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 the rate of any particular basepair substitution observed between two distant relatives in the class mammalia, *H. Sapiens* and *S. paradoxus*, is identical for 5'-UTR exons and lncRNA regions. This suggests that the evolutionary pressure on these regions is agnostic to the basepair composition of changes, an interesting conclusion given the disparate biological mechanisms of these genetic elements.
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. While protein coding regions are typically most meaningfully compared at the codon or amino acid level, genetic elements like 5'-UTR (untranslated region) exons are interesting to compare at a basepair granularity because many of their attributes are emergent properties of their basepair content (Statello *et al.*, 2021). For instance, 5'-UTR regions have been shown to be GC-rich, which is responsible for many important secondary structures that contribute to their influence on translation (Babendure *et al.*, 2006). The more enigmatic long-noncoding RNA (lncRNA), on the other hand, also shown to be a prolific genetic regulator via a variety of different mechanisms, is intriguing for undergoing gene-like processing by cellular machinery and the role of basepair content and primary sequence in function is less clear (Chakraborty *et al.*, 2014).

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 potentially disparate biological consequences of their sequence content 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 SNP Extraction

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 used this utility in a modified form by requiring a SNP to be reported even when there was no difference between the aligned basepairs. All code is available on our GitHub. The regions of interest are those described in the previous section.

2.3 Sampling Regions of Interest

The optimization of our scoring matrices was a very computationally demanding 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 21 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. To produce our final sequences for analysis, all aligned 5'-UTR exons were concatenated into a single sequence and all lncRNA were concatenated into another one (see independence assumptions in the next section). These concatenated sequences became the input for the next stage of our analysis.

3 Sequence Processing

3.1 Maximizing Likelihood Model

We begin by making the assumption that basepair substitutions are independent of one another because it allows us to model the divergence of two sequences as resulting entirely from an accumulation of single point mutations. This is a highly simplified model of the genome that allowed us to examine substitution differences at a basepair level.

Thus, the question of finding substitution matrix $\mathbf{A}$ that captures the probabilities of each basepair substitution between the sequences x and y can be formulated as maximizing the following likelihood function:

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

Where T represents the number of iterations in which a given symbol may have transformed. The optimal $\mathbf{A}$ then represents the one that maximizes the probability that sequence x becomes sequence y.

3.2 Optimization With Constraints

We began with an initial guess for $\mathbf{A}$ with all entries equal to 0.25, representing uniform probability for each transformation, then used the implementation of the Trust-Region Constrained Algorithm in the python library `scipy.optimize` to optimize our likelihood function subject to constraints (Lalee *et al.*, 1998). The constraints on $\mathbf{A}$ were those generally required by stochastic matrices: in particular, the rows of $\mathbf{A}$ must sum to 1 and each entry in $\mathbf{A}$ must be a valid probability value between 0 and 1.

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

3.3 Modifications to the Algorithm

Finding an optimal $\mathbf{A}$ given the parameter T proved highly problematic. Specifically, when the likelihood function involved raising $\mathbf{A}$ to any power other than 1, the added complexity made it difficult for the algorithm to converge to an optimal solution. Many other basic optimization methods, such as stochastic gradient descent, proved infeasible because the stochastic matrix constraints are a key property of this problem. Approaches such as differential evolution showed some promise but proved computationally intractable. Because the scope of our analysis was already limited to two species, we altered our approach and solved for the optimized $\mathbf{A}^T$, instead of solving for $\mathbf{A}$. This made comparing the sequence substitution rates between the two species a far less computationally demanding problem.

In this way our function becomes

$$f(x, y, \boldsymbol{A^T}) = \prod_i \boldsymbol{A^T}[x_i][y_i] \tag{2}$$

4 Results

The matrices are presented in Table 1.

5 Discussion and 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. Unfortunately, relying on this annotation data is itself a different source of bias as sequence similarity is a driving force in the alignment process in the first place. For example, it is very possible many many SNP-heavy regions of interest were not aligned precisely because they were SNP-heavy. The elucidation of lncRNA is also still somewhat ambiguous, making our analysis potentially sensitive to the selection of the lncRNA database.

Furthermore, our analysis relied on the calculation of 'exponentiated' substitution matrices ($\mathbf{A}^T$), rather than the true scoring matrices ($\mathbf{A}$) 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. Developing an efficient optimization procedure for a likelihood function that includes time would also make expanding our results to more than two species more tenable.

Despite these limitations, our investigation does produce the biologically meaningful suggestion that mutation in 5'-UTR exons and

lncRNA, at least 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. It may suggest that the evolutionary pressure on these regions is agnostic to the basepair composition of genetic changes, which is interesting considering the potentially very different implications of sequence variation within each of these molecule classes. Most importantly, our analysis demonstrates a concrete use of a massive pan-genome alignment to examine evolutionary pressure at a basepair level. Ideally, this type of analysis could eventually be extended to all 240 Zoonomia species, with each pair of species having a unique value of T, and could lead to many interesting avenues of research into how DNA changes over time.

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

Babendure, J. *et al.* (2006). Control of mammalian translation by mRNA structure near caps. *RNA*, **12**(5), 851–861.

Chakraborty, S., Deb, A., Maji, R. K., Saha, S., and Ghosh, Z. (2014). LncRBase: an enriched resource for lncRNA information. *PLoS One*, **9**(9), e108010.

Genereux, D. *et al.* (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, **587**(7833), 240–245.

Hickey, G. *et al.* (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, **29**(10), 1341–1342.

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

Lee, B. *et al.* (2022). The UCSC genome browser database: 2022 update. *Nucleic Acids Res.*, **50**(D1), D1115–D1122.

Statello, L. *et al.* (2021). Gene regulation by long non-coding rnas and its biological functions. *Nature Reviews Molecular Cell Biology*, **22**(2), 96–118.

Volders, P. *et al.* (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.*, **47**(D1), D135–D139.

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