Journal of Biological Research

Bollettino della Società Italiana di Biologia Sperimentale

eISSN 2284-0230 - pISSN 1826-883
https://www.pagepressjournals.org/index.php/jbr/index

J Biol Res 2026 [Online ahead of print]

To cite this Article:
Alaqeeli O. **Trinucleotide Repeat (TNR) expansions mining algorithm based on the Apriori algorithm.** *J Biol Res* doi: 10.4081/jbr.2026.14607

Submitted: 25 October 2025
Accepted: 4 June 2026
Early access: 7 August 2026

Trinucleotide Repeat (TNR) expansions mining algorithm based on the Apriori algorithm

Omar Alaqeeli

Department of Computer Science, Saudi Electronic University, Riyadh, Saudi Arabia

Correspondence: Omar Alaqeeli, Department of Computer Science, Saudi Electronic University, Riyadh, Saudi Arabia. Tel.: 966565781953. E-mail: o.alaqeeli@seu.edu.sa

Key words: Apriori algorithm; DNA mining; genetic disease; trinucleotide; codon.

Abstract

The Apriori algorithm is a data mining algorithm used exclusively for identifying associations between itemsets in a series of business transactions. In this paper, we modify this algorithm to identify associations between trinucleotide repeat expansions in DNA, which are known for their associations with specific genetic diseases. The proposed modification here is to treat all possible forms of DNA trinucleotides as item sets and the list of all DNA segments as a series of transactions. In the first iteration of the algorithm, the number of occurrences of every single trinucleotide is counted across all DNA segments. Then, the results are filtered out according to a preset threshold. In the second iteration, the number of occurrences of every pair of trinucleotides is counted, and similarly, the results are filtered out according to a threshold. In the third iteration, the number of occurrences of every trio of trinucleotides is handled similarly. The algorithm continues until no results are produced. Lastly, the association rules are formed, yielding the conclusion of the existence of a certain genetic disease will also mean an existence of another genetic disease.

Introduction

Trinucleotide repeat expansions[1–3] are short segments of gene sequences whose frequency is proven to be associated with genetic diseases.[4–6] An individual with a specific range of trinucleotide repeats, caused by inheritance or mutation,[7,8] will probably develop a neurological disorder in a later stage of his or her life. The existence of such trinucleotide repeats, and if their RNA is a coding

RNA,[9–11] will disturb the production of proteins, thus they are abnormally expressed, which will eventually lead to the development of the disorder. These types of disorder are usually characterized by symptoms of cognitive, movement, and psychiatric impairments. The most well-known cases of these genetic disorders are Huntington's disease,[12,13] Fragile X syndrome,[14,15] Myotonic dystrophy,[16,17] and Spinocerebellar ataxias (SCAs).[18,19]

A trinucleotide consists of 3 out of 4 nucleotides: Adenine, Thymine, Guanine, and Cytosine. There are 4^3 (64) possible formations of trinucleotides. The repetitions of some of these formations are proven to be associated with certain genetic diseases. For example, Huntington's disease, a neurodegenerative disorder,[20-22] is characterized by the repetition of CAG[23-25] (Cytosine, Adenine, Guanine) form. When it exceeds a threshold (40 or more consecutive repeats), an individual will develop the disorder at a later stage of his or her life.

The Apriori algorithm[26–28] is a data mining recursive algorithm[29–31] that is used to identify frequency and associations between itemsets that appear in a series of transactions. It is almost exclusively used for mining business transactions. The traditional Apriori algorithm works as follows: In the first iteration, the algorithm iterates over all transactions and scans the list of itemsets to find the frequency of appearance of each item, which is known as the "Support Count", as well as the Support Percent-age. Assuming that T is the set of all transactions, $T = \{T_1, T_2, \ldots, T_n\}$, N is the total number of all transactions, and X is an itemset. The Support Count is defined as $\text{supp count}(X) = |\{T_i \in N \mid X \subseteq T_i\}|$, that is, the number of appearances of X in all transactions. The Support Percentage is defined as

$$\text{supp}(X) = \frac{\text{supp count}(x)}{N} \times 100\%$$

The list of support for all items will be refined based on a preset threshold, so an item with a support percentage of less than 60%, for example, will be eliminated. Determining the threshold is based on certain criteria decided on an individual basis and set up by the algorithm implementer. In the second iteration, the algorithm scans for the appearance of each distinguished pair of items and creates the support table for that pair, that is, the frequency of appearance of the pair in each transaction. Again, the support table for all possible pairs will be refined according to the preset threshold, and any pair that has a support percentage below the threshold will be omitted from the table. In the third iteration, the support table for three items is created, and so on. The algorithm continues iterating until the list of itemsets in the support table remains unchanged.

The final step of the algorithm is establishing the association rules to determine the confidence of the existence between individual items, pairs, or more items.. Assuming that X and Y are itemsets, and $X \rightarrow Y$ is the association rule. The confidence is defined as:

$$\text{supp}(X) = \frac{\text{supp count}(X \cup Y)}{\text{supp count}(X)}$$

Thus, the existence of a specific item or a specific pair of items means that there is, for example, a 65% chance of another specific item or pair to exist.

Several adaptations and improvements have been made to the Apriori algorithm. One improvement is to resolve the issue of over-scanning the database of transactions and producing large sets of candidates by using mapping and further pruning of the itemsets list to improve the efficiency of the algorithm.[32] In another improvement, this problem was approached by restructuring the storage of the database and reducing the number of itemsets to avoid unnecessary candidate sets that need to be added to the itemsets list.[33] In another optimization, the algorithm was adapted to solve the problem of frequent itemset counting at its initial iteration by shortening itemset list with effective pruning.[34] In another literature, the Apriori algorithm was integrated with other outdated open source tools to improve its performance.[35] Moreover, one adaptation of the algorithm is combing the classic Apriori algorithm with prefix tree technique. The purpose is to detect motifs in a series of DNA sequences.[36] Nevertheless, all these efforts are related to the optimization of the efficiency of the algorithm. None was specifically related to the domain of trinucleotide repeat expansions, which makes our modification of the Apriori algorithm unique in its nature.

The current form of the Apriori algorithm is not applicable in the domain of DNA sequences simply because identifying the frequency of a single nucleotide does not have any phenotypic meaning.[37–39] In the following sections, we will explain how we modified the conventional Apriori algorithm to be able to identify trinucleotide repeat expansions and to find the associations between different genetic disorders that are caused by trinucleotide repeat expansions in DNA.

Materials and Methods

The methodology of our modified Apriori algorithm is to look at the DNA segment as a collection of trinucleotides rather than a collection of single nucleotides. Since DNA sequences are composed of a series of bases (A, G, T, and C), there are 64 possible formations of trinucleotides. The set of all

of these formations is F = {ACT, ATA, TGT, ACA, CAA, GAA, CTC, TCG, CCA, GGT, GTC, TCC, GTG, CCG, TTT, CGC, TTA, GTT, AGT, TCT, ATC, CGT, TCA, GGG, TAG, GAG, TGA, ACC, GCC, GGC, AGG, GGA, ATG, AAC, AAT, CTA, CGA, GCA, CAG, CTG, GAC, TAC, ATT, CTT, ACG, GAT, AGA, TGG, TTC, TTG, GTA, CGG, AAA, CCC, TAA, AGC, GCT, CAT, CAC, TAT, AAG, TGC, CCT, GCG}. These forms represent itemsets while DNA segments represent transactions (a DNA segment here could be an Exome, an Exon, an Intron, or even an entire chromosome. So, in the case of Huntington's disease, for example, the group of DNA segments can be a group of Hunt-ington loci HTT (Huntingtin) genes (or IT15) or a group of complete chromosome 4 where the *HTT* gene is located. As a preprocessing step, the algorithm iterates over the four nucleotides (A, G, T, and C) to create the set of all possible forms of trinucleotides. In the first iteration of the algorithm, each individual form (item) in the DNA sequence (transaction) is counted, and the table of support and percentage of each form is created. Any trinucleotide with a support less than the threshold will be omitted from the table. Since our purpose is to look for associations between genetic diseases, the threshold should be set to include the lowest number of pathogenic trinucleotide repeats that are proven to be associated with a genetic disease. For example, in Spinocerebellar ataxia type 6 (SCA6), the lowest number of pathogenic CAG trinucleotide repeats is 21, and at this point an individual will develop speech and muscle movement impairment. Thus, if a trinucleotide's support is less than 21, then that trinucleotide will be omitted from the support table.

In the second iteration, the table of support and percentage of pairs of trinucleotides is created, and then any pair with a support that is less than the previously determined threshold will be omitted. In the third iteration, the table of support for three trinucleotides is created and refined. The algorithm continues until only itemsets with supports of equal to or more than threshold remain.

Let us assume that we have the following four DNA segments:

AGGAGGAGGGATGAGGAGCCCGCCCTCCTGGGCTTTAGTGGCGCTTAGGAGCCTGGGAGCT
TCCCGAAAAAAAAAACCCCCCCCCCCCCTTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAGGGG
GGGGGGGGGGG
AGGAGGAGGGATGAGGAGCCCGCCCTCCTGGGCTTTAGTGGCGCTTAGGAGCCTGGGAGCT
TCCCGAAAAAAAAAACCCCCCCCCCCCCTTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAGGCG
GCGGCGGCGGC

AGGAGGAGGGATGAGGAGCCCGCCCTCCTGGGCTTTAGTGGCGCTTAGGAGCCTGGGAGCT
TCCCGAAAAAAAAAACCCCCCCCCCCCCTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAGGCG
GCGGCGGC
AGGAGGAGGGATGAGGAGCCCGCCCTCCTGGGCTTTAGTGGCGCTTAGGAGCCTGGGAGCT
TCCCGAAAAAAAAAACCCCCCCCCCCCCTTTTTTTTTTTTTTTAAAAAAAAAAAAGGCGGCGGCGGC
GGCGGCGGC

For demonstration purposes, we will consider any trinucleotide that has five or more repeats in an individual segment as significant, and it will be added to the support and percentage table. Also, we will assume that the threshold of the support percentage is 50%. In the first iteration of the algorithm, our support and percentage (Table 1) will be as follows: CGG, GCG, and GGG have support of less than 50% and appear in only one segment; thus, these trinucleotides will be omitted. In the second iteration, the support of possible pairs of trinucleotides (Table 1) is created. The pairs AAA, GGC, and GGC, TTT have supports of less than 50% thus they will be omitted. At the third iteration of three possible trinucleotides, AAA, GGC, and TTT, is created. But this trio contains a subset that was found infrequent, AAA and GGC, thus it is going to be pruned. The only remaining set of trinucleotides with a support of above the threshold is the pair of AAA and TTT (in the second iteration), which has a support of 3 and a support percentage of 75%.
This finding is interpreted as there is an association between the existence of AAA and the existence of TTT trinucleotides. An additional step in the algorithm is to calculate the confidence of this existence, which is defined by the formula:

$$\text{Confidence}(\text{AAA} \rightarrow \text{TTT}) = \frac{\text{Support of AAA \& TTT}}{\text{Support of AAA}} = \frac{3}{3} = 100\%$$

Thus, whenever the trinucleotide AAA is found in a DNA segment, TTT will also be found. On the other hand, if we, for example, calculate the confidence of the existence of GGC and TTT, then it will be only a 50% chance for that coexistence;

$$\text{Confidence}(\text{GGC} \rightarrow \text{TTT}) = \frac{\text{Support of GGC \& TTT}}{\text{Support of GGC}} = \frac{1}{2} = 50\%$$

Applying the algorithm on real a world example, we collected 30 datasets (Sequence Read Archive Study ID: SRP305710, Gene Expression Omnibus accession ID: GSE166567) from the NCBI (The National Center for Biotechnology Information). The datasets are for the *HTT* gene (Huntington's gene) and consist of two groups: 20 pathological and 10 control. The total size of all datasets is 18.03G bases.

The threshold of a significant repeat expansion40 to be considered pathogenic is any expansion that is greater than 26 trinucleotide repeats. We will also set the threshold for the support percentage as 60% or above.

Results

In the first iteration of the algorithm, the Table 2 of support count and support percentage for individual trinucleotides is created (note that some of the expansions yield for more than one trinucleotide. For example, the expansion of GAAGAAGAA yields either GAA, AAG, or AGA). CAA/ACA, GGA/GAG, and TGA/GAT/ATG appear once, CAG/AGC/GCA appears 4 times, and TTT has a very low support percentage; thus all of them will be omitted. In the second iteration, the table of support percentage will be as shown in Table 3.

The frequency of GAA/AGA/AAG with TTC/TCT/CTT is 56.7% so this association will be omitted from the table. In the third iteration, where the association of three trinucleotides is examined, the output Table 4 of the algorithm will be as follow:

The only remaining trios of trinucleotides that have support above the threshold are GAA/AGA/AAG, AAA, GGG, and AAA, GGG, TTC/TCT/CTT. The algorithm will continue until no more support above the threshold.

We can use Table 1 to find the association between two or three trinucleotides. If we want to calculate the confidence of the existence of a pair of trinucleotides, say AAA and TTC/TCT/CTT, then:

$$\text{Confidence}(\text{AAA} \rightarrow \text{TTC/TCT/CTT}) = \frac{\text{Support of AAA \& TTC/TCT/CTT}}{\text{Support of AAA}} = \frac{18}{30} = 60\%$$

That is, if AAA exists, then there is a 60% chance that GAA/AGA/AAG will also exist. For the confidence in the existence of AAA and GGG,

$$\text{Confidence}(\text{AAA} \rightarrow \text{GGG}) = \frac{\text{Support of AAA \& GGG}}{\text{Support of AAA}} = \frac{21}{30} = 70\%$$

and for GAA/AGA/AAG and AAA, it is

$$\text{Confidence}(\text{GAA/AGA/AAG} \rightarrow \text{AAA}) = \frac{\text{Support of GAA/AGA/AAG \& AAA}}{\text{Support of GAA/AGA/AAG}} = \frac{20}{20} = 100\%$$

For trio trinucleotides (Table 1), finding the confidence of the existence of one trinucleotide will also means the existence of a pair of trinucleotides, thus

$$\text{Confidence}(\text{AAA} \rightarrow \text{GGG, TTC/TCT/CTT}) = \frac{\text{Support of AAA \& GGG, TTC/TCT/CTT}}{\text{Support of AAA}} = \frac{18}{30}$$

$$= 60\%$$

The proposed algorithm

Algorithm: Trinucleotide repeat expansions specific apriori algorithm.
Input: A list of DNA segments.
Output: A list of strongly associated trinucleotide repeat expansions.

1: F is the empty set of all possible trinucleotide formations
2: N = {A, C, G, T }, L is a list of DNA sequences
3: r is the threshold for the minimum trinucleotide repeat to be considered as signif-
icant, s is the threshold of support count
4: T is an empty table of support count and support percentage

5: **while** length of F ≤ 64 **do**

6: Randomize a trinucleotide from the set N

7: **if** trinucleotide ∉ F **then**

8: F ← F + trinucleotide

9: **end if**

10: **end while**

11: **for** i to length(F) **do**

12: **for** j to length(L) **do**

13: Find n (the highest trinucleotide repeats) of F_i in L_j

14: **if** n ≥ r **then**

15: T ←T + F_i

16: T ← T + n ▷ Add support count of F_i

17: **end if**

18: T ← T + total of all F occurrences/length(L) ▷ Add support % of F_i

19: **end for**

20: **end for**

21: **for** k in T **do**

22: **if** T_k's support_count ≤ s **then**

23: Omit T_k and its supersets from T

24: **end if**

25: **end for**

26: Repeat steps 21 - 25 for the pair T_k & $T_{k...m}$ support, then T_{k+1}, $T_{k...m}$ support . . . then for the trio T_k, T_{k+1} & $T_{k...m}$ support . . . etc.

Discussion

Modifying the Apriori algorithm expands its scope to larger data pools. The algorithm has the ability to search large databases for associations between any trinucleotides. It can also be adjusted to search for tetranucleotide, pentanucleotide, hexanucleotide, or more nucleotide repeats. We only have to adjust the possible number of formations. For tetranucleotide, there will be 4^4 possible forms, for pentanucleotide, there are 4^5, and for hexanucleotide, there are 4^6. Abstractly, 4^x where x is the number of repeats required.

However, one limitation of our modified algorithm is the challenge of determining the threshold for the support count and percentage. There is no clear guideline for choosing the appropriate threshold, as this is the case for all domains where the Apriori algorithm is applied. Another limitation is the determination of the number of repeats for an expansion to be considered abnormal. The threshold varies widely among genetic diseases, which makes it difficult to deduce a firm conclusion.

Conclusions

The current form of the Apriori algorithm is not applicable to the domain of DNA sequences simply because identifying the frequency of a single nucleotide does not have any phenotypic or pathogenic meaning.

The key modification of the Apriori algorithm here is to treat every "significant" trinucleotide repeat as an item in a transaction (DNA segments). The level of significance of the repeat expansions depends on the genetic diseases being mined. For example, in Spinocerebellar ataxia Type 6, 21 is considered pathogenic, while in Fragile X syndrome, the expansions must be more than 230 trinucleotides for pathogenic repeats.

In the case study of the *HTT* gene (Huntington's gene)[20,23] previously discussed, GAA/AGA/AAG appears with TTC/TCT/CTT but the support is only 56.7%, while GAA/AGA/AAG with AAA is 66.7%. The association here is stronger therefore required higher attention. This study will be extensively investigated in a separate paper.

References

[1] Mitas M. Trinucleotide repeats associated with human disease. Nucleic Acids Res 1997;25:2245–53.

[2] Kovtun I, McMurray C. Features of trinucleotide repeat in-stability *in vivo*. Cell Res 2008;18:198–213.

[3] Samadash-wily G, Raca G, Mirkin S. Trinucleotide repeats affect DNA replication *in vivo*. Nat. Genet 1997;17:298–304.

[4] Lieberman A, Shakkottai V, Albin R. Polyglutamine repeats in neurodegenerative diseases. Annu Rev Pathol 2019;14:1–27.

[5] Everett C, Wood N. Trinucleotide repeats and neurodegenerative disease. Brain 2004;127:2385–405.

[6] La Spada A, Paulson H, Fischbeck K. Trinucleotide repeat expansion in neurological disease. Ann Neurol 1994;36:814–22.

[7] Pearson C, Edamura K, Cleary J. Repeat in-stability: mechanisms of dynamic mutations. Nat Rev Genet 2005;6:729–42.

[8] Richards R. Dynamic mutations: a decade of unstable expanded repeats in human genetic disease. Hum Mol Genet 2001;10:2187–94.

[9] Sarkar S. Decoding "coding": Information and DNA. BioScience 1996;46:857–64.

[10] Nogueira F, De Rosa J, Vicente E, et al. RNA expression profiles and data mining of sugarcane response to low temperature. Plant Physiol 2003;132:1811–24.

[11] Fu X. Non-coding RNA: a new frontier in regulatory biology. Natl Sci Rev 2014;1:190–204.

[12] Gusella J, MacDonald M, Ambrose C, Duyao M. Molecular genetics of Huntington's disease. Arch Neurol 1993;50:1157–63.

[13] Imarisio S, Carmichael J, Korolchuk V, et al. Huntington's disease: from pathology and genetics to potential therapies. Biochem J 2008;412:191–209.

[14] Santoro M, Bray S, Warren S. Molecular mecha-nisms of fragile x syndrome: A twenty-year perspective. Annu Rev Pathol 2012;7:219–45.

[15] Jin P, Warren S. Understanding the molecular basis of fragile x syndrome. Hum Mol Genet 2000;9:901–8.

[16] Turner C, Hilton-Jones D. The myotonic dystrophies: diagnosis and management. J Neurol Neurosurg Psychiatry 2010;81:358–67.

[17] Lee J, Cooper T. Pathogenic mechanisms of myotonic dystrophy. Biochem Soc Trans 2009;37:1281–6.

[18] Manto M. The wide spectrum of spinocerebellar ataxias (scas). Cerebellum 2005;4:2–6.

[19] Koeppen A. The pathogenesis of spinocerebellar ataxia. Cerebellum 2005;4:62–73.

[20] Illarioshkin S, Igarashi S, Onodera O, et al. Trinucleotide re-peat length and rate of progression of Huntington's disease. Ann Neurol 1994;36:630–5.

[21] Kieburtz K, MacDonald M, Shih C, et al. Trinucleotide repeat length and progression of illness in Huntington's disease. J Med Genet 1994;31:872–4.

[22] Lee J, Conrad A, Epping E, et al. Effect of trinucleotide repeats in the Huntington's gene on intelligence. eBioMedicine 2018;31:47–53.

[23] Andrew S, Goldberg Y, Kremer B, et al. The relationship between trinucleotide (cag) repeat length and clinical features of Huntington's disease. Nat Genet 1993;4:398–403.
[24] Rosenblatt A, Kumar B, Mo A, et al. Age, cag repeat length, and clinical progression in Huntington's disease. Mov Disord 2012;27:272–6.
[25] Bañez-Coronel M, Porta S, Kagerbauer B, et al. A pathogenic mechanism in Huntington's disease involves small cag-repeated RNAs with neurotoxic activity. PLoS Genet 2012;8:1–15, 02.
[26] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. SIGMOD Rec 1993;22:207–16.
[27] Dongre J, Prajapati G, Tokekar S. The role of apriori algorithm for finding the association rules in data mining. Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) 2014; pp. 657–60.
[28] Yuan X. An improved apriori algorithm for mining association rules. AIP Conference Proceedings 2017;1820:080005.
[29] Larose D. Introduction to Data Mining, chapter 1, John Wiley & Sons, Ltd, 2004; pp. 1–26.
[30] Jackson J. Data mining; a conceptual overview. Commun Assoc Inf Syst 2002;8:19.
[31] Zhang S, Zhang C, Yang Q. Data preparation for data mining. Appl Artif Intell 2003;17:375–81.
[32] Yuan X. An improved apriori algorithm for mining association rules. AIP Conference Proceedings 2017;1820:080005.
[33] Du J, Zhang X, Zhang H, Chen L. Research and improvement of apriori algorithm. Proceedings of the 2016 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) 2016; pp. 117–21.
[34] Perego R, Orlando S, Palmerini P. Enhancing the apriori algorithm for frequent set counting. In Kambayashi Y, Winiwarter W, Arikawa M, eds. Data Warehousing and Knowledge Discovery. Berlin Heidelberg, Springer Nature. 2001; pp. 71–82.
[35] Aflori C, Craus M. Grid implementation of the apriori algorithm. Adv Eng Softw 2007;38:295–300.
[36] Al-Turaiki I, Badr G, Mathkour H. Trieamd: a scalable and efficient apriori motif discovery approach. Int J Data Min Bioinform 2015;13:13–30.
[37] Alaqeeli O. A comparison of dropout rate of three commonly used sin-gle cell RNA-sequencing protocols. Biotechnol Biotechnol Equip 2024;38:237837.

[38] Alaqeeli O,Alturki R. Evaluating the performance of the generalized linear model (glm) r package using single-cell RNA-sequencing data. Appl Sci 2023;13:11512.

[39] Schalling M, Hudson T, Buetow K, Housman D. Direct detection of novel expanded trinucleotide repeats in the human genome. Nat Genet 1993;4:135–9.

[40] Richards R, Sutherland G. Dynamic mutation: possible mechanisms and significance in human disease. Trends Biochem Sci 1997;22:432–6.

Table 1. Frequency of individual trinucleotides through algorithm iterations (iter).

Iter	Trinucleotide	Count	Percentage
	AAA	3	75
	CGG	1	25
1	GCG	1	25
	GGC	2	50
	GGG	1	25
	TTT	3	75
	AAA, GGC	1	25
2	GGC, TTT	1	25
	AAA, TTT	3	75
3	AAA, GGC, TTT	1	25

Table 2. Frequency of individual trinucleotides.

	Trinucleotide	Support Count	Support Percentage
1	CAA/ACA	1	3.3
2	GAA/AGA/AAG	20	66.7
3	AAA	30	100
4	GGA/GAG	1	3.3
5	TGA/GAT/ATG	1	3.3
6	CAG/AGC/GCA	4	13.3
7	TTC/TCT/CTT	18	60
8	GGG	21	70
9	TTT	11	36.7

Table 3. Frequency of pairs of trinucleotides.

	Trinucleotides	Support Count	Support Percentage
1	GAA/AGA/AAG ↔ AAA	20	66.7
2	GAA/AGA/AAG ↔ TTC/TCT/CTT	17	56.7
3	GAA/AGA/AAG ↔ GGG	20	66.7
4	AAA ↔ TTC/TCT/CTT	18	60
5	AAA ↔ GGG	21	70
6	TTC/TCT/CTT ↔ GGG	18	60

Table 4. Frequency of trios of trinucleotides.

	Trinucleotides	Support Count	Support Percentage
1	GAA/AGA/AAG ↔ AAA ↔ GGG	20	66.7
2	GAA/AGA/AAG ↔ AAA ↔ TTC/TCT/CTT	17	56.7
3	GAA/AGA/AAG ↔ GGG ↔ TTC/TCT/CTT	17	56.7
4	AAA ↔ GGG ↔ TTC/TCT/CTT	18	60

Conflict of interest: the author declares no conflict of interest.

Ethics and consent to participate: not applicable.

Data availability: datasets used in this research can be found at: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA701282 The result of processing these datasets can be found at: https://github.com/Omar-Alaqeeli/TNR-AprioriAlgorithm

Funding: no funding received for this research.