Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Кафедра обчислювальної математики факультету кібернетики

Випускна кваліфікаційна робота бакалавра
на тему

Числові методи аналізу ДНК

Виконав студент 4-го курсу
Товт Аттіла Аттілович

Науковий керівник
доктор фіз.-мат. наук, професор
Клюшин Дмитро Анатолійович

Київ 2015

Зміст

1 Вступ

З розвитком біотехнологій все більше зразків послідовностей ДНК вдається отримати. Кількість послідовностей росла експонентно на протязі минулих двадцяти років. Послідовність ДНК складається з чотирьох різних нуклеотидів: аденін(A), цитозин(C), гуанін(G) і тимін(T). Вона містить багато біологічної, фізіологічної й хімічної інформації, через що стало дуже важливо аналізувати генетичні послідовності. Було запропоновано багато обчислювальних і статистичних методів для порівняння біологічних послідовностей. Не зважаючи 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.

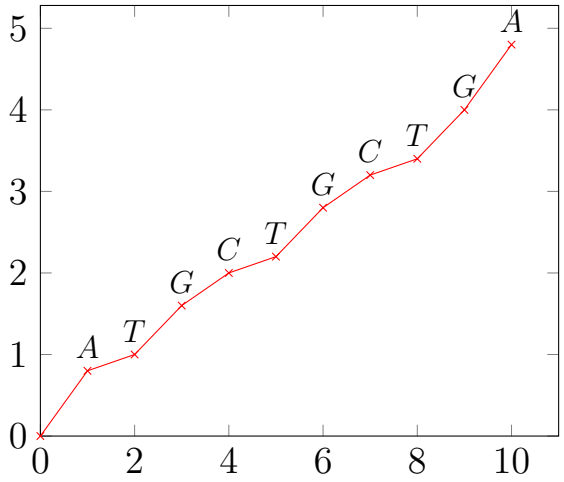

Рис. 1: Графічне представлення послідовності ATGCTGCTGA

На Рис. 1 показано графічне представлення ДНК послідовності "ATGCTGCTGA". Таким чином ми отримуємо взаємно однозначну 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 нуклеотиду на графіку ДНК, $\overrightarrow{y_i}$ відповідає вибору y-координати при i-тому нуклеотиді 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 центр мас і тензори моменту інерції.

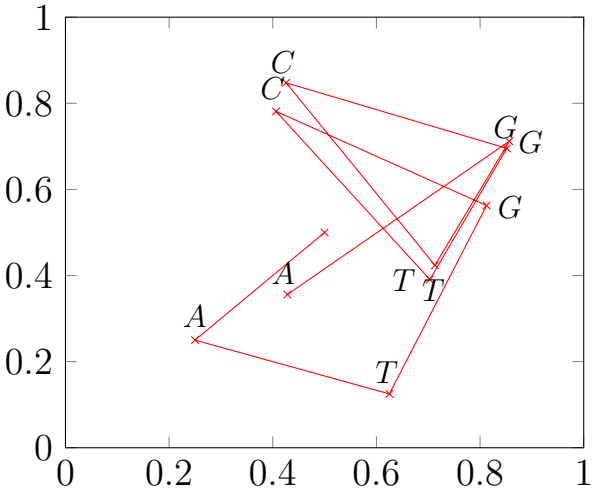

Рис. 2: Графічне представлення послідовності ATGCTGCTGA

У [3] використовуються нова область фізики, відома як, "нелінійна динаміка "хаотичні динамічні системи або просто "хаос". Насправді, ця ітеративна процедура з'явилася 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 послідовність можна розглядати, як рядок складений з чотирьох літер A, C, G і T, то наступні точки ставляться у відповідність чотирьом нуклеотидам: 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. 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 одновимірних, двовимірних і тривимірних просторах. Основна ідея така сама, як в ітеративній процедурі описаній у [3]. У $2D$-просторі алгоритми збігаються. Відмінність тривимірного простору у тому, що тут нуклеотидам ставляться у відповідність 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 допомогою них послідовностей ДНК різних видів. Послідовності ДНК можна взяти з ГенБанку(www.ncbi.nlm.nih.gov/genbank/).

4 Алгоритм розв'язання

Подаємо відрізок ДНК у вигляді числової послідовності, за допомогою одного з методів описаного нижче. Вважаємо, що отримана числова послідовність є вибіркою, якогось неперервного розподілу і, використовуючи p-статистики, порівнюємо вибірки отримані з геномних послідовностей різних видів.

4.1 Методи 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 двовимірну послідовність. Всі елементи послідовності різні. Для побудови p-статистик використовуємо еліпси Петуніна.

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. Отримуємо одновимірну послідовність, елементи якої можуть повторюватися, тому застосовуємо модифіковані p-статистики.

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 відповіднім векторам. Точки через які ми проходимо утворюють послідовність, причому точка стільки разів зустрічається у послідовності, скільки разів ми в неї потрапили. Таким чином кожна точка може зустрічатися більше одного разу. Через що, необхідно застосовувати модифіковані p-статистики разом з еліпсами Петуніна.

Метод 4 Розташовуємо нуклеотиди 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. Тут, аналогічно до попереднього методу, можуть виникнути елементи, які зустрічаються більше одного разу, і тому необхідно використати модифіковані p-статистики. Послідовність двовимірна, тому потрібно використати еліпси Петуніна.

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

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

Отримуємо двовимірну послідовність, і оскільки всі p_i додатні, то елементи не повторюються, через це, при побудові p-статистик використовуємо еліпси Петуніна.

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 $p_i = (x_i, y_i)$. Тут послідовність одномивірна, але може мати повторення, тому застосовуємо модифіковані p-статистики.

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 пів відстань до числа яке відповідає i-тому нуклеотиду. Отримуємо одновимірну послідовність, яка може мати повторення, тому застосовуємо модифіковані p-статистики.

4.2 p-статистики

Позначимо через H гіпотезу про рівність неперервних функцій розподілу $F_G(u)$ і $F_{G'}(u)$ генеральних сукупностей G і G' відповідно. Нехай $x = (x_1, ..., x_n) \in G$ і $x' = (x'_1, ..., x'_m) \in G'$, $x_{(1)} < ... < x_{(n)}$, $x'_{(1)} < ... < x'_{(n)}$ - порядкові статистики. 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 $\left(x_{(i)}, x_{(j)}\right)$, тобто $A_{ij}^{(k)} = \left\{x'_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$. Якщо $F_G(u) = F_{G'}(u)$ (тобто $G = G'$) імовірність цієї події обчислюється за формулою:

$$P\left(A_{ij}^{(k)}\right) = P\left(x'_k \in \left(x_{(i)}, x_{(j)}\right)\right) = p_{ij}^{(n)} = \frac{j-i}{n+1} = \frac{q}{n+1}, q = j - i$$

Покладемо

$$p_{ij}^{(1)} = \frac{h_{ij}^{(n)} m + g^2/2 - g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2}, \tag{1}$$

$$p_{ij}^{(2)} = \frac{h_{ij}^{(n)} m + g^2/2 + g\sqrt{h_{ij}^{(n)}(1-h)m + g^2/4}}{m + g^2}, \tag{2}$$

де $h_{ij}^{(n)}$ — частота події $A_{ij}^{(n)}$ в m випробуваннях. Величина g визначає рівень значущості довірчого інтервалу $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$; в силу правила 3σ при $g = 3$ рівень значущості цього інтервалу не перевищує 0.05.

Позначимо через N кількість всіх довірчих інтервалів $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ $(N = n(n-1)/2)$ і L - кількість тих інтервалів $I_{ij}^{(n,m)}$, які містять ймовірності $p_{ij}^{(n)}$. Покладемо $h^{(n,m)} = \rho\left(F^*, F^{*'}\right) = \rho(x, x') = \frac{L}{N}$. Оскільки $h^{(n,m)}$ – частота випадкової події $B = \left\{p_{ij}^{(n)} \in I_{ij}^{(n,m)}\right\}$, що має імовірність $p(B) = 1 - \beta$, то, покладаючи в формулі $h_{ij}^{(n,m)} = h^{(n)}, m = N$ і $g = 3$, ми отримаємо довірчий інтервал $I^{(n,m)} = \left(p^{(1)}, p^{(2)}\right)$ для імовірності $p(B)$. Статистику $h^{(n)}$ називатимемо p-статистикою. Вона є мірою близькості $\rho(x, x')$ між вибірками x і x'.

4.3 Модифікована p-статистика

Нехай $x = (x_1, x_2, ..., x_n)$ — вибірка, отримана з генеральної сукупності G, що має функцію розподілу $F(u)$, за допомогою простого випадкового вибору. Атомом вибірки x будемо називати вибіркове значення x_k , що зустрічається у вибірці x більше одного разу: $x_k = x_{k_1} = ... = x_{k_i}, k, k_1, ..., k_i \in \{1, 2, ..., n\}$. Кратністю $t(x_k)$ довільного вибіркового значення x_k називатимемо кількість його повторень у виборці x. Таким чином, атоми — це вибіркові значення, кратність яких перебільшує одиницю.

Нехай $x_{(1)} \leq ... \leq x_{(n)}$ - варіаційний ряд вибірки x. Тоді згідно [7]:

$$p_{ij} = p\left(A_{ij}\right) = p\left(x^* \in (x_{(i)}, x_{(j)})\right) \approx \gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1}, \tag{3}$$

$$\gamma_l = \gamma\left(x_{(l)}\right) = \frac{t(x_{(l)}) - 1}{n+1},$$

де x^* вибіркове значення із генеральної сукупності G , що не залежить від вибірки x.

Позначимо через H гіпотезу про рівність неперервних функцій розподілу $F_G(u)$ и $F_{G'}(u)$ генеральних сукупностей G і G' відповідно. Нехай $x = (x_1, ..., x_n) \in G$ і $x' = (x'_1, ..., x'_m) \in G'$, $x_{(1)} \leq ... \leq x_{(n)}$, $x'_{(1)} \leq ... \leq x'_{(n)}$ - порядкові статистики. Причому, вибірки x і x' містять атоми. 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 $\left(x_{(i)}, x_{(j)}\right)$, тобто $A_{ij}^{(k)} = \left\{x'_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$. Якщо $F_G(u) = F_{G'}(u)$ (тобто $G = G'$) імовірність цієї події обчислюється за формулою 3.

За формулами 1 і 2 обчислюємо $p_{ij}^{(1)}$ і $p_{ij}^{(2)}$, відповідно. Позначимо через N кількість всіх довірчих інтервалів $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ $(N = n(n-1)/2)$ і L - кількість тих інтервалів $I_{ij}^{(n,m)}$, які містять ймовірності $p_{ij}^{(n)}$. Покладемо $h^{(n,m)} = \rho\left(F^*, F^{*'}\right) = \rho(x, x') = \frac{L}{N}$. Оскільки $h^{(n,m)}$ – частота випадкової події $B = \left\{p_{ij}^{(n)} \in I_{ij}^{(n,m)}\right\}$, що має імовірність $p(B) = 1 - \beta$, то, покладаючи в формулі $h_{ij}^{(n,m)} = h^{(n)}, m = N$ і $g = 3$, ми отримаємо довірчий інтервал $I^{(n,m)} = \left(p^{(1)}, p^{(2)}\right)$ для імовірності $p(B)$. Статистику $h^{(n)}$ називатимемо модифікованою p-статистикою.

4.4 Еліпсоїд Петуніна

Розглянемо множину двовимірних точок M_n:

$$M_n = \{\overrightarrow{x_1}, ..., \overrightarrow{x_n}\}, \overrightarrow{x_i} = (x_i, y_i)$$

Побудуємо опуклу оболонку точок $M_n = \{(x_1, y_1), ..., (x_n, y_n)\}$. Знайдемо вершини опуклої оболонки (x_k, y_k) і (x_l, y_l), які лежать на діаметрі опуклої оболонки, тобто найбільш віддалені одна від одної вершини. З'єднаємо точки (x_k, y_k) і (x_l, y_l) відрізком L. Знайдемо вершини (x_r, y_r) і (x_q, y_q) опуклої оболонки найбільш віддалені від L. З'єднаємо точки (x_r, y_r) і (x_q, y_q) відрізками паралельними L_1 і L_2, паралельними відрізку L. Проведемо через точки (x_k, y_k) і (x_l, y_l) відрізки L_3 і L_4, перпендикулярно до відрізка L. Перетин відрізків L_1, L_2, L_3 і L_4 утворюють прямокутник П, сторони якого мають довжину a і b.(Рис. 3)

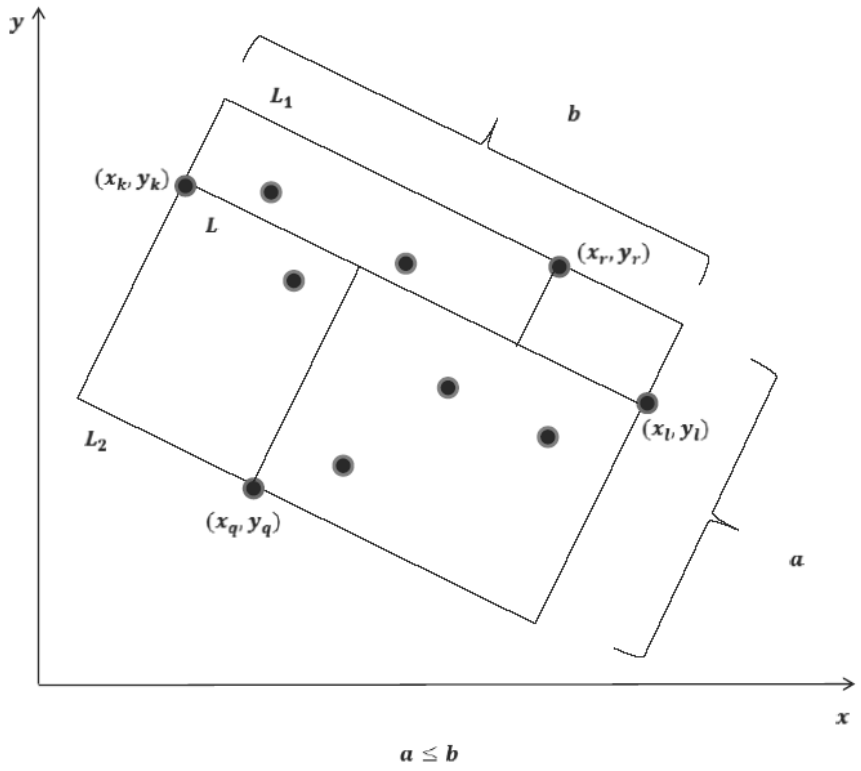

Рис. 3: Прямокутник Петуніна

Будемо вважати, що $a \leq b$. Переведемо лівий нижній кут прямокутника у початок нової системи координат з осями Ox' і Oy' за допомогою паралельного переносу і повороту. Точки $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ перейдуть в точки$(x'_1, y'_1), (x'_2, y'_2), ..., (x'_n$ Відобразимо точки $(x'_1, y'_1), (x'_2, y'_2), ..., (x'_n, y'_n)$ в точки $(\alpha x'_1, y'_1), (\alpha x'_2, y'_2), ..., (\alpha x'_n, y'_n)$, де $\alpha = \frac{a}{b}$. В результаті отримаємо сукупність точок, які належать квадрату S.

Порахуємо центр (x'_0, y'_0) квадрата S і знайдемо відстані $r_1, r_2, ..., r_n$ від нього до кожної точки $(\alpha x'_1, y'_1), (\alpha x'_2, y'_2), ..., (\alpha x'_n, y'_n)$. Найбільше число $R = max(r_1, r_2, ..., r_n)$ визначає круг з центром у точці (x'_0, y'_0) і радіусом R.

В результаті всі точки $(\alpha x'_1, y'_1), (\alpha x'_2, y'_2), ..., (\alpha x'_n, y'_n)$ всередині круга з радіусом R. Розтягуючи цей круг вздовж осі Ox' з коефіцієнтом $\beta = \frac{1}{\alpha}$ і виконуючи обернені перетворення повороту і переносу, отримаємо еліпс Петуніна (Рис. 4).

У m-вимірному просторі на першому кроці знайдемо вершини опуклої оболонки $\overrightarrow{x_k}$ і $\overrightarrow{x_l}$, що лежать на діаметрі випуклої оболонки. З'єднаємо точки $\overrightarrow{x_k}$ і $\overrightarrow{x_l}$ відрізком L. Повернемо і перенесемо систему координат, щоб діаметр опуклої оболонки лежав на осі Ox'_1. Побудуємо найменший прямокутний паралелепіпед, що містить точки $\overrightarrow{x_1}, ..., \overrightarrow{x_n}$.

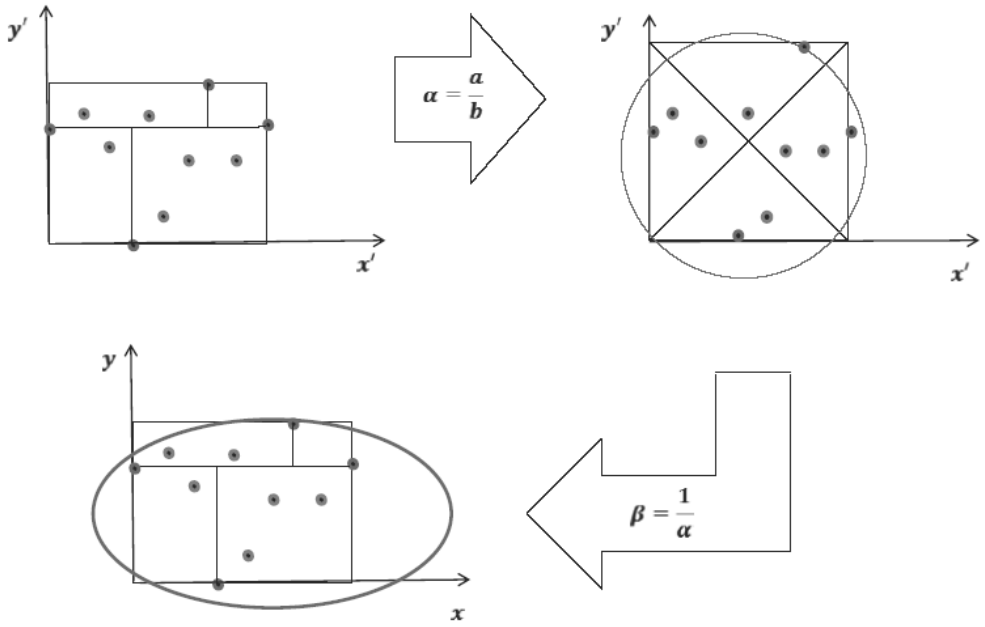

Рис. 4: Побудова еліпса Петуніна

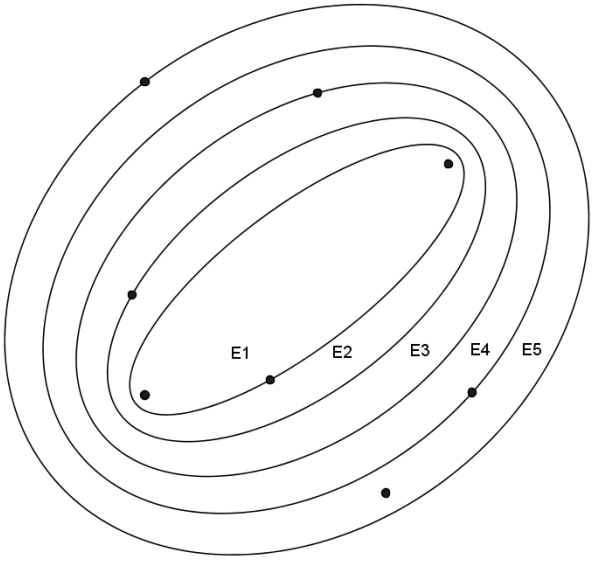

Рис. 5: Вкладені еліпси Петуніна

Зжимаючи прямокутний паралелепіпед, Відобразимо точки у гіперкуб. Знайдемо центр $\overrightarrow{x_0}$ гіперкуба і порахуємо відстані $r_1, r_2, ..., r_n$ від нього до кожної точки. Знайдемо найбільше число $R = max(r_1, r_2, ..., r_n)$ і побудуємо гіперкулю з центром в точці $\overrightarrow{x_0}$ і радіусом R. Застосовуючи до цієї гіперкулі обернені операції розтягування, повороту і переносу, отримаємо еліпсоїд Петуніна в m-вимірному просторі.

Якщо побудувати серію вбудованих еліпсоїдів, то на кожному з них буде лежати по одній точці, тобто відбувається їх ранжування (Рис. 5).

При побудові p-статистики варіаційному ряду вибірок $\overrightarrow{x}_{(1)} \preceq \overrightarrow{x}_{(2)} \preceq ... \preceq \overrightarrow{x}_{(n)}$ поставимо у відповідність послідовність вкладених еліпсоїдів $E_{(1)} \subset E_{(2)} \subset ... \subset E_{(n)}$. Ймовірність того, що елемент $\overrightarrow{x}$ із генеральної сукупності G задовольняє умові $\overrightarrow{x}_{(i)} \preceq \overrightarrow{x} \preceq \overrightarrow{x}_{(j)}$, рівна ймовірності потрапити між еліпсами $E_{(i)}$ і $E_{(j)}$, тобто $\frac{j-i}{n+1}$. Ця умова дозволяє побудувати p-статистику для багатовимірного випадку.

Таким чином будуємо p-статистику як завжди, тільки подія $A_{ij}^{(k)}$ буде полягати в тому, що x'_k попаде в область $E_{(j)} \setminus E_{(i)}$.

5 Результати

Нижче представлені результати застосування одного з методів числового представлення ДНК і p-статистик.

В таблицях представлені результати, порівнюючи послідовностей ДНК джунглевих кур(gallus), пацюка(rat), кролика(rabbit), людини(human), качки(duck) і горили(gorilla).

Геномні послідовності були взяті з ГенБанку (www.ncbi.nlm.nih.gov/genbank/):

(gi|126165289|ref|NM_001081704.1| Gallus gallus hemoglobin, beta (HBE), mRNA),

rat(gi|56251|emb|X06701.1| Rat beta-globin gene),

rabbit(gi|1489|emb|V00882.1| Rabbit (O. cuniculus) gene for beta-globin),

human(gi|34190730|gb|BC047343.2| Homo sapiens cDNA clone IMAGE:5177205),

duck(gi|483525371|gb|KB742400.1| Anas platyrhynchos breed Pekin duck unplaced genomic scaffold scaffold30, whole genome shotgun sequence),

gorilla(gi|401612384|ref|NW_004008464.1| Gorilla gorilla gorilla unplaced genomic scaffold, gorGor3.1 Primary Assembly unplaced10718_1_913, whole genome shotgun sequence)

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0089988408	0.00899884	0.00450450	0.01364569	0.01810953
rat	-	-	0.00869077	0.00633129	0.00797394	0.00879594
rabbit	-	-	-	0.00618809	0.00996201	0.01100094
human	-	-	-	-	0.00595238	0.00441478
duck	-	-	-	-	-	0.01754385
gorilla	-	-	-	-	-	-

Табл. 1: p-статистики при представленні ДНК методом 1

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0322127997	0.03896447	0.01770280	0.07635287	0.09231692
rat	-	-	0.30298690	0.07789030	0.06677661	0.05957562
rabbit	-	-	-	0.05915402	0.09385886	0.08013153
human	-	-	-	-	0.03564335	0.03172258
duck	-	-	-	-	-	0.72325906
gorilla	-	-	-	-	-	-

Табл. 2: p-статистики при представленні ДНК методом 2

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0183905103	0.01839051	0.00921658	0.02752176	0.04045058
rat	-	-	0.03392735	0.01003638	0.02044937	0.01820214
rabbit	-	-	-	0.01050304	0.02044937	0.01820214
human	-	-	-	-	0.00820510	0.00912197
duck	-	-	-	-	-	0.03174404
gorilla	-	-	-	-	-	-

Табл. 3: p-статистики при представленні ДНК методом 3

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.2065361072	0.20121814	0.18472535	0.28257377	0.32460903
rat	-	-	0.15437512	0.12880921	0.22095923	0.20541976
rabbit	-	-	-	0.12979891	0.21980265	0.21225572
human	-	-	-	-	0.16494656	0.16270584
duck	-	-	-	-	-	0.22192357
gorilla	-	-	-	-	-	-

Табл. 4: p-статистики при представленні ДНК методом 4

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.0089988408	0.00901917	0.00450450	0.01366603	0.01821121
rat	-	-	0.00876779	0.00635178	0.00800546	0.00881996
rabbit	-	-	-	0.00620181	0.00999747	0.01100814
human	-	-	-	-	0.00594646	0.00446042
duck	-	-	-	-	-	0.01758709
gorilla	-	-	-	-	-	-

Табл. 5: p-статистики при представленні ДНК методом 5

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.4182884917	0.46787871	0.18442031	0.35209362	0.46492994
rat	-	-	0.99500072	0.34450510	0.50635038	0.43533710
rabbit	-	-	-	0.37162247	0.52875309	0.60944450
human	-	-	-	-	0.81119662	0.38868026
duck	-	-	-	-	-	0.49541423
gorilla	-	-	-	-	-	-

Табл. 6: p-статистики при представленні ДНК методом 6

-	gallus	rat	rabbit	human	duck	gorilla
gallus	-	0.59803144	0.56985540	0.90713399	0.88856689	0.94818294
rat	-	-	0.99777945	0.56046845	0.39722931	0.40133020
rabbit	-	-	-	0.42243421	0.38002829	0.37859812
human	-	-	-	-	0.59812857	0.77452633
duck	-	-	-	-	-	0.95201619
gorilla	-	-	-	-	-	-

Табл. 7: p-статистики при представленні ДНК методом 7

6 Висновки

Довжина послідовностей ДНК досягає 10^5, а запропоновані алгоритми будують послідовності довжини яких пропорційні довжині вхідної послідовності. Алгоритм побудови p-статистики досить повільний і не може мати практичного застосування при довжині вибірки $> 10^4$, або прискорення самого алгоритму. Через це, методи порівняння геномних послідовностей, які були тут розглянуті можуть мати практичне значення лише для відносно коротких послідовностей ДНК.

Також варто проаналізувати отримані міри близькості між геномними послідовностями і загальноприйнятими уявленнями про міжвидову близькість. Методи чисельного представлення ДНК у комбінації з p-статистикою дають здебільшого суперечливі результати, якщо припускати, що міра близькості між генами видів має відображати близькість видів при науковій класифікації.

Зокрема, логічно припусти, що до людини найближче має були горила, в той час, як всі запропоновані методи з цим не погоджуються. Аналогічно жоден з методів не погоджується з тим, що до горили найближче має бути людина. При цьому більшість з методів сходяться на думці, що джунглеві кури близькі до горили, ближче ніж до качки, хоча вони з одного класу.

Дані результати можуть говорити або про те, що наші уявлення про зв'язок між видовою збіжністю і геномними послідовностями неправильний, або що вибрані методи порівняння послідовностей ДНК не мають практичного застосування, або неправильно були вибрані послідовності ДНК і можливо перед порівнянням, потрібно робити ще якусь попередню обробку ДНК.

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