Лекція 15. Розпізнавання раку молочної залози на підставі фрактального аналізу інтерфазних ядер букального епітелію

Вступ

Задача ранньої діагностики раку молочної залози на фоні зростання захворюваності і смертності від цього виду раку і в Україні, і в усьому світі, безперечно є однією з найважливіших. Для того щоб діагностувати онкологічні захворювання на ранній стадії потрібен автоматизований метод скринінгу — масового обстеження великих груп населення з метою виявлення раку на ранній, безсимптомній стадії. Ефективний тест скринінгу повинен мати високу чутливість і специфічність, тобто чітко відокремлювати практично здорових людей від людей, що належать до групи ризику.

Стандартна діагностика раку молочної залози — клінічне обстеження, маммографія, аспіраційна біопсія (Breast Triple Assessment), — гарантує високу точність, але досягти її можна лише завдяки досить небезпечним для пацієнта інвазивним процедурам, які передбачають радіоактивне опромінення і травмування пухлини, до того ж за умови, що злоякісний процес досяг стадії, на якій пухлину можна помітити на рентгенівському знімку і потрапити у неї тонкою голкою. Застосування таких методів для скринінгу є недостатньо ефективним. Отже, другою важливою властивістю ефективного методу скринінгу має бути неінвазивність і нешкідливість, тобто відсутність травмування пухлини та інших небезпечних факторів.

Одним із перспективних напрямків пошуку ефективного методу скринінгу є виявлення пухлинно-асоційованих змін у клітинах, які віддалені від пухлини і легко доступні для дослідження. Ці методи розподіляються на дві групи: методи першої групи націлені на дослідження клітин, що перебувають у безпосередній близькості до пухлини, але не є пухлинними [1],[2], друга група складається із методів, що вивчають пухлинно-асоційовані зміни у клітинах, віддалених від пухлини, зокрема, в букальному епітелії (слизовій оболонці щоки) [3],[4].

Нещодавні роботи, проведені в США [5], прямо продемонстрували виникнення пухлинно-асоційованих змін, які проявляються у пошкодженні ДНК розташованих далеко від пухлини клітин в мишах, хворих на M05076 саркому, B16 меланому і COLON26 карциному. Найбільш імовірною причиною пухлинно-асоційованих змін у віддалених від пухлини тканинах вважається реакція імунної системи на наявність пухлини.

З іншого боку, у 2009 році широкого визнання набула робота [6], в якій колективу авторів вдалося довести, що ДНК в ядрі клітини упакована як фрактальна глобула, тобто скручена за тривимірною кривою Пеано.

Таким чином, виникла ідея об'єднати два фактори — наявність пухлинно-асоційованих змін у букальному епітелії і фрактальну природу упаковки ДНК в хроматині — і на їх підставі розробити новий класифікатор, який би відділяв здорових людей від хворих на рак молочної залози або фіброаденоматоз за допомогою обробки цифрової фотографії ядра клітини.

1. Метод і матеріали

Предметом дослідження були три групи людей: контрольна група (29 людей), група хворих на рак молочної залози (68 пацієнтів), група хворих на фіброаденоматоз (33 пацієнта), діагноз яких підтверджений гістологічним шляхом.

Матеріалом для дослідження були зіскоби епітеліоцитів слизової оболонки ротової порожнини з середньої глибини шипуватого шару, отримані після сушки при кімнатній температурі, фіксації у суміші Нікіфорова та гістохімічної реакції Фельгена з холодним гідролізом в 5 н HCl протягом 15 хвилин при температурі t = 21-22 °C.

Однією із головних вимог, які потрібно задовольнити при фіксації параметрів цифрової фотографії ядра — інваріантність відносно повороту сканограми, оскільки орієнтація ядра на предметному склі може бути абсолютно випадковою. Для того щоб отримати інваріантність відносно повороту ми пропонуємо застосувати фрактальні криві [7], тобто зчитувати значення RGB кольорів пікселів сканограми не построчково, а вздовж кривих Гільберта і Серпинського відповідно [8]. Таким чином, ми позбавляємося від залежності від орієнтації ядра на фотографії і можемо розглядати наше зображення як вектор, а не як матрицю.

Цифрові фотографії, піддані аналізу, мали розмір 128 на 128 пікселів, що дозволяє зчитувати всі пікселі зображення, проходячи за кривою Гільберта 7-го порядку і кривою Серпинського. Таким чином було отримано 3 масиви чисел для 3-ох компонент кольору (Red Green Blue) для кожної кривої.

Втім ці залежності можна трактувати, як часові ряди, що виникають під час блукання частинки вздовж кривої, набуваючи випадкового заряду, що дорівнює значенню відповідного компонента кольору — червоного, синього і зеленого. Тепер до цих функцій можна використовувати математичний апарат, розроблений для часових рядів.

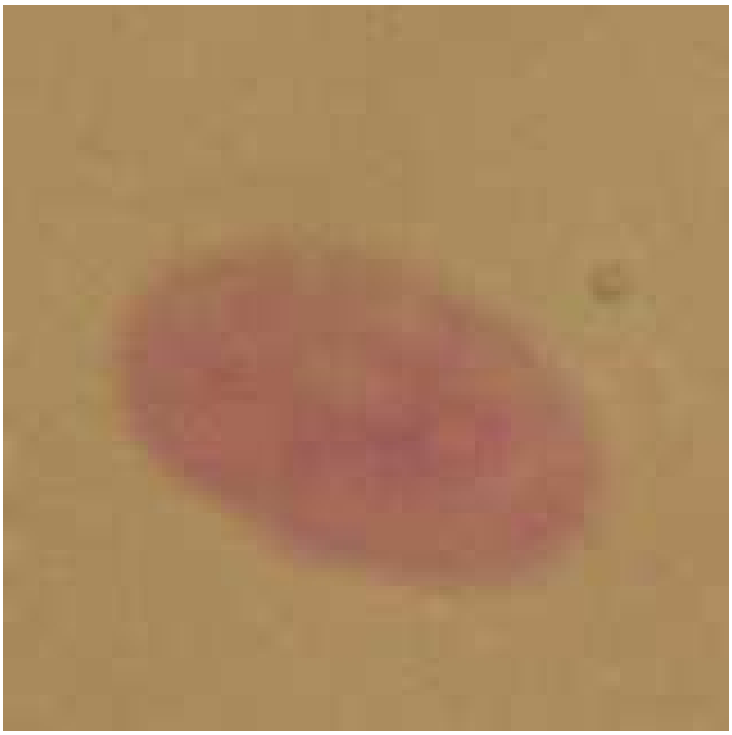

Рис. 1. Фотографія ядра клітини після забарвлення за Фьольгеном

З огляду на гіпотезу про фрактальну природу розподілу хроматину, для аналізу часових рядів було обрано коефіцієнт Хьорста, який пов'язаний із фрактальною розмірністю D формулою $H = 2 - D$. Показник Хьорста обчислюється за наступним алгоритмом [9].

1. Обчислюється відхилення значень часового ряду від середнього значення протягом певного періоду

$$\delta_{m,N} = \sum_{i=1}^{m} (x_i - \bar{x}_N),$$

де N — довжина періода, що змінюється від 2 до довжини часового ряду, m — верхня межа сумування, що змінюється від 1 до $N-1$, x_i — елемент часового ряду, $\bar{x}_N$ — середнє значення часового ряду на протязі поточного періоду. Отримаємо $N-1$ значень $\delta_{2,N}, ..., \delta_{N-1,N}$.

2. Обчислюємо розмах відхилення часового ряду

$$R = \max_{m=2,...,N} \delta_{m,N} - \min_{m=2,...,N} \delta_{m,N}.$$

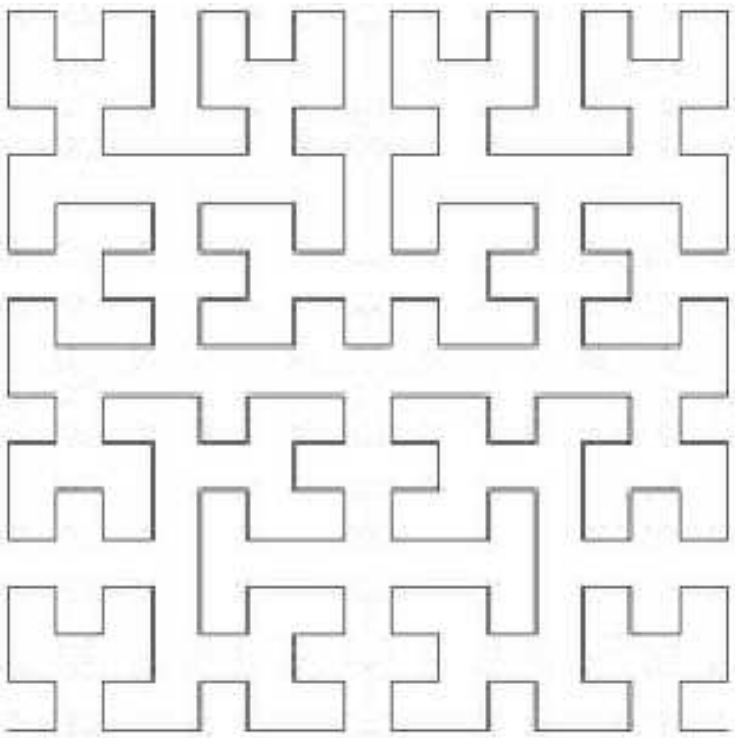

Рис. 2. Крива Гільберта 3-го порядка [8]

3. Нормуємо розмах, обчислюючи

$$Q = \frac{R}{s},$$

де s — стандартне відхилення часового ряду.
4. Обчислюємо $\lg Q$ і $\lg N$ і будуємо лінійну апроксимацію графіку залежності $\lg Q$ від $\lg N$
5. Обчислюємо показник Хьорста, який є тангенсом кута нахилу прямої, що апроксимує залежність $\lg Q$ від $\lg N$.

Методи засновані на цьому показнику мають мінімальні припущення щодо системи, що вивчається, а також є надзвичайно стійкими. Показник Хьорста характеризує хаотичність розподілу елементів часового ряду:

1. Якщо $0 < H < 0,5$, то ряд є ергодичним, тобто якщо система демонструє ріст в попередній проміжок часу, то скоріш за все в наступний момент часу почнеться спад, і навпаки.
2. Якщо $H = 0,5$, то ряд є хаотичним, тобто значення ряду не мають впливу на подальші значення.

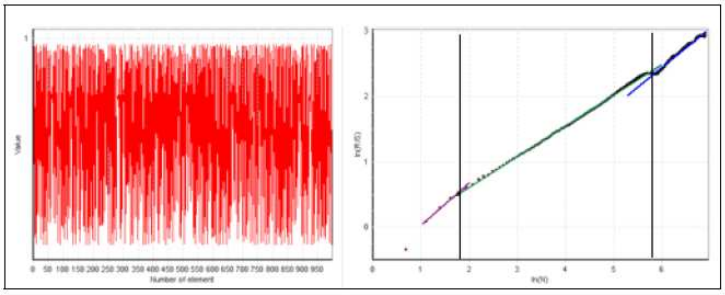

Рис. 3. Приклад логістичного часового ряду і графіку коефіцієнта Хьорста [9]

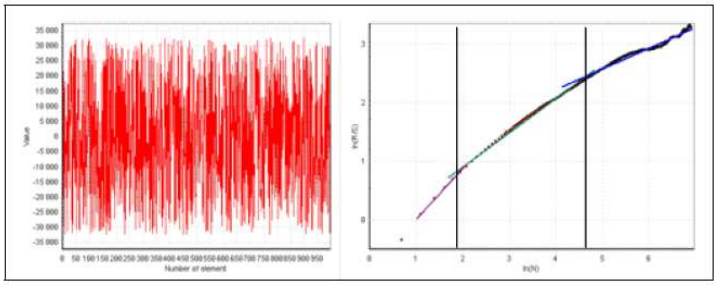

Рис. 4. Приклад випадкового часового ряду і графіку коефіцієнта Хьорста [9]

3. Якщо $0,5 < H < 1,0$, то ряд є трендостійким, тобто якщо ряд зростає або спадає в попередній проміжок часу, то вірогідно, що він продовжить цю тенденцію певний час в майбутньому.
4. Якщо $H > 1$, то мова йде про випадковий процес з фрактальним часом. Відбуваються незалежні стрибки амплітуди, розподілені за Леві, за час, визначений величиною стрибка, і зростаючий разом із ним.

Для кожної клітини були пораховані значення H для часових рядів, утворених червоним, зеленим та синім компонентом. Для кожного пацієнта порахували такі показники (назвемо їх першою групою):

1. Середнє значення коефіцієнта Хьорста для кожного пацієнта.
2. Максимальне значення коефіцієнта Хьорста кожного пацієнта.
3. Мінімальне значення коефіцієнта Хьорста для кожного пацієнта.

З'ясувавши за допомогою кривої Гільберта (див. табл. 1), що інформативним є саме синій колір, на другому етапі для кожного пацієнта по синій компоненті по всім клітинам ми обчислили такі показники (назвемо їх другою групою).

1. Дисперсія для значень коефіцієнта Хьорста по всім клітинам пацієнта.
2. Сума квадратів відхилень значень коефіцієнта Хьорста по всім клітинам пацієнта від середнього значення коефіцієнта.
3. Квантиль порядку 0,75.
4. Медіана значень коефіцієнта Хьорста по всім клітинам пацієнта.
5. Середнє гармонічне значень коефіцієнта Хьорста по всім клітинам пацієнта.
6. Середнє геометричне значень коефіцієнта Хьорста по всім клітинам пацієнта.
7. Середнє урізане значення коефіцієнта Хьорста по всім клітинам пацієнта(враховуючи всі, окрім 5%)
8. Середнє арифметичне значення коефіцієнта Хьорста по всім клітинам пацієнта.
9. Ексцес значень коефіцієнта Хьорста по всім клітинам пацієнта.

Результати

За шкалою коефіцієнта Хьорста за компонентами кольору RGB (червона, зелена, синя) маємо такі результати.

Таблиця 1. — Характеристики часового ряду, побудованого за кривими Гільберта і Серпинського (чисельник — за кривою Гільберта, знаменник — за кривою Серпинського)

Червона компонента			
Ряд/діагноз	Здорові	Рак	Фіброаденоматоз
Хаотичний	0/0	0/0	0/0
Ергодичний	0/0	0/0	0/0
Трендостійкий	29/29	54/68	27/33
Фрактальний	0/0	14/0	16/0
Зелена компонента			
Ряд/діагноз	Здорові	Рак	Фіброаденоматоз
Хаотичний	0/0	0/0	0/0
Ергодичний	0/0	0/0	0/0
Трендостійкий	2/28	6/68	6/33
Фрактальний	27/1	62/68	27/33
Синя компонента			
Ряд/діагноз	Здорові	Рак	Фіброаденоматоз
Хаотичний	0/0	0/0	0/0
Ергодичний	0/0	0/0	0/0
Трендостійкий	29/29	43/68	24/33
Фрактальний	0/0	29/0	9/0

Ці результати свідчать про існування певного зсуву за червоною та синьою компонентами: трендостійкі часові ряди характерні для всіх здорових і більшості хворих на фіброаденоматоз та рак, але частина хворих із останніх двох груп потрапляє до "екстремальної"категорії випадкових процесів із фрактальним часом. Втім, слід зазначити, що за двовибірковими Z-критерієм і χ^2-критерієм із рівнями значущості 0,05 ця різниця не є статистично значущою. За зеленою компонентою такого розділення взагалі не існує. Отже, простій класифікації за видом часового ряду ці захворювання не піддаються, хоча гіпотеза про наявність зсуву дає підстави для застосування більш складних методів дискримінантного аналізу. Для класифікації було обрано метод дерев класифікації CART (classification and regression trees) і синю компоненту [10]. Зазначимо, що часові ряди, отримані за кривою Серпинського, майже усі (за винятком одного із 130) виявилися трендостійкими за усіма кольоровими компонентами.

Уведемо такі позначення: N_1 — кількість здорових пацієнтів, N_2 — кількість пацієнтів на рак, N_3 — кількість пацієнтів, хворих на фіброаденоматоз, M_1 — кількість правильно класифікованих здорових пацієнтів, M_2 — кількість правильно класифікованих пацієнтів, хворих на рак, M_3 — кількість правильно класифікованих пацієнтів, хворих на фіброаденоматоз, Sp — специфічність бінарної класифікації, Sc — чутливість бінарної класифікації, $Sp_i,\ i = 1, 2, 3$ — специфічність класифікації i-го діагнозу при класифікації трьох груп, $Sc_i,\ i = 1, 2, 3$ — чутливість класифікації i-го діагнозу при класифікації трьох груп, P — точність класифікації.

Чутливість і специфічність є кількісними характеристиками бінарної класифікації, які визначають долю позитивних результатів серед хворих та долю негативних результатів серед здорових відповідно. Наприклад, при тестуванні N_1 здорових пацієнтів і N_2 хворих на рак, обчислюються за формулами $Sp = \frac{M_1}{N_1}$ і $Sc = \frac{M_2}{N_2}$. При класифікації трьох груп хворих необхідно ввести попарні специфічність і чутливість. Відповідно, отримуємо $Sp^{(i)} = \frac{M_i}{N_i}$ і $Sc^{(i)} = \frac{M_k+M_l}{N_k+N_l},\ k, l = 1, 2, 3,\quad k, l \neq i$. Точність дорівнює долі правильно класифікованих серед усіх пацієнтів, тобто для бінарного випадку вона дорівнює $P = \frac{M_1+M_2}{N_1+N_2}$, а для випадку класифікації трьох класів $P = \frac{M_1+M_2+M_3}{N_1+N_2+N_3}$.

Результати класифікації для першої групи показників. При класифікації здорових та хворих на рак молочної залози методом крос-валідації з використанням кривих Гільберта і Серпинського отримано такі результати (нульовою гіпотезою вважалося припущення, що у пацієнта рак).

Точність класифікації за кривою Гільберта склала 63,08%, а за кривою Серпинського — 67,69%. Як бачимо, сканування за кривими Гільберта і Серпинського у сполученні із класифікацією трьох захворювань має посередню точність і не дозволяє відокремити фіброаденоматоз.

Оскільки відрізнити рак від фіброаденоматозу за допомогою описаного вище методу на вдалося, на другому етапі класи хворих на рак і фіброаденоматоз були об'єднані в спільний клас і проведено класифікацію здорових та хворих пацієнтів. Ця ситуація є типовою для скринінгу, коли серед великої популяції людей треба розпізнати групу ризику, не проводячи диференційну діагностику хвороб. У цьому випадку було взято всіх (29) здорових пацієнтів і половину об'єднаного класу (51 пацієнт). Вибір хворих пацієнтів робився випадковим чином. Це було зроблено з міркувань великої кількості хворих пацієнтів і малої здорових.

Таблиця 2. — Попарна специфічність, чутливість і точність диференціальної діагностики за першою групою показників

Здорові і рак		
Показник/Вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	68,96	93,10
Чутливість, %	92,65	98,52
Точність, %	85,57	96,90
Здорові і фіброаденоматоз		
Вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	58,62	96,55
Чутливість, %	66,67	96,97
Точність, %	62,90	96,77
Рак і фіброаденоматоз		
Вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	15,15	95,59
Чутливість, %	77,94	00,00
Точність, %	57,43	64,36

Таблиця 3. — Специфічність і чутливість диференціальної діагностики за першою групою показників

Крива Гільберта			
Показник/діагноз	Здорові	Рак	Фіброаденоматоз
Специфічність, %	72,41	89,71	0,00
Чутливість, %	60,39	33,87	84,54
Крива Серпинського			
Показник/діагноз	Здорові	Рак	Фіброаденоматоз
Специфічність, %	79,31	95,59	0,0
Чутливість, %	64,36	37,10	90,72

Таблиця 4. — Попарна специфічність, чутливість і точність скринінгу за першою групою показників

Здорові і хворі (рак або ФАМ)		
Показник/вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	86,21	89,66
Чутливість, %	72,00	98,00
Точність, %	77,22	94,97

Результати класифікації для другої групи показників. Наведемо аналогічні результати, отримані для описових статистик розподілу коефіцієнтів Хьорста.

Таблиця 5. — Попарна специфічність, чутливість і точність диференціальної діагностики за другою групою показників

Здорові і рак		
Показник/вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	72,41	86,21
Чутливість, %	85,29	94,12
Точність, %	81,44	91,75
Здорові і фіброаденоматоз		
Показник/вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	86,21	96,55
Чутливість, %	81,82	93,94
Точність, %	83,87	95,16
Рак і фіброаденоматоз		
Показник/вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	18,18	100,00
Чутливість, %	80,88	0,0
Точність, %	60,40	67,33

Як бачимо, попарна класифікація здорових і пацієнтів, хворих на рак або фіброаденоматоз, дозволяє побудувати досить точні моделі

на основі кривих Гільберта і Серпинського. Водночас, диференціальна діагностика раку і фіброаденоматозу за допомогою цих кривих є неможливою.

Таблиця 6. — Специфічність і чутливість диференціальної діагностики за другою групою показників

Крива Гільберта			
Показник/діагноз	Здорові	Рак	Фіброаденоматоз
Специфічність, %	65,52	88,40	0,00
Чутливість, %	60,39	30,65	82,47
Крива Серпинського			
Показник/діагноз	Здорові	Рак	Фіброаденоматоз
Специфічність, %	86,21	94,12	0,00
Чутливість, %	63,37	40,32	91,75

Точність класифікації за кривою Гільберта склала 61,54%, а за кривою Серпинського — 68,46%. Як бачимо, сканування за кривою Гільберта у сполученні із класифікацією трьох захворювань, як і в попередньому випадку, має середню точність і не дозволяє відокремити фіброаденоматоз. Водночас, застосування кривої Серпинського дозволяє отримати тест, який можна вважати відмінніи.

При класифікації здорових людей та пацієнтів, хворих на рак або фіброаденоматоз, за другою групою показників, було отримано такі результати.

Таблиця 7. — Попарна специфічність, чутливість і точність скринінгу за другою групою показників

Здорові і хворі		
Показники/вид кривої	Крива Гільберта	Крива Серпинського
Специфічність, %	89,66	96,55
Чутливість, %	78,43	90,00
Точність, %	82,50	92,41

Як бачимо, класифікація здорових людей і пацієнтів, хворих на рак або фіброаденоматоз, яка лежить в основі скринінгу, дозволяє побудувати відмінні за точністю моделі на основі кривих Гільберта і Серпинського.

Висновки

За допомогою статистичного аналізу часових рядів вдалося виявити пух-линно-асоційовані зміни фрактальної структури хроматину у жінок, хворих на фіброаденоматоз і рак молочної залози. Ця відмінність

означає, що інформація, "закодована"у клітинах, має певний коефіцієнт "хаотичності", причому, у здорових клітин частіше спостерігається "хаотичність", ніж у хворих (коефіцієнт Хьорста ближче до 0,5). Побудовані моделі класифікації мають дуже добру (крива Гільберта) та відмінну (крива Серпинського) точність, що дозволяють виявляти наявність у організмі людини злоякісної або доброякісної пухлини під час скринінгу. Водночас вони не можуть застосовуватись для диференційної діагностики раку і фіброаденоматозу, і цю тему необхідно дослідити іншими методами.

Література

1. Susnik B. Malignancy-associated changes in the breast: changes in chromatin distribution in epithelial cells in normal-appearing tissue adjacent to carcinoma / B. Susnik, A. Worth, J. LeRiche, B. Palcic // Analytical and Quantitative Cytology and Histology. — 1995. — 17(1). — P. 62 – 68.
2. Mairinger T. Nuclear chromatin texture analysis of nonmalignant tissue can detect adjacent prostatic adenocarcinoma / Mairinger T., G. Mikuz, A. Gschwendtner // The Prostate. — 1999. — 41(1). — P. 12 – 19.
3. Us-Krasovec M. Malignancy associated changes in epithelial cells of buccal mucosa: a potential cancer detection test / M. Us-Krasovec, J. Erzen, M. Zganec M. et. al. // Anal Quant Cytol Histol. — 2005 Oct. — 27(5). — P. 254—262.
4. Andrushkiw R.I. Computer-aided cytogenetic method of cancer diagnosis / R.I. Andrushkiw, N.V. Boroday, D.A. Klyushin, Yu.I. Petunin. — New Jork: Nova Publishers, 2007.
5. Redon C.E. Tumors induce complex DNA damage in distant proliferative tissues in vivo // C. Redon et al. / Proceedings of the National Academy of Sciences. — October 19, 2010. — 107(42) — P. 17992–17997.
6. Lieberman-Aiden E. Comprehensive mapping of long-range interactions reveals folding principles of the human Genome / E. Lieberman-Aiden, N.L. van Berkum // Science. — 2009. — 326. — P. 289–293.
7. Nikolaou N. Color image retrieval using a fractal signature extraction technique / N. Nikolaou, N. Papamarkos // Engineering Applications of Artificial Intelligence. — 2002. — 15(1). — P. 81–96.
8. Sagan H. Space-filling curves / H. Sagan — Springer-Verlag: New York-Berlin, 1994.
9. Бутаков В. Оценка уровня стохастичности временных рядов произвольного происхождения при помощи показателя Хёрста / В. Бутаков, А. Граковский // Computer Modelling and New Technologies. — 2005. — 9(2). — P.27–32.
10. Breiman L. Classification and regression trees. Monterey / L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone — CA: Wadsworth and Brooks/Cole Advanced Books and Software, 1984.