



Универзитет „Св. Кирил и Методиј“ во Скопје
**ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И
КОМПЈУТЕРСКО ИНЖЕНЕРСТВО**

Проект по предметот Инженерска Математика

Тема: Фуријеви трансформации во геномиката

Изработено од: Мила Ѓуровска 231116

Скопје, Јануари 2025

Вовед

Геномиката е научна дисциплина која го проучува геномот, односно целокупниот сет на ДНК кај организмите, вклучувајќи ги сите гени, за разлика од генетиката која ги проучува само оние кои произведуваат протеини. Оваа дисциплина е од огромно значење за разбирање на генетските основи на животот, откривање на болести, развој на нови терапии и подобрување на култури и екосистеми. Во последниве децении, со напредокот на технологијата, геномските податоци станаа достапни во огромни количини, што создаде потреба од ефикасни методи за нивна анализа и обработка. Еден од таквите методи се Фуриевите трансформации, кои се широко користени во обработката на сигнали и податоци, а во последно време се користени и во биологијата.

Фуриевите трансформации

Фуриевите трансформации се математички техники кои се користат за претворање на сигнали или функции од временски домен во фреквенциски домен. Во геномиката, Фуриевите трансформации се применуваат на секвенците на ДНК, кои се составени од низи на нуклеотиди (А, Т, С, Г). Постојат различни видови на Фуриевите трансформации, кои се разликуваат според видот на сигналот (континуиран или дискретен) и начинот на нивна пресметка. Тука ќе ги разгледаме обичните (континуирани) Фуриевите трансформации, брзите Фуриевите трансформации (FFT) и дискретните Фуриевите трансформации (DFT).

Континуираната Фуриева трансформација (CTFT) се применува на континуирани сигнали, односно сигнали кои се дефинирани за секој момент во времето. Оваа трансформација претвора временски сигнал во фреквенциски домен. Ваквите трансформации се карактеризираат со формулата

$$F\{f(t)\} = \int_{-\infty}^{\infty} e^{-i\omega t} f(t) dt,$$

каде $f(t)$ е оригиналниот сигнал во временски домен, $F\{f(t)\}$ е трансформираниот сигнал во фреквенциски домен, ω е фреквенцијата, а i е имагинарната единица.

Дискретната Фуриева трансформација (DFT) е верзија на Фуриевата трансформација која се применува на дискретни сигнали, односно сигнали кои се дефинирани само на одредени

временски интервали. DFT претвора дискретен временски сигнал во дискретен фреквенциски домен.

Брзата Фуриева трансформација (FFT) е ефикасен алгоритам за пресметување на DFT. FFT го намалува бројот на пресметковни операции од $O(n^2)$ (кај DFT) до $O(n \log n)$, што ја прави многу побрза и практична за големи множества на податоци. Дискретната и брзата Фуриева трансформација се претставени со формулата

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{-i(2\pi/N)kn},$$

каде $F(k)$ е амплитуда на компонентата на фреквенција, $f(n)$ е низата на нуклеотидни вредности (репрезентација на секвенцата), N е вкупен број на точки во низата (должина на секвенцата), $e^{-i(2\pi/N)kn}$ е комплексна експоненцијална функција што ги претставува синусоидните компоненти.

Примена на Фуриеви трансформации во геномиката

Компресија на податоци

Геномските податоци се огромни и често бараат ефикасна компресија за заштеда на простор и подобрување на брзината на обработка. Фуриевата трансформација може да се користи за претворање на податоците во фреквенциски домен, што овозможува подобра компресија. Компресираните податоци може да се репрезентираат со минимален губиток на биолошка значајност, што е од голема важност за зачувување на информацијата.

Идентификација на повторувачки секвенци

Повторувачките секвенци во ДНК, или мотиви, се кратки низи на нуклеотиди кои се повторуваат повеќепати во геномот. Тие имаат значајна улога во:

- *Регулација на генската експресија*: Контролираат кога, каде и колку активно ќе се изразат гените.
- *Формирање на хромозомската структура*: Помагаат во организирањето на ДНК во компактна форма, што е критично за процеси како репликација и делба на клетките.

ДНК или дезоксирибонуклеинска киселина е составена од четири азотни бази:

- Аденин (A): Секогаш се спарува со Тимин (T) преку две водородни врски.
- Тимин (T): Постои само во ДНК и секогаш се спарува со Аденин (A).
- Цитозин (C): Секогаш се спарува со Гуанин (G) преку три водородни врски.
- Гуанин (G): Формира пар со Цитозин (C).

Фуриевата трансформација ја конвертира секвенцата на азотните бази (A, T, C, G) во фреквенциски домен, каде што повторувачките секвенци се појавуваат како фреквенциски пикови. Ова овозможува лесна идентификација на овие секвенци, што е важно за разбирање на нивната функција во геномот.

Филтрирање на шум

Податоците добиени од ДНК секвенционирање често содржат шум поради грешки при мерење или контаминација. Со премин во фреквенциски домен, шумот (кој често се наоѓа на високи фреквенции) може да се отстрани, додека важните сигнали (кои се наоѓаат на пониски фреквенции) остануваат недопрени. Ова е особено корисно за подобрување на точноста на анализата на геномските податоци.

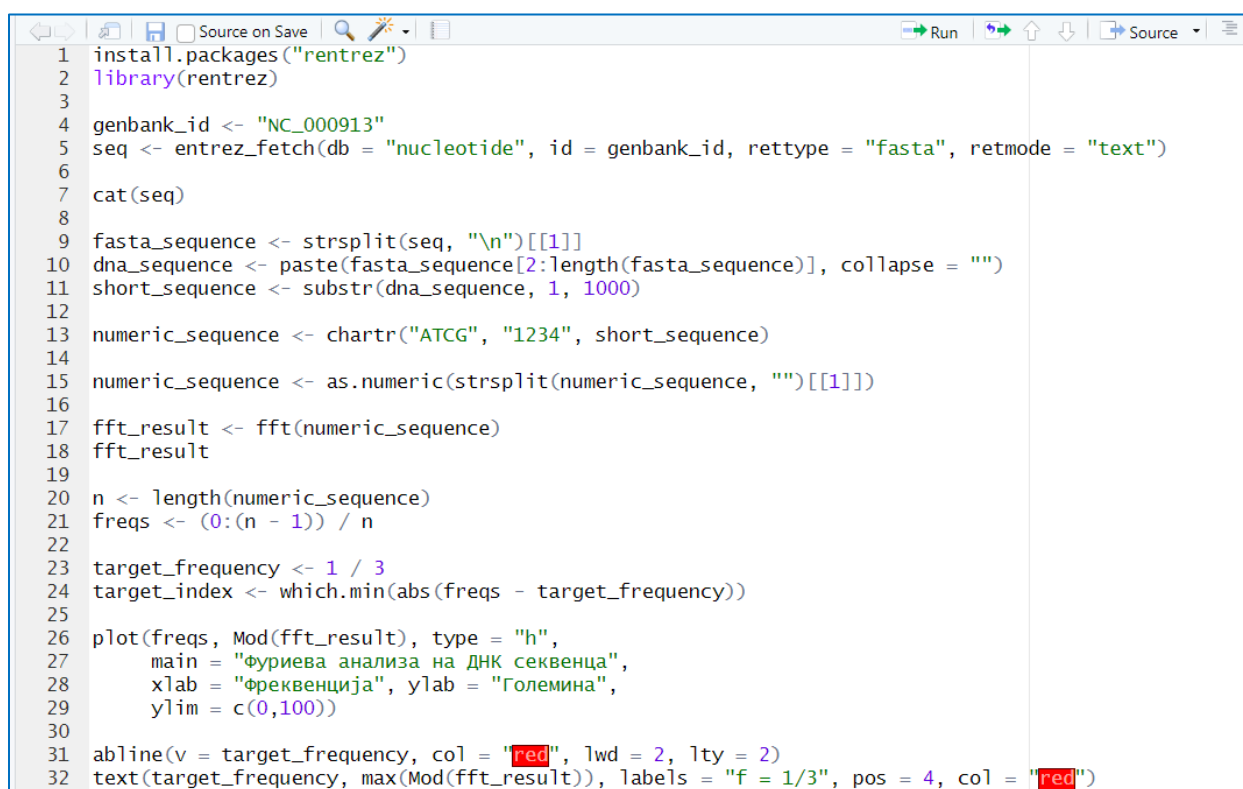
Анализа на кодирачки региони

Кодирачките региони во ДНК често покажуваат три-базна периодичност, што е резултат на начинот на кој генетскиот код ги претвора нуклеотидите во аминокиселини. Секоја група од три бази (кодон) кодира една аминокиселина. Кога Фуриевата трансформација се применува на низа од бази (претставени како нумерички вредности), се појавува пик на фреквенција $f = 1/3$. Овој пик е карактеристичен за кодирачките региони, бидејќи ги рефлектира правилните интервали (периодичност) во секвенцата.

Некодирачките региони се делови од ДНК кои немаат информации за синтеза на протеини. Иако некогаш се нарекувале „junk DNA“, сега е познато дека овие региони имаат регулаторни или структурни функции. Некодирачките региони не покажуваат три-базна периодичност, што овозможува лесно да се разликуваат од кодирачките региони.

Пример за употреба на Фуриеве трансформации во геномиката во R

Програмскиот јазик R е широко користен за анализа на геномски податоци. Во R, Фуриевите трансформации може да се применат со користење на функции како што е `fft()` (брза Фуриева трансформација). Оваа функција може да се користи за претворање на геномските секвенци во фреквенциски домен, што овозможува идентификација на повторувачки секвенци, филтрирање на шум и анализа на кодирачки региони. Следната програма е наменета за да покаже дали ДНК секвенцата на *Escherichia coli* има кодирачки региони (Слика 1).



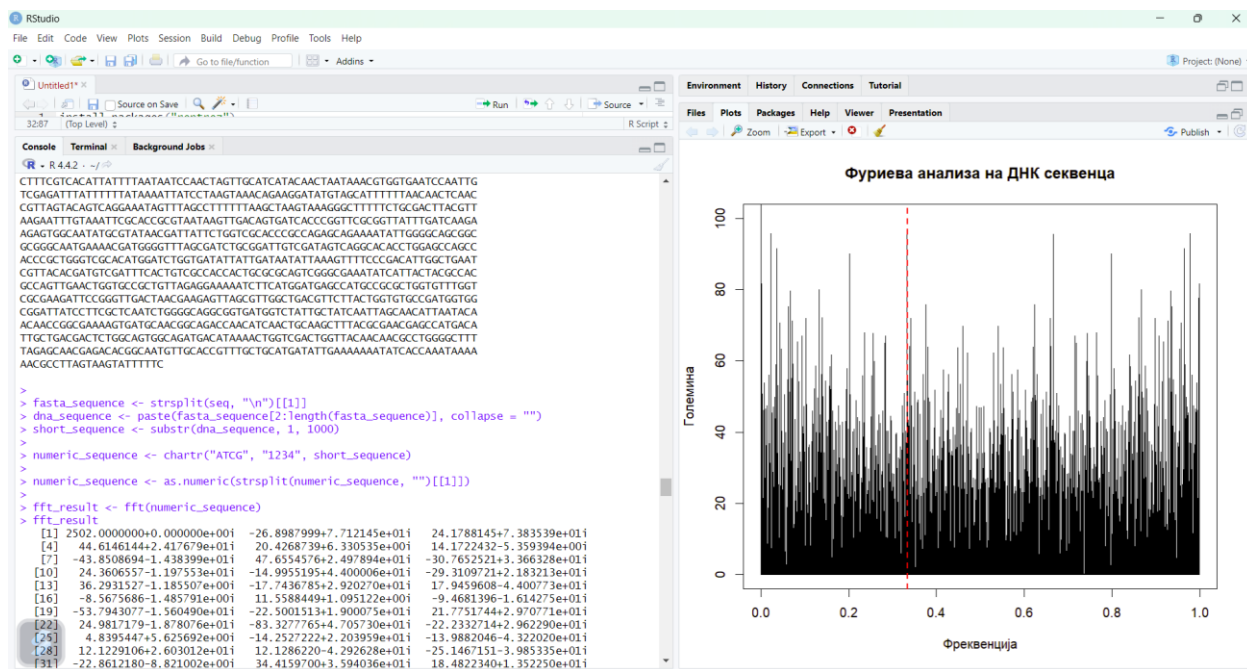
```
1 install.packages("rentrez")
2 library(rentrez)
3
4 genbank_id <- "NC_000913"
5 seq <- entrez_fetch(db = "nucleotide", id = genbank_id, rettype = "fasta", retmode = "text")
6
7 cat(seq)
8
9 fasta_sequence <- strsplit(seq, "\n")[[1]]
10 dna_sequence <- paste(fasta_sequence[2:length(fasta_sequence)], collapse = "")
11 short_sequence <- substr(dna_sequence, 1, 1000)
12
13 numeric_sequence <- chartr("ATCG", "1234", short_sequence)
14
15 numeric_sequence <- as.numeric(strsplit(numeric_sequence, "")[[1]])
16
17 fft_result <- fft(numeric_sequence)
18 fft_result
19
20 n <- length(numeric_sequence)
21 freqs <- (0:(n - 1)) / n
22
23 target_frequency <- 1 / 3
24 target_index <- which.min(abs(freqs - target_frequency))
25
26 plot(freqs, Mod(fft_result), type = "h",
27      main = "Фуриева анализа на ДНК секвенца",
28      xlab = "Фреквенција", ylab = "Големина",
29      ylim = c(0,100))
30
31 abline(v = target_frequency, col = "red", lwd = 2, lty = 2)
32 text(target_frequency, max(Mod(fft_result)), labels = "f = 1/3", pos = 4, col = "red")
```

Слика 1

Чекори за анализа на кодирачки региони во ДНК секвенца на *Escherichia coli*:

1. Инсталација и вчитување на библиотеката *rentrez*, која овозможува пристап до базите на податоци на NCBI (National Center for Biotechnology Information).
2. Се презема ДНК секвенцата од *GenBank* преку функцијата `entrez_fetch()` каде се специфира идентификаторот на организмот кој сакаме да го проучуваме, и враќа низа од податоци.

3. За полесно прегледување на резултатите и во интерес на времето, се крати оваа ДНК секвенца на 1000 примероци.
4. Оваа низа од "ATCG" се претвора во низа од броеви и потоа се применуваат брзи Фуриеви трансформации врз неа.
5. Се визуелизираат резултатите од честотата на фреквенциите преку график, каде е истакната црвена линија на фреквенцијата $1/3$.



Слика

2

Резултатите од извршување на оваа програма се дадени на Слика 2, каде од левата страна е покажан дел од ДНК секвенцата на *Escherichia coli* и ДНК секвенцата по извршување на Фуриеви трансформации, каде реалниот дел претставува амплитудата на фреквенцијата, а имагинарниот е фазата, додека од десната страна е прикажан графикот. Со оглед на тоа дека на фреквенцијата $1/3$ има пик, може да се заклучи дека оваа ДНК секвенца најверојатно има кодирачки региони, но за да се знае со сигурност треба да се направат повеќе испитувања.

Заклучок

Фуриевите трансформации се моќна алатка во геномиката, која овозможува ефикасна обработка и анализа на огромните количини на геномски податоци. Тие се користат за компресија на податоци, идентификација на повторувачки секвенци, филтрирање на шум и анализа на кодирачки региони. Со продолжениот развој на технологијата и методите за

анализа, Фуриевите трансформации ќе продолжат да играат клучна улога во идните истражувања во геномиката.