JOURNAL OF
EMERGING INVESTIGATORS

Article

A novel approach to determine which organism best displays Gijswijt's Sequence in its genome

Archit Chaturvedi[1], Wendy Slijk[1]
[1] Canyon Crest Academy, San Diego, California

SUMMARY

The genomes of organisms are filled with hidden mathematical sequences present in the sequence of the nitrogenous bases that make up DNA. Using mathematical and computational tools, many of these mathematical sequences can be deciphered from the genetic code. Gijswijt's Sequence is a mathematical integer sequence named after Dion Gijswijt and is a self-describing sequence where each term counts the maximum number of repeated blocks of numbers in the sequence that precedes the term. Therefore, we are interested in finding the organism that best displays Gijswijt's Sequence in its genome, or has a DNA sequence somewhere in its genome that is very similar to the integer sequence. We hypothesized that there is an organism that displays Gijswijt's Sequence multiple times in its genome with significant matches to the input DNA sequences, and therefore, is the organism that best displays the sequence in its genome. We applied Gijswijt's Sequence to the different permutations of the four nitrogenous bases present in DNA and using the Basic Local Alignment Search Tool (BLAST) through BioPython in the Python programming language, we concluded that the common carp best displays Gijswijt's Sequence in its genome.

INTRODUCTION

Gijswijt's Sequence is a mathematical integer sequence that can be defined recursively as:

$$a(1) = 1$$
$$a(n+1) = k$$

where the sequence starts with $a(1) = 1$ and goes on to xy^k. Furthermore, x and y are strings of integers, where x is the initial string of integers and can be empty, y is a nonempty string, and k is as large as possible. Then, the next term is k (1, 2). Each term of the sequence denotes the maximum run of terms of any length. The first 28 terms of Gijswijt's Sequence go as follows: 1, 1, 2, 1, 1, 2, 2, 2, 3, 1, 1, 2, 1, 1, 2, 2, 2, 3, 2, 1, 1, 2, 1, 1, 2, 2, 2, 3 (3). The main focus of this study is not the mathematical concept of Gijswijt's Sequence in itself, but rather applying the sequence to the nitrogenous bases present in DNA.

The four nitrogenous bases of DNA are adenine (A), thymine (T), cytosine (C), and guanine (G). When DNA is assembled, the sequence of these nitrogenous bases in DNA leads to the formation of an organism's genome. The sequence of these nitrogenous bases in an organism is referred to as the organism's DNA sequence. This ultimately gives rise to the main question of this study: "Which organism best displays Gijswijt's Sequence in its genome?" The hypothesis of this study is that there is an organism that displays Gijswijt's Sequence multiple times in its genome with significant matches to the input DNA sequences. The organism that best displays the integer sequence in its genome is one that meets the criteria that is used to subjectify this "best" organism. This is particularly important, as it would decipher a mathematical sequence from an organism's genome, and further studies can use the methods of this study to explore the presence of other integer sequences in the genomes of organisms. The BLAST biotechnology tool allows us to compare the DNA sequence that is input to a large database of DNA sequences present in different genomes of many organisms. The expected (E)-value displayed based on a matching result can be used to determine the significance of the match in a BLAST search. A lower E-value indicates that the match is more significant. The Python programming language can be used to run a BLAST search for a specific DNA sequence. These DNA sequences are taken out of the NCBI database of the genome assemblies of many different organisms (4, 5). Using the Bio.Blast package in BioPython, a user can run the BLAST search for their DNA sequence, and the Python interpreter would display resulting matches along with their

Figure 1: A flow chart to briefly outline the overall design of the study.

E-values (6). The final conclusion of this study is the organism that best displays the sequence in its genome, based on the data analysis that is performed in this study.

RESULTS

The experimental aspect of this study was to run a BLAST search on each of the permutations of the nitrogenous bases of DNA to which Gijswijt's Sequence is applied to **(Figure 1)**. It should be noted that in the permutations, no base of DNA was repeated twice. This gives a set of data values that includes the organisms output by the BLAST searches that are run for the input DNA sequences, along with the E-values of each match. The organisms that have a matching DNA sequence for a specific permutation were recorded in a data table, along with the data values that represent the E-values collected (7-12) **(Table 1)**.

The criteria that we used to determine the organism that *best* displays the integer sequence in their genome has two factors: The E-value is less than 0.1 and as close to zero as possible, and the organism displays the sequence for more than one permutation.

It should be noted that if an organism displayed the integer sequence multiple times in its genome with a higher E-value than another organism that displayed it fewer times with a lower E-value, then as long as the E-value is less than 0.1, the organism that displays the sequence more times displays Gijswijt's Sequence in its genome with a better match.

Cyprinus carpio displays Gijswijt's Sequence for four different permutations in its genome, while the other organisms present only display the sequence for one single permutation **(Figure 2)**. Out of the four permutations for which *C. carpio* displays the sequence, three of the permutations lead to matches that have E-values less than 0.1, and two of these permutations lead to matches that have E-values less than 0.01, meaning that the matches are significant **(Table 1)**. Therefore, *C. carpio* best matches the criteria described above, and we conclude that *C. carpio*, also known as the common carp, is indeed the organism that best displays Gijswijt's Sequence in its genome.

Permutation of Bases	Organism	E-value
CGTA	*C. carpio*	0.00118489
GTAC	*C. carpio*	0.00413566
TACG	*C. carpio*	0.0503827
ACGT	*C. carpio*	0.175853
ACGT	*D. elegans*	0.175853
CATG	*D. willistoni*	0.175853
GACT	*E. affinis*	0.00413566
GCAT	*P. reticulata*	0.0503827
CAGT	*S. viridis cultivar*	0.175853

Table 1: Results of the BLAST searches, for all of the organisms that displayed the sequence for particular permutations of the nitrogenous bases, along with the E-values of the matches. The permutations that do not result in any matches with genomes from other organisms were excluded from Table 1, as they were not significant to the results of this study.

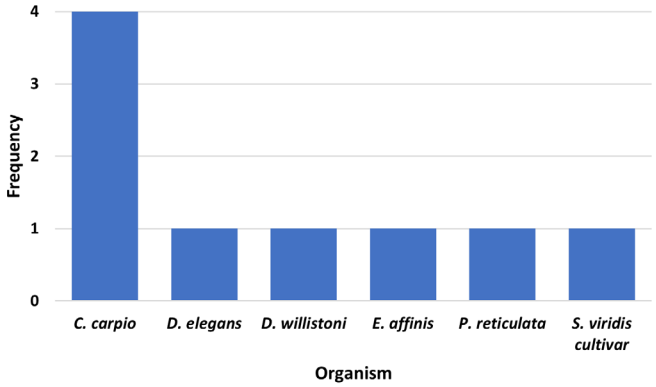

Figure 2: A bar graph to represent the frequency of Gijswijt's Sequence in each organism's genome, based on the results output by the Python interpreter.

DISCUSSION

Based on the experiment and data analysis conducted in this study, we concluded that the common carp best displays Gijswijt's Sequence in its genome. The final result also supports the hypothesis, as the common carp displays the sequence multiple times in its genome, based on the permutations of the nitrogenous bases, and the matches are significant (E-value less than 0.01).

Numerical Representation

There are a couple of advantages to the method we applied in this study. In this study, Gijswijt's Sequence is applied to the different permutations of the nitrogenous bases present in DNA. The permutations of the bases of DNA allow for a proper method to apply Gijswijt's Sequence to genomics. There are definitely other ways to arrange the nitrogenous bases of DNA, to which Gijswijt's Sequence can be applied to. But applying the sequence to the permutations of the bases, such that none of the bases are repeated, is an extremely feasible way of doing so, since in terms of combinatorics and probability theory, the permutations of A, T, C, and G, allow for a proper mathematical way to arrange the nitrogenous bases of DNA, and based on this effective arrangement, Gijswijt's Sequence can be applied. It should also be noted that all four bases of DNA are used the same number of times in each permutation, which prevents any bias in terms of how many times a certain base is used when applying the terms of Gijswijt's Sequence to the permutations.

Other Findings

Although we determined the common carp best displays the integer sequence somewhere in its genome, we found that other organisms also display the sequence in their genome.

From the experimental results, the following organisms also display the sequence in their genome with a relatively low E-value (8-12): *D. willistoni*, *D. elegans*, *S. viridis cultivar*, *E. affinis*, and *P. reticulata*.

Sources of Bias

A notable source of bias in this study is that we used 28 terms from Gijswijt's Sequence. If fewer terms were used, the BLAST search may have resulted in more matches, and if more terms were used, then the BLAST search may have resulted in fewer matches. Another potential source of bias is the E-value threshold that is used in the Python program. While no organisms that display the sequence have an E-value that is greater than 0.2, having an E-value threshold of 0.5 may result in matches that are not very significant. However, this does not affect any of the data that is collected in this study, because all values were less than 0.2. A third potential source of bias in this study is the criteria used to assess the collected data. While this leads to the desired conclusion of this study, the criteria in itself is relatively subjective, and can differ from person to person. It should also be noted that BLAST in itself does not give a perfect alignment, and a "-" denotes regions of the matching genome where the alignments are different. Therefore, the matched sequences in the organisms' genomes do not give a perfect representation of Gijswijt's Sequence as described in this study, but rather give a close representation of the sequence.

Pandoravirus Salinus

Another point to be discussed about the data collected in this study is the exclusion of one of the genomes that was output from the BLAST search (13). The genome of *Pandoravirus salinus* also displays the sequence with an E-value of 0.175853 for the TGCA permutation, meaning that the match is relatively significant. However, a virus does not display all of the properties of life, and therefore, the virus is excluded from **Table 1**, due to the fact that it does not classify as an organism.

Future Applications

There are many future areas of study relating to this study. It should be noted that Gijswijt's Sequence is an extremely slow progressing integer sequence, with the number 5 appearing as the ten to the ten to the 23rd term of the sequence. Other studies can also be conducted to determine which organisms best display other slow-progressing integer sequences in their genome, such as the Kolakoski Sequence, which is in nature, extremely similar to Gijswijt's Sequence (1). Another area of study is to look closer at the common carp's genome and try to find other mathematical sequences present in its genome. There are many ways that other studies can branch from this study, but overall, they would all relate to deciphering the mathematical integer sequences present in the genomes of different organisms. Another future direction to go from this study is to determine which organism best displays Gijswijt's Sequence in its genome for more than 28 terms, where the number of terms is a multiple of 4. Perhaps having more terms could result in a match that has a lower E-value, meaning that the matched organism would display the sequence with even greater accuracy and significance than that of a match based on 28 terms.

MATERIALS AND METHODS

Permutations of Nitrogenous Bases

The first step in applying Gijswijt's Sequence to the nitrogenous bases was to determine all of the possible permutations for the nitrogenous bases themselves. It should also be noted that in these permutations, no base was repeated twice. In other words, each of A, T, C, and G was only present once in each permutation. As there are four different nitrogenous bases, the total number of permutations is equal to 4! or 24 total permutations.

Applying Gijswijt's Sequence

For each of these permutations, the first 28 terms of Gijswijt's Sequence were applied to each of the permutations for the nitrogenous bases described above. We used the first 28 terms of the sequence because 28 terms provides a relatively large sequence size, from which the BLAST search can be performed upon without any conflicts pertaining to the E-value threshold of the BLAST program, and 28 is a multiple of 4, which prevents any bias relating to the four nitrogenous bases. In other words, all of A, T, C, and G are used the same number of times (7 per permutation) based on the terms of Gijswijt's Sequence. Take for example ATCG. When Gijswijt's Sequence is applied, the resulting DNA sequence that is input for the BLAST search is: ATCCGATTCCGGAAATCGGATCCGGAATTTCCGATTCG AATTCCGGG, where the 28 terms of the sequence are split into groups of four terms, and each nitrogenous base is repeated a certain number of times, based on the term in each grouping respectively. The terms are applied based on the order of the nitrogenous bases described by the permutation itself. The implementation of Gijswijt's Sequence described above is performed for each of the 24 permutations of the nitrogenous bases. The resulting DNA sequences are each 47 bases long.

Python Program

Next, we wrote a Python program to perform the BLAST searches required for this study. We used Python to conduct the BLAST searches because the results were output in a specific format that were simple to understand and record. Also, the E-values that are output by the Python interpreter include more significant figures, providing a better representation for the match. Two files were used in the program for this study. The first file was called *main.py*, and it contained the code that ran the BLAST search on the DNA sequence. The second file was called *dnaseq.fa*. This file contained the DNA sequences that the BLAST search was

```
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML
fasta_string = open("dnaseq.fa").read()
result_handle = NCBIWWW.qblast("blastn", "nt",
fasta_string)
blast_record = NCBIXML.read(result_handle)
E_VALUE_THRESH = 0.5
for alignment in blast_record.alignments:
 for hsp in alignment.hsps:
  if hsp.expect < E_VALUE_THRESH:
   print("****Alignment****")
   print("sequence:", alignment.title)
   print("length:", alignment.length)
   print("e value:", hsp.expect)
   print(hsp.query)
   print(hsp.match)
   print(hsp.sbjct)
```

Figure 3: Python code used for BLAST searches to identify organisms iwth genomes containing Gijswijt's Sequence.

performed on and was written in the FASTA format.

Using the Bio.Blast package present in BioPython (14), a piece of code was written that performs the BLAST search on the DNA sequences, and also outputs the results of the search in the desired format. Writing the Python program used in this study was relatively simple, and overall follows the same structure for any program that would run a BLAST search on a *FASTA* file. The Python version used for this code was *Python 3.8.2* (15). **Figure 3** shows the code that performed the BLAST searches for the DNA sequences and output the results.

The E-value threshold for the Python program was set to 0.5, in order to allow the interpreter to output a higher number of significant matches for organisms that display the integer sequence somewhere in their genome. The word size and match/mismatch score parameters were set to their default values of 28 and 1, -2 respectively. Matches with an E-value of 0.1 to 0.5 were also recorded **(Table 1)** and were included when counting the total number of times an individual organism displays the integer sequence in its genome **(Figure 2)**.

Output and Data Analysis

The output of the Python program has the name of the organism, the E-value of the match, as well as the particular DNA sequence to which the match is made. The output is easy to understand, and all of the permutations that result in a match display results in a similar format. The output results were recorded for each of the permutations that result in an organism that matches the DNA sequence **(Table 1)**. The permutations that do not result in any matches were denoted by "No matches found" and were not included in the final data table shown in **Table 1**. Using this data, along with the criteria discussed in the Results section of the paper, we determined which organism best displays the sequence in its genome.

ACKNOWLEDGMENTS

I would like to thank my family for supporting me throughout my research. I would also like to acknowledge the NCBI databases for containing the data that is used across this study. And finally, I would also like to acknowledge the *repl.it* online IDE, as the Python program used in this study is coded using the IDE.

Received: January 1, 2021
Accepted: December 19, 2021
Published: January 29, 2022

REFERENCES

1. Van de Bult, Fokko J., *et al. A Slow-Growing Sequence Defined by an Unusual Recurrence,* 22 Feb. 2006, pp. 1–24.
2. Sloane, N. J. A. *Seven Staggering Sequences.* 3 Apr. 2006, pp. 3–5.
3. Gijswijt, Dion. "A090822." OEIS, *oeis.org/A090822.*
4. "BLAST: Basic Local Alignment Search Tool." *National Center for Biotechnology Information,* U.S. National Library of Medicine, blast.ncbi.nlm.nih.gov/blast.cgi.
5. *National Center for Biotechnology Information,* U.S. National Library of Medicine, www.ncbi.nlm.nih.gov./.
6. "Biopython - Overview of BLAST." *Tutorialspoint,* www.tutorialspoint.com/biopython/biopython_overview_of_blast.htm.
7. "*Cyprinus Carpio* (ID 10839)." *National Center for Biotechnology Information,* U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genome/10839?genome_assembly_id=291130.
8. "Drosophila Willistoni Annotation Report." *National Center for Biotechnology Information,* U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genome/annotation_euk/Drosophila_willistoni/101.
9. "*Drosophila Elegans* Annotation Report." *National Center for Biotechnology Information,* U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/genome/annotation_euk/Drosophila_elegans/101.
10. "*Eurytemora Affinis* Strain: Atlantic Clade (ID 203087) - BioProject - NCBI." *Eurytemora Affinis Annotation Report,* 13 Mar. 2014, www.ncbi.nlm.nih.gov/bioproject/203087.
11. "*Poecilia Reticulata* Annotation Report." *Poecilia Reticulata Annotation Report,* 24 June 2014, www.ncbi.nlm.nih.gov/genome/annotation_euk/Poecilia_reticulata/100.
12. "Setaria_viridis_v2.0 - Genome - Assembly - NCBI." *www.ncbi.nlm.nih.gov,* www.ncbi.nlm.nih.gov/assembly/GCF_005286985.1/.
13. "ViralProj215788 - Genome - Assembly - NCBI." *www.ncbi.nlm.nih.gov,* www.ncbi.nlm.nih.gov/assembly/GCF_000911955.1/.
14. Cock, P.J. et al., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25(11), pp.1422–1423

15. "Python Release Python 3.8.2." *Python.org,* www.python.org/downloads/ release/python-382/.