The International EFAL-IT BLOG
Information Technology *innovations in Economics, Finance, Accounting, and Law*
Volume 3 – Issue 2/2022 – Bacau (Romania)

Dubai, 14.02.2022

Revisiting the number of substitutions in Tandem Repeats

AUTHOR
Dr Atheer Matroud
Senior Lecturer – De Montfort University Dubai, Dubai, UAE

ABSTRACT

In this paper, the expected number of substitutions occurring in tandem repeats is investigated.
Single point substitutions happen in the duplication history of tandem repeats. These single substitutions produce a set of variants that we observe at the current time. The number of substitutions that occurred in the past is unknown, a lower bound on the number of substitutions is the length of the Steiner tree of the variants, which can be approximated by the length of the minimum spanning tree of the variants' graphs. However, there might be some parallel substitutions that happened in the duplication history where an existing variant is created again. In this section, we calculate the likelihood $P(k,i)$ that i observed substitutions are the result of $k \ge i$ substitutions on a motif of size n. There are $3n$ possible substitutions, where n is the length of the motif.
A substitution can either be a new substitution or it can be parallel (it duplicates an existing substitution); in the second case the number of substitutions does not increase. If we observe i substitutions after k substitutions have occurred, then the $k-$th substitution produced either a new substitution (with probability $\frac{3n-(i-1)}{3n}$), or reproduced an existing substitution (a parallel substitution with probability $\frac{i}{3n}$). Thus the probability $P(k,i)$ of observing i substitutions after $k>0$ substitutions can be calculated using the recursive formula
$$P(k,i)=P(k-1,i-1)\times \frac{3n-(i-1)}{3n}+P(k-1,i)\times \frac{i}{3n},$$
$$\text{ for } k>0,i>0,$$
with initial values
$$P(0,1)=1, P(0,i)=0, P(k,0)=0 \text{ for } i \neq 1.$$
Figure 2.2 plots $P(k,i)$ for $n=11$ and $i=19$, as is the case for the tandem motifs in NZ1 [1]. This suggest that the number of parallel substitutions in NZ1 is more likely to be in the range 5--13.

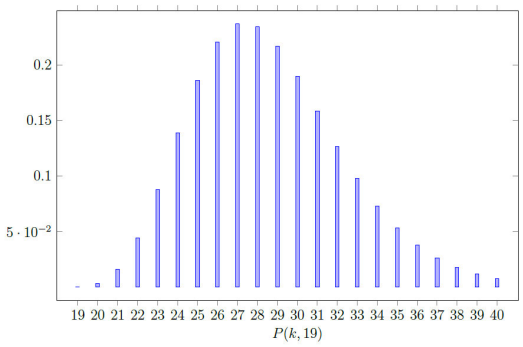

Figure 2.2: The likelihood distribution of the number of substitutions to produce 19 variants of a motif of length 11 bp.

REFERENCES

- Matroud, Atheer, Christopher Tuffley, and Michael D. Hendy. "An asymmetric alignment algorithm for estimating ancestor-descendant edit distance for tandem repeats." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021).
- Matroud, A. A., M. D. Hendy, and C. P. Tuffley. "NTRFinder: an algorithm to find nested tandem repeats." *arXiv preprint arXiv:1006.1730* (2011).
- MATROUD, AA, MD HENDY, and CP TUFFLEY. "NTRFINDER: AN ALGORITHM TO FIND NESTED TANDEM REPEATS."
- Matroud, Atheer. *Nested tandem repeat computation and analysis*. Diss. Ph. D. Thesis, Massey University, 2013.
- Matroud, A. A., M. D. Hendy, and C. P. Tuffley. "NTRFinder: an algorithm to find nested tandem repeats." *arXiv preprint arXiv:1006.1730* (2011).
- Matroud, Atheer A., Christopher P. Tuffley, and Michael D. Hendy. "An algorithm to solve the motif alignment problem for approximate nested tandem repeats in biological sequences." *Journal of Computational Biology* 18.9 (2011): 1211-1218.
- Matroud, Atheer A., Michael D. Hendy, and Christopher P. Tuffley. "An algorithm to solve the motif alignment Problem for approximate nested tandem repeats." *RECOMB International Workshop on Comparative Genomics*. Springer, Berlin, Heidelberg, 2010.
- Matroud, Atheer A., Michael D. Hendy, and Christopher P. Tuffley. "An algorithm to solve the motif alignment Problem for approximate nested tandem repeats." *RECOMB International Workshop on Comparative Genomics*. Springer, Berlin, Heidelberg, 2010.
- Matroud, Atheer A., Christopher P. Tuffley, and Michael D. Hendy. "An algorithm to solve the motif alignment problem for approximate nested tandem repeats in biological sequences." *Journal of Computational Biology* 18.9 (2011): 1211-1218.
- Matroud, Atheer A., et al. "A comparison of three heuristic methods for solving the parsing problem for tandem repeats." *Brazilian Symposium on Bioinformatics*. Springer, Berlin, Heidelberg, 2012.