the set of all combinations of pairs (i.e., 780 pairs formed by two sites across the 40 site long sequence).

The first order analysis, which includes calculating covariances and variances of nucleotides at each site, revealed a number of highly correlated sites (absolute values of covariances being greater than .05) sequestered in the 5' region of STAR. In order to test the hypothesis that sites that are correlated (Fig 5) are interacting in their effect on the phenotype, we built second order polynomials for each pair of sites across the STAR sequence (with 780 unique pairs in total). See the corresponding supplementary section that shows an example of how this is done with a different set of sites.

After building second order polynomials for each pair of sites across the STAR sequence, regressions of the phenotype onto each pair of sites were computed. For a given pair of sites, this results in a 4x4 matrix, with 16 total values that each correspond to a given nucleotide at the first site and another nucleotide at the second site. Fig 8 shows an example of this for sites 3, 5 and 21. The regressions shown in this figure are scaled by the absolute value of the largest regression across all combinations of pairs in the sequence.

To determine the effect of increasing distance between a pair of sites on the phenotype, we took the distances between the sites and plotted them against the absolute values of the maximum regressions (these are regressions of the phenotype onto two sites at a time). Fig 9(a) shows the sampling distribution of these regressions. It can be seen that a few regressions exist at the very tail end of the distribution. These regressions are also shown as a cloud of points that appear in the top part of Fig 9(b). To visualize which sites made up these site pairs and which nucleotides at these sites correspond to the high positive or negative regressions, a plot similar to the covariance

figure (Fig. 4) was constructed (Fig. 9c). For example, the blue curve connecting 'A' at site 5 with 'C' at site 15 means that having that combination contributes substantially to the OFF value, independently of the individual contributions of these sites by themselves. This visualization shows that there is no tendency for strongly interacting sites to be adjacent. In addition, there does not appear to be high regressions of the phenotype onto site pairs in the 5' part of the sequence (as predicted by the covariance structure in Fig 5).

Out of the 14 site pairs shown in Fig 9, there are 4 pairs that are between sites in the 5' and 3' regions. This potentially supports the hypothesis mentioned earlier which states that possible binding between these parts of the sequence would allow the hairpin region of the RNA to stay intact (by not binding with the hairpin) and pursue its function of transcription termination of the downstream gene. However, in addition to this long-range interaction, there are a cluster of strongly interacting sites in the middle portion of the sequence, between sites 11 and 20. This includes site 17 which has the most interactions (a 'C' at site 17 interacting with four other sites). This suggests that there might exist some level of potential intramolecular binding in the middle section of the sequence.

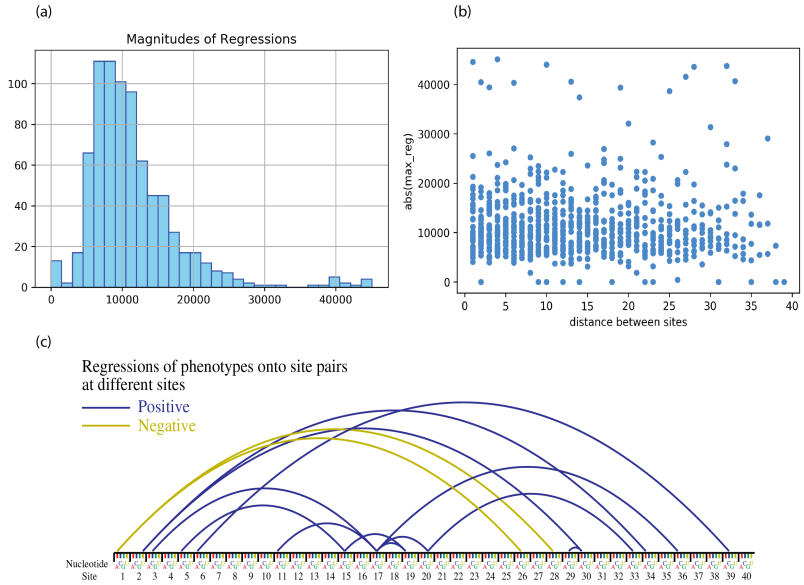

Fig 9. Regressions of the phenotype onto 780 site pairs along the STAR sequence. (a) shows the sampling distribution of absolute values of the largest regressions of the phenotype onto combination of two sites. Notable regressions are those at the tail end of the distribution, (> 35,000). (b) Relationship between distances across STAR sites and absolute values of the largest regressions. (c) To further zoom into the 14 pairs of sites with notable regressions, as depicted in (a) and (b), curves are drawn between the interacting nucleotide at one site and the corresponding nucleotide at another site. This indicates a more global structure across the STAR sequence, instead of a case in which regressions of the phenotype onto pairs of sites in the 5' region of the sequence are greater than those in the 3' region.

Discussion

In this work, we propose a novel mathematical tool to describe sites along biological sequences as vectors and quantify sequence-function relationships by projecting phenotypes onto the sequence space. Given a set of sequences and corresponding phenotypic data for each sequence, tensor-based orthogonal polynomials can be constructed based on the actual variation in sequences. The regression of phenotypes onto these polynomials can quantify not only the effects of different nucleotides at individual sites, but also the effect on phenotype of combinations of nucleotides at different sites. To show proof of concept, this method was applied to a case of synthetic RNA regulatory sequences, described in previously published work [12], that were experimentally constructed with the goal of identifying how sequence structure and design motifs affect RNA regulatory activity.

Application of our method to the case of these regulatory RNAs showcase that even though a cluster of sites near the 5' end are correlated with other sites throughout the sequence, there is no obvious preference for correlations concentrated at the 3' end. After identifying sites along the STAR sequence that are correlated with each other, we built second and third order orthogonal polynomials that quantified the effect of the phenotype on two-way and three-way combinations of sites (Supplementary Figs S3-S8).

In order to test whether the 5' region contained combinations of sites that had a greater effect on the phenotype than those at the 3' end, we built second order polynomials of combinations of two pairs of sites across the STAR sequence. Since the sequence is 40 sites long, there were 780 unique combinations. When projecting the phenotype onto these pairs of sites, we assessed the impact of distance between the pair of sites on the degree to which they influence phenotype, and whether combinations of sites in the 5' region would have a greater effect on phenotype. As seen in Fig 9, there seems to be a global structure to the interacting site pairs across the sequence that have high regressions. In particular, site 17, which is located approximately in the middle of the linear 40-site region, connects with four other sites, two upstream and two downstream.

The interaction data (Fig 9) does show some long range interactions, but it also shows a cluster of interacting sites in the middle. Furthermore, the pattern of interaction between sites in their impact on phenotype is not predicted by the nucleotide correlations between sites. Interactions at the 3' and 5' ends of the site, combined with those occurring in the middle of the sequence, suggest the possibility of competing intramolecular interactions and secondary structure formation in the absence of the target RNA. The authors utilized NUPACK aiming to minimize competing intramolecular interactions, however, our analyses indicate that there exist substantial competing intramolecular interactions that can interfere with the intermolecular interaction between the STAR:target complex. This aspect was not predicted sufficiently by the NUPACK algorithm. This presents an exciting opportunity to use novel computational and mathematical design approaches to inform experimental data and use the results to refine the design approach.

While direct Watson-Crick binding of the STAR to the target is a critical component of the STAR function, the function of this molecule will also likely be impacted by its potential to form secondary structures in the absence of the target RNA. Some level of secondary structure could be beneficial, perhaps by providing the STAR a level of stability from spontaneous degradation or nuclease-mediated degradation whereas other

structures may negatively impact its stability for the same reasons. Additionally, a high amount of structural potential may create too much intramolecular binding and not allow for the intermolecular binding between the STAR and the target. However, for other reasons that might not be easily predicted, a certain level of STAR structure might be beneficial in the intermolecular binding.

In conclusion, our methods provide a mathematical tool to find patterns in sequence data and to quantify the effect of the corresponding phenotype on the underlying sequence structure. Using this vector-based orthogonal polynomial approach, we can not only look at global patterns of sequence structure but can also identify the nucleotide state at each given site and how this affects phenotype at first and higher order levels. While we have given proof of concept of this approach using an example of regulatory RNAs, this method can be applied to other questions that aim to understand sequence-phenotype interactions such as transcription factor binding sites and how their sequence composition gives rise to different TF binding energies along with applications in synthetic biology that aim to understand the relationships between the underlying RNA sequence and corresponding secondary structure [18]- [19].

Supporting information

S1 Appendix. Supplementary Information Supplementary figures and information.

S2 Appendix. Supplementary Methods Mathematical background on tensor-valued orthogonal polynomials for sequences.

Acknowledgments

The authors would like to thank Saugato Rahman Dhruba, Olga Botvinnik, and Vijayanta Jain for their insightful comments and ideas. We'd like to thank the authors of the STAR RNA paper, including James Chappell, for doing the initial work that we used to apply our methods to and for stimulating discussions and comments. We'd like to acknowledge the TTU CISER program and the TTU Department of Biological Sciences for their support throughout the time this work was being done.

References

1. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007;8,995–1005. doi:10.1038/nrm2281.
2. Kel A, Goessling E, Reuter I, Cheremushkin E, Kel-Margoulis O, Wingender E. Match(TM) : A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 2003;31, 3576-3579.
3. Ho-Sui S, Mortimer J, Arenillas D, Brumm J, Walsh C, Kennedy B, et al. oPOSSUM: Identification of over-represented transcription factor binding sites in co-expressed genes. Nucleic Acids Res. 2005;33(10):3154-64.
4. Moss W. Methods in Enzymology, Volume 530 # 2013 Elsevier Inc. ISSN 0076-6879. doi:10.1016/B978-0-12-420037-1.00001-4.
5. Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinformatics. 2004. doi:10.1186/1471-2105-5-104.
6. Wakeman C, Winkler W, Dann C. Structural features of metabolite-sensing riboswitches. Trends in Biochemical Sciences. 2007. doi:10.1016/j.tibs.2007.08.005.
7. Chappell J, Watters K, Takahashi M, Lucks J. A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future. Current Opinion in Chemical Biology. 2015. doi:10.1016/j.cbpa.2015.05.018.
8. Mutalik V, Qi L, Guimaraes J, Lucks J, Arkin A. Rationally designed families of orthogonal RNA regulators of translation. Nat Chem Biol. 2012. doi: 10.1038/nchembio.919.

9. Lucks J, Qi L, Mutalik V, Wang D, Arkin A. Versatile RNA-sensing transcriptional regulators for engineering genetic networks. Proc Natl Acad Sci USA. 2011;108(21):8617–8622. doi:10.1073/pnas.1015741108.

10. Zeng Z, Huang B, Huang S, et al. The development of a sensitive fluorescent protein-based transcript reporter for high throughput screening of negative modulators of lncRNAs. Genes Dis. 2018;5(1):62–74. doi:10.1016/j.gendis.2018.02.001.

11. Rice S. Universal rules for the interaction of selection and transmission in evolution. Philosophical Transactions of the Royal Society B. 2020. doi:10.1098/rstb.2019-0353.

12. Chappell J, Westbrook A, Verosloff M et al. Computational design of small transcription activating RNAs for versatile and dynamic gene regulation. Nat Commun. 2017; 8,1051.

13. Krebs J, Goldstein E, Kilpatrick T. Lewin's Genes XII. 12th ed. Jones & Bartlett Learning; 2018.

14. Zadeh J, Steenberg C, Bois J, et al. NUPACK: analysis and design of nucleic acid systems. J Comput Chem. 2011. 32:170–173, 2011.

15. Caldelari I, Chao Y, Romby P, and Vogel J. RNA-Mediated Regulation in Pathogenic Bacteria. Cold Spring Harb Perspect Med 2013;3:a010298. doi:10.1101/cshperspect.a010298

16. Takahashi M, Lucks J. A modular strategy for engineering orthogonal chimeric RNA transcription regulators. Nucleic acids research. 2013. doi:10.1093/nar/gkt452.

17. Bervoets, Indra & Charlier, Daniel. Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. 2019. FEMS Microbiology Reviews. doi:10.1093/femsre/fuz001/5306444.

18. Le D & Shimko T et al. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1715888115.

19. Singh J, Hanson J, Paliwal K et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun 10, 5407. 2019. doi:10.1038/s41467-019-13395-9.