

**دانشگاه اصفهان**

**دانشکده مهندسی کامپیوتر**

گزارش پروژه

**DNA**

****

پدیدآورنده:

**محمد ­امین کیانی**

**4003613052**

دانشجوی کارشناسی، دانشکده‌ی کامپیوتر، دانشگاه اصفهان، اصفهان،

Aminkianiworkeng@gmail.com

استاد: جناب اقای دکتر نقش‌نیلچی

نیمسال اول تحصیلی 04-1403

فهرست مطالب

[مستندات 3](#_Toc187111225)

[1-مسئله و تحلیل کلی آن: 3](#_Toc187111226)

[2- توضیح کد برنامه: 6](#_Toc187111227)

[3- تحلیل نتایج: 10](#_Toc187111228)

[4- خروجی‌ها و نمودارها: 14](#_Toc187111229)

[5- مراجع 19](#_Toc187111230)

# مستندات

## 1-مسئله و تحلیل کلی آن:

DNA (دئوکسی ریبونوکلئیک اسید) مولکولی است که اطلاعات ژنتیکی موجودات زنده را ذخیره می‌کند. این مولکول از واحدهایی به نام **نوکلئوتید** تشکیل شده است. هر نوکلئوتید شامل سه بخش اصلی است:

1. یک گروه فسفات.
2. یک قند پنج‌کربنی (دئوکسی‌ریبوز).
3. یک باز نیتروژنی که می‌تواند یکی از چهار نوع زیر باشد:
   * آدنین (A)
   * تیمین (T)
   * گوانین (G)
   * سیتوزین (C)

**دی‌ان‌ای** از دو رشته مکمل تشکیل شده است که بازهای نیتروژنی آن‌ها با پیوندهای هیدروژنی به هم متصل هستند. بازهای A و T با دو پیوند هیدروژن[ی و باز](https://www.ncbi.nlm.nih.gov/genbank/)های G و C با سه پیوند هیدروژنی جفت می‌شوند.

**ژن**، واحد اصلی اطلاعات ژنتیکی در DNA است. هر ژن دنباله‌ای از نوکلئوتیدهاست که اطلاعات لازم برای ساخت پروتئین‌ها را فراهم می‌کند. ژن‌ها به صورت کدهای سه‌تایی نوکلئوتیدها (کدون‌ها) خوانده می‌شوند و هر کدون یک اسید آمینه خاص را مشخص می‌کند.

**پایداری و مقاومت حرارتی DNA**

پایداری و مقاومت حرارتی DNA به ساختار شیمیایی آن وابسته است. یکی از مهم‌ترین معیارهای بررسی این ویژگی‌ها، **درصد GC** است:

* **GC Content** به درصد بازهای گوانین (G) و سیتوزین (C) در دنباله DNA اشاره دارد.
* بازهای G و C به دلیل سه پیوند هیدروژنی که دارند، پایداری بیشتری نسبت به بازهای A و T (که فقط دو پیوند هیدروژنی دارند) ایجاد می‌کنند.
* **GC بالا** (بیش از 50%) نشان‌دهنده پایداری و مقاومت حرارتی بالاتر است.

**معرفی موجود : Sorex roboratus**

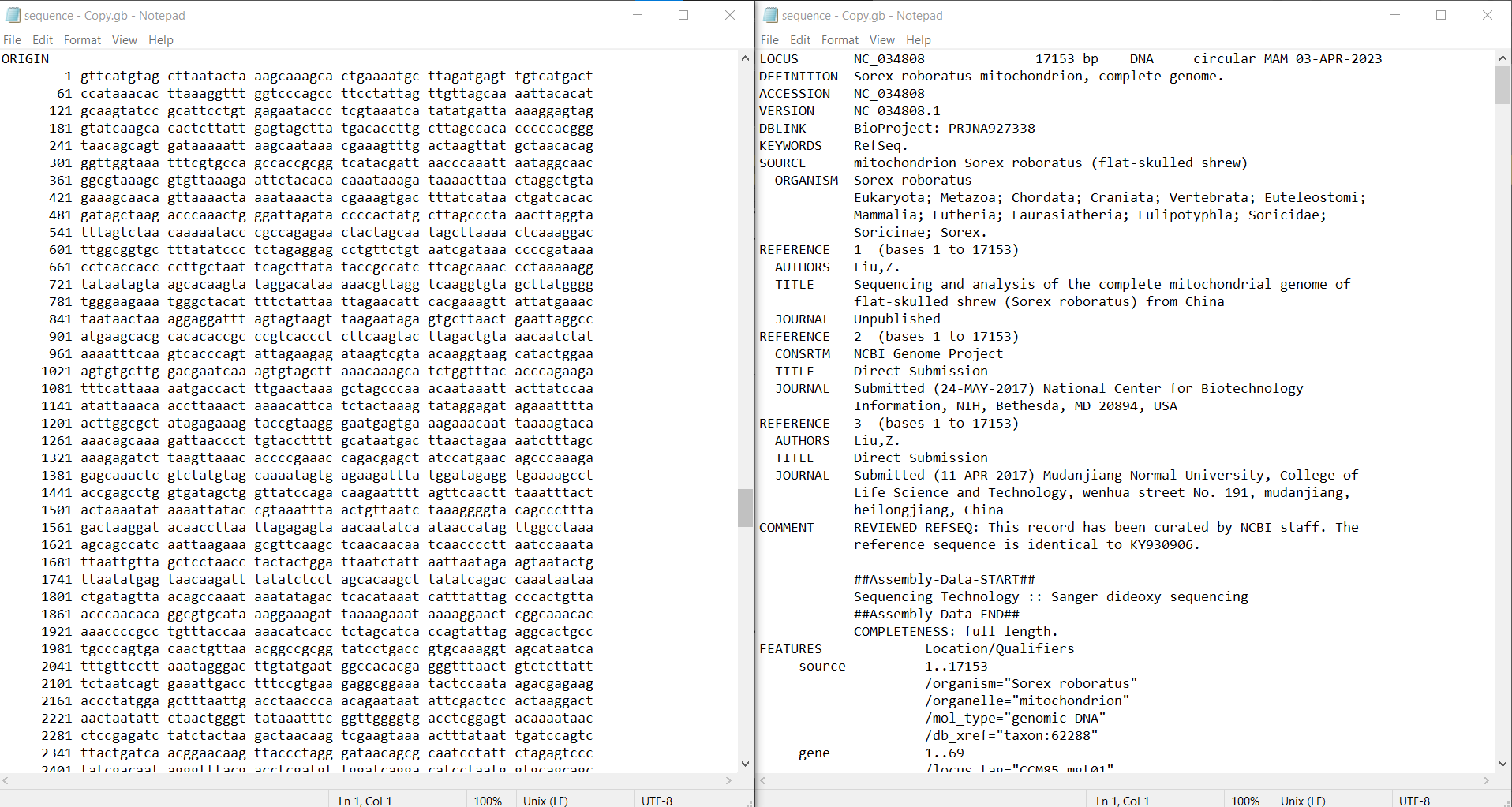
گونه‌ای از پستانداران کوچک و متعلق به خانواده **حشره‌خواران (Soricidae)** است. این موجودات به دلیل اندازه کوچک، زیستگاه‌های خاص، و نقش کلیدی در اکوسیستم‌های زمینی مورد توجه زیست‌شناسان هستند.

**ژنوم میتوکندری:**

ژنوم میتوکندری موجودات شامل اطلاعات ژنتیکی است که برای عملکرد میتوکندری‌ها لازم است. میتوکندری، اندامکی است که مسئول تولید انرژی سلول از طریق فرآیند تنفس سلولی می‌باشد. در بیشتر جانداران، ژنوم میتوکندری ساختاری حلقوی داشته و شامل ژن‌های زیر است:

* ژن‌های کدکننده پروتئین‌های مرتبط با زنجیره انتقال الکترون.
* ژن‌های کدکننده tRNA و rRNA برای سنتز پروتئین.

ژنوم میتوکندری به دلیل اندازه کوچک و سرعت تکامل سریع، یکی از بهترین منابع برای مطالعات ژنتیکی و فیلوژنتیکی است.



**توضیحات دنباله DNA**

ژنوم میتوکندری Sorex roboratus شامل حدود 17,000 جفت باز (bp) است که تمام ژن‌های مورد نیاز برای عملکرد میتوکندری را در بر می‌گیرد. دنباله DNA شامل نواحی کدکننده ژن‌ها و نواحی غیرکدکننده (کنترلی) است.

**اهمیت درصد GC در ژنوم میتوکندری**

درصد GC به دلیل تاثیر آن بر پایداری ساختار DNA از اهمیت ویژه‌ای برخوردار است:

* بازهای **گوانین (G)** و **سیتوزین (C)** دارای سه پیوند هیدروژنی هستند که باعث افزایش پایداری حرارتی DNA می‌شود.
* در ژنوم میتوکندری، درصد GC معمولاً پایین‌تر از ژنوم هسته‌ای است و این به دلیل ساختار و عملکرد متفاوت آن است.

**مراحل انجام پروژه:**

**1- دریافت دنباله DNA از GenBank**

**GenBank**  یک پایگاه داده زیستی شامل اطلاعات مربوط به توالی‌های DNA است که برای دریافت دنباله باید مراحل زیر را طی کنیم:

1. به وب‌سایت [GenBank](https://www.ncbi.nlm.nih.gov/genbank/) مراجعه می‌کنیم.
2. ژن مورد نظر خود را جستجو می‌کنیم (**Sorex roboratus mitochondrion**).
3. دنباله ژن را انتخاب کرده و فایل آن را با فرمت .gb دانلود می‌کنیم. ژنوم میتوکندری Sorex roboratus را می‌توان از پایگاه داده GenBank دریافت کرد و فایل با فرمت .gb حاوی اطلاعات کامل ژنوم است.

**2- تحلیل دنباله DNA با Python**

به کمک زبان پایتون انجام موارد زیر را انجام می‌دهیم:

**1. شمارش نوکلئوتیدها**

نوکلئوتیدهای ساده شامل A، T، G و C هستند. برنامه تعداد هر یک از این نوکلئوتیدها را محاسبه کرده و درصد آنها را نسبت به کل دنباله گزارش می‌کند.

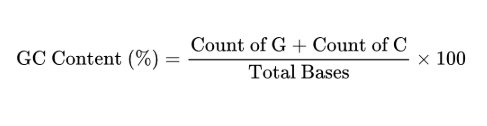
**2. شمارش دی‌نوکلئوتیدها و تری‌نوکلئوتیدها**

* **دی‌نوکلئوتید**: دنباله‌های دو حرفی مانند AT، GC و غیره.
* **تری‌نوکلئوتید**: دنباله‌های سه حرفی مانند ATG، CGT و غیره.

برنامه تعداد این کلمات را محاسبه کرده و درصد آنها را گزارش می‌کند.

**3.محاسبه درصد GC**

فرمولGC Content درصد بازهای G و C در کل دنباله است که به صورت زیر محاسبه می‌شود:



این معیار برای تحلیل پایداری و مقاومت حرارتی استفاده می‌شود که دو حالت دارد:

* + - GC بالا (بیش از 50%): پایداری و مقاومت حرارتی بیشتر به دلیل وجود 3 پیوند هیدروژنی.
    - GC پایین (کمتر از 50%): پایداری کمتر.

## 2- توضیح کد برنامه:

در این بخش، کدی که با استفاده از زبان Python نوشته‌ایم، تحلیل دنباله DNA را انجام می‌دهد. هر بخش کد به شرح زیر می‌باشد:

!pip install Bio

from Bio import SeqIO

from collections import Counter

import pandas as pd

* **کتابخانه Bio:** پس از نصب، آن را برای خواندن فایل GenBank که دنباله DNA را شامل می‌شود، استفاده می‌کنیم.
* **کتابخانه Counter:** برای شمارش فراوانی نوکلئوتیدها و کلمات دو و سه حرفی.
* **کتابخانه Pandas:** برای ساخت و نمایش جداول نتایج.

**تابع اصلی تحلیل دنباله DNA:**

def DNA\_Analyzer(GB\_file\_path):

    # load from GenBank

    record = SeqIO.read(GB\_file\_path, "genbank")

    dna\_sequence = str(record.seq).upper()

* : file\_pathمسیر فایل GenBank شامل دنباله DNA
* : SeqIO.readدنباله DNA را از فایل می‌خواند.
* : dna\_sequenceدنباله DNA به صورت رشته بزرگ (uppercase)

**بررسی طول دنباله**

    if len(dna\_sequence) < 2000:

        raise ValueError(".طول دنباله دی‌ان‌ای باید حداقل 2000 نوکلئوتید باشد")

* طبق صورت سوال پروژه بررسی می‌کند که دنباله حداقل 2000 نوکلئوتید باشد و اگر کمتر باشد، خطا می‌دهد زیرا دنباله‌ی مورد بررسی باید طویل باشد.

**شمارش نوکلئوتیدهای ساده**

    # شمارش نوکلئوتیدهای ساده

    simple\_counts = Counter(dna\_sequence)

    simple\_percentages = {nuc: count / len(dna\_sequence) \* 100 for nuc, count in simple\_counts.items()}

* Counter : تعداد هر نوکلئوتید را می‌شمارد.
* سپس درصد هر نوکلئوتید محاسبه می‌شود.

**شمارش دی‌نوکلئوتیدها**

    # شمارش دی‌نوکلئوتیدها

    di\_nucleotides = [dna\_sequence[i:i+2] for i in range(len(dna\_sequence) - 1)]

    di\_counts = Counter(di\_nucleotides)

    di\_percentages = {dinuc: count / len(di\_nucleotides) \* 100 for dinuc, count in di\_counts.items()}

* تمام ترکیبات دو حرفی در دنباله استخراج می‌شوند.
* سپس Counter تعداد هر ترکیب دو حرفی را می‌شمارد.

**شمارش تری‌نوکلئوتیدها**

    # شمارش تری‌نوکلئوتیدها

    tri\_nucleotides = [dna\_sequence[i:i+3] for i in range(len(dna\_sequence) - 2)]

    tri\_counts = Counter(tri\_nucleotides)

    tri\_percentages = {trinuc: count / len(tri\_nucleotides) \* 100 for trinuc, count in tri\_counts.items()}

* تمام ترکیبات سه حرفی استخراج و شمارش می‌شوند.

**ساخت جداول**

    # ساخت جداول

    simple\_table = pd.DataFrame({

        "Nucleotide": list(simple\_counts.keys()),

        "Count": list(simple\_counts.values()),

        "Percentage": list(simple\_percentages.values())

    })

    di\_table = pd.DataFrame({

        "Di-nucleotide": list(di\_counts.keys()),

        "Count": list(di\_counts.values()),

        "Percentage": list(di\_percentages.values())

    })

    tri\_table = pd.DataFrame({

        "Tri-nucleotide": list(tri\_counts.keys()),

        "Count": list(tri\_counts.values()),

        "Percentage": list(tri\_percentages.values())

    })

* جدولی شامل تعداد و درصد نوکلئوتیدها را مطابق فرمت خواسته شده، می‌سازد.

**محاسبه درصد GC**

    # محاسبه GC

    gc\_content = (simple\_counts.get('G', 0) + simple\_counts.get('C', 0)) / len(dna\_sequence) \* 100

    return simple\_table, di\_table, tri\_table, gc\_content

* تعداد بازهای G و C جمع می‌شود.
* درصد GC محاسبه می‌شود.
* تابع درنهایت تمامی خروجی ها را برمی‌گرداند.

**بکارگیری تابع و نمایش خروجی**

* # استفاده از تابع
* GB\_file\_path = "sequence.gb"
* simple\_table, di\_table, tri\_table, gc\_content = DNA\_Analyzer(GB\_file\_path)
* # نمایش نتایج
* print("جدول نوکلئوتیدهای ساده:")
* print(simple\_table)
* print("\nجدول دی‌نوکلئوتیدها:")
* print(di\_table)
* print("\nجدول تری‌نوکلئوتیدها:")
* print(tri\_table)
* print(f"\nدرصد GC: {gc\_content:.2f}%")  #High GC content (>50%) indicates greater stability and heat resistance.

## 3- تحلیل نتایج:

**تحلیل ژنوم Sorex roboratus**

ژنوم میتوکندری Sorex roboratus اطلاعات ارزشمندی برای تحلیل‌های ژنتیکی و فیلوژنتیکی فراهم می‌کند. تحلیل درصد GC و ترکیبات نوکلئوتیدی، ویژگی‌های پایداری و مقاومت حرارتی این ژنوم را برجسته می‌کند. نتایج حاصل از این پروژه می‌تواند در تحقیقات زیست‌شناسی و پزشکی کاربرد داشته باشد.

تذکر: عبارت **bps**  مخفف **base pairs**  (جفت بازها) است و به تعداد بازهای آلی A، T، G، C در یک توالی DNA اشاره دارد.

1. **جدول نوکلئوتیدهای ساده**

خروجی جدول نشان می‌دهد که چه میزان از ژنوم توسط هر نوکلئوتید A، T، G، C تشکیل شده است.

|  |  |  |
| --- | --- | --- |
| **Nucleotide** | **Count** | **Percentage** |
| G | 2245 | 13.09% |
| T | 5047 | 29.42% |
| C | 4182 | 24.38% |
| A | 5679 | 33.11% |

**A و T فراوان‌ترین نوکلئوتیدها هستند:**

نوکلئوتیدهای آدنین (A) و تیمین (T) با مجموع 62.53% بیشترین بخش از توالی را تشکیل داده‌اند. این موضوع نشان‌دهنده درصد پایین GC در این ژنوم است.

**C و G درصد کمتری دارند:**

نوکلئوتیدهای گوانین (G) و سیتوزین (C) جمعاً 37.47% از توالی را تشکیل می‌دهند. این مقدار کمتر نسبت به A و T نشان‌دهنده محتوای GC پایین‌تر و پایداری حرارتی متوسط ژنوم است

1. **جدول دی‌نوکلئوتیدها و تری‌نوکلئوتیدها**

این جداول توزیع ترکیبات دو و سه‌حرفی را نشان می‌دهد که اطلاعات مهمی درباره ساختار ژنومی فراهم می‌کند.

**جدول دی‌نوکلئوتیدها**

|  |  |  |
| --- | --- | --- |
| **Di-nucleotide** | **Count** | **Percentage** |
| GT | 589 | 3.43% |
| TT | 1479 | 8.62% |
| TC | 1133 | 6.61% |
| CA | 1397 | 8.14% |
| AT | 1691 | 9.86% |
| ... | ... | ... |

**فراوانی بالای دی‌نوکلئوتیدهای AT و TA :**

دو نوکلئوتید AT (9.86%) و TA (10.74%) بیشترین فراوانی را دارند. این موضوع با درصد بالای A و T در ژنوم همخوانی دارد و نشان‌دهنده وجود مناطق غنی از A و T در توالی است.

**GC کمتر از AT :**

درصد دی‌نوکلئوتیدهای GC (3.28%) نسبت به AT بسیار پایین‌تر است. این نشان‌دهنده توزیع نامتقارن بازها و محتوای پایین GC است.

**فراوانی دی‌نوکلئوتیدهای تکراری AA و TT :**

وجود درصد بالای AA (10.31%) و TT (8.62%) نشان می‌دهد که ژنوم شامل مناطق تکراری است که ممکن است به تنظیم بیان ژن کمک کند.

**جدول تری‌نوکلئوتیدها**

|  |  |  |
| --- | --- | --- |
| **Tri-nucleotide** | **Count** | **Percentage** |
| GTT | 161 | 0.94% |
| TTC | 369 | 2.15% |
| TCA | 369 | 2.15% |
| CAT | 411 | 2.40% |
| ATG | 214 | 1.25% |
| ... | ... | ... |

**فراوانی متعادل تری‌نوکلئوتیدها:**

ترکیبات مختلف به‌طور متعادل توزیع شده‌اند. این تنوع می‌تواند نشان‌دهنده عملکردهای مختلف ژنوم و توالی‌های رمزگذاری‌کننده برای پروتئین‌های متنوع باشد.

**ترکیبات آغازگر کدون (ATG) :**

ATG که کدون آغازگر بسیاری از پروتئین‌هاست، با درصد 1.25% به فراوانی در ژنوم وجود دارد و این منطقی است، زیرا این ترکیب نشان‌دهنده مناطق کدکننده ژن است.

**توالی‌های غنی از AT :**

بسیاری از تری‌نوکلئوتیدهای غالب مانند GTT و CAT شامل A یا T هستند که نشان‌دهنده غلبه این بازها در ساختار ژنوم است.

1. **درصد GC**

درصد GC برای ژنوم میتوکندری Sorex roboratus تخمین زده می‌شود. این مقدار معمولاً کمتر از 50% است و نشان‌دهنده پایداری متوسط در شرایط دمایی مختلف است. درصد GC پایین (کمتر از 50%) نشان می‌دهد که توالی ژنوم عمدتاً غنی از بازهای A و T است. بازهای GC به دلیل وجود سه پیوند هیدروژنی، پایداری بیشتری نسبت به بازهای AT (که دو پیوند هیدروژنی دارند) به DNA می‌دهند. بنابراین، ژنومی با محتوای GC پایین‌تر پایداری حرارتی کمتری دارد.

**چرا به این مقدار رسیدیم؟**

* ژنوم میتوکندری معمولاً درصد GC کمتری نسبت به ژنوم هسته‌ای دارد. دلیل این امر وابستگی به عملکرد خاص میتوکندری و همچنین محدودیت‌های تکاملی مرتبط با آن است.

**4- پایداری و مقاومت در برابر گرما**

**1 . پایداری ژن:**

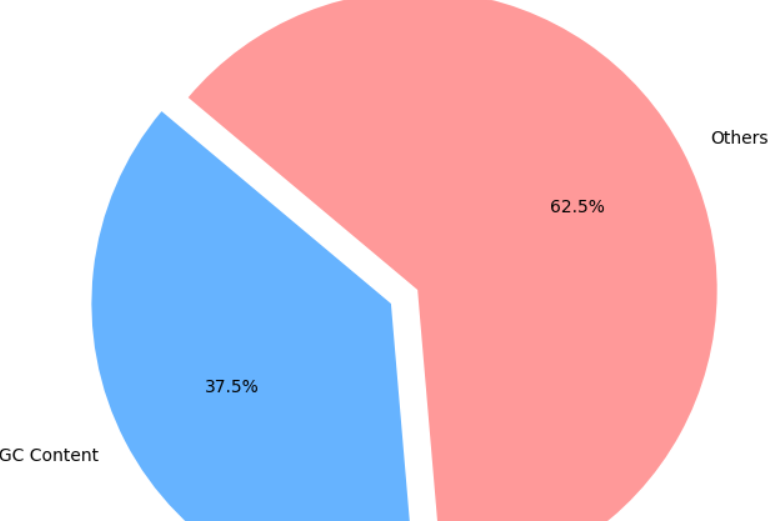
* محتوای GC پایین نشان‌دهنده پایداری متوسط است. چنین ژنومی در برابر تغییرات محیطی و شیمیایی پایداری کمتری نسبت به توالی‌هایی با محتوای GC بالا دارد.

**2 . مقاومت در برابر گرما:**

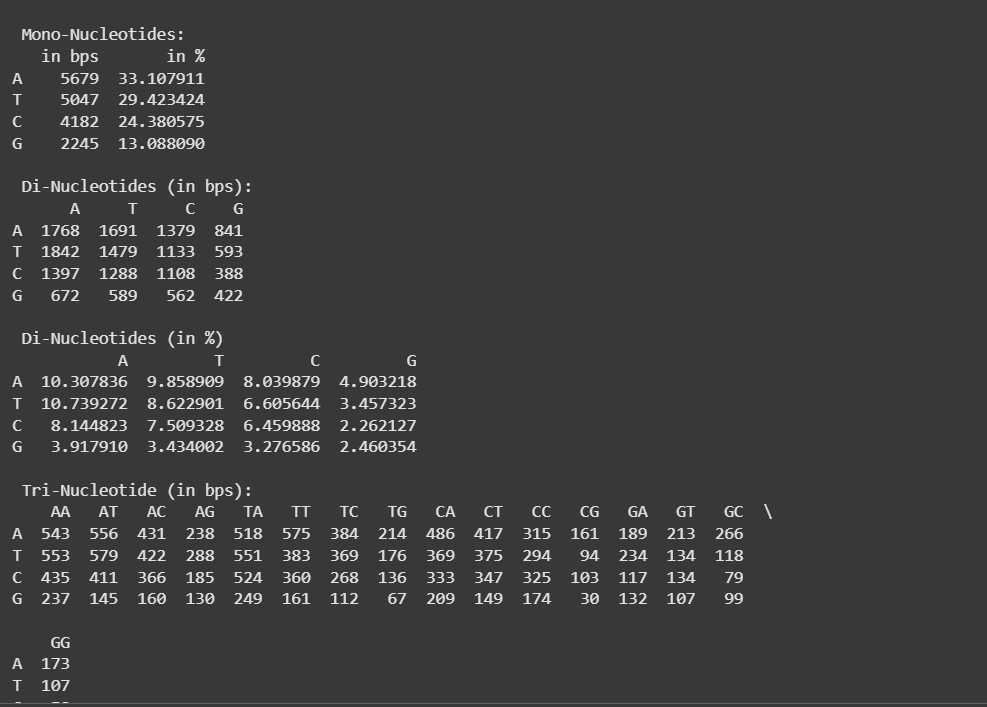
* DNA با محتوای GC بالا، مقاومت حرارتی بیشتری دارد، زیرا پیوندهای هیدروژنی بیشتری در بازهای G و C وجود دارد.
* برای ژنوم *Sorex roboratus* با GC = 37.47%، مقاومت در برابر گرما در سطح متوسط قرار دارد. احتمالاً در دماهای بالای معمول، DNA این ژنوم ناپایدار می‌شود.

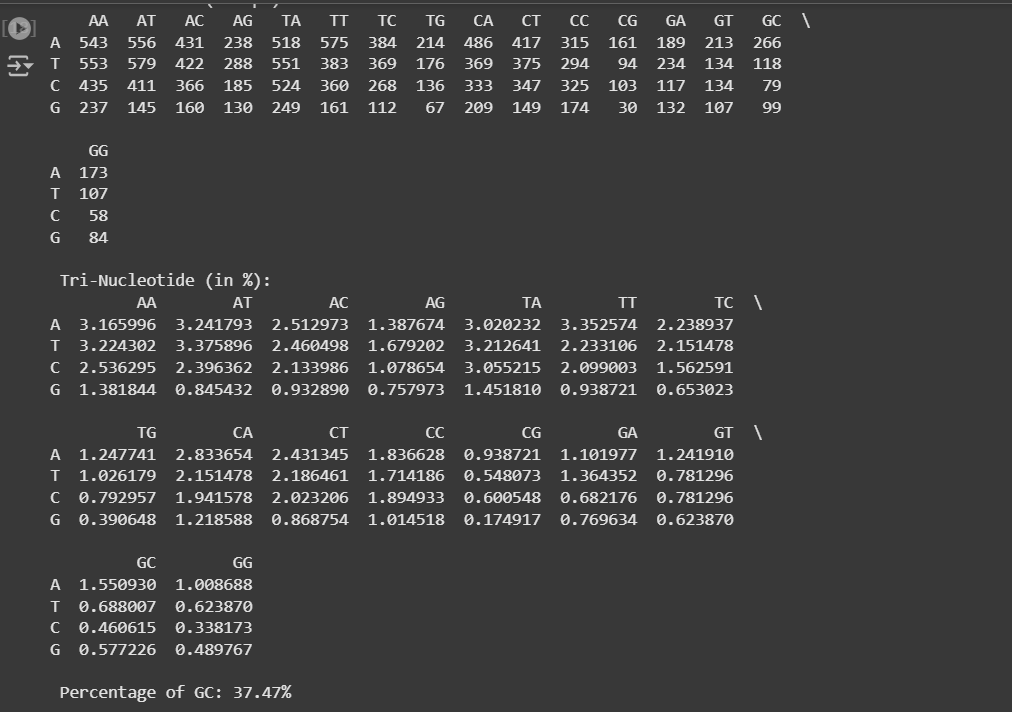
**5- نتیجه‌گیری کلی**

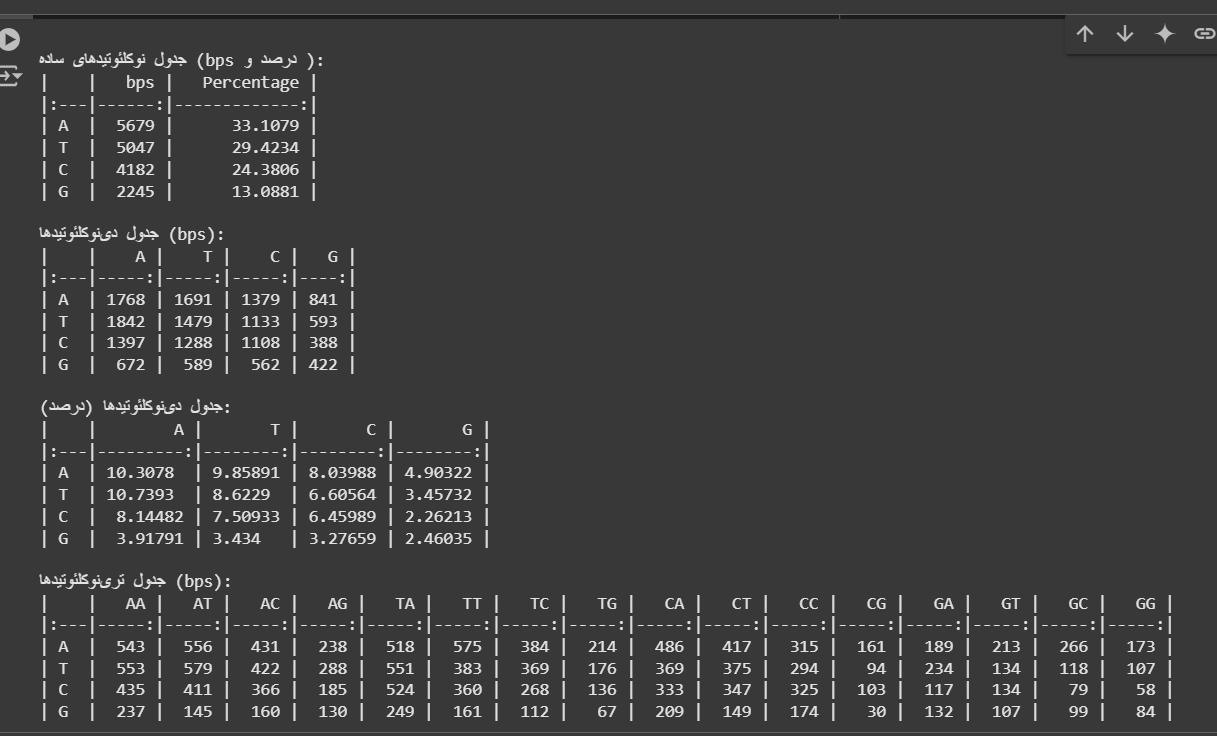
1. ژنوم میتوکندری *Sorex roboratus* غنی از A و T است که نشان‌دهنده مناطق خاص ژنتیکی و تکاملی این موجود است.
2. درصد GC پایین (37.47%) نشان می‌دهد که این ژنوم مقاومت حرارتی بالایی ندارد و احتمالاً در دماهای بالا ناپایدار می‌شود.
3. این اطلاعات می‌تواند در مطالعات تکاملی و همچنین بررسی ویژگی‌های زیستی این موجود مورد استفاده قرار گیرد.

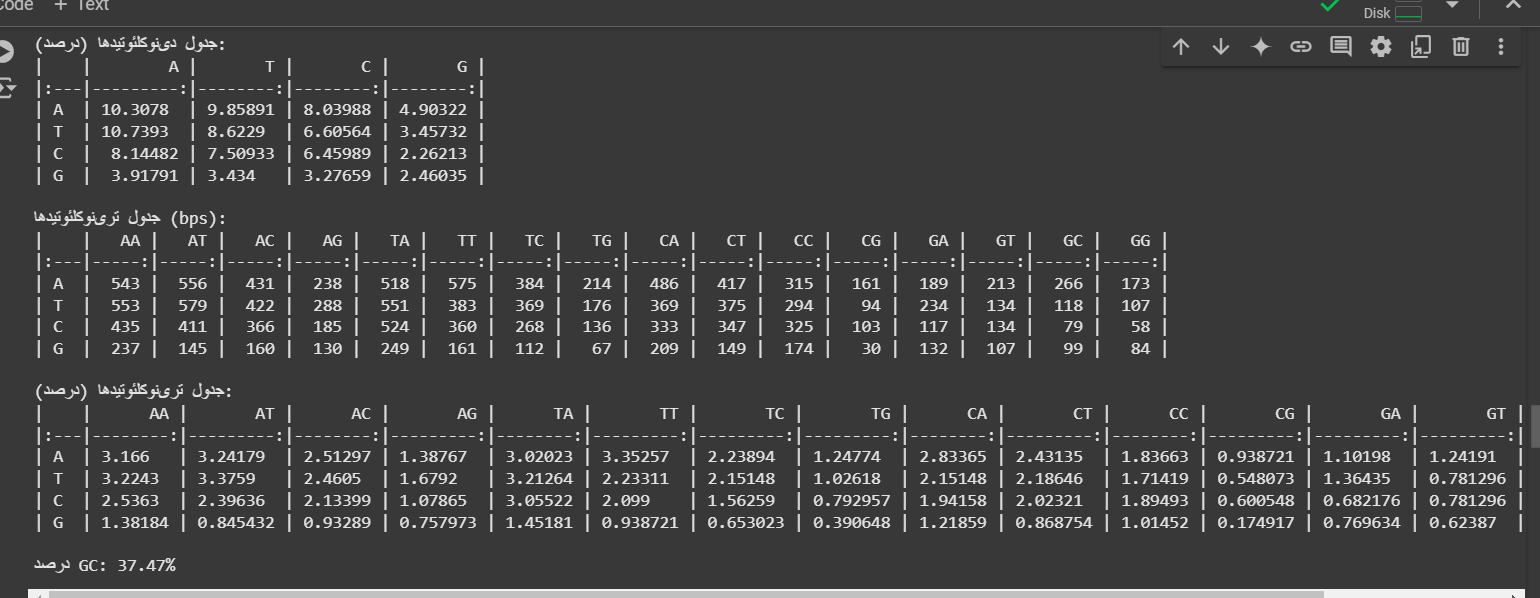


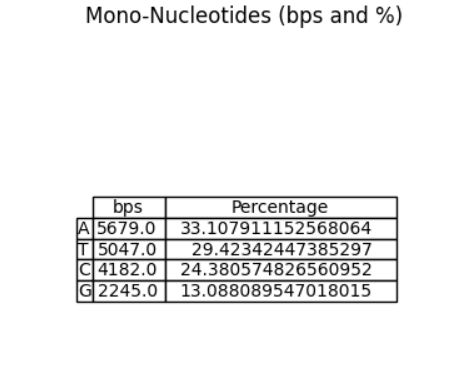
## 4- خروجی‌ها و نمودارها:

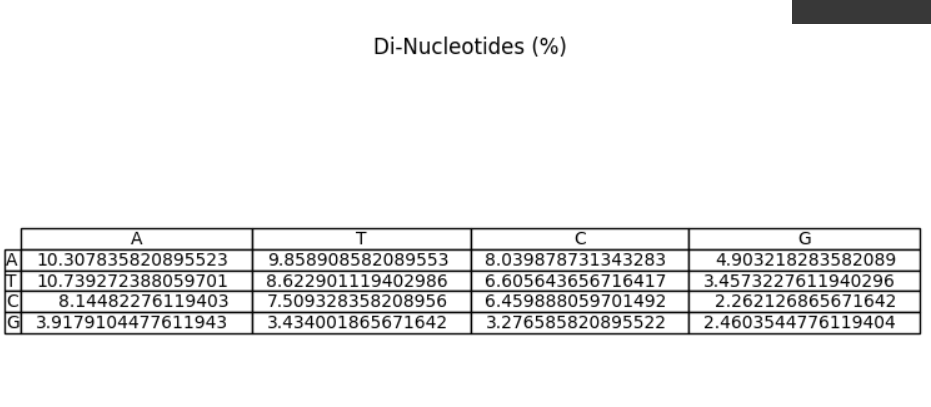
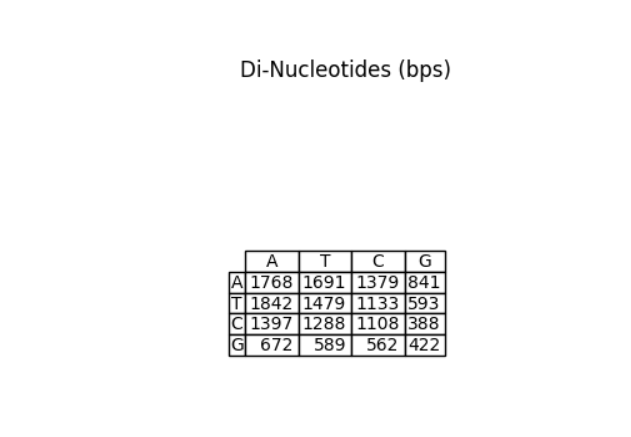
****

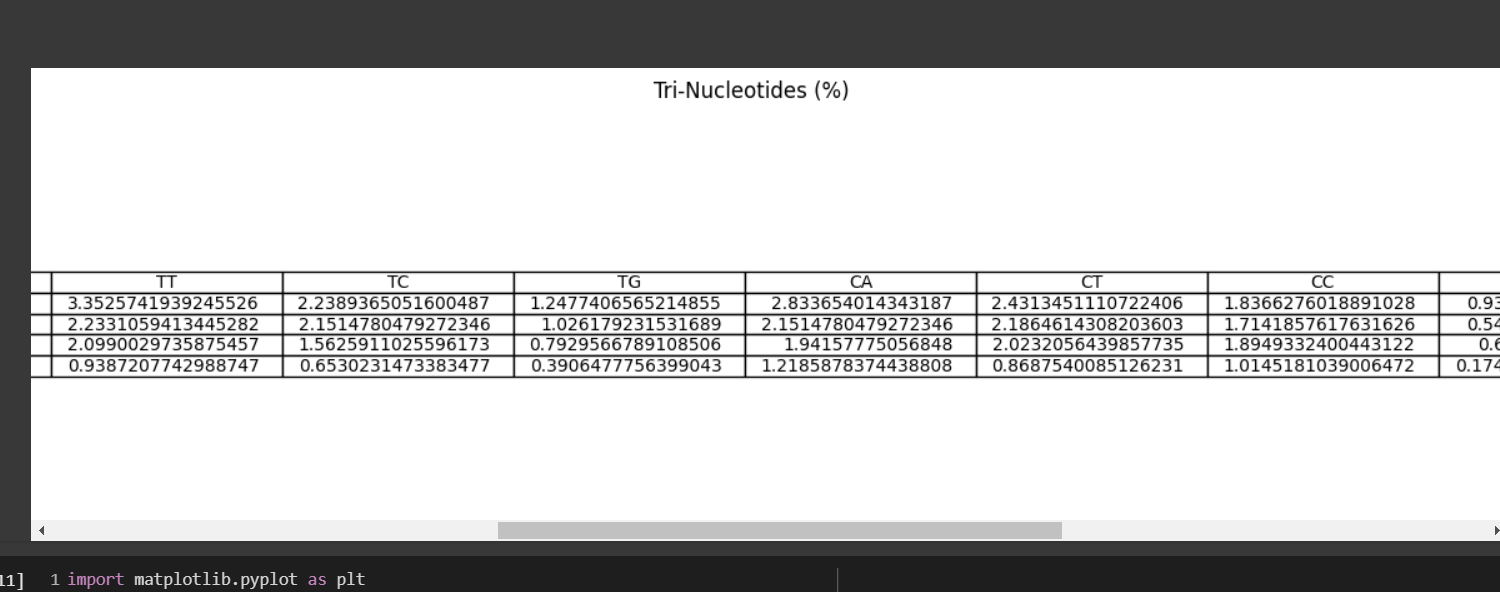
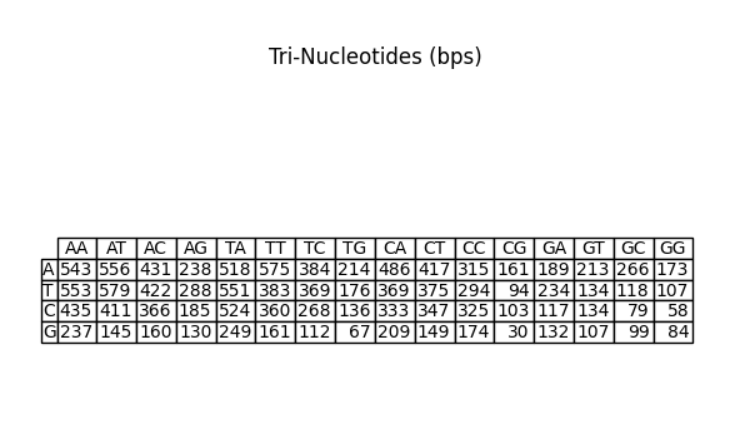


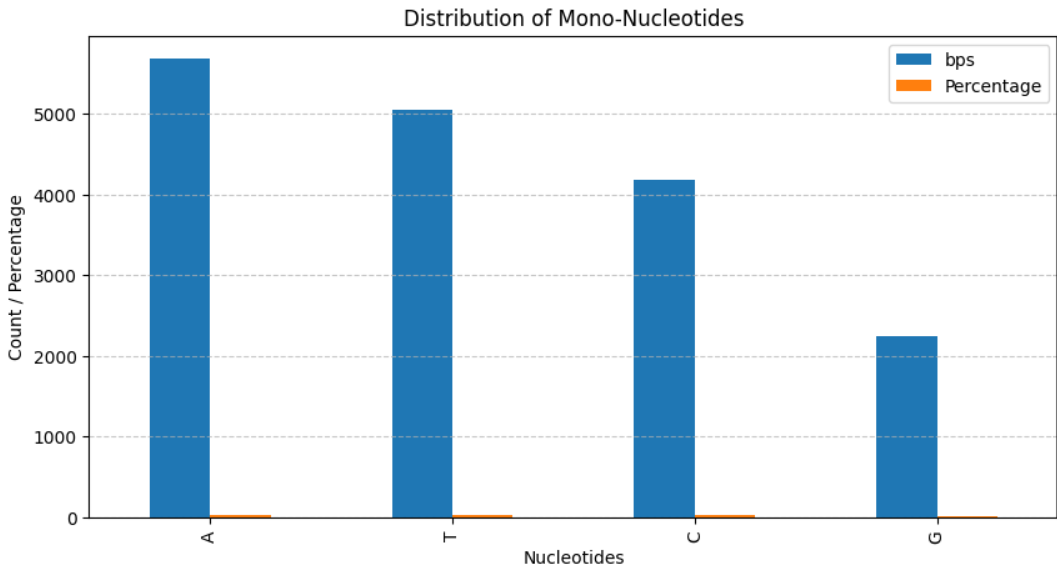


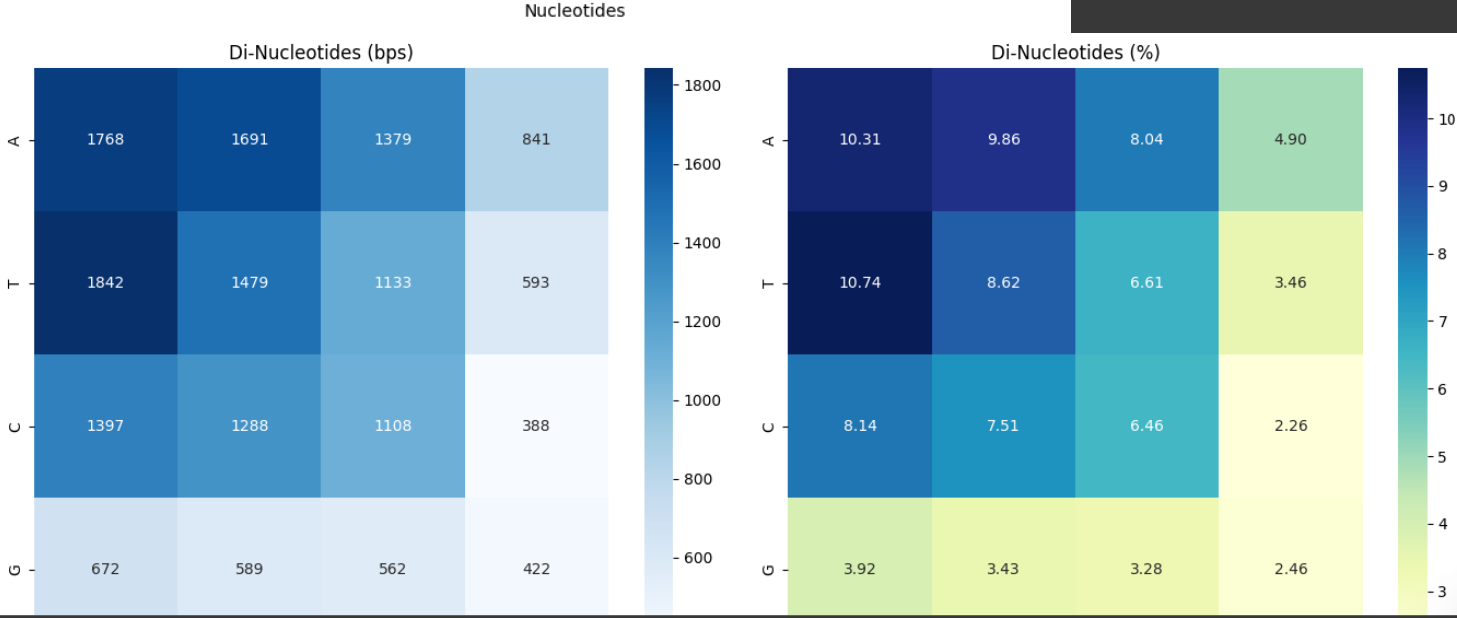


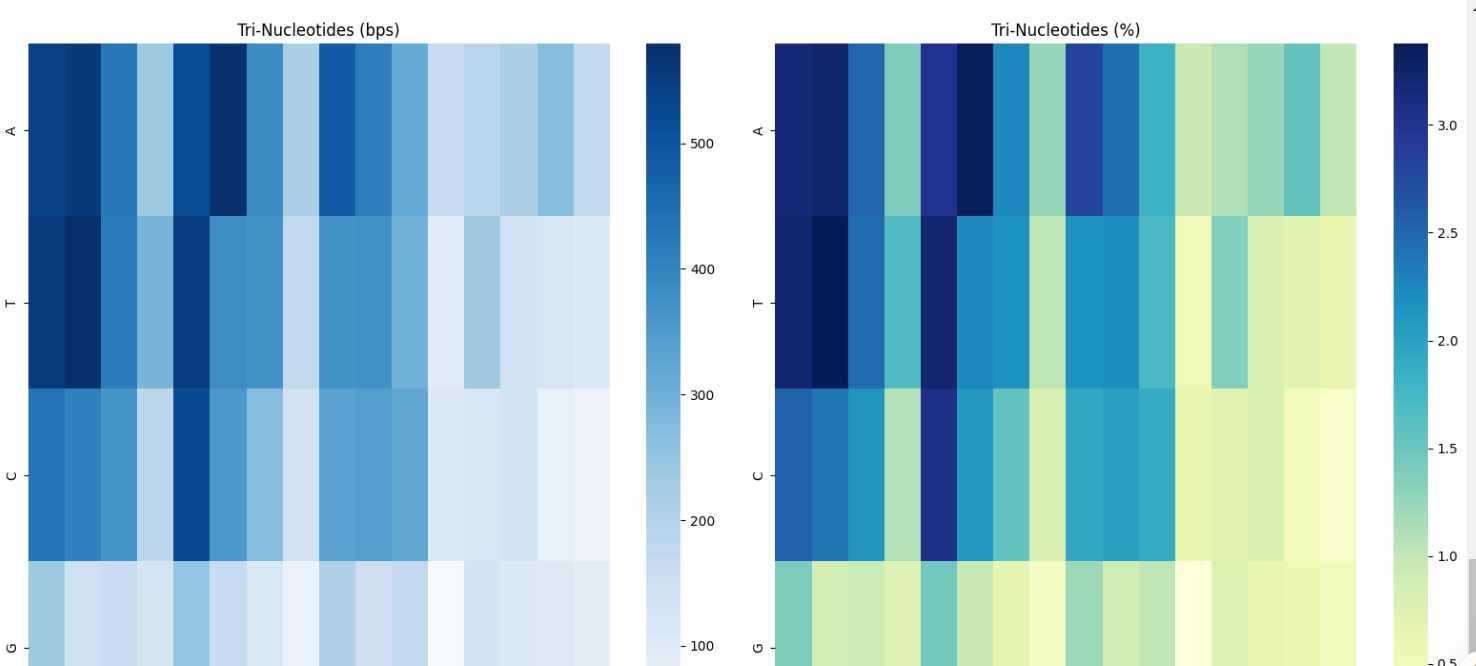












## 5- مراجع

[https://www.ncbi.nlm.nih.gov/](%20https://www.ncbi.nlm.nih.gov/%20/)

[https://www.ncbi.nlm.nih.gov/nuccore/NC\_034808.1](https://www.ncbi.nlm.nih.gov/tools/primer-blast/primertool.cgi?ctg_time=1736087630&job_key=Mjjs1BXgGEg_dh1zEBM5QWoIKHNHGzNuRg&CheckStatus=Check)

کد را با دنباله‌ی بالا به عنوان انتخاب اصلی اجرا کردیم و به 37% رسیدیم که یعنی ناپایدارتر است.

<https://www.ncbi.nlm.nih.gov/nuccore/AL035632.2>

کد را با دنباله‌ی بالا به عنوان تست دیگر مثال‌ها نیز اجرا کردیم و به 53% رسیدیم که یعنی این دنباله پایدارتر است.

<https://colab.research.google.com/>

<https://www.researchgate.net/publication/318192460_Sequencing_and_analysis_of_the_complete_mitochondrial_genome_of_flat-skulled_shrew_Sorex_roboratus_from_China>

<https://github.com/M-Amin-Kiani/BioInfo/tree/main/research>

لینک گیت‌هاب بالا حاوی کدها و ژن بانک مربوطه برای اجرای کد پروژه به همراه تمامی گزارش‌های این درس است.

