Журнал обчислювальної та прикладної математики **2017, №2(125)** Journal of Numerical & Applied Mathematics

УДК 528.854
MSC 68T10

FRACTAL ANALYSIS OF MALIGNANCY-ASSOCIATED CHANGES OF INTERPHASE NUCLEI OF BUCCAL EPITHELIUM

Dmitry Klyushin[1], Dmitry Shervarly, Chan Kha Wu, Kateryna Golubeva[1], Natalia Boroday[2]

[1]Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, E-mail: klyushin@unicyb.kiev.ua

[2]Институт экспериментальной патологии, онкологии и радиобиологии им. Р. Е. Кавецкого НАН Украины, Киев, Украина

ФРАКТАЛЬНЫЙ АНАЛИЗ ОПУХОЛЬ-АССОЦИИРОВАННЫХ ИЗМЕНЕНИЙ ИНТЕРФАЗНЫХ ЯДЕР БУККАЛЬНОГО ЭПИТЕЛИЯ

Д. А. Клюшин[1], Д. Г. Шерварлы, Чан Ха Ву, Е. Н.Голубева[1], Н. В. Бородай[2]

[1]Факультет компьютерных наук и кибернетики, Киевский национальний университет имени Тараса Шевченко, Киев, Украина, E-mail: klyushin@unicyb.kiev.ua

[2]Институт экспериментальной патологии, онкологии и радиобиологии им. Р. Е. Кавецкого НАН Украины, Киев, Украина

Abstract. We have analyzed malignancy-associated changes of fractal structure of chromatin in buccal epithelium in patients with fibroadenomatosis and breast cancer. We have developed a screening method that provides high accuracy, sensitivity and specificity. The method is based on calculating the Minkowski dimension of digital image obtained by highlighting the blue component of digital image of DNA stained by Felgen.
Keywords: malignancy-associated changes, decision tree proximity measure, fractal analysis, Minkowski dimension.

Резюме. Проанализированы опухоль-ассоциированные изменения фрактальной структуры хроматина буккального эпителия у женщин, болеющих фиброаденоматозом и раком молочной железы. Разработан метод скрининга, обеспечивающий высокую точность, чувствительность и специфичность. Метод основан на вычислении размерности Минковского цифрового изображения, полученного путем выделения синей компоненты цифрового изображения ДНК, окрашенного по Фельгену.
Ключевые слова: опухоль-ассоциированные изменения, дерево решений, мера близости, фрактальный анализ, размерность Минковского.

1. Вступ

Одним из основных методов борьбы с раком молочной железы является скрининг — обследование населения с целью выявления опухоли на ранней стадии. Поскольку скрининг носит массовый характер, он должен иметь высокую чувствительность и специфичность, быть простым в применении и безвредным для людей. Стандартные методы диагностики рака молочной железы — клиническое обследование, маммография и аспирационная биопсия, образующие так называемый „золотой стандарт“, позволяют поставить диагноз с точностью до 99 %. Тем не менее, радиоактивное облучение и потенциальное травмирование потенциально злокачественной опухоли во время биопсии несет с собой риск для здоровья пациента. Таким образом, задача разработки быстрого, неинвазивного и точного метода скрининга все еще остается актуальной.

Выявление опухоль-ассоциированных изменений интерфазных ядер буккального эпителия у больных раком молочной железы открыло возможность для разработки безвредных и точных методов диагностики рака (см. описание проблемы в работе [1]). В последнее время пристальное внимание исследователей привлекли фрактальные свойства распределения хроматина в ядрах клеток человека [2]. Эти идеи легли в основу работ [3], [4], в которых предложены новые методы скрининга рака молочной железы. В этих работах исследовались ряды, возникающие при обходе изображений клеток вдоль фрактальных кривых Гильберта, Серпинского и Пеано. Естественным продолжением этих исследований является прямой анализ фрактальной размерности изображений, не связанный с выбором направления обхода.

2. Метод и материалы

Для исследования были случайным образом выбраны три группы людей: контрольная группа (29 здоровых женщин), группа больных раком молочной железы (68 пациенток) и група больных фиброаденоматозом (33 пациентки), диагноз которых верифицирован с помощью гистологического исследования.

Предметом анализа являлись фотографии соскобов эпителицитов слизистой оболочки ротовой полости, взятые со средней глубины шиповатого слоя, зафиксированные в смеси Никифорова и окрашенные по Фельгену. Каждое изображение представляет собой квадрат 160×160 пикселей. Каждой пациентке соответствует набор от 20 до 30 фотографий.

3. Дерево решений

В работах [3], [4] исследовались коэффициенты Херста временных рядов, возникающих при обходе изображений по фрактальным кривым. Это показатель, тесно связанный с фрактальной размерностью изображения, оказался весьма информативным, поэтому возникла идея непосредственно исследовать фрактальную размерность, вычислив ее не по Херсту, а по Минковскому. Поскольку одним из результатов работ [3], [4] было выявление

высокой информативности синей компоненты изображения ДНК интерфазных ядер буккального эпителия, окрашенных по Фельгену, фрактальная размерность Минковского вычислялась только для синей компоненты.

Пусть $N(\epsilon)$ — минимальное количество шаров радиуса ϵ, которые необходимы для покрытия компактного множества A. Тогда, если существует предел

$$dim_M(A) = -\lim_{\epsilon \to \infty} \frac{logN(\epsilon)}{log(\epsilon)},$$

он называется размерностью Минковского, или фрактальной размерностью [5].

На первом этапе для вычисления фрактальной размерности применим модифицированный метод box-counting, позволяющий максимально учесть информацию, содержащуюся в изображении. Отличие модифицированного метода box-counting от классического состоит в том, что при наложении сетки квадратов на изображение считается не количество квадратов, в которые попали пиксели изображения, а средняя яркость пикселей, попадающих в каждый квадрат. Таким образом, для каждого квадрата в методе box-counting вычисляется весовой множитель

$$h = \frac{255 - I_c}{255},$$

где I_c — средняя яркость пикселей в квадрате. Теперь заменим величину $N(r)$ — сумму квадратов со стороной r, необходимых для покрытия множества, на $M(r)$, — сумму весовых множителей квадратов со стороной r, необходимых для покрытия множества. Тогда формула для определения фрактальной размерности примет вид

$$d' = \frac{log(M(r_1)) - log(M(r_2))}{log\frac{1}{r_1} - log\frac{1}{r_2}}.$$

Для построения деревьев классификации использовался метод CART [6]. На втором этапе для каждой пациентки вычислялись следующие статистические показатели размерности Минковского:

1) дисперсия;
2) сумма квадратов отклонений от среднего значения;
3) квантиль порядка 0,75;
4) медиана;
5) среднее гармоническое;
6) среднее геометрическое;
7) среднее усеченное (учитывая все значения, кроме 5%);
8) среднее арифметическое;
9) эксцесс.

Для разделения на здоровых и больных (рак молочной железы и фиброаденоматоз) используем классификатор, построенный на базе синего компонента цвета и модифицированного метода box-counting. Проведя кросс-валидацию, получаем следующие результаты.

Точность $= \frac{29+97}{29+101} = 96,92\%$.
Специфичность $= \frac{29}{29} = 100,00\%$.
Чувствительность $= \frac{97}{101} = 96,04\%$.

Для сравнения применим для разделения на здоровых и больных (рак молочной железы и фиброаденоматоз) классификатор, построенный на базе синей компоненты цвета и классическом методе box-counting. Проведя кроссвалидацию получаем следующие результаты.

Точность $= \frac{29+93}{29+101} = 93,85\%$.
Специфичность $= \frac{29}{29} = 100,00\%$.
Чувствительность $= \frac{93}{101} = 91,08\%$.

4. Мера близости и метод ближайших соседей

Фрактальный анализ изображений позволяет осуществить классификацию не только пациентов в целом, но и отдельных клеток. Благодаря этому возникает возможность оценить характеристики распределений фрактальной размерности в группах здоровых людей, больных раком молочной железы и фиброаденоматозом.

В этом случае исходный набор данных рассматривался как совокупность 6751 фотография интерфазных ядер буккального эпителия (всего 20253 фотографии в формате RGB), взятых у 130 пациентов, указанных в разделе „Метод и материалы“. Таким образом, каждый пациент был представлен выборкой значений размерности Минковского интерфазных ядер взятых у него клеток интерфального эпителия, измеренных для каждого из трех каналов — синего, зеленого и красного.

Для сравнения выборок, состоящих из значений размерности Минковского соответствующих фотографий, применялась мера близости Петунина [7]. Пусть $x = (x_1, x_2, ..., x_n) \in G$ и $y = (y_1, y_2, ..., y_m) \in H$ — две выборки, извлеченные из генеральных совокупностей G и H соответственно. Предположим, что $F_G(x) \equiv F_H(x)$ и построим вариационный ряд $x_1 \leq ... \leq x_n$. Обозначим как $A_{ij}^{(k)} = \left\{x_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$ событие, состоящее в том, что выборочное значение y_k попадает между порядковыми статистиками $x_{(i)}$ и $x_{(j)}$.

В соответствии с теоремой Хилла

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

Построим доверительный интервал для вероятности события $A_{ij} : k$ по формулам Вильсона

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

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

где $h_{ij}^{(n,m)}$ — частота события $A_{ij}^{(n)}$ в m испытаниях. Вычислим доверительный интервал $I_{ij} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ с уровнем значимости, определяемым параметром g. В частности, при $g = 3$ уровень значимости этого интервала не превышает 0,05. Вычислим $N = \frac{n(n-1)}{2}$ — общее количество интервалов, образованных порядковыми статистиками, и L — количество интервалов I_{ij}, содержащих вероятности p_{ij}. Тогда $h = \rho(x, y) = \frac{L}{N}$ — мера близости между виборками x и y. Положив $h_{ij} = h, m = N$ и $g = 3$, получаем доверительный интервал $I = \left(p^{(1)}, p^{(2)}\right)$ для вероятности $p(B)$.

Для тестирования выборки x из обучающего набора выбирались k ближайших (в смысле меры близости) соседей. Выборка x классифицировалась по доминантному классу этих k елементов. Параметр $k = 9$ был определен эмпирически.

Для разделения здоровых и больных (на рак молочной железы и фиброаденоматоз) используем классификатор, построенный на базе синего компонента цвета. Получаем следующие результаты:

Точность $= \frac{28+93}{29+101} = 93,1\,\%$

Специфичность $= \frac{28}{29} = 96,6\,\%$

Чувствительность $= \frac{93}{101} = 92,1\,\%$

5. Выводы

Подтверждена значимость синего компонента изображения для выявления контрольной группы и группы риска рака молочной железы и фиброаденоматоза при скрининге. Предложена модификация метода box-counting, обеспечивающая высокую точность, чувствительность и специфичность, а также метод скрининга, основанный на мере близости Петунина. Показано, что модифицированный метод имеет более высокую точность, а метод, основанный на мере близости Петунина, обеспечивает высокую чувствительность и специфичность.

Литература

1. Andrushkiw R. I., Boroday N. V., Klyushin D. A., Petunin Yu. I. Computer-aided cytogenetic method of cancer diagnosis — New Jork: Nova Publishers, 2007.
2. Lieberman-Aiden E., N. L. van Berkum Comprehensive mapping of long-range interactions reveals folding principles of the human Genome // Science. — 2009. — 326. — P. 289–293.
3. Клюшин Д. А., Шерварли Д. Г., Присяжна М. В., Бородай Н. В., Белоусова О. А. Новий метод скринінгу раку молочної залози на підставі фрактального аналізу інтерфазних ядер букального епітелію // Журнал обчислювальної та прикладної математики. — 2011. — №1(104). — С. 68–76.
4. Klyushin D. A., Shervarly D. G., Golubeva E.N., Boroday N. V., Belousova E. A. Screening of breast cancer using peano curve // Журнал обчислювальної та прикладної математики. — 2012. — №4(110). — С. 117–121.
5. Кроновер Р. М. Фракталы и хаос в динамических системах. Основы теории. — Москва: Потсмаркет, 2000. — 352 с.

6. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. — Chapman & Hall/CRC, 1984. — 368 p.
7. Клюшин Д. А., Петунин Ю. І. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками // Укр. матем. журн. — 2003. — Т. 5, №2. — С. 147–163.

Поступила 4.09.2017