Міністерство освіти і науки України
Департамент науки і освіти Харківської облдержадміністрації
Харківське територіальне відділення МАН України

Відділення: Фізики і астрономії
Секція: Теоретична фізика

Кореляційні характеристики РНК коронавірусів

Роботу виконав:
Пасько Павло Григорович,
учень 11 класу Харківського
Навчально-виховного комплексу
№45 «Академічна гімназія»
Харківської міської ради
Харківської області

Наукові керівники:
Сергій Сергійович Мельник,
старший научний співробітник
відділу теоретичної фізики
інституту радіофізики та електро-
ніки ім. О.Я. Усикова,
кандидат фізико-математичних
наук

Іврій Ілля Леонідович,
учитель фізики Харківського
навчально-виховничого
комплексу №45
«Академічна гімназія» Харківської
міської ради Харківської області,
спеціаліст вищої категорії,
«Відмінник освіти України»

Харків – 2020

Автор роботи: Пасько Павло Григорович

Метою дослідження є визначення статистично-інформаційних сигнатур геномних послідовностей РНК коронавірусів, зокрема вірусу SARS-CoV-2, з використанням технік математичного аналізу символьних послідовностей. Зазвичай базовими статистичними об’єктами є стандартні сигнатури: характеристики вірусів, такі як взаємна інформація, ентропія та кореляційна функція. В роботі пропонується використання моделі адитивного марківського ланцюга вищого порядку для апроксимації послідовності генома, в рамках якої враховується далека взаємодія між нуклеотидами геному. Це дозволяє отримати в явному вигляді вирази для попарних ймовірностей, кореляторів, їх перетворень Фур’є та інших характеристик, що спрощує вирішення оберненої задачі – визначення функції пам’яті в моделі Маркова для даного геному. Отриману функцію пам’яті можна потім використовувати в якості нової сигнатури генома.

ЗМІСТ

ВСТУП

Останнім часом проводиться багато досліджень статистичних властивостей послідовностей ДНК та РНК. Вивчення нуклеотид-нуклеотидних кореляцій може виявляти особливості геномної первинної структури. Кореляції є важливою статистичної мірою, оскільки вони здатні розкривати приховані зв'язки між різними елементами складної системи. Отже, вони можуть допомогти зрозуміти структуру і функції досліджуваної системи.

Геноми можна розглядати як лінійні послідовності з чотирьох нуклеотидів: аденіну (A), гуаніну (G), цитозину (C) і тиміну (T) чи урацилу (U). Це дозволяє трактувати послідовності геномів як символьні послідовності і з використанням математичних методів отримувати різні чисельні дані для символьної послідовності. Ці дані називаються сигнатурами генома.

Геномні сигнатури є компактними математичними відображеннями ДНК/РНК послідовностей. Вони характеризують послідовності таким чином, який дає змогу розкрити особливості, специфічно притаманні організму, з якого була отримана ДНК/РНК, і які можуть бути корисними для подальшого аналізу.

Найпростіший спосіб використання сигнатур – це визначення простої, обчислювально ефективної міри відмінності, яка відображає різницю та взаємозв'язки між геномними послідовностями. Такі міри відмінності використовуються для класифікації хромосом за походженням, для поділу і кластеризації підтипів вірусів, а також для класифікації фрагментів ДНК та РНК, вони можуть розглядатися як система швидкої ідентифікації, зокрема, в небезпечних для здоров'я ситуаціях.

Одна з сигнатур, що використовуються для аналізу послідовностей, близькими до проводимих досліджень, є матриця парних кореляційних функцій. За допомогою такої числової сигнатури можна ввести відстані між послідовностями, а потім вивчати філогенетичні зв'язки, будувати філогенетичні дерева з використанням алгоритмів кластеризації, вирішувати питання класифікації.

Однією з відомих математичних моделей, що застосовується для роботи з послідовностями ДНК/РНК в рамках стохастичного підходу (коли послідовність розглядається як стаціонарний випадковий дискретний процес), є модель Маркова. У цій моделі ймовірність появи даного нуклеотиду не залежить від інших нуклеотидів, розташованих поза контекстом цього нуклеотиду, тобто процес має обмежену пам’ять. Цей контекст може мати різну довжину, що призводить до марківських моделей різного порядку.

РОЗДІЛ 1.
РІЗНОВИДНІСТЬ ЛАНЦЮГІВ МАРКОВА

Для того, щоб визначити випадковий ланцюг (або будь-який інший випадковий об'єкт) ми повинні задати ймовірність реалізації різних станів системи. Для ланцюга, це ймовірності набування різних даних символів. Крім того, треба задати ймовірності для всіх пар символів на різних відстанях, для всіх трійок і так далі. Треба зауважити, що ймовірності для пар символів, що знаходяться на однаковій відстані, але розташовані в іншому місті в ланцюзі не обов'язково мають бути рівними (в такому випадку ланцюг не є однорідним). Такий прямолінійний метод визначення випадкового ланцюга є громіздким, тому що в загальному випадку потрібно прописати велику кількість ймовірностей для різних багатосимвольніх слів. Інший підхід до випадкових ланцюгів заснований на використанні умовних ймовірностей. Визначимо ланцюг Маркова за допомогою такого підходу.

Ланцюг Маркова порядку N – це послідовність випадкових змінних $a_i, i \in \mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$, яка має наступну властивість: ймовірніть символу a_i мати певне значення, за умовою, що усі попередні значення символів фіксовані, залежить лише від значеннь N попередніх символів.

$$P(a_i = a|..., a_{i-2}, a_{i-1}) = P(a_i = a|a_{i-N}, ..., a_{i-2}, a_{i-1}) \quad (1.1)$$

Також N називають довжиною пам'яті. Далі ми будемо використовувати запис $T_{N,i}$ для відрізку з N символів перед a_i тобто для набору символів $a_{i-N}, ..., a_{i-1}$. Ланцюг Маркова є однорідною послідовністю, тому що умовна ймовірність (1.1) не залежить від i. Вона залежить лише від N попередніх символів. Якщо умовна ймовірність не залежить від порядку символів в функції умовної ймовірності, то такий ланцюг Маркова називають переставним.

Важливою властивістю корельованної послідовності є ергодичність. Починаючи з якоїсь комбінації символів та будуючи ланцюг відповідно до умовних ймовірностей, маємо ненульову ймовірність отримати будь-яку іншу комбінацію після будови деяких символів. Це, зокрема, значить, що ланцюг є однорідним при умові, що функція умовної ймовірності не залежить від і явно. Далі

вважатимемо, що ці умови буде виконано.

Важливим класом випадкових ланцюгів є бінарні послідовності. Якщо кожен символ a_i може приймати лише два значення, s_0 та s_1, тоді таку послідовність називають бінарною. Зручно змінити можливі значення a_i на 0 та 1, використовуючи лінійне перетворення:

$$a_i := \frac{a_i - s_0}{s_1 - s_0}$$

Окремий та не менш важливий клас – аддитивний бінарний ланцюг Маркова. Умовна ймовірність цих ланцюгів описується наступною формулою:

$$P(a_i = 1 | a_{i-N}, ..., a_{i-2}, a_{i-1}) = \overline{a} + \sum_{r=1}^{N} F(r)(a_{i-r} - \overline{a}) \quad (1.2)$$

Де $F(r)$ – функція пам'яті, $\overline{a}$ – середня кількість одиниць в послідовності. Аддитивність такого ланцюга передбачає, що попередні символи впливають на ймовірність генерації поточного незалежним чином.

Функція $P(a_i = 1 | T_{N,i})$ містить повну інформацію про кореляційні властивості ланцюга Маркова. Загалом, функція кореляції та інші моменти вживаються як базові характеристики для опису корельованих випадкових систем. Тим не менше, корелятор враховує не тільки взаємозв'язок елементів a_i та a_{i+r}, але й непряму взаємодію через інші елементи.

РОЗДІЛ 2.
СТАТИСТИЧНІ ХАРАКТЕРИСТИКИ ВИПАДКОВИХ ЛАНЦЮГІВ

Разом із умовною ймовірністю, яка вказує прямий зв'язок між символами, існує корелятор високого порядку. Корелятор s-го порядку задається наступною формулою:

$$K_s(r_1, r_2, ..., r_{s-1}) = \overline{(a_0 - \overline{a})(a_{r_1} - \overline{a})(a_{r_1+r_2} - \overline{a})...(a_{r_1+...+r_{s-1}} - \overline{a})} \quad (2.1)$$

тут запис $\overline{(...)}$ – статистичне середнє значення по всіх реалізаціях випадкової послідовності. Формально, функція К має залежити від s аргументів(індексів символів), але Марківський ланцюг, який розглядається, є стаціонарним. Тоді функція К залежить від s-1 аргументів (різниця індексів сусідніх символів $r_1, r_2, ..., r_{s-1}$). Корелятори вказують на неявні взаємодії символів. Якщо ми вкажемо значення всіх кореляторів всіх порядків, ми визначимо повністю випадкову послідовність. Тому, задача пошуку кореляторів високого порядку дуже важлива. Бінарна кореляція, яка відповідає за статистичний зв'язок пар елементів ланцюга є особливо важливою. Вона задається наступним чином:

$$K(r) = K_2(r) = \overline{(a_0 - \overline{a})(a_r - \overline{a})} = \overline{a_0 a_r} - \overline{a}^2 \quad (2.2)$$

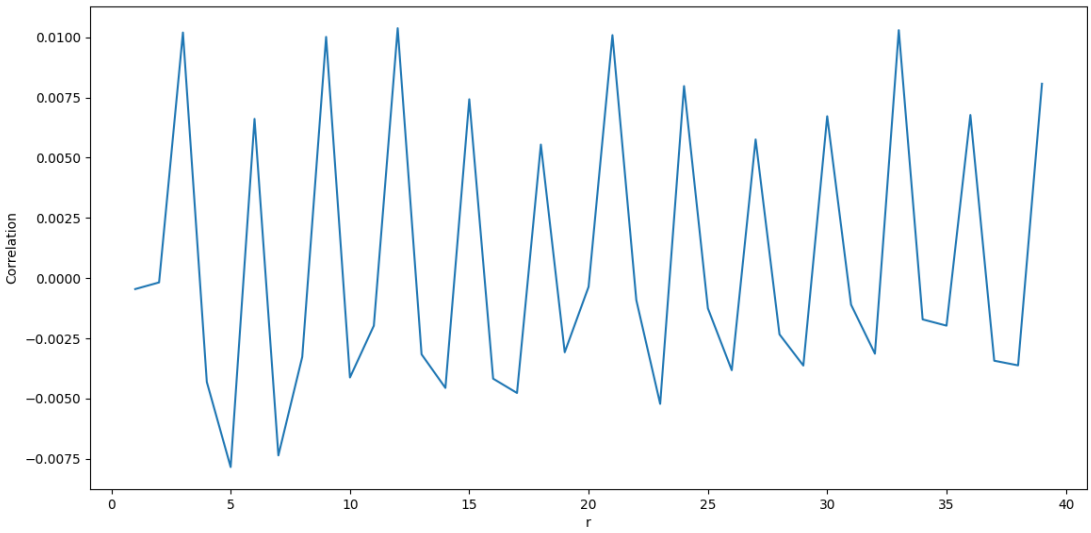

Рис. 1. Кореляційна функція для огрубленої послідовності РНК SARS-CoV-2, де символи A та G замінені на 1, а C та T – на 0

Помітно, що графік має періодичність, рівну трьом. Це пояснюється триплетністю генетичного коду. Корелятор має наступний сенс: якщо $K(r) > 0$ то на цій відстані послідовність має персистентну поведінку, тобто якщо $a_i = A$, то ймовірність того, що a_{i+r} теж буде A росте. В іншому випадку, коли $K(r) < 0$ ймовірність того, що $a_{i+r} = A$ падає, за умови, що $a_i = A$. За визначенням, корелятор є функцією парною, тобто K(-r) = K(r), а дисперсія випадкової змінної a_i для бінарної послідовності – K(0) = $\overline{a}(1 - \overline{a})$

Для вивчення статистичні властивості Марківських ланцюгів, розрахуємо розподіл $W_L(k)$ слів довжини L за кількістю одиниць у такому слові.

$$k_i(L) = \sum_{l=1}^{L} a_{i+l} \quad (2.3)$$

Та дисперсію величини k,

$$D(L) = \overline{k^2} - \overline{k}^2 \quad (2.4)$$

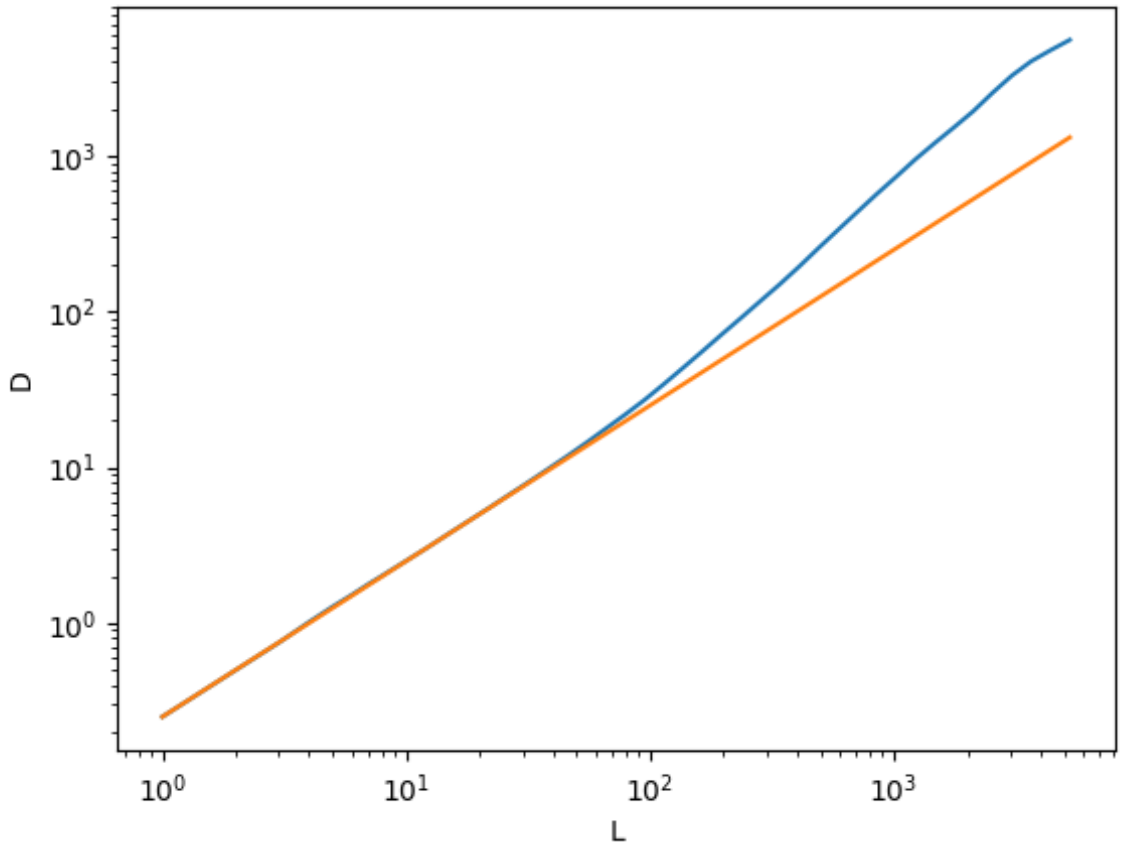

Рис. 2. Дисперсія для огрубленої послідовності РНК SARS-CoV-2, де символи A та G замінені на 1, а C та T – на 0(синім). Помаранчевий графік – дисперсія некорельованного процесу. Графік побудований в подвійному логарифмічному маштабі.

Якщо ланцюг є некорельованим, приходимо до результату Броуновської дифузії,

$$D(L) = \overline{a}(1-\overline{a})L \text{ (2.5)}$$

Видно, що на Рис.2 на відстанях до 10^2 дисперсія майже не відрізняється від дисперсії, притаманній некорельованій послідовності, а на більших відстанях графік йде вище помаранчевої лінії, тобто проявляється персистентна поведінка. Якщо би синій графік був би нижче ніж помаранчевий, тоді поведінка на таких відстанях була б антиперсистентною.

Як слідує з (2.2) та (2.4) дисперсія пов'язана з бінарною кореляцією за наступним рівнянням:

$$K(r) = \tfrac{1}{2}(D(r-1) + D(r) + D(r+1)) \text{ (2.6)}$$

або

$$K(r) = \frac{1}{2}\frac{d^2 D(r)}{dr^2} \quad (2.7)$$

в непреревному наближенні.

РОЗДІЛ 3.
ФУНКЦІЯ ПАМ'ЯТІ В БІНАРНИХ МАРКОВСЬКИХ ЛАНЦЮГАХ

Функція пам'яті $F(r)$ описує силу впливу попереднього символу a_{i-r} на символ a_i. Додатні значення функції пам'яті спричинені персистентною дифузією де попередні зміщення Броуновської частки в якомусь напрямку викликають свої послідовне зміщення у тому ж напрямі. Негативні значення функції пам'яті відповідають антиперсистентній дифузії, де зміни напрямку руху більш ймовірні. Із [3] отримуємо наступну систему лінійних рівнянь, вирішив яку можна отримати функцію пам'яті для всіх відстастаней для яких перед цим рахувалась функція кореляції:

$$K(r) = \sum_{r'=1}^{N} F(r')K(r - r'), r \geq 1 \text{ (3.1)}$$

Ці результати можна розуміти інтуїтивно. Функція пам'яті F(r) характеризує прямий вплив символів на відстані r. У контрасті, корелятор також засвідчує неявний вплив. Кожний доданок в (3.1) є "впливом"символів на відстані $(r - r')$ з "інтенсивністю"$F(r')$.

Також, функція пам'яті входить до рівняння умовної ймовірності(1.2). Тобто, якщо нам відомі значення функції для всіх відстаней, ми можемо побудувати(сгенерувати) послідовність. В цьому вона відрізняється від корелятора.

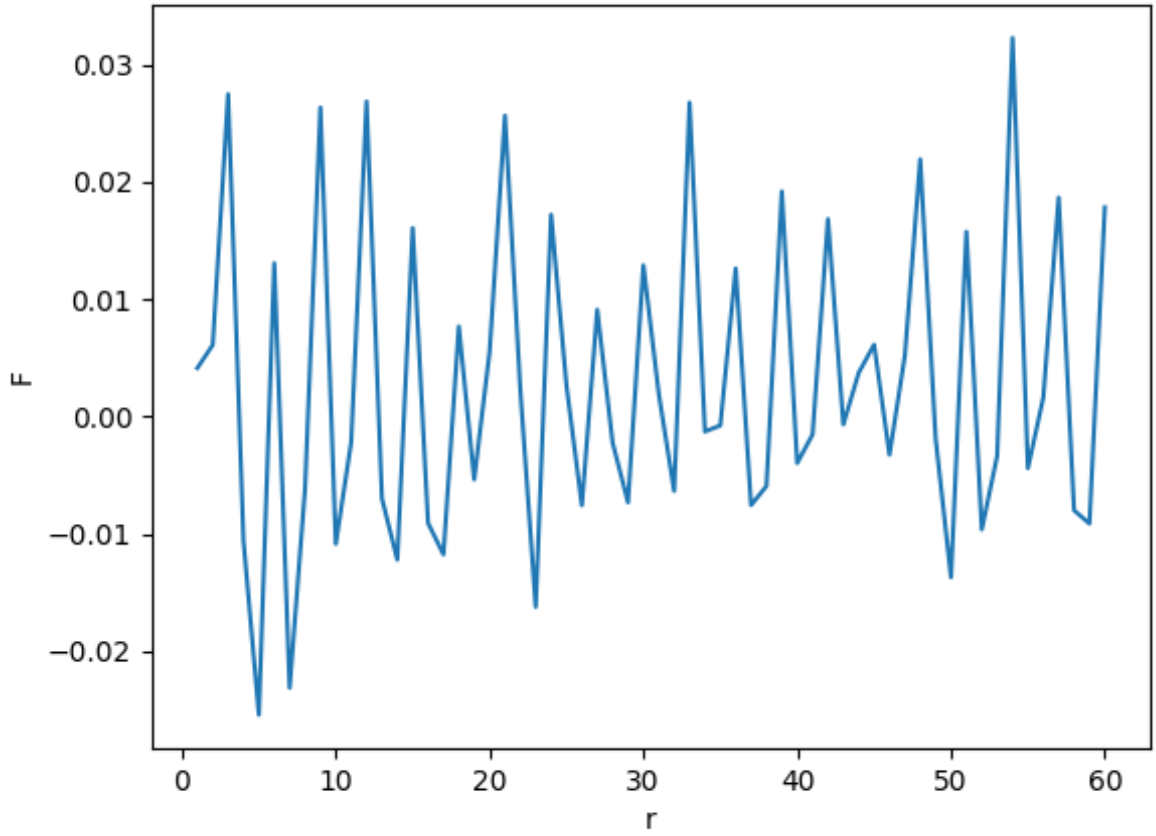

Рис. 3. Функція пам’яті огрубленної послідовності РНК SARS-CoV-2, A та G переходять в 1, інші символи в 0

Помітимо, що графік функції пам’яті має періодичність, дорівнюючу 3, що, як і в випадку з корелятором, можна пов’язати з триплетною структурою геному.

РОЗДІЛ 4.
ОТРИМАНІ СИГНАТУРИ РІЗНИХ КОРОНАВІРУСІВ

За допомогою розробленої комп'ютерної програми на Python (яку можна знайти за посиланням: https://github.com/DummyEgg/corelationsRNA) були знайдені деяки сигнатури різніх коронавірусів:

4.1. MERS-CoV

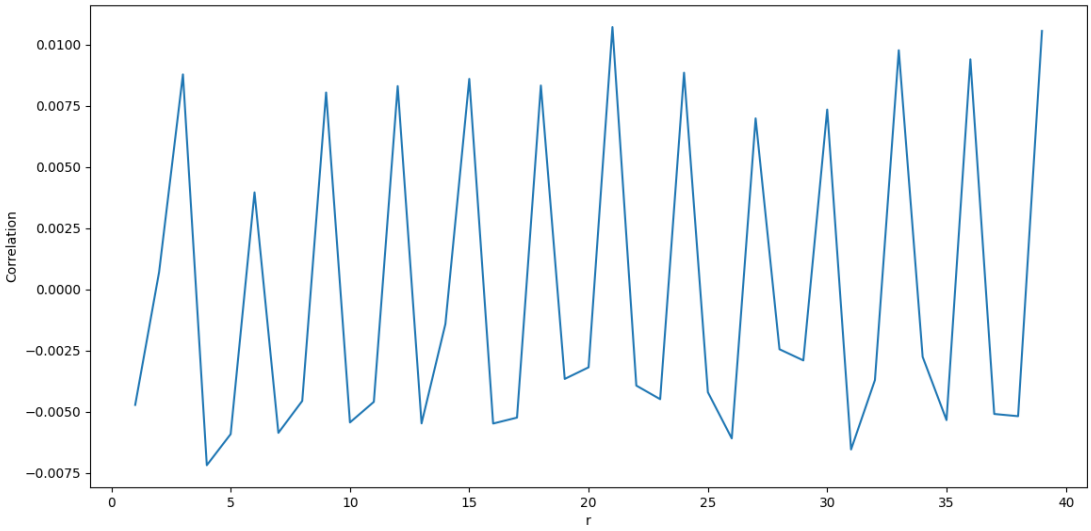

Рис. 4. Корелятор огрубленної послідовності РНК MERS-CoV, A та G переходять в 1, інші символи в 0

Як і у випадку з SARS-CoV-1, графік має періодичність 3, що пов'язано з триплетною структурою генкоду.

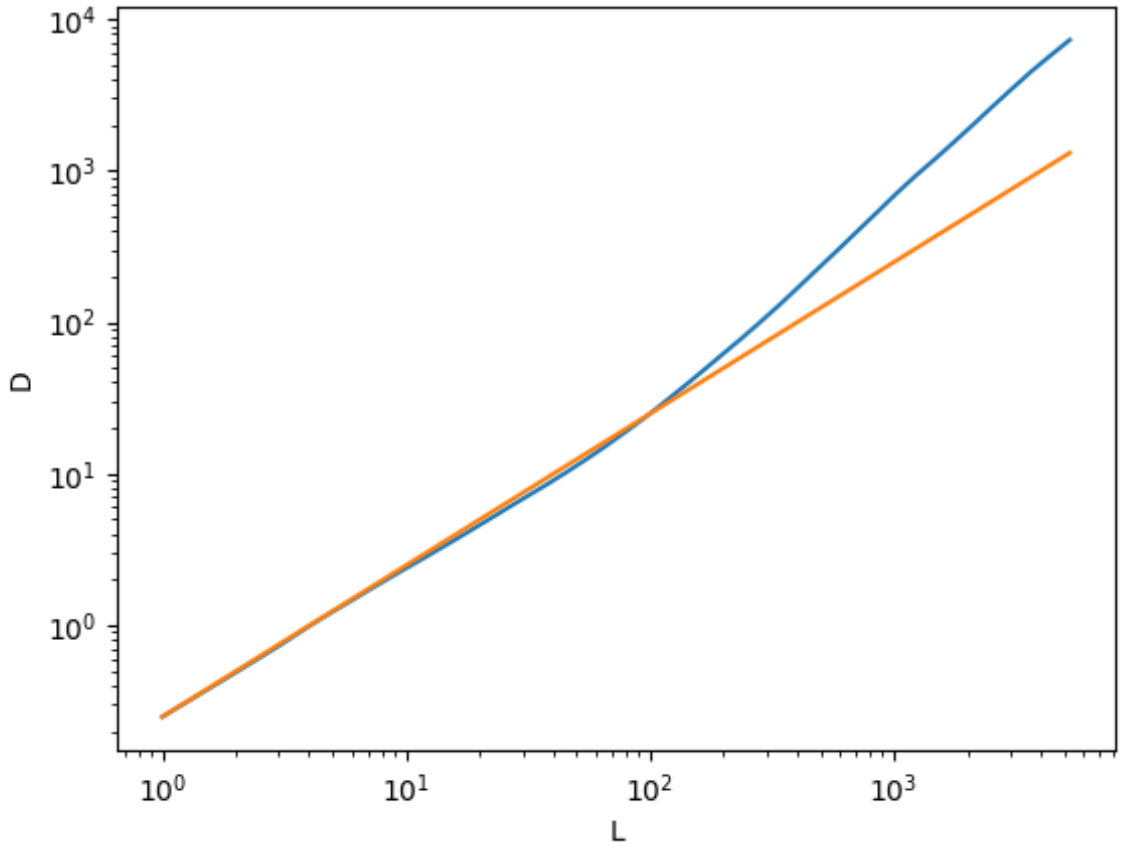

Рис. 5. Дисперсія огрубленної послідовності РНК MERS-CoV(синім).Помаранчевий графік – дисперсія некорельованного процесу. A та G переходять в 1, інші символи в 0

Порівнюючи Рис. 5 та Рис. 2 видно, що на між 10^1 та 10^2 синій графік проходить трішки нижче ніж помаранчева лінія, а потім, як і у випадку з SARS-CoV-2, йде набагато вищче.

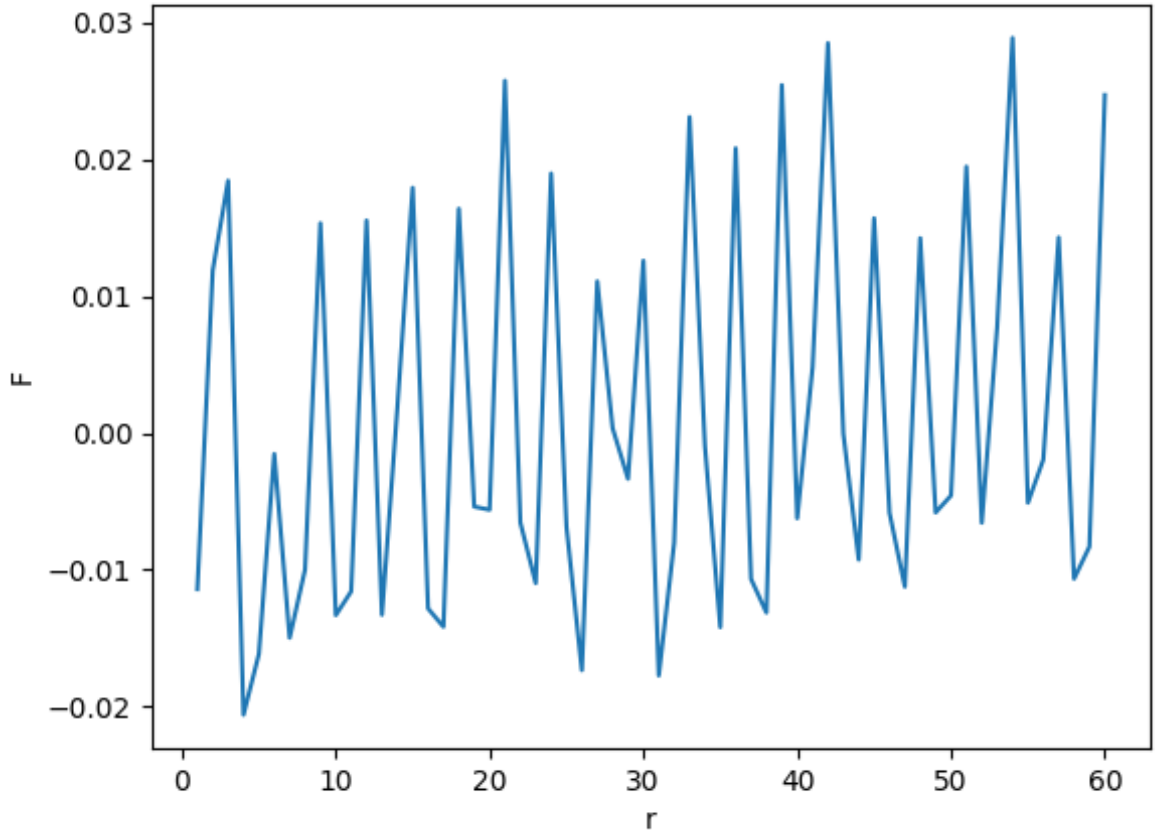

Рис. 6. Функція пам’яті огрубленної послідовності РНК MERS-CoV, A та G переходять в 1, інші символи в 0

4.2. SARS-CoV-1

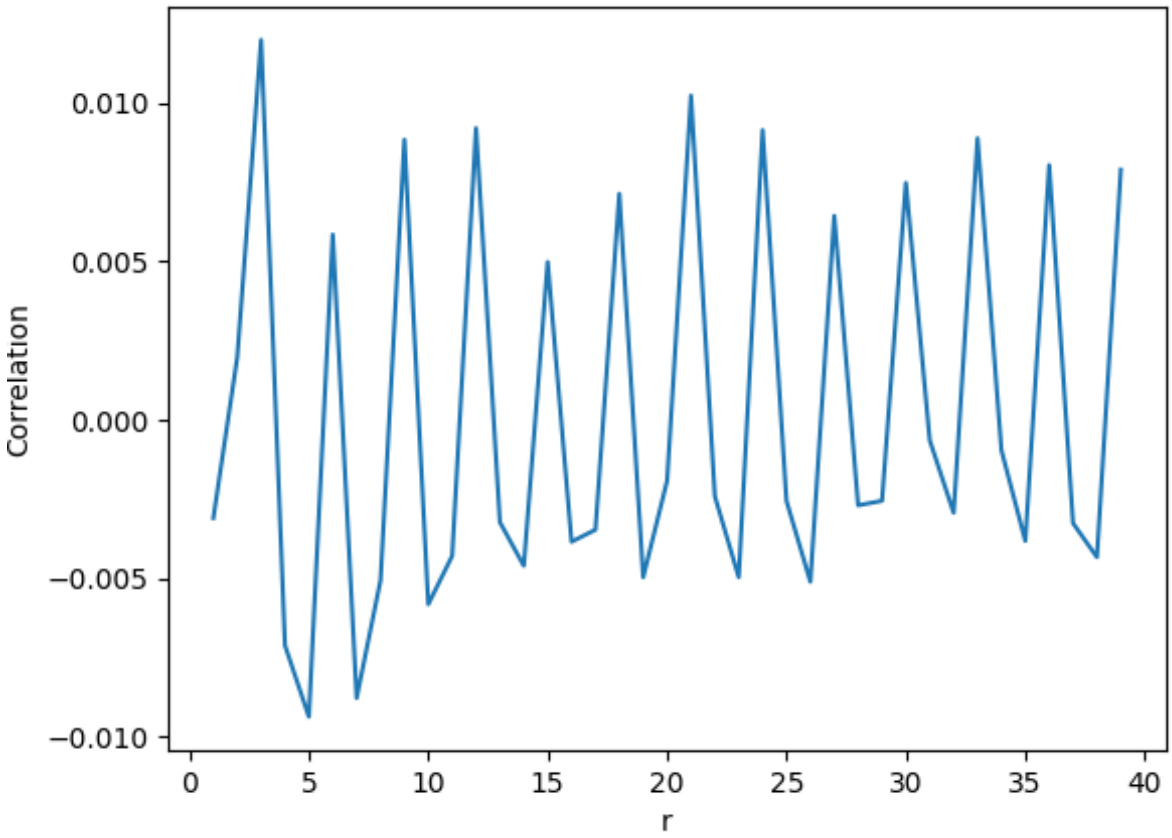

Рис. 7. Корелятор огрубленної послідовності РНК SARS-CoV-1, A та G переходять в 1, інші символи в 0

Як і в двох попередніх випадках, графік має періодичність 3 з тих же причин.

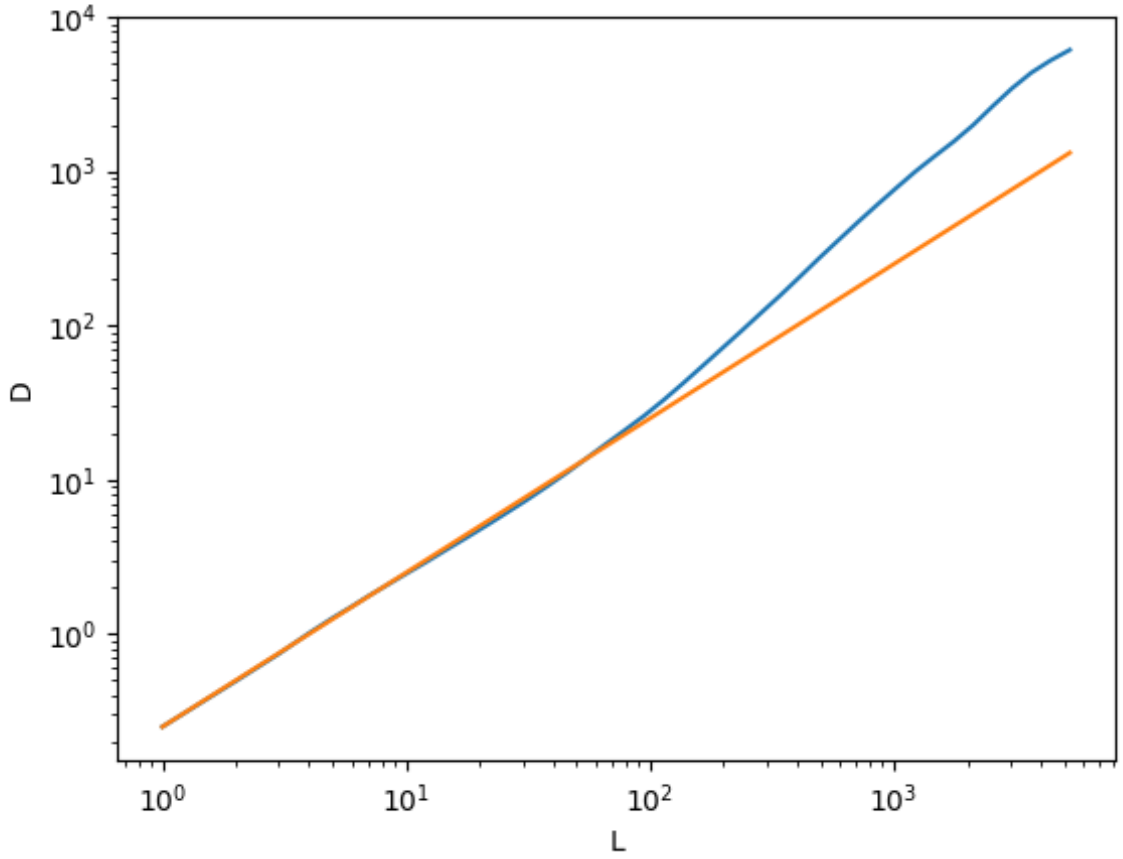

Рис. 8. Дисперсія огрубленної послідовності РНК SARS-CoV-1(синім). Помаранчева графік– дисперсія некорельованого процесу. A та G переходять в 1, інші символи в 0

Як і в випадку на Рис. 5, на відстанях між 10^1 та 10^2 синій графік проходить трішки нижче ніж помаранчевий. Потім, як і в двух попередніх випадках, графік на більших відстанях проходить вище лінії.

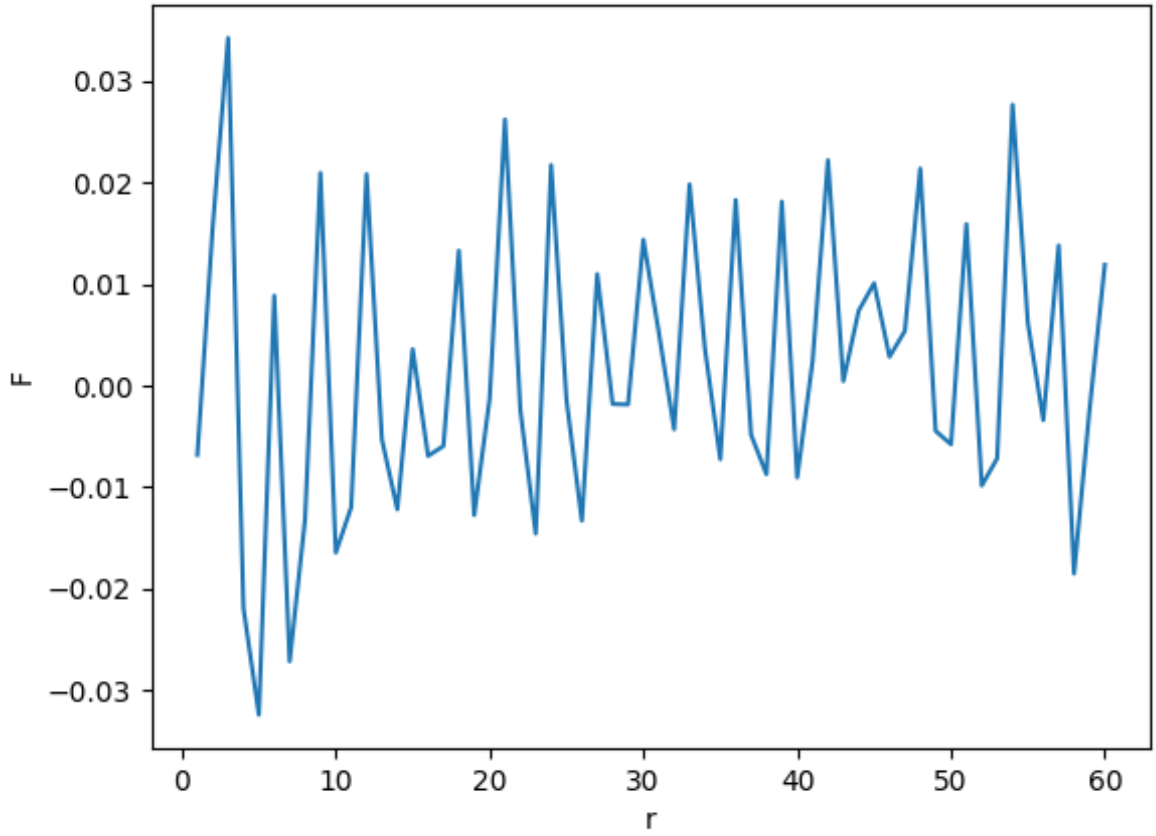

Рис. 9. Функція пам’яті огрубленної послідовності РНК SARS-CoV-1, A та G переходять в 1, інші символи в 0

РОЗДІЛ 5.
ВИСНОВКИ

В ході роботи були отримані сигнатури вірусу SARS-CoV-2 та MERS-CoV, що описують далекі кореляції, присутні в геномах вірусів. До них належать функція пам’яті випадкової системи та матриця парних кореляційних функцій. Ці сигнатури чисельним чином характеризують далекий взаємовплив нуклеотидів один на одного, який простягається на всю довжину геномної послідовності. Такі далекі кореляції раніше майже не досліджувалися, хоча вони, поряд з досліджуваними ближніми кореляціями, в істотній мірі впливають на властивості вірусу та на регуляторні функції його геному.

Також, в ході роботи була розроблена методика, котра може допомогти в задачах визначення міри відмінності вірусів за вибраними параметрами, класифікації фрагментів ДНК та РНК, швидкої ідентифікації в небезпечних для здоров’я ситуаціяї, тощо. Передбачається зіставлення сигнатури SARS-CoV-2 та інших коронавірусів с їх біологічними і клінічними проявами та перехресно зіставити сигнатури інших вірусів.

Таким чином, отримані внаслідок виконання роботи дані дадуть унікальну додаткову інформацію про структуру вірусів, яка в подальшому може бути використана вірусологами, клініцистами та епідеміологами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. O.V Usatenko, S.S. Apostolov, Z.A. Mayzelis, S.S. Melnik. Random Finite-Valued Dynamical Systems
2. S.S. Melnik, O.V. Usatenko. Entropy and long-range memory in random symbolic additive Markov chains
3. S.S. Melnyk, O.V. Usatenko, V.A. Yampol'skii, S.S. Apostolov, Z.A. Maiselis. Memory functions and correlations in additive binary
Markov chains
4. R. G. Jahn, B. J. Dunne, R. D. Nelson, Y. H. Dobyns, and G. J. Bradish. Correlations of Random Binary Sequences with Pre-StatedOperator Intention: A Review of a 12-Year Program