Зміст

1 Вступ

З розвитком біотехнологій все більше зразків послідовностей ДНК вдається отримати. Кількість послідовностей росла екпотенціально на протязі минулих двадцяти років. Полідовність ДНК складається з чотирьох різних нуклеотидів: аденін(A), цитозин(C), гуанін(G) і тимін(T). Вона містить багато біологічної, фізіологічної і хімічної інформації, через що стало дуже важливо аналізувати генетичні послідовності. Було запропоновано багато обчислювальниз і статистичних методів для порівняння біологічних послідовностей. Незважаючи на це, тема порівняння послідовностей залишається актуальною і на цей час. Існуючі методи можна розділити на групуючі і не групуючі.

Групуючі методи використовують динамічне програмування, за допомогою регресії знаходять оптимальне групування за допомогою присвоєння рахунку до різних можливих групувань і вибирають групування з найбільшим рахунком.

Серед всіх існуючих не групуючих методів порівняння біологічних послідовностей, графічне представлення забезпечує простий спосіб перегляду, сортування та порівняння генних структур. Мета графічного подання це відображення послідовності ДНК або білка графічно, так що ми можемо легко візуально визначити наскільки схожі або наскільки відрізняються послідовності. Звичайно, тільки візуального порівняння послідовностей недостатньо для подальшого дослідження. Потрібний більш точний спосіб порівняння.

У даній роботі будуть розглянуті основні методи представлення послідовності ДНК у числовому вигляді, а також спроба застосувати p-статистики як міри близькості між ними. Чисельні послідовності, які отримуються за допомогою одного з описаних нижче алгоритмів розглядаються як вибірки деякого неперервного розподілу. Далі ми використовуємо p-статистики, як міри близькості між розподілами.

2 Огляд літератури

Описано багато методів числового представлення генетичних послідовностей. Тут ми розглянемо тільки ті, які здаються найбільш перспективними. У [1] розглядається наступний спосіб подання ДНК. Чотирьом нуклеотидам A, G, C і T ставляться у відповідність вектори: A $(1, 0.8)$, G $(1, 0.6)$, C $(1, 0.4)$, T $(1, 0.2)$. Елементи послідовності ми отримуємо, сумуючи вектори, що ставляться у відповідність нуклеотидам з послідовності.

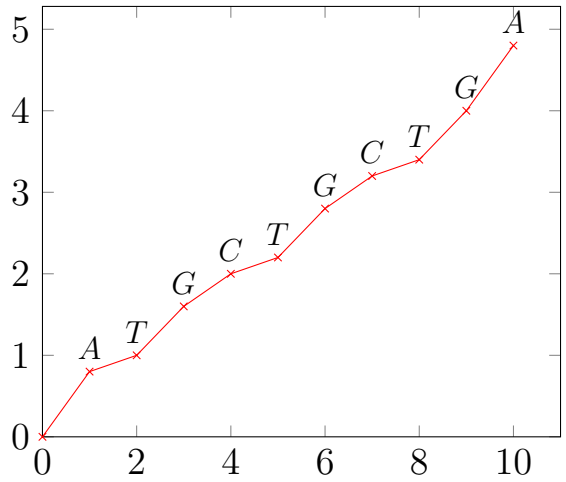

Рис. 1: Графічне представлення послідовності ATGCTGCTGA

На Рис. 1 показано графічне представлення ДНК послідовності "ATGCTGCTGA". Таким чином ми отримуємо взаємнооднозначну відповідність між послідовність нуклеотидів і отриманими точками.

Після цього у відповідність послідовності ДНК довжини n, ставиться у відповідність розподіл ймовірностей $(p_1, p_2, ..., p_n)$,

$$\frac{x_i - \overrightarrow{y_i}}{\frac{1}{2}n(n+1) - y_n},$$

де (x_i, y_i) відповідає позиції i-того нуклиотиду на графіку ДНК, $\overrightarrow{y_i}$ відповідає вибору y-координати при i-тому нуклиотиді у графічному представленні. Далі у цій статті доводиться, що це дійсно буде дискретним розподілом ймовірностей, а далі

використовується розбіжність Кульбака-Лейблера або відносна ентропія.

У [2] описується метод графічного представлення послідовностей ДНК. Тут використовуються блукання у $2D$-просторі. Починають з точки $(0, 0)$. Потім, в залежності від послідовності рухаємося у одному з чотирьох напрямків. Напрямки співвідносяться з нуклеотидами наступним чином: A$=(-1, 0)$, G$=(1, 0)$, C$=(0, 1)$, T$=(0, -1)$. Зрозуміло, що блукаючи таким чином, точки будуть повторюватися, тому якщо ми потрапили в точку m раз, ми присвоюємо їй вагу m. Для порівняння ДНК використовуються характеристики отриманих точок, такі як ценрт масс і тензори моменту інерції.

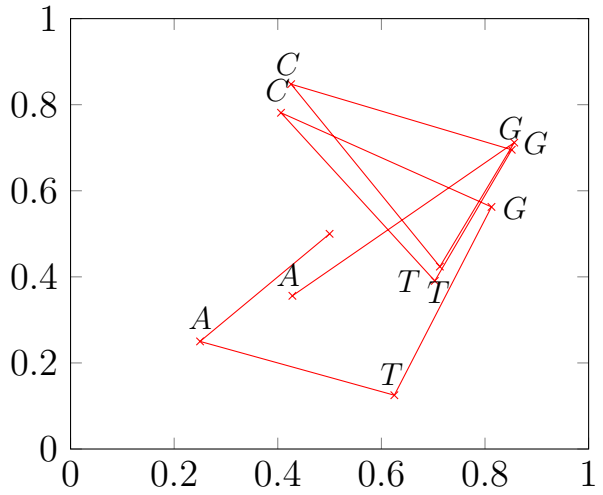

Рис. 2: Графічне представлення послідовності ATGCTGCTGA

У [3] використовуються нова область фізики, відома як, "нелінійна динаміка "хаотичні динамічні системи або просто "хаос". Насправді, ця ітеративна процедура з'вилася у статистичній механіці, зокрема в теорії хаосу. Простір можна розглядати як безперервну систему посилань, в якій всі можливі послідовності будь-якої довжини займають унікальне положення. Позиція отримується за допомогою чотирьох можливих нуклеотидів, які розглядаються як точки на квадраті зі стороною 1. Оскільки, формально генетичну послідовнясть можна розглядати, як рядок складений з чотирьох літер A, C, G і T, то наступні точки ставляться у відповідність чотирьом нуклуотидам: A$=(0, 0)$, G$=(1, 1)$, C$=(0, 1)$, T$=(1, 0)$. Координати послі-

довності рахуються ітеративно, рухаючись на половину відстані між попередньою позицією і точкою квадрата, якій відповідає наступний нуклеотид у напрямку цієї точки. Наприклад, якщо G наступний нуклеотид, то наступна точка буде по середині відрізка, що з'єднує попередню точку і $(1, 1)$. Ітеративну процедуру можна задати наступним чином:

$$p_i = p_{i-1} - 0.5(p_{i-1} - g_i)$$

$$i = 1, ..., n; p_0 = (0.5, 0.5),$$

де g_i - координати, що відповідають i-тому нуклеотиду, n - довжина послідовності ДНК. На Рис. 2 показано ламану утворену точками p_i для послідовності "ATGCTGCTGA".

Кожній точці ставлять у відповідність число:

$$z_i = x_i + y_i,$$

де x_i, y_i це x-координата і y-координата точки p_i. Далі розглядають чисельні характеристики послідовності z_i, зокрема середнє, часткове середнє, стандартне відхилення.

У [4] представленні методи кодування ДНК послідовностей у одновирних, двомимірних і тривимірних просторах. Основна ідея така сама, як в ітеративній процедурі описаній у [3]. У $2D$-просторі алгоритми співпадають. Відмінність тривимірного простору у тому, що тут нуклеотидам ставляться у відподність точки, що є вершинами тетраедра, а у одновимірному просторі нуклеотидам T,G ставиться у відповідність 1, а A,C ставиться у відповідність -1.

У [5] вводиться поняття p-статистики, як міри близькості між неперервними розподілами. У [6] це поняття розширюється на багатовимірні розподіли, а у [7] вводиться модифікована p-статистика, яку можна застосовувати до вибірок з повтореннями.

3 Постановка задачі

Використовуючи відомі алгоритми числового представлення послідовності ДНК, знайти такі, які допускають використання p-статистик. Перевірити доцільність застосування p-статистик до знайдених алгоритмів шляхом порівняння за допомою них послідовностей ДНК різних видів. Послідовності ДНК можна взяти з ГенБанку(www.ncbi.nlm.nih.gov/genbank/).

4 Алгоритм розв'язання

Тут розглядаються 7 різних методів числового подання ДНК.

Метод 1 Кожному нуклеотиду A, G, C, T ставимо у відповідність вектори $(1, 0.8)$, $(1, 0.6)$, $(1, 0.4)$, $(1, 0.2)$. Починаючи з точки $(0, 0)$ рухаємось у напрямку векторів. Точки через які ми проходимо утворюють послідовність.

Метод 2 Спочатку використаємо попередній метод щоб отримати числову послідовність (x_i, y_i). Далі використаємо наступну формулу для обчислення результуючої послідовності:

$$\frac{x_i - \overrightarrow{y_i}}{\frac{1}{2}n(n+1) - y_n},$$

де $\overrightarrow{y_i}$ це y-компонента вектора, що відповідає i-тому нуклеотиду при використанні методу 1, n це розмір ДНК послідовності.

Метод 3 Нуклеотидам A, G, C, T ставимо у відповідність вектори $(-1, 0)$, $(1, 0)$, $(0, 1)$, $(0, -1)$. Починаємо з точки $(0, 0)$ і рухаємось по відповідним векторам. Точки через які ми проходимо утворюють послідовність, причому точка стільки разів зустрічається у послідовності, скільки разів ми в неї потрапили.

Метод 4 Розташовуємо нуклуотиди у вершинах квадрата зі стороною 1: A=$(0,0)$, G=$(1,1)$, C=$(0,1)$, T=$(1,0)$. Координати послідовності рахуються ітеративно, рухаючись на половину відстані між попередньою позицією і точкою квадрата, якій відповідає наступний нуклеотид у напрямку цієї точки. Ітеративну процедуру можна задати наступним чином:

$$p_i = p_{i-1} - 0.5(p_{i-1} - g_i)$$

$$i = 1, ..., n; p_0 = (0.5, 0.5),$$

де g_i - координати, що відповідають i-тому нуклеотиду, n - довжина послідовності ДНК.

Метод 5 Використовуємо попередній метод, щоб отримати послідовність p_i, отримуємо результуючу, як суму всіх попередніх:

$$z_i = \sum_{j=1}^{i} p_i$$

Метод 6 Отримуємо за допомогою методу 4 послідовність p_i і, щоб отримати результуючу послідовність, кожній точці ставимо у відповідність число:

$$z_i = x_i + y_i,$$

де $p_i = (x_i, y_i)$.

Метод 7 Нуклеотидам A,C ставимо у відповідність -1, а нуклеотидам T,G ставимо у відповідність 1. Починаючи з точки 0 рухаємось ітеративно:

$$p_i = p_{i-1} - \frac{(g_i - p_{i-1})}{2} sign(g_i)$$

де g_i число яке відвідає i-тому нуклеотиду. Тобто ми, подібно до методу 4, рухаємося на піввідстань до числа яке відповідає i-тому нуклеотиду.

Введемо означення p-статистики.

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0089988408	0.0089988408	0.0045045045	0.0136456999	0.0181095317
rat	-	-	0.0086907785	0.0063312938	0.0079739443	0.0087959493
rabbit	-	-	-	0.0061880905	0.0099620119	0.0110009416
human	-	-	-	-	0.0059523810	0.0044147883
duck	-	-	-	-	-	0.0175438596
gorilla	-	-	-	-	-	-

Табл. 1: Результати p-статистик при представленні ДНК методом 1

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0322127997	0.0389644724	0.0177028044	0.0763528766	0.0923169219
rat	-	-	0.3029869070	0.0778903070	0.0667766113	0.0595756231
rabbit	-	-	-	0.0591540274	0.0938588610	0.0801315309
human	-	-	-	-	0.0356433537	0.0317225841
duck	-	-	-	-	-	0.7232590650
gorilla	-	-	-	-	-	-

Табл. 2: Результати p-статистик при представленні ДНК методом 2

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0183905103	0.0183905103	0.0092165899	0.0275217614	0.0404505888
rat	-	-	0.0339273529	0.0100363808	0.0204493793	0.0182021462
rabbit	-	-	-	0.0105030456	0.0204493793	0.0182021462
human	-	-	-	-	0.0082051022	0.0091219711
duck	-	-	-	-	-	0.0317440415
gorilla	-	-	-	-	-	-

Табл. 3: Результати p-статистик при представленні ДНК методом 3

Позначимо через H гіпотезу про рівність неперервних функцій розподілу генеральних сукупностей G и G відповідно. Нехай порядкові статистики, де G i G — емпіричні генеральні сукупності, породжені гіпотетичними генеральними сукупностями G и G. Припустимо, що F G (u) = F G (u) . Позначимо через A ij (k) , k 1,2,..., m випадкову подію, яка полягає в тому, що x k потрапляє в інтервал x (i) , x (j) ,

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.2065361072	0.2012181482	0.1847253574	0.2825737702	0.3246090334
rat	-	-	0.1543751287	0.1288092140	0.2209592377	0.2054197652
rabbit	-	-	-	0.1297989157	0.2198026513	0.2122557215
human	-	-	-	-	0.1649465645	0.1627058473
duck	-	-	-	-	-	0.2219235795
gorilla	-	-	-	-	-	-

Табл. 4: Результати p-статистик при представленні ДНК методом 4

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0089988408	0.0090191772	0.0045045045	0.0136660362	0.0182112135
rat	-	-	0.0087677960	0.0063517816	0.0080054696	0.0088199689
rabbit	-	-	-	0.0062018164	0.0099974780	0.0110081474
human	-	-	-	-	0.0059464699	0.0044604254
duck	-	-	-	-	-	0.0175870948
gorilla	-	-	-	-	-	-

Табл. 5: Результати p-статистик при представленні ДНК методом 5

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.4182884917	0.4678787139	0.1844203120	0.3520936286	0.4649299412
rat	-	-	0.9950007244	0.3445051081	0.5063503885	0.4353371079
rabbit	-	-	-	0.3716224769	0.5287530934	0.6094445086
human	-	-	-	-	0.8111966236	0.3886802627
duck	-	-	-	-	-	0.4954142354
gorilla	-	-	-	-	-	-

Табл. 6: Результати p-статистик при представленні ДНК методом 6

тобто A ij (k). Якщо F G (u) = F G (u) (тобто G G) імовірність цієї події обчислюється за формулою (9).Вісник Київського університету Покладемо де h ij (n) — частота події A ij (n) в m випробуваннях. Величина g визначає рівень значущості довірчого інтервалу I ij (n , m) при g =3 рівень значущості цього інтервалу не перевищує 0,05. I ij (n , m) Позначимо через N кількість всіх довірчих інтервалів . Оскільки h (n , m) – частота випадкової події B p ij , що має імовірність p (B) , містять імовірності p ij (n) . Покладемо h (n , m) , x то, покладаючи в

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.5980314400	0.5698554085	0.9071339963	0.8885668965	0.9481829459
rat	-	-	0.9977794554	0.5604684515	0.3972293154	0.4013302012
rabbit	-	-	-	0.4224342110	0.3800282940	0.3785981246
human	-	-	-	-	0.5981285762	0.7745263350
duck	-	-	-	-	-	0.9520161988
gorilla	-	-	-	-	-	-

Табл. 7: Результати p-статистик при представленні ДНК методом 7

формулі (11) h ij (n , m) h (n) , m N i g 3 , ми отримаємо довірчий інтервал I (n , m) p (1) , p (2) для імовірності p (B) . Статистику h (n) називатимемо модифікованою p-статистикою. Вона є шуканою мірою близькості x , x між виборками x и x .

Література

[1] Chenglong Yu, Mo Deng, Stephen S.-T. Yau, DNA sequence comparison by a novel probabilistic method, Information Sciences 181 (2011) 1484–1492

[2] Dorota Bielinska-Waz, Timothy Clark, Piotr Waz, Wiesław Nowak, Ashesh Nandy, 2D-dynamic representation of DNA sequences, Chemical Physics Letters 442 (2007) 140–144

[3] Wei Deng and Yihui Luan, Hindawi Publishing Corporation, Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation, Abstract and Applied Analysis, Volume 2013, Article ID 926519, 6 pages, http://dx.doi.org/10.1155/2013/926519

[4] Jure Zupan and Milan Randic, Algorithm for Coding DNA Sequences into "Spectrum-like" and "Zigzag" Representations, J. Chem. Inf. Model. 2005, 45, 309-313

[5] Д. А. Клюшин, Ю. И. Петунин, Непараметрический Критерий Эквивалентности Генеральных Совокупностей, основанный На Мере Близости Между Выборками, ДК 519.21

[6] Д. А. Клюшин, М. В. Присяжная, Многомерное ранжирование с помощью эллипсов Петунина, Журнал обчисл. та прикл. матем. № 4(114) 2013, стор. 1-7, УДК 519.71

[7] Дмитро А. Клюшин, Міра близькості між виборками, що містять атоми, Вісник Київського університету, Серія: фізико-математичні науки, 2005, 3, УДК 519.9