УДК 616-006.04; 577.212.3; 57.087.1 DOI: 10.26565/2313-6693-2022-45-02

ЕНТРОПІЯ ПОСЛІДОВНОСТЕЙ ДНК І СМЕРТНІСТЬ ПАЦІЄНТІВ З ЛЕЙКЕМІЄЮ

Мартиненко О. В.[A,C,D,E,F]***, Пастор К. Д.***[A,B,E,F]***, Фрід С. А.***[B,F]***, Гіл Д. Р.***[B,F]***,***
Малярова Л. В.[E,F]

A – концепція та дизайн дослідження; B – збір даних; C – аналіз та інтерпретація даних; D – написання статті; E – редагування статті; F – остаточне затвердження статті

Вступ. Дезоксирибонуклеїнова кислота (ДНК) не є випадковою послідовністю чотирьох комбінацій нуклеотидів: комплексні огляди літератури переконливо показують довго- та короткодіапазонні кореляції в ДНК, періодичні властивості та кореляційну структуру послідовностей. Методи теорії інформації, зокрема інформаційна ентропія, мають на увазі кількісну оцінку обсягу інформації, що міститься в послідовностях. Зв'язок між ентропією та виживанням пацієнтів широко поширений у деяких галузях медицини та медичних дослідженнях, таких як: кардіологія, неврологія, хірургія, травма. Таким чином, існує необхідність реалізації переваг методів теорії інформації для дослідження взаємозв'язку між смертністю певної категорії пацієнтів та ентропією їх послідовностей ДНК.

Мета. Надати надійну формулу для точного розрахунку ентропії для коротких послідовностей ДНК і показати, як використовувати запропонований аналіз ентропії для вивчення смертності хворих на лейкемію.

Матеріали і методи. Використовувалась база даних пацієнта з лейкемією Барселонського університету (UB) з 117 знеособленими записами, які складаються з наступного: дата діагнозу пацієнта, дата смерті пацієнта, діагнози лейкемії, послідовність ДНК пацієнта. Середній час смерті пацієнта після встановлення діагнозів: 99 ± 77 місяців. Формальними характеристиками послідовностей ДНК в БД UB хворих на лейкемію є: середня кількість ДНК основ $N = 496 \pm 69$; min (N) = 297 основ; max (N) = 745 основ.

Була запропонована узагальнена форма оцінювача ентропії (*EnRE*) для коротких послідовностей ДНК та продемонстровані ключові ознаки *EnRE*.

Аналіз виживання був проведений за допомогою статистичного пакета IBM SPSS Statistics 27 методами Каплана-Мейєра та регресії Кокса.

Результати. Точність запропонованої формули для розрахунку ентропії була перевірена для різних відрізків часових рядів і різних типів випадкових розподілів з відомими теоретичними значеннями ентропії. Показано, що у всіх випадках для $N = 500$ відносна похибка при розрахунку точного значення ентропії не перевищує 1 %, при цьому величина кореляції не гірше 0,995.

Код алфавіту початкової послідовності ДНК був перетворений в числовий код основ, з використанням правила оптимізації, щоб отримати тільки одне мінімальне і симетричне числове декодування близько нуля, що дає мінімум для стандартного відхилення *EnRE* і коефіцієнт варіації.

Ентропія *EnRE* була розрахована для хворих на лейкемію після оптимального цілочисельного декодування в двох спостереженнях: 2 групи, розділені медіаною *EnRE* = 1,47, та 2 групи, що належать до 1-го ($EnRE \leq 1.448$) та 4-го квартилів ($EnRE \geq 1.490$). Результат аналізу виживання Каплана-Мейєра та моделювання виживання Кокс-регресій статистично значущі з $p < 0,05$ для груп поділених медіаною і з $p < 0,005$ для груп, що уособлюють 1-й та 4-й квартілі. Небезпека смерті для пацієнта з *EnRE* нижче медіани в 1,556 рази більше, ніж у пацієнта з *EnRE* понад медіаною та небезпека смерті для пацієнта 1-го ентропійного квартиля (найнижчий *EnRE*) в 2,143 рази більше, ніж у пацієнта 4-го ентропійного квартиля (найвищий *EnRE*).

Висновки. Перехід від розширених (медіальних) до менших (квартільних) груп пацієнтів з більшою різницею у *EnRE* підтвердив унікальне значення ентропії послідовностей ДНК для визначення смертності пацієнтів з лейкемією. Це значення статистично доведено підвищенням небезпеки для хворих на лейкемію з меншою ентропією послідовностей ДНК: більша різниця *EnRE* означає збільшення ризику смерті та скорочення тривалості життя після діагнозу в групах пацієнтів з меншою ентропією послідовностей ДНК.

КЛЮЧОВІ СЛОВА: *ентропія, послідовності ДНК, смертність, лейкемія*

ІНФОРМАЦІЯ ПРО АВТОРІВ

Мартиненко Олександр Віталійович, д.фіз-мат.н., професор, професор кафедри гігієни та соціальної медицини Харківського національного університету імені В. Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022, e-mail: Alexander.v.martynenko@karazin.ua, ORCID ID: https://orcid.org/0000-0002-0609-2220.

Пастор Ксав'є Дюран, доктор медицини, професор, керівник відділення медичної інформатики, клініка Університету Барселони, вул. Віллароель 170, Барселона, Іспанія, 08036. e-mail: xpastor@clinic.cat, ORCID: 0000-0001-8267-7151

Фрід Сантьяго Андрес, доцент кафедри фундаментальної клініки, медичний факультет, Університет Барселони, вул. Казанови 143, Барселона, Іспанія, 08036, e-mail:frid@clinic.cat, ORCID: 0000-0001-8400-5770

Гіл Рожас Джессика, дата менеджер, відділення медичної інформатики, клініка Університету Барселони, вул. Віллароель 170, Барселона, Іспанія, 08036, e-mail:jegil@clinic.cat. ORCID: 0000-0002-7690-7288

Малярова Людмила Володимирівна, асистент кафедри гігієни та соціальної медицини Харківського національного університету імені В. Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022, e-mail: l.v.maliarova@karazin.ua, https://orcid.org/0000-0002-7902-7016

Для цитування:

Мартиненко ОВ, Пастор КД, Фрід СА, Гіл ДР, Малярова ЛВ. ЕНТРОПІЯ ПОСЛІДОВНОСТЕЙ ДНК І СМЕРТНІСТЬ ПАЦІЄНТІВ З ЛЕЙКЕМІЄЮ. Вісник Харківського національного університету імені В. Н. Каразіна. Серія «Медицина». 2022;45:12–23. **DOI: 10.26565/2313-6693-2022-45-02**

ВСТУП

Дезоксирибонуклеїнова кислота (ДНК) не є випадковою послідовністю чотирьох комбінацій нуклеотидів (А – аденін, С – цитозин, Г – гуанін, Т – тимін): комплексні огляди літератури [1, 2] переконливо показують довго- та короткодіапазонні кореляції в ДНК, періодичні властивості та кореляційну структуру послідовностей. Методи теорії інформації мають на увазі кількісну оцінку обсягу інформації, що міститься в послідовностях. Інформаційна ентропія Клода Шеннона була одним з перших інформаційних заходів для досліджуваних послідовностей ДНК [3]. В даний час реалізації ентропії Шеннона продовжують залишатися дуже успішними для аналізу вірусної РНК, таких як SARS-COV-2 [4]. Ґрунтовний огляд різних підходів до ентропії «для виявлення формальних звэязків між генетичним різноманіттям та потоком інформації» був наведений в [5] і досконалий огляд [6] демонструє сучасний стан реалізацій теорії інформації для аналізу «експресії генів та транскриптоміки, порівняння послідовностей без вирівнювання, секвенування та виправлення помилок, картування зв'язків між геномами та генами, метаболічних мереж та метаболоміка, а також аналіз послідовності білків, структури та взаємодії».

З іншого боку, зв'язок між ентропією та виживанням пацієнтів широко поширений у деяких галузях медицини та медичних дослідженнях. Надамо деякі приклади:

1. ***Кардіологія***: Використовується апроксімаційна ентропія на основі варіабельності серцевого ритму (ВСР) для прогнозування раптової серцевої смерті, оцінки впливу специфічних фармакологічних засобів на ВСР [7]; застосовується ентропія на основі електрокардіограм Холтера (ЕКГ) і частоти серцевих скорочень (ЧСС) нормальної серцевої динаміки і тих, що мають різний ступінь гострих серцевих патологій [8]; пацієнти після інфаркту міокарда, які перенесли пізній гадоліній покращений магнітний резонанс серця (МР) з похідною ентропією тканин МР-візуалізації. За пацієнтами спостерігалася відповідна імплантована кардіовертерно-дефібриляторна терапія та смертність [9].
2. ***Неврологія***: досліджено зв'язок ентропії серцевого ритму (ЕСР) зі смертністю після внутрішньомозкової кровотечі [10];
3. ***Хірургія (загальна анестезія)***: Моніторинг ентропії передбачає використання електроенцефалографії (ЕЕГ) для оцінки глибини загальної анестезії у хірургічних пацієнтів [11];
4. ***Травма***: Для категоризації травми за допомогою ентропії необхідно розглянути основну ентропію хворобливості осіб, до якої додається ентропія травми, яка потім може призвести до смерті [12]; показана цілочисельна багатомасштабна ентропія (MSE) частота серцевих скорочень (HR), як показник складності, що прогнозує смерть у довго тривалому термені. MSE частоти серцевих скорочень протягом декількох годин після госпіталізації прогнозує смерть, що настає через кілька

днів [13]; У цьому дослідженні вимірювали ентропію Шеннона і Цалліс задля температурних сигналів у когорті критично хворих пацієнтів. Зменшені вейвлети ентропії температурних сигналів Шеннона і Цалліса може доповнюватися послідовною оцінкою органної недостатності в прогнозуванні смертності [14].

Таким чином, випливає, що існує необхідність реалізації переваг методів теорії інформації для дослідження взаємозв'язку між смертністю певної категорії пацієнтів та ентропією їх послідовностей ДНК. Мета даної роботи – надати надійну формулу для точного розрахунку ентропії для коротких послідовностей ДНК і показати, як використовувати існуючий аналіз ентропії для вивчення смертності хворих на лейкемію.

МАТЕРІАЛИ ТА МЕТОДИ

Ми використовували базу даних пацієнтів з лейкемією Барселонського університету (UB) з 117 знеособленими записами, які складаються з наступного: дата діагнозу пацієнта, дата смерті пацієнта, діагнози лейкемії, послідовність ДНК пацієнта. Середній час смерті пацієнта після встановлення діагнозів: 99 ± 77 місяців. Формальними характеристиками послідовностей ДНК в БД UB хворих на лейкемію є: середня кількість ДНК основ N = 496 ± 69; min (N) = 297 основ; max (N) = 745 основ.

Статистично ДНК здебільшого близька до рівномірного розподілу, але має абсолютно різні неоднорідні частотні патерни, наприклад, базові фрікени *мітохондріону людини* (16 569 основ) становлять А – 31 %, С – 31 %, Г – 13 %, Т – 25 % або *екзони глобіну плода людини* (882 основи) становлять А – 24 %, С – 25 %, Г – 28 %, Т – 22 % [15]. Ми показали порівняння реальної послідовності ДНК і змодельованої рівномірним розподілом на рисунках 1.a. і 1.b.:

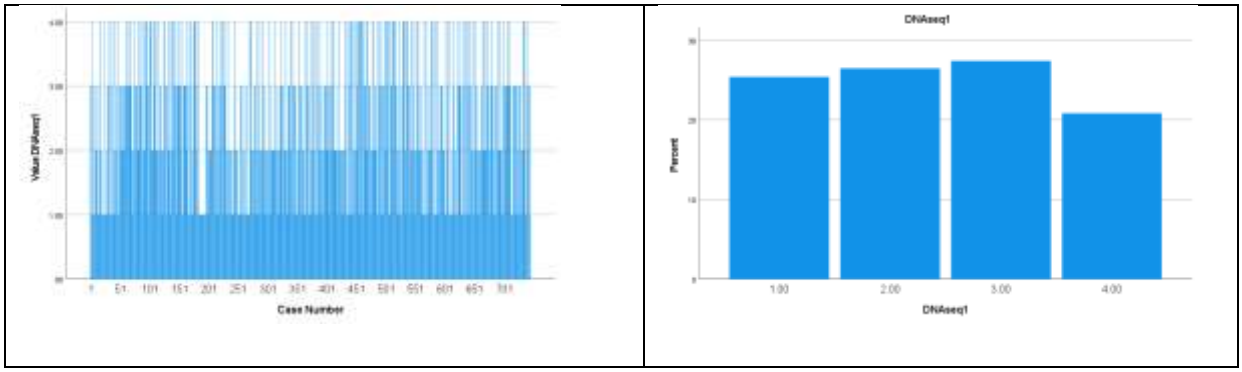

Рис. 1.a. Реальна послідовність ДНК пацієнта, N = 745 основ (UB DB пацієнтів з лейкемією).
Fig. 1a. Real patient DNA sequence, N = 745 bases (UB leukemia patient DB)

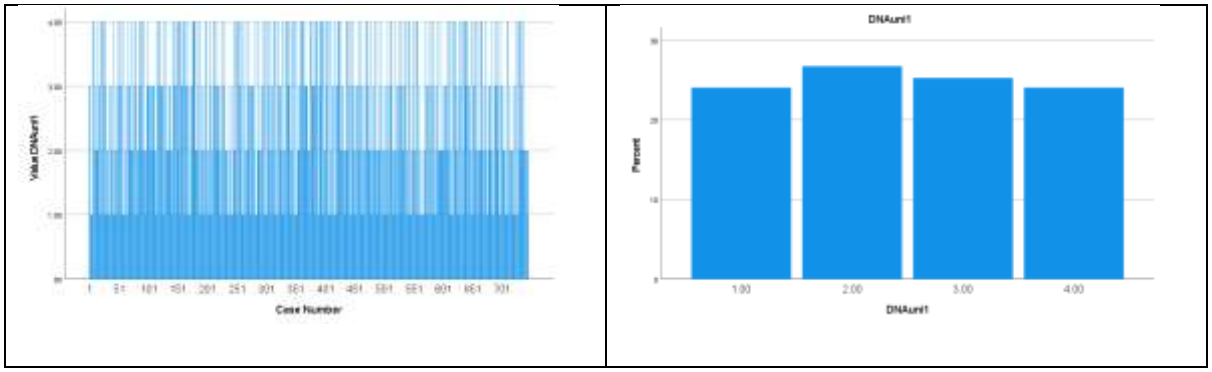

Рис. 1.b. Змодельована послідовність 4 елементів за рівномірним розподілом, N = 745
Fig. 2.b. Simulated 4 elements sequence by Uniform distribution, N = 745

Ми розрахували ентропію обмежених часових рядів за оригінальною формулою, запропонованою Клодом Шенноном в 1948 році [16] і вона тут називається Емпірична ентропія (*EnEmp*) через обмеження часових рядів:

$$EnEmp = -\sum_{i=1}^{N} P(x_i) ln(P(x_i)) \qquad 1$$

Проблемою використання формули (1) на практиці є:

- нечутливість до зміни положень нуклеотидів в послідовності ДНК. Є чутливість тільки до зміни бази;
- низька точність для невеликої кількості точок в ряді (наприклад, коли $N < 1000$);
- повільна сходимість до точного значення зі збільшенням довжини послідовності.

Показана в табл.2 залежність точності обчислення ентропії за формулою (1) від довжини ряду для окремих видів розподілу випадкової величини. Точні значення ентропії для пов'язаних розподілів наведені в Tab.1.

Таблиця 1
Table 1

Різні розподіли ймовірностей і відповідна ентропія [17]

Various probability distributions and correspondent Entropy

Розподіл	**Функція ймовірності**	**Ентропія (*En*, нат)**
Нормальний розподіл	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$En = ln\left(\sqrt{2\pi\sigma^2}\right)$
Рівномірний розподіл	$f(x) = \frac{x}{b-a}$	$En = ln(b-a)$
Експоненціальний розподіл	$f(x) = \lambda\, exp(-\lambda x)$	$En = 1 - ln(\lambda)$
Логарифмічно-нормальний розподіл	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left(-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right)$	$En = \mu + ln\left(\sqrt{2\pi\sigma^2}\right)$
Розподіл Парето	$f(x) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}}$	$En = ln\left(\frac{x_m}{\alpha}\right) + 1 + \frac{1}{\alpha}$

Таблиця 2
Table 2

Залежність від тривалості ряду точності оцінки ентропії та кореляція за змодельованими параметрами розподілу для різних розподілів ймовірностей

Dependence from the length of time series of Entropy estimation accuracy and Correlation along simulated distribution parameters for various probability distributions

Розподіл	Довжина зразка	Емпірична ентропія (*EnImp*)		Робастний оцінювач ентропії (*EnRE*)	
		Точність (відносна похибка, %)	Кореляція	Точність (відносна похибка, %)	Кореляція
Рівномірний розподіл (a = 0; b = 4)	N = 100	6.92	0.978	4.71	0.991
	N = 500	4.11	0.988	0.57	0.997
	N = 1000	3.83	0.999	0.11	0.998
Нормальний розподіл (M = 1000; $\sigma = 100 - 200$)	N = 100	7.74	0.994	1.95	0.995
	N = 500	1.83	0.997	0.35	0.998
	N = 1000	0.91	0.999	0.16	0.999
Експоненціальний розподіл ($\lambda = 0.0001 - 0.0011$)	N = 100	46.24	0.452	0.77	0.993
	N = 500	28.38	0.903	0.25	0.997
	N = 1000	19.31	0.950	0.06	0.999
Лог-нормальний розподіл ($\mu = 7$; $\sigma = 0.002 - 0.012$)	N = 100	3.69	0.980	3.38	0.986
	N = 500	1.17	0.997	0.49	0.997
	N = 1000	0.80	0.999	0.22	0.999
Розподіл Парето (α = 2; s = 1000 – 2000)	N = 100	32.68	0.589	1.01	0.997
	N = 500	17.78	0.867	0.35	0.998
	N = 1000	14.75	0.946	0.12	0.999

Випадок рівномірного розподілу (показаний на рис. 1.a, b) отримує особливу увагу, але інші розподіли також враховуються як необхідні приклади використання числових формул для аналізу обмеженого ряду, оскільки не завжди вдається точно зіставити спостережувану послідовність ДНК з деяким фіксованим випадковим розподілом. Можна визнати неможливість застосування формули (1) до короткого часового ряду $N < 1000$. Тому, здавалося б, необхідна розробка формули для точного вимірювання ентропії для невеликої довжини послідовностей ДНК.

На початку минулого століття італійський професор статистики Коррадо Джині запропонував спосіб вимірювання нерівності між значеннями частотного розподілу (коефіцієнт Джині) [18]:

$$G = \frac{1}{2N^2M}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\left|x_i - x_j\right|\right), \qquad (2)$$

де M – середнє значення x. Коефіцієнт Джині виявився дуже популярним в економіці та соціології, і є спроби застосувати його і до інших областей, включаючи аналіз ВСР [19]. Коефіцієнт Джині є екземпляром узагальненого індексу нерівності [20], а його альтернатива, як міра відхилення від балансу – узагальнений індекс ентропії – виводиться з теорії інформації як міра надмірності в даних [21]. Відомі обмеження при використанні коефіцієнта Джині для аналізу даних: залежність від адитивної зміни середнього; малий відбір істотно зменшує величину коефіцієнта і т. і.

Тому після аналізу відомих визначень мір відхилення від рівноваги і ступеня порядку була запропонована узагальнена форма оцінювача ентропії (*EnRE*) для часових рядів в [22] і наступним запропонована для послідовностей ДНК:

$$EnRE = ln\left(\frac{A}{N^{l/2}}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\left(|B_i - B_j||B_j - MD|\right)^{1/k}}{(D_{ij})^{m/2}}\right)\right), \qquad (3)$$

де MD — медіана послідовності для чисельно закодованих основ B; D_{ij}-відстань між B_i і B_j; A, l, m, k – оціночні коефіцієнти. Умови для пошуку коефіцієнтів A, l, m, k наступні:

1) точне наближення для відомих розподілів випадкової величини;

2) незалежність *EnRE* від N для початкових часових рядів і для серій після сортування;

3) незалежність *EnRE* від адитивної зміни середнього значення.

Після чисельних досліджень, остаточні результати яких представлені в табл. 2, були знайдені наступні значення коефіцієнтів: $l = 3$, $m = 1$, $k = 2$. Виділимо деякі ключові ознаки запропонованої узагальненої форми *EnRE* і коефіцієнтів:

1) форма запису (3) і знайдені коефіцієнти l, m, k забезпечують незалежність від адитивної зміни середніх рядів і від величини виділення N для базових рядів і для рядів після сортування;

2) значення *EnRE* чутливе до структурних змін рядів, таких як, наприклад, сортування, яке збільшує ступінь порядку послідовно, зменшуючи *EnRE*;

3) значення *EnRE* чутливе до зміни положення нуклеотидів у послідовності ДНК;

4) коефіцієнт переналагодження A самостійно може знадобитися для знаходження кращого значення *EnRE* в іншому діапазоні зміни параметрів різних випадкових розподілів, що завжди можна зробити за допомогою методу найменших квадратів.

Аналіз виживання був проведений за допомогою статистичного пакета IBM SPSS 27.

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

Точність. Перш за все, перевіримо точність запропонованої формули (3) для розрахунку ентропії: Табл. 2 надає значення *EnRE* для різних відрізків часових рядів і різних типів випадкових розподілів, а для кожного з цих результатів наводяться значення похибок при обчисленні ентропії порівняно з точними значеннями при зміні параметрів розподілу. Зауважимо, що у всіх випадках для $N = 500$ відносна похибка при розрахунку точного значення ентропії не перевищує 1 %, при цьому величина кореляції не гірше 0,995; при рівномірному і нормальному розподілі відносна похибка для довжини часових рядів N = 500÷1000 менше 0,6 %, а кореляція становить близько 0,998.

Оптимальне кодування послідовності ДНК. Код алфавіту початкової послідовності ДНК повинен бути перетворений в числовий код основ, але така перестановка є довільною. Тому ми використали принцип максимальної ентропії, щоб уникнути такої довільності. Одночасно оптимальне чисельне перекодування має давати мінімальне значення для стандартного відхилення та коефіцієнт варіації для розрахункової ентропії послідовностей ДНК, оскільки у нас однорідна група пацієнтів. Крім того, це правило оптимізації зменшило самовплив дисперсії числового декодування. У табл. 3 ми наведемо різні числові розшифровки послідовностей ДНК та їх значення ентропії, стандартні відхилення та коефіцієнти варіації.

Таблиця 3
Table 3

Числове кодування послідовностей ДНК і відповідне *EnRE*, стандартне відхилення і коефіцієнт варіації *EnRE*

Numerical decoding of DNA sequences and correspondent *EnRE*, standard deviation and coefficient of variation of *EnRE*

Цілочисельний код ДНК	**Середня ентропія *EnRE***	**Стандартне відхилення *EnRE***	**Коефіцієнт варіації (CV)**
A=1,C=2,G=3,T=4 or A=4,C=3,G=2,T=1	1.205	0.030	0.025
A=2,C=1,G=3,T=4 or A=3,C=4,G=1,T=2	1.254	0.036	0.029
A=3,C=4,G=2,T=1	1.241	0.034	0.027
A=1,C=4,G=3,T=2	1.235	0.043	0.035
A=1,C=3,G=4,T=2	1.221	0.040	0.033
A=1,C=3,G=2,T=4	1.211	0.033	0.027
A=1,C=2,G=4,T=3	1.223	0.039	0.032
A=-2, C=-1, G=1, T=2 (дзеркальна симетрія за модулем)	1.430	0.023	0.016
A=-1, C=-2, G=1, T=2 (трансляційна симетрії за модулем)	**1.470**	**0.022**	**0.015**

Можна стверджувати, що за властивостями *EnRE* будь-яка симетрична зміна цілочисельного декодування дає однакове значення *EnRE* (див. перші два рядки табл. 3); тільки одне мінімальне і симетричне числове декодування близько нуля дає максимум ентропії і мінімум для стандартного відхилення *EnRE* і коефіцієнт варіації (виділений жирним шрифтом в табл. 3). Таким чином, числове декодування довільності було прибрано тільки однією можливою цілочисельною комбінацією.

Смертність пацієнтів з лейкемією. Ентропія *EnRE* була розрахована для всіх хворих на лейкемію після оптимального цілочисельного декодування, створеного групою симетрії трансляції (A = -1, C = -2, G = 1, T = 2).

А. Медіанні групи. Всі пацієнти були розділені медіаною *EnRE* = 1,47 на 2 групи:

1. Група ‘1’, 58 пацієнтів, *EnRE* нижче медіани;
2. Група ‘2’, 59 пацієнтів, *EnRE* вище медіани.

Результат аналізу виживання Каплана-Мейєра наведено на рис. 2. Всі загальні порівняння показують статистичну значимість: p = 0,015 для Log Rank (Мантель-Кокс); p = 0,002 для Бреслоу (узагальнений Вілкоксон); p = 0,003 для Tarone-Ware. Середні та медіани для хворих на лейкемію на час виживання наведені в таблиці 4.

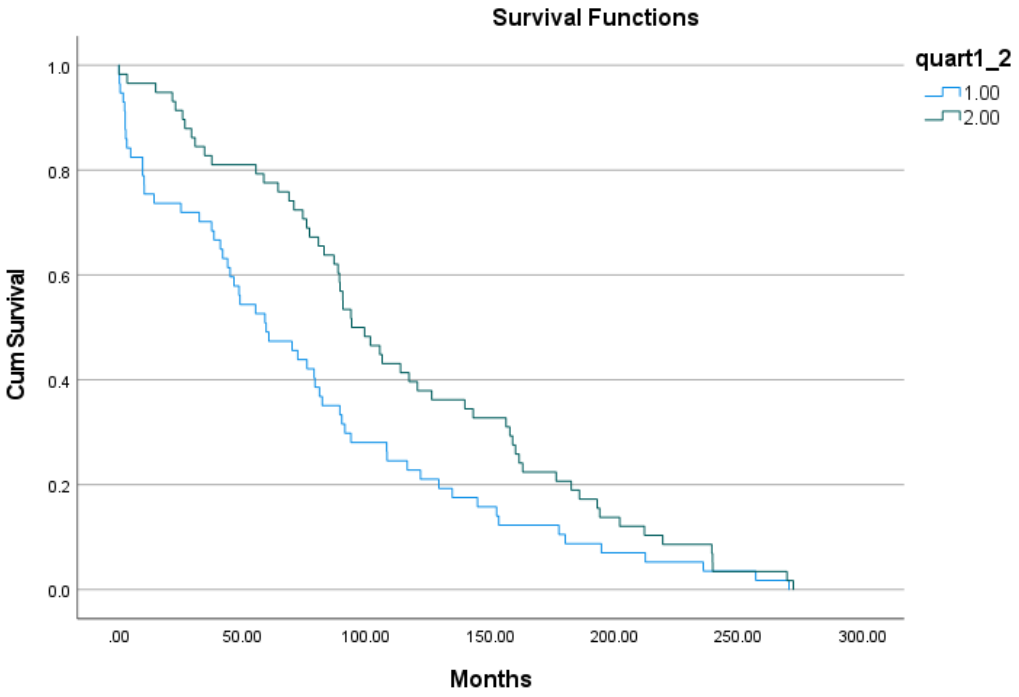

Рис. 2. Криві виживання Каплана-Мейєра для медіанних груп

Fig. 2. Kaplan-Meier survival plot for median groups

Таблиця 4
Table 4

Середні та медіани для часу виживання (медіанні групи)
Means and Medians for Survival Time (median groups)

	Середні[a]				Медіани			
			95% Confidence Interval				95% Confidence Interval	
quart1_2	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
1.00	76.409	9.208	58.361	94.456	59.367	12.743	34.391	84.343
2.00	114.301	9.257	96.157	132.445	93.867	9.477	75.291	112.442
Overall	95.520	6.738	82.313	108.726	82.767	5.802	71.395	94.138

a. Оцінка обмежена найбільшим часом виживання, якщо вона піддається цензурі

Результат моделювання виживання Кокс-регресій наведено на рис. 3. Всі загальні порівняння показують статистичну значимість: p = 0,015 для Омнібусного тесту модельних коефіцієнтів; p = 0,016 для змінних у рівнянні з ‘-2 Log Likelihood’=862.2.

Значення *Exp(B)* для *модельної змінної* показує, що небезпека смерті для пацієнта з *EnRE* нижче медіани в 1,556 рази більше, ніж у пацієнта з *EnRE* понад медіаною.

Б. 1-й і 4-й квартілі. Сформовано 2 групи хворих відповідно до їх приналежності до1-го та 4-го квартилів:

3. Група ‘1’, 29 пацієнтів, $EnRE \leq 1.448$, тобто нижче 1-го квартиля;

4. Група ‘4’, 29 пацієнтів, $EnRE \geq 1.490$, тобто вище 4-го квартиля.

Результат аналізу виживання Каплана-Майєра наведено на рис. 4. Всі загальні порівняння показують статистичну значимість: p = 0,005 для Log Rank (Мантель-Кокс); p = 0,003 для Бреслоу (узагальнений Вілкоксон); p = 0,003 для Tarone-Ware. Середні та медіани для виживання хворих на лейкемію наведені в таблиці 5.

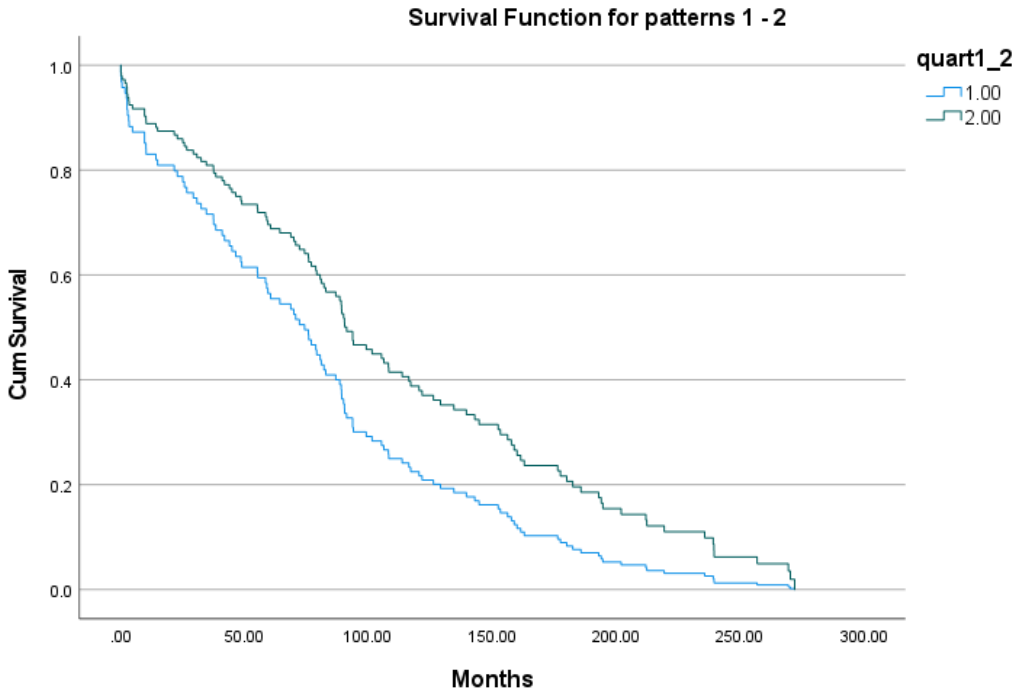

Рис. 3. Кокс-регресії виживаності для медіанних груп.
Fig. 3. Cox Regressions survival plot for median groups

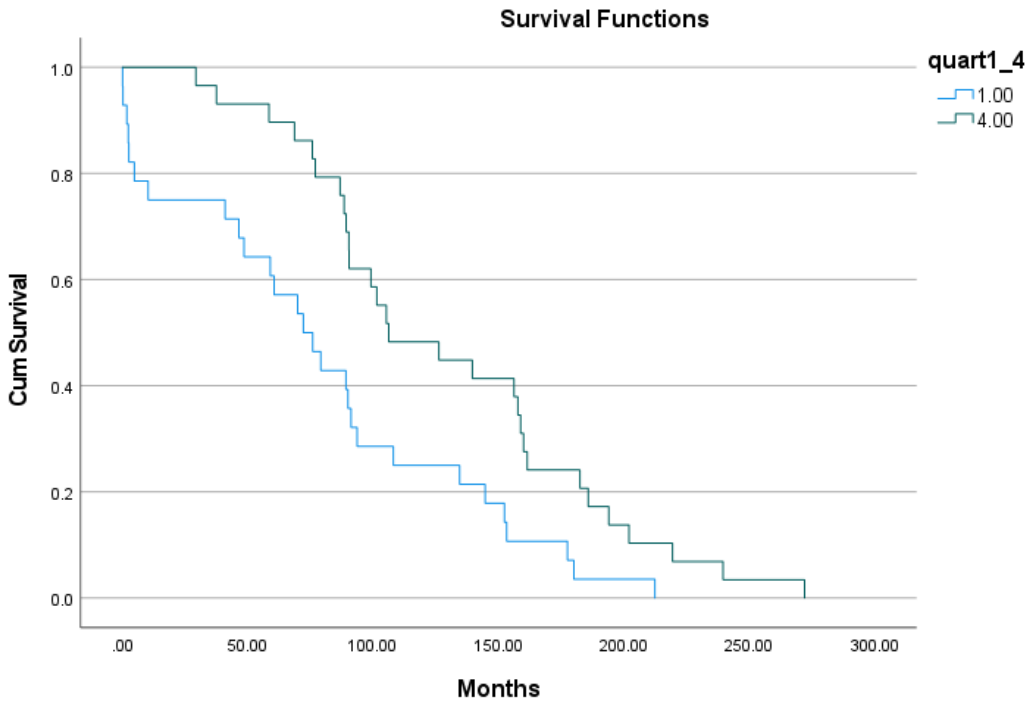

Рис. 4. Криві виживання Каплана-Майєра (1-й і4-й квартилі)
Fig. 4. Kaplan-Meier survival plot (1st and 4th quarterlies)

Таблиця 5
Table 5

Середні та медіани для виживання (1-й та4-й квартали)
Means and Medians for Survival Time (1st and 4th quarterlies)

	Середні[a]				Медіана			
			95% Confidence Interval				95% Confidence Interval	
quart1_4	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
1.00	78.600	11.619	55.826	101.374	72.133	12.325	47.977	96.290
4.00	129.603	11.329	107.398	151.809	106.233	22.156	62.808	149.659
Overall	104.549	8.731	87.436	121.663	90.367	5.338	79.904	100.829

a. Оцінка обмежена найбільшим часом виживання, якщо вона піддається цензурі

Результат моделювання виживання Кокс-регресій наведено на рис. 5. Всі загальні порівняння показують статистичну значимість: p = 0,005 для Омнібусного тесту модельних коефіцієнтів; p = 0,005 для змінних у рівнянні з '-2 Log Likelihood'=345.4.

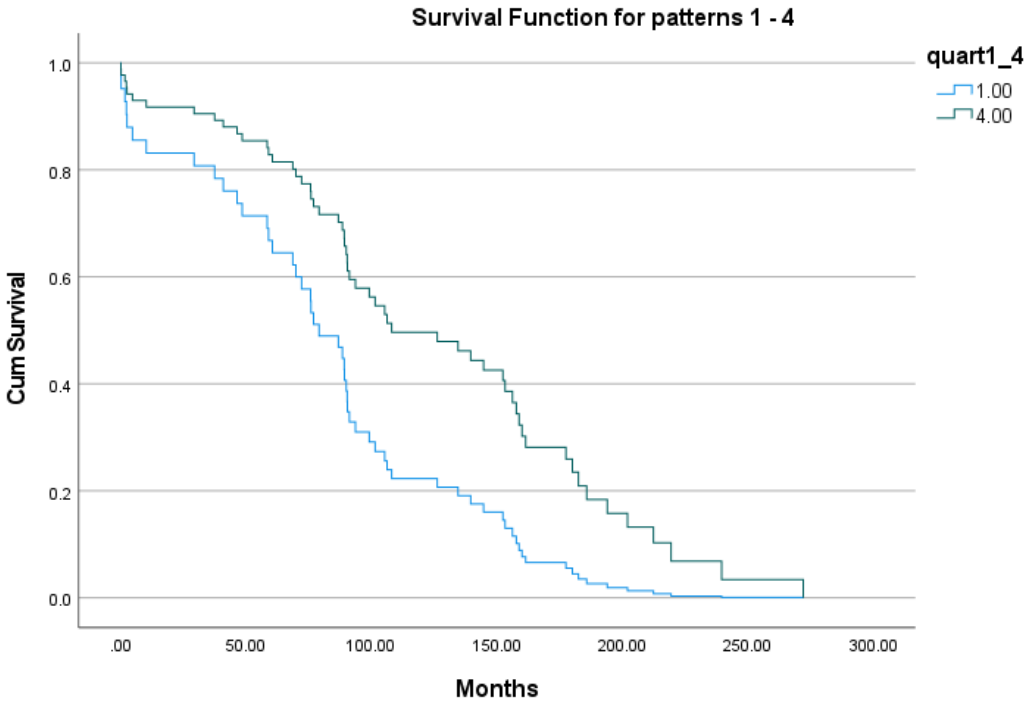

Рис. 5. Кокс-регресії криві виживання (1-й і 4-й квартилі)
Fig. 5. Cox Regressions survival plot (1st and 4th quarterlies)

Значення *Exp(B)* для *модельної змінної* показує, що небезпека смерті для пацієнта 1-го ентропійного квартиля (найнижчий *EnRE*) дорівнює 2,143 рази більше, ніж у пацієнта 4-го ентропійного квартиля (найвищий *EnRE*).

ВИСНОВКИ

Узагальнена форма оцінювача ентропії (3), яка була ефективно використана для обчислення значень ентропії для різноманітних випадкових розподілів (табл. 1–таб. 2), запропонована в даній роботі для послідовностей ДНК невеликої довжини (N < 1000).

Параметри в узагальненому вигляді для робастного оцінювача ентропії (3) були виведені з наступних критеріїв:

1. точне наближення для деяких відомих функцій розподілу ймовірностей;

2. незалежність *EnRE* від *N* для початкових часових рядів і для серій після сортування;

3. незалежність *EnRE* від адитивної зміни середнього значення.

Важливими характеристиками знайденої узагальненої форми *EnRE* і коефіцієнтів є:

1. форма запису (3) і знайдені коефіцієнти *l, m, k* забезпечують незалежність від адитивної зміни середніх рядів і від величини виділення *N* для базових рядів і для рядів після сортування;

2. значення *EnRE* чутливе до структурних змін рядів, таких як, наприклад, сортування, яке збільшує ступінь порядку послідовно, зменшуючи *EnRE*;

3. значення *EnRE* чутливе до зміни положення нуклеотидів у послідовності ДНК;

4. коефіцієнт переналагодження *A* самостійно може знадобитися для знаходження кращого значення *EnRE* в іншому діапазоні зміни параметрів різних випадкових розподілів, що завжди можна зробити за допомогою методу найменших квадратів.

Використовуючи запропоновану узагальнену форму робастного оцінювача ентропії (3) для бази даних пацієнтів з лейкемією UB, продемонстровано використання *EnRE* з короткими послідовностями ДНК для аналізу смертності пацієнта:

А. Групи, розділені за медіаною. Обидва аналізи – аналіз виживання Каплана-Мейєра та моделювання виживання Кокс-регресій, показали статистично значущі результати для $p < 0,05$. Значення *Exp(B)* для *змінної моделі* регресій Кокса показує, що небезпека смерті для пацієнтів з *EnRE* нижче медіани в 1,556 рази вище ніж для пацієнтів з *EnRE* вище за медіану. Середній час після діагнозу до смерті у 1,496 рази більше для пацієнтів з *EnRE* понад медіану порівняно з пацієнтами з *EnRE* нижче медіани.

Б. Групи хворих формуються з 1-го і 4-го квартилів. Обидва аналізи – аналіз виживання Каплана-Мейєра та моделювання виживання Кокс-регресій, показали статистично значущі результати для $p < 0,005$. Значення *Exp(B)* для *модельної змінної* показує, що небезпека смерті для пацієнтів 1-го квартиля ентропії (найнижчий *EnRE*) в 2,143 рази більше, ніж у пацієнта 4-го квартиля ентропії (найвищий *EnRE*). Середній час після діагнозу до смерті у 1,649 рази більше для пацієнтів 4-го квартиля ентропії порівняно з пацієнтами 1-го квартиля ентропії.

Таким чином, перехід від розширених до менших груп пацієнтів з більшою різницею у *EnRE* підтвердив унікальне значення ентропії послідовностей ДНК для визначення смертності пацієнтів з лейкемією. Це значення статистично доведено підвищенням небезпеки для хворих на лейкемію з меншою ентропією послідовностей ДНК.

Майбутнє продовження сучасних досліджень полягає у включенні більш різних груп пацієнтів для дослідження виживання пацієнта у зв'язку з послідовністю ентропійної ДНК, а також із залученням інших методів фрактального аналізу послідовностей ДНК, таких як фрактальна розмірність або оборотність послідовностей.

СПИСОК ЛІТЕРАТУРИ

1. Li WT. The study of correlation structures of DNA sequences: a critical review. Comput. Chem. 1997; 21 (4): 257–271. DOI: 10.1016/s0097-8485(97)00022-3
2. Damasevicius R. Complexity estimation of genetic sequences using information-theoretic and frequency analysis methods. Informatica. 2010; 21 (1): 13–30. DOI: 10.15388/Informatica.2010.270
3. Rowe GW, Trainor LEH. On the informational content of viral DNA. J. Theoretical Biology. 1983; 101: 151–170. DOI: 10.1016/0022-5193(83)90332-6
4. Vopson MM, Robson SC. A new method to study genome mutations using the information entropy. Physica A. 2012;1-9. DOI: 10.1016/j.physa.2021.126383
5. Sherwin WB. Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography. Entropy. 2010;12:1765-1798. DOI:10.3390/e12071765

6. Chanda P, Costa E, Hu J, Sukumar S, Van Hemert J, Walia R. Information Theory in Computational Biology: Where We Stand Today. Entropy. 2020;22:627-637. DOI: 10.3390/e22060627
7. Villareal RP, Liu BC, Massumi A. Heart rate variability and cardiovascular mortality. Curr Atheroscler Rep. 2002; 4: 120–127. DOI: 10.1007/s11883-002-0035-18
8. Rodríguez J, Correa C, Ramírez L. Heart dynamics diagnosis based on entropy proportions: Application to 550 dynamics. Revista Mexicana de Cardiología. 2017; 28 (1): 10–20.
9. Androulakis AFA, Zeppenfeld K, Paiman EHM, Piers SRD, Wijnmaalen AP, Siebelink HJ, Sramko M, Lamb HJ, van der Geest RJ, de Riva M, Tao Q. Entropy as a Novel Measure of Myocardial Tissue Heterogeneity for Prediction of Ventricular Arrhythmias and Mortality in Post-Infarct Patients. JACC Clin Electrophysiol. 2019 Apr;5 (4): 480–489. DOI: 10.1016/j.jacep.2018.12.005. Epub 2019 Feb 27. PMID: 31000102.
10. Sykora M, Szabo J, Siarnik P, Turcani P, Krebs S, Lang W, Czosnyka M, Smielewski P. Heart rate entropy is associated with mortality after intracereberal hemorrhage. Journal of the Neurological Sciences. 2020: 418: 117033, ISSN 0022-510X, 1–5: DOI: 10.1016/j.jns.2020.117033
11. Matsuda E. Entropy Monitoring in Patients Undergoing General Anesthesia. Am J Nurs. 2017 Mar;117(3):62. DOI: 10.1097/01.NAJ.0000513290.22001.8d
12. Neal-Sturgess C. The Entropy of Morbidity Trauma and Mortality. Arxiv Cornell University. Med. Physics. 2010; 1–20. DOI: 10.48550/arxiv.1008.3695
13. Norris PR, Anderson SM, Jenkins JM, Williams AE, Morris JAJr. Heart rate multiscale entropy at three hours predicts hospital mortality in 3,154 trauma patients. Shock. 2008 Jul; 30 (1): 17–22. DOI: 10.1097/SHK.0b013e318164e4d0
14. Papaioannou VE, Chouvarda IG, Maglaveras NK, Baltopoulos GI, Pneumatikos IA. Temperature multiscale entropy analysis: a promising marker for early prediction of mortality in septic patients. Physiol Meas. 2013 Nov;34(11):1449-66. DOI: 10.1088/0967-3334/34/11/1449
15. Weir BS. Statistical analysis of molecular genetic data. IMA J. of Math. Applied in Medicine and Biology. 1985; 2:1–39.
16. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27 (3): 379–423. DOI:10.1002/j.1538-7305.1948.tb01338.x
17. Lazo A, Rathie P. On the entropy of continuous probability distributions. IEEE Transactions on Information Theory. 1978;24(1). DOI:10.1109/TIT.1978.1055832
18. Gini C, Ottaviani G. Università di Roma. Memorie Di Metodologia Statistica. Roma: E.V. Veschi; 1955.
19. Sánchez-Hechavarría M.E. and etc. Introduction of Application of Gini Coefficient to Heart Rate Variability Spectrum for Mental Stress Evaluation. Arq Bras Cardiol. 2019; [online].ahead print, PP.0-0. DOI: 10.5935/abc.20190185
20. Firebaugh G. Empirics of World Income Inequality. American Journal of Sociology. 1999; 104 (6): 597–1630. DOI:10.1086/210218
21. Shorrocks AF. The Class of Additively Decomposable Inequality Measures. Econometrica. 1980; 48 (3): 613–625. DOI: 10.2307/1913126
22. Martynenko A, Raimondi G, Budreiko N. Robust Entropy Estimator for Heart Rate Variability. Klin. Inform. Telemed. 2019; 14 (15): 67–73. DOI: 10.31071/kit2019.15.06

ENTROPY OF DNA SEQUENCES AND LEUKEMIA PATIENTS MORTALITY

***Martynenko O. V.*[A,C,D,E,F], *Pastor X. D.*[A,B,E,F], *Frid S. A.*[B,F], *Gil J. R.*[B,F], *Maliarova L. V.*[E,F]**

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation; D – writing the article; E – critical revision of the article; F – final approval of the article.

Introduction. Deoxyribonucleic acid (DNA) is not a random sequence of four nucleotides combinations: comprehensive reviews [1, 2] persuasively shows long- and short-range correlations in DNA, periodic properties and correlations structure of sequences. Information theory methods, like Entropy, imply quantifying the amount of information contained in sequences. the relationship between entropy and patient survival is widespread in some branches of medicine and medical researches: cardiology, neurology, surgery, trauma. Therefore, it appears there is a necessity for implementing advantages of information theory methods for exploration of relationship between mortality of some category of patients and entropy of their DNA sequences.

Aim of the research. The goal of this paper is to provide a reliable formula for calculating entropy accurately for short DNA sequences and to show how to use existing entropy analysis to examine the mortality of leukemia patients.

Materials and Methods. We used University of Barcelona (UB) leukemia patient's data base (DB) with 117 anonymized records that consists: Date of patient's diagnosis, Date of patient's death, Leukemia diagnoses, Patient's DNA sequence. Average time for patient death after diagnoses: 99 ± 77 months. The formal characteristics of DNA sequences in UB leukemia patient's DB are: average number of bases N = 496 ± 69; min (N) = 297 bases; max(N) = 745 bases.

The generalized form of the Robust Entropy Estimator (*EnRE*) for short DNA sequences was proposed and key *EnRE* futures was showed.

The Survival Analysis has been done using statistical package IBM SPSS 27 by Kaplan-Meier survival analysis and Cox Regressions survival modelling.

Results. The accuracy of the proposed *EnRE* for calculating entropy was proved for various lengths of time series and various types of random distributions. It was shown, that in all cases for $N = 500$, relative error in calculating the precise value of entropy does not exceed 1 %, while the magnitude of correlation is no worse than 0.995.

In order to yield the minimum *EnRE* standard deviation and coefficient of variation, an initial DNA sequence's alphabet code was converted into an integer code of bases using an optimization rule for only one minimal numerical decoding around zero.

Entropy *EnRE* were calculated for leukemia patients for two samples: 2 groups divided by median *EnRE* = 1.47 and 2 groups of patients were formed according to their belonging to 1st (*EnRE* ≤ 1.448) and 4th (*EnRE* ≥ 1.490) quartiles. The result of Kaplan-Meier survival analysis and Cox Regressions survival modelling are statistically significant: $p < 0,05$ for median groups and $p < 0,005$ for patient's groups formed of 1st and 4th quartiles. The death hazard for a patient with *EnRE* below median is 1.556 times that of a patient with *EnRE* over median and that the death hazard for a patient of 1st entropy quartile (lowest *EnRE*) is 2.143 times that of a patient of 4th entropy quartile (highest *EnRE*).

Conclusions. The transition from widen (median) to smaller (quartile) patients' groups with more *EnRE* differentiation confirmed the unique significance of the entropy of DNA sequences for leukemia patient's mortality. This significance is proved statistically by increasing hazard and decreasing of average time of death after diagnoses for leukemia patients with lower entropy of DNA sequences.

KEY WORDS: *entropy, DNK sequence, patients surviving, leukemia*

INFORMATION ABOUT AUTHORS

Martynenko Oleksandr Vitalyevich, D.Sc., Ph.D., Full Professor, Department of Hygiene and Social Medicine, School of Medicine, V. N. Karazin Kharkiv National University, 6, Svobody sq., Kharkiv, Ukraine, 61022. e-mail: Alexander.v.martynenko@karazin.ua, ORCID ID: https://orcid.org/0000-0002-0609-2220.

Pastor Xavier Duran, Doctor of Medicine and Surgery, University Professor, Department of Surgery and Medical-Surgical, University of Barcelona, Chief of Medical Informatics Unit, Hospital Clinic, 170, Villarroel st. Barcelona, Spain, 08036. e-mail: xpastor@clinic.cat, ORCID: 0000-0001-8267-7151

Frid Santiago Andres, MD, M.Sc., Medical Associated Professor, Department of Clinical Foundations, School of Medicine, Universitat de Barcelona, 143, Casanova st., Barcelona, Spain, 08036. Chief of Area of Projects and Developments, Medical Informatics Unit, Hospital Clínic de Barcelona, 170, Villarroel st., Barcelona, Spain, 08036. e-mail:frid@clinic.cat, ORCID: 0000-0001-8400-5770

Gil Rojas Jessyca, MSc, Data Manager, Medical Informatics Unit, Hospital Clínic de Barcelona, 170, Villarroel st., Barcelona, Spain, 08036. e-mail:jegil@clinic.cat. ORCID: 0000-0002-7690-7288

Maliarova Liudmila Volodimirivna, Assistant, Department of hygiene and social medicine, School of Medicine, V. N. Karazin Kharkiv National University, 6, Svobody sq., Kharkiv, Ukraine, 61022, e-mail: l.v.maliarova@karazin.ua, https://orcid.org/0000-0002-7902-7016

For citation:

Martynenko OV, Pastor XD, Frid SA, Gil JR, Maliarova LV. ENTROPY OF DNA SEQUENCES AND LEUKEMIA PATIENTS MORTALITY. The Journal of V. N. Karazin Kharkiv National University. Series «Medicine». 2022; 45:12–23. **DOI: 10.26565/2313-6693-2022-45-02.**

Conflicts of interest: *author has no conflict of interest to declare.*
Конфлікт інтересів: *відсутній.*

Отримано: 12.10.2022
Прийнято до друку: 20.11.2022
Received: 10.12.2022
Accepted: 11.20.2022