Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 02 May 2022
Application Notes Category

OXFORD

Computational Genomics

Nucleotide substitution rates are the same in 5'-UTR exons and lncRNAs for two distant members of the mammalian family tree

Joshua Morin-Baxter [1,*], Talya Koschitzky [1] and Robert Becker [1,*]

[1]Computational Genomics Class, Columbia University, New York, 10027, United States

Abstract

Motivation: A surplus of new genomic sequencing data in recent years invites the application of old tools like substitution matrices to brand new questions. In particular, the Zoonomia project provides many new genome assemblies and alignments for the mammalian family tree. The value of pan-genomic datasets like these is becoming apparent when it comes to answering biologically relevant questions about evolutionary pressure on organisms at the basepair level.
Results: We demonstrate how the pan-genome alignment from the Zoonomia project can be used to infer that rates for every possible transition or transversion of nucleotides is essentially identical between 5'-UTR exons and lncRNA regions among two distant relatives in the class mammalia. This suggests that the evolutionary pressure on these regions is agnostic to the basepair composition of changes, an interesting conclusion given the disparate biological roles these two genomic elements play.
Availability and Supplementary Information: All data, including supplementary data, along with all code is available on our GitHub at https://github.com/joshuamb/ZoonomiaCG.

1 Introduction

Scoring matrices, used to encode the likelihood that a given symbol in one biologically significant sequence transforms into a different symbol, usually through mutation, are at the heart of many important sequencing applications. They are often the front-line tool when it comes to comparing important biological strings such as DNA, RNA, and proteins. With a surplus of new sequencing data in recent years comes a new opportunity to revisit many well-worn scoring matrices in many different fields as well as to develop new ones which previously were not possible. In particular, the Zoonomia project (Genereux *et al.*, 2020) recently provided the sequencing community with genome assemblies for over 140 previously unsequenced vertebrate species and a pan-genome alignment of over 240 species, primarily in the class mammalia. Here, we demonstrate a simple use of this self-described genomic 'multitool' to investigate the per-basepair substitution rates between 5'-UTR exons and lncRNA in two relatively distant members of the mammalian family tree, *Homo sapiens* and *Solenodon paradoxus*. We show that the substitution rates per nucleotide are identical in both types of genomic regions despite the very different biological roles of these sequence types and the many years of evolution separating *S. paradoxus* and *H. sapiens*.

2 Sequence Acquisition

2.1 5'-UTR and lncRNA sequeneces from *H. sapiens*

Human genomic data with carefully curated annotation of genomic regions, such as 5'UTR exons and lncRNA, is widely available from many different sources (Zerbino *et al.* (2020)). While Zoonomia is a step toward this level of annotation in other species, for most non-model organisms like *S. paradoxus*, curated annotation data is unavailable and must be investigated *de novo*. For this reason, we started our investigation with human 5'-UTR regions downloaded from the UCSC Table Browser (Lee *et al.* (2022), track: NCBI Refseq) and human lncRNA sequences from the LNCipedia high-confidence set (Volders *et al.* (2019)). The Zoonomia pan-genome alignment provided a direct mapping from these regions in *H. sapiens* to those in *S. paradoxus*.

2.2 Modification of SNP extraction utility

The Zoonomia pan-genome alignment is 804GB and is provided in the specialized HAL format, designed for storing and analyzing multiple genome alignments as described by Hickey *et al.* (2013). Exploration of this data is most easily done via the use of haltools by the same authors. In particular, we used their single-nucleotide polymorphism (SNP) extraction utility, which queries the pan-genome alignment for basepairs that have been aligned between two species of interest but which are not a match for one another, all over a genomic region of interest. Because the optimization

of scoring substitution matrices also requires sequence information about aligned basepairs that *are* a match for one another, we were required to modify this utility to output all aligned basepairs, whether they were SNPs or not. The regions of interest are those described in the previous section.

2.3 Sampling regions of interest

The actual optimization of our scoring matrices was computationally a very difficult task. For this reason, we generally did not fit substitution matrices to the entire human lncRNA or 5'-UTR exon space. Instead, we performed a number of optimizations on samples selected uniformly from the set of lncRNA or 5'-UTR exons and confirmed that the matrices produced were relatively invariant to changes in samples or sample size (data not shown; see supplemental data files on our Github). The specific substitution matrices provided in this paper are taken from a sample of 20 lncRNA regions (134,681 basepairs, 21.3% aligned) and 202 5'-UTR exons (101,731 basepairs, 7.35% aligned). Considering only the aligned basepairs within each genomic region type, the ratio of SNP to exact match is similar at 1:3.42 (5'-UTR exons) and 1:3.5 (lncRNA), respectively.

3 Optimization Algorithm

3.1 General Algorithm

If we make the assumption that basepair transitions are independent of one another, the question of finding substitution matrix **A** that captures the probabilities of each symbol transition between the sequences x and y is a question of optimizing the following function:

$$f(x, y, \mathbf{A}, T) = \prod_i \mathbf{A}^T[x_i][y_i] \quad (1)$$

Where T represents the number of iterations in which a given symbol may have transformed. We began with an initial guess for A with all matrix entries equal to 0.25, representing uniform probability for each transition. We then utilized the Trust-Region Constrained Algorithm (Lalee *et al.*, 1998). This allowed us to optimize our likelihood function subject to constraints. Our constraints on A were that the rows of A must sum to 1, and each entry in A must be between 0 and 1 to represent a probability. We used the python library scipy.optimize to implement this algorithm. For our analysis, we further considered individual 5'-UTR exons to be independent and x_i was created by concatenating the aligned sequences of each randomly sampled exon. The same was done for lncRNA.

3.2 Modifications to the Algorithm

We discovered that optimizing $\mathbf{A}^T$ to be problematic when A is raised to a high power. Specifically we found that our toy examples outputted expected behavior when A was not raised to a power or was raised to a low power, but when T was large, the added complexity made it difficult for the algorithm to find an optimal solution in practice. We attempted to implement other basic optimization methods in place of the Trust Region Constrained Algorithm (such as stochastic gradient descent), but those methods did not work well as this is fundamentally a constrained optimization problem; without the constraints, all of the values in the matrix A will simply blow up.

Finding the optimal $\mathbf{A}^T$ is far more tractable than finding **A**. In this way our function becomes

$$f(x, y, \mathbf{AT}) = \prod_i \mathbf{AT}[x_i][y_i] \quad (2)$$

5'-UTR Exon Regions

	A	G	C	T
A	0.76	0.06	0.12	0.06
G	0.10	0.63	0.06	0.22
C	0.21	0.06	0.63	0.09
T	0.07	0.13	0.05	0.75

Long Non-coding RNA

	A	G	C	T
A	0.76	0.05	0.12	0.07
G	0.10	0.61	0.05	0.24
C	0.23	0.05	0.62	0.10
T	0.07	0.12	0.05	0.77

Table 1. Substitution scoring matrices for UTR and lncRNA

Fig. 1. Caption, caption.

4 Final Matrices and Discussion

The matrices are presented in Table 1.

5 Conclusions

Our study has several major limitations. One is the limited scope of the sequences and species we compared: Our analysis examined only a very small portion of the genome of only two candidate species. It is also worth noting that our annotation data is one-way: all the regions we investigated were annotated as human 5'-UTR exons or human lncRNA and did not consider whether these annotations applied to *S. paradoxus*, though the per-basepair alignment data provided by the Zoonomia pan-genome alignment provides some justification for this selection.

Furthermore, our analysis relied on the calculation of 'exponentiated' substitution matrices ($\mathbf{A}^T$), rather than the true scoring matrices (**A**) to reduce the computational costs and increase the efficacy of our optimization algorithms. This makes our results useful for demonstrating our negative results but not for performing actual sequence comparisons. Figuring out an efficient way to optimize the likelihood function that includes time would also aid with the limitation of performing the analysis on only two species. Ideally, this analysis could be performed across all 240 Zoonomia species, with each pair of species having a unique value of T.

Despite these limitations, our investigation does produce the biologically meaningful suggestion that mutation in 5'-UTR exons and lncRNA, within the class mammalia, is not differentiated by the type of substitutions that are most likely to occur. If corroborated in larger studies, this has interesting implications for the evolutionary pressure that these regions experience. In particular, it may suggest that the evolutionary pressure on these regions is agnostic to the basepair composition of genetic changes. This is an interesting conclusion given the disparate biological roles of these two genomic elements. TODO WHAT ARE THESE IMPLICATIONS

Acknowledgements

Thanks to Dr. Pe'er, Philippe Chlenski, and Ziyuan Jiang for their help this semester. Thanks also to our classmates for their support and feedback as well as we developed this project.

Funding

This work has been supported by Columbia University.

References

Genereux, D. P., Serres, A., Armstrong, J., Johnson, J., Marinescu, V. D., MurÃ©n, E., Juan, D., Bejerano, G., Casewell, N. R., Chemnick, L. G., Damas, J., Di Palma, F., Diekhans, M., Fiddes, I. T., Garber, M., Gladyshev, V. N., Goodman, L., Haerty, W., Houck, M. L., Hubley, R., Kivioja, T., Koepfli, K.-P., Kuderna, L. F. K., Lander, E. S., Meadows, J. R. S., Murphy, W. J., Nash, W., Noh, H. J., Nweeia, M., Pfenning, A. R., Pollard, K. S., Ray, D. A., Shapiro, B., Smit, A. F. A., Springer, M. S., Steiner, C. C., Swofford, R., Taipale, J., Teeling, E. C., Turner-Maier, J., Alfoldi, J., Birren, B., Ryder, O. A., Lewin, H. A., Paten, B., Marques-Bonet, T., Lindblad-Toh, K., Karlsson, E. K., and Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, **587**(7833), 240–245.

Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, **29**(10), 1341–1342.

Lalee, M., Jorge, N., and Plantenga, T. (1998). On the implementation of an algorithm for large-scale equality constrained optimization. *SIAM J. OPTIM*, **8**(3), 682 – 706.

Lee, B. T., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Muthuraman, P., Nassar, L. R., Nguy, B., Pereira, T., Perez, G., Raney, B. J., Rosenbloom, K. R., Schmelter, D., Speir, M. L., Wick, B. D., Zweig, A. S., Haussler, D., Kuhn, R. M., Haeussler, M., and Kent, W. J. (2022). The UCSC genome browser database: 2022 update. *Nucleic Acids Res.*, **50**(D1), D1115–D1122.

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.*, **47**(D1), D135–D139.

Zerbino, D. R., Frankish, A., and Flicek, P. (2020). Progress, Challenges, and Surprises in Annotating the Human Genome. *Annual review of genomics and human genetics*, **21**, 55–79. Edition: 2020/05/18.