http://dx.doi.org/10.35596/1729-7648-2020-18-8-21-28

Оригинальная статья
Original paper

УДК 004.93'1;004.932; 004.8.032.26;616-006

РАЗРАБОТКА АЛГОРИТМА ПОИСКА ОПУХОЛЕВЫХ ОБЛАСТЕЙ НА ОСНОВЕ ОБРАБОТКИ ПОЛНОСЛАЙДОВЫХ ГИСТОЛОГИЧЕСКИХ ИЗОБРАЖЕНИЙ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

РЯБЦЕВА С.Н.[1], КОВАЛЕВ В.А.[2], МАЛЫШЕВ В.Д.[2], СЕМЕНИК И.А.[1], ДЕРЕВЯНКО М.А.[1], МОСКАЛЕНКО Р.А.[3], ДОВБЫШ А.С.[3], САВЧЕНКО Т.Р.[3], РОМАНЮК А.Н.[3]

[1]*Институт физиологии Национальной академии наук Беларуси (г. Минск, Республика Беларусь)*

[2]*Объединенный институт проблем информатики Национальной академии наук Беларуси (г. Минск, Республика Беларусь)*

[3]*Сумский государственный университет (г. Сумы, Украина)*

Поступила в редакцию 11 ноября 2020

Аннотация. Анализ полнослайдовых изображений рака молочной железы является крайне трудоемким процессом. Гистологические полнослайдовые изображения обладают рядом особенностей, затрудняющих их разработку: высокая степень разнообразия тканей как на одном изображении, так и между различными изображениями, иерархичность, большой объем графической информации и различные артефакты. В ходе научной работы проведена обработка полнослайдовых изображений ткани рака молочной железы, что включало нормализацию распределения цвета на полнослайдовых гистологических изображениях и выделение области изображения, на которой располагается изучаемый образец ткани, чтобы уменьшить время работы остальных алгоритмов и не анализировать области полнослайдового изображения с фоном. Также разработан и реализован алгоритм поиска похожих для полуавтоматического выделения опухолевых участков с помощью различных дескрипторов изображений.

Ключевые слова: полнослайдовые изображения, рак молочной железы, обработка, поиск похожих.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Благодарности. Работа выполнена в рамках задания ГКНТ по конкурсу совместных Белорусско-Украинских научно-технических проектов «Разработать автоматизированную программу дифференциальной диагностики новообразований молочной железы с морфометрической оценкой рецепторного статуса раковых клеток» при поддержке Белорусского фонда фундаментальных исследований.

Для цитирования. Рябцева С.Н., Ковалев В.А., Малышев В.Д., Семеник И.А., Деревянко М.А., Москаленко Р.А., Довбыш А.С., Савченко Т.Р., Романюк А.Н. Разработка алгоритма поиска опухолевых областей на основе обработки полнослайдовых гистологических изображений рака молочной железы. Доклады БГУИР. 2020; 18(8): 21-28.

DEVELOPMENT OF NEOPLASTIC REGION SELECTION ALGORITHM BASED ON BREAST CANCER WHOLE SLIDE IMAGE

SVETLANA N. RJABCEVA[1], VASILII A. KOVALEV[2], VALERY D. MALYSHEV[2], IRINA A. SIAMIONIK[1], MARYNA A. DEREVYANKO[1], ROMAN A. MOSKALENKO[3], ANATOLII C. DOVBYSH[3], TARAS R. SAVCHENKO[3], ANATOLII N. ROMANIUK[3]

[1]*Institute of Physiology of National Academy of Sciences of Belarus (Minsk, Republic of Belarus)*

[2]*United Institute of Informatics Problem of National Academy of Sciences of Belarus (Minsk, Republic of Belarus)*

[3]*Sumy State University (Sumy, Ukraine)*

Submitted 11 November 2020

Abstract. Analysis of breast cancer whole-slide image is an extremely labor-intensive process. Histological whole slide images have the following features: a high degree of tissue diversity both in one image and between different images, hierarchy, a large amount of graphic information and different artifacts. In this work, pre-processing of breast cancer whole-slide tissue image was carried out, which included normalization of the color distribution and the image area selection. We reduced the operating time of the other algorithms and excluded areas of breast cancer whole-slide tissue with a background to analyze. Also, an algorithm for finding similar neoplastic regions for semi-automatic selection using various image descriptors has been developed and implemented.

Keywords: whole-slide image, breast cancer, processing, similarity search.

Conflict of interests. The authors declare no conflict of interests.

Gratitude. The work was carried out as part of the assignment of joint Belarusian-Ukrainian projects of State Committee for Science and Technology “To develop an automated program for breast cancer differential diagnosis with morphometric assessment of cancer cells receptor status” with the support of the Belarusian Foundation for Basic Research.

For citation. Rjabceva S.N., Kovalev V.A., Malyshev V.D., Siamionik I.A., Derevyanko M.A., Moskalenko R.A., Dovbysh A.S., Savchenko T.R., Romaniuk A.N. Development of neoplastic region selection algorithm based on breast cancer whole slide image. Doklady BGUIR. 2020; 18(8): 21-28.

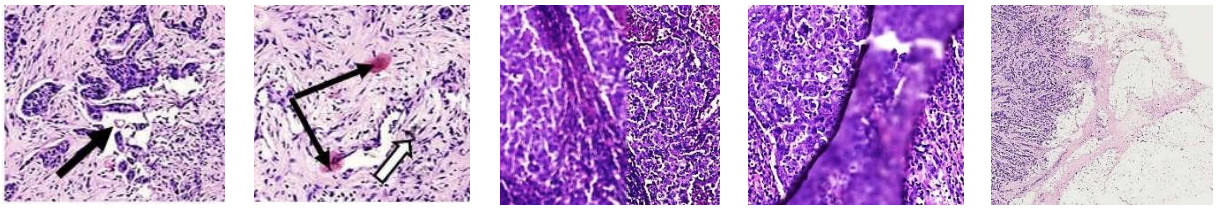
Введение

В Беларуси ежегодно выявляют около 4000 новых случаев рака молочной железы у женщин. Злокачественные эпителиальные опухоли молочной железы женского населения составляют 5–6 % от всех патологических образований данной локализации. Пятилетняя выживаемость на первой стадии процесса составляет 85–90 %, на второй – 70–80 %, на третьей – 35–50 %. Назначение эффективной терапии базируется на данных гистологической верификации диагноза с определением рецепторного статуса опухолевых клеток. При адекватном ведении пациентов с данной патологией шанс на выздоровление возрастает. Значительные трудности в морфологической диагностике нозологических форм рака молочной железы, субъективизм оценки патологических изменений, отсутствие достаточного опыта зачастую становятся причиной неверной диагностики. Достоверность гистологической верификации рака молочной железы зависит от опыта, знаний, готовности самосовершенствоваться и изучать специализированную литературу врача-морфолога. Встречая трудные диагностические случаи, патологоанатом обычно консультируется со своими коллегами и, ссылаясь на материалы, изложенные в учебниках, атласах, статьях из журналов, выставляет окончательный диагноз. Нельзя также отрицать и долю субъективизма при постановке гистологического заключения. Учитывая вышеизложенные

факты, вопрос создания высокотехнологичных (с применением передовых информационных технологий) экспертных систем, аккумулирующих опыт и знания передовых врачей-диагностов, является актуальным не только для врачей-патологоанатомов, но и для каждой женщины, ожидающей достоверную морфологическую верификацию злокачественного процесса молочной железы. Таким образом, разработка и внедрение автоматизированной дифференциально-диагностической системы поддержки принятия решений в диагностике новообразований молочной железы являются актуальными. Цель данного исследования – разработка методов, алгоритмов и базовых элементов программного комплекса для автоматического поиска опухолевых участков для предварительной диагностики новообразований молочной железы.

Экспериментальная часть

В данной работе использовались полнослайдовые изображения фрагментов инвазивного протокового рака молочной железы, окрашенные гематоксилином и эозином, полученные путем сканирования образцов тканей в светлом поле на гистологическом сканере Zeiss Axioscan с объективом 20×. Трудности обработки полнослайдовых изображений связаны с различными артефактами, которые могут появиться как в процессе подготовки образцов ткани, так и в процессе сканирования. К артефактам, возникающим при технической подготовке гистологических срезов, относят: разрывы ткани, вариабельность цветовой окраски гематоксилином и эозином, недостаточная промывка от красителей, деформация гистологических срезов, инородные предметы под покровным стеклом (рис. 1).



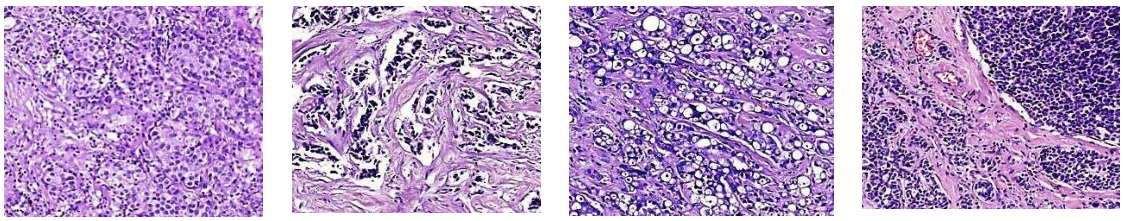
a *b* *c* *d* *e*

Рис. 1. Примеры артефактов после технической подготовки и сканирования гистологических препаратов: *a* – разрыв ткани (стрелка); *b* – остатки розового красителя – эозина (черная стрелка) и инородные предметы (белая стрелка); *c, d* – нарушение фокусировки; *e* – наличие фона после сканирования гистологического среза (увеличение объектива 2×), сканер Zeiss Axioscan, окраска гематоксилином и эозином, увеличение объектива 20×

Fig. 1. Examples of artifacts after technical preparation and scanning of slides: *a* – black arrow shows a tear in tissue; *b* – black arrow shows eosin spots, white arrow highlights foreign object under cover slip; *c, d* – folds in the tissue in defocusing; *e* – the presence of a background after scanning (a 2× objective), scanner Zeiss Axioscan, H&E, a 20×objective

Другие артефакты могут проявиться при сканировании изображений [1]. Для получения изображения сканер делает снимок на максимальном увеличении, а потом из полученных снимков собирает полнослайдовое изображение, поэтому могут проявиться такие артефакты, как разница в освещении, фокусе, проблемы со сдвигом камеры на правильное расстояние (рис. 1). Также трудности обработки полнослайдовых гистологических изображений рака молочной железы связаны с разнообразием клеточной компоновки – вариабельностью гистологических форм данной злокачественной опухоли (рис. 2).

Высокая вероятность наличия артефактов разного характера, вариабельность клеточной компоновки в опухоли в совокупности сильно усложняют применение методов глубокого обучения на таких изображениях. В дополнение к этому высокое разрешение полнослайдовых изображений делает невозможным применение методов глубокого обучения «напрямую». Таким образом, в большинстве существующих решений используется разделение полнослайдовых изображений на небольшие участки [2].



a *b* *c* *d*

Рис. 2. Варианты гистологического строения инвазивного протокового рака молочной железы (*a–d*), полнослайдовые изображения при увеличении объектива 20×, сканер Zeiss Axioscan, окраска гематоксилином и эозином

Fig. 2. Variants of histological structure of invasive ductal breast cancer (*a–d*), parts of whole slide tissue, scanner Zeiss Axioscan, H&E, a 20×objective

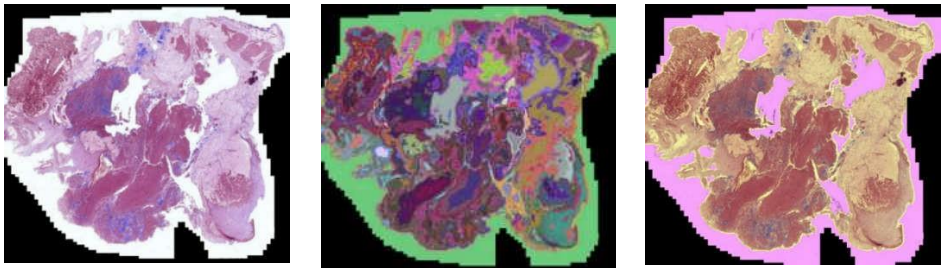
Основным методом нормализации цветовой гаммы на полнослайдовых изображениях был выбран обобщенный алгоритм, который включал этапы: а) исключение неподходящих для определения цветовой схемы областей (например, фон, на котором отсутствует образец ткани); б) перевод пространства цветового формата данных гистологических изображений RGB (RGB – Red Green Blue) в некоторое другое пространство; в) выделение опорных векторов, линейная комбинация которых определяет данное пространство; г) обратное преобразование этих векторов в RGB пространство для того, чтобы получить основные цвета, комбинации которых присутствуют на изображении; д) перевод изображения в цвета, определенные выше; е) замена основных цветов на эталонные, заданные до начала работы алгоритма; ж) обратное преобразование из значений концентрации основных цветов и RGB значений эталонных цветов в RGB пространство.

В данной работе было опробовано два подхода к нормализации цветового распределения. В обоих использовали преобразование RGB значений пикселей изображения в значения оптической плотности по формуле [3] $OD = -\log_{10}(I)$.

На изображении не рассматривались пиксели с слишком низкой оптической плотностью, которые считались фоном на анализируемых изображениях. В первом алгоритме использовали разложение формата кодирования цвета в изображениях SVD (Singular Value Decomposition) с последующей нормализацией по длине векторов. Дальнейшие операции выполнялись с углами между векторами и направлениями SVD разложения [3]. Во втором алгоритме вместо SVD разложения была применена ковариационная матрица RGB каналов изображения, где вектора получали из собственных векторов этой матрицы [4]. Углы считали между векторами, определяемыми координатами в новом пространстве, и полученными собственными векторами. В качестве основных векторов брались 1-й и 99-й перцентили полученного распределения углов векторов относительно найденных базисных векторов. Таким образом, цветовая схема всех изображений соответствовала одинаковым базовым цветам первоначальных химических маркеров. Далее применялись алгоритмы нормализации изображений, а именно эквализация гистограммы яркости изображения и удаление 1-го и 99-го перцентилей интенсивностей пикселей изображения, для того, чтобы частично избежать влияния различных артефактов изображений. Стоит отметить, что эти два алгоритма нормализации применялись к каналу Y в YCbCr представлении изображений (сингулярное разложение YCbCr – способ кодирования цвета, где Y обозначает яркость, Cb и Cr – синяя и красная цветоразностные компоненты), так как нормализуется только контрастность изображения, а цветовая схема должна оставаться такой же.

Вторым необходимым компонентом перед использованием алгоритма выделения опухолевого участка на полнослайдовом изображении было отделение области изображения, на которой изображен рассматриваемый образец ткани, от основного фона. В данной работе использовался ряд алгоритмов для того, чтобы достичь такого результата. Сначала была использована область полнослайдового изображения порядка нескольких мегапикселей, покрывающая весь слайд, и к ней применялся слабый эффект размытия для того, чтобы повысить гладкость и стабильность регионов, получаемых с помощью алгоритма заливки методом «наводнения». Такой способ содержит множество параметров, изменение которых позволяет регулировать количество областей. Эти параметры включают размер минимальной

области, связность, максимально допустимую разницу между интенсивностями пикселей в пределах одного региона и т. д. Кроме этого, важно, какой канал изображения используется. В данный момент в основном используется S канал HSV (HSV – Hue Saturation Brightness) представления изображения, показывающий степень насыщенности цвета в каждом пикселе, что отлично подходит для нахождения фона изображения. Алгоритм последовательно заливает регионы из свободных, не залитых точек изображений, при этом каждый пиксель может принадлежать только одному региону. Далее вычислялись различные дескрипторы для найденных областей. Дескрипторы включали в себя гистограммы каналов изображения в различных представлениях (RGB и HSV), а также норму отличия от направления единичного вектора, например, в RGB это значение говорит о том, насколько серый цвет записан в данном пикселе. Для того чтобы оптимизировать время, затрачиваемое на алгоритм кластеризации, вместо самого дескриптора использовали результат его преобразования с помощью метода главных компонент. Это позволило сократить количество элементов дескриптора до 16 и значительно ускорить вычисления. Применяя алгоритм средних для кластеризации полученных регионов с помощью вычисленных дескрипторов, получили три группы регионов. Далее регионы одной группы объединяли и вычисляли средние значения S канала пикселей этих регионов. Эти значения использовали для того, чтобы определить, какие из этих трех регионов принадлежат фону, а какие нет. В итоге получен алгоритм, достаточно надежный для применения на изображениях, используемых в данной работе. Шаги алгоритма представлены на рис. 3.



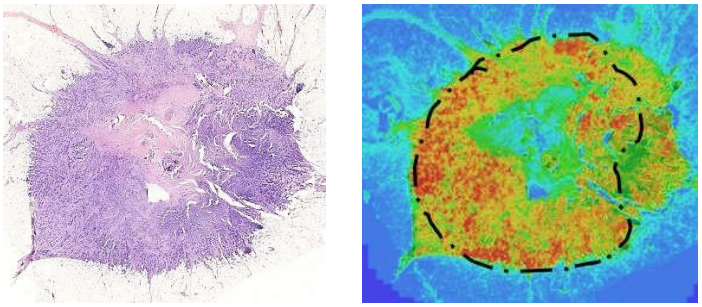
a *b* *c*

Рис. 3. Пошаговые результаты алгоритма выделения областей в полнослайдовом изображении:
a – оригинальное изображение (гематоксилин и эозин); *b* – найденные регионы;
c – объединенные регионы с выделенной областью с образцом ткани
Fig. 3. Step-by-step results of algorithm for selecting regions in a whole-slide image:
a – original whole-slide image (hematoxylin and eosin); *b* – found regions; *c* – combined regions with a selected area with a tissue sample

Результаты и их обсуждение

В данной работе разработан полуавтоматический алгоритм выделения опухолевой области на полнослайдовых гистологических изображениях инвазивного протокового рака молочной железы. Основная идея алгоритма заключается в поиске похожих опухолевых участков с помощью различных дескрипторов и метрик. Сам алгоритм включает следующие шаги: 1) загрузка изображения; 2) разделение изображений на небольшие квадратные регионы заданного размера; 3) вычисление дескрипторов всех регионов и сохранение их в отдельном файле; 4) ручное выделение небольшой части опухолевой области; 4) вычисление дескриптора выделенной области; 5) вычисление расстояний по выбранной метрике между дескриптором выделенной области и дескрипторами регионов изображения; 6) нахождение регионов с минимальным расстоянием до выбранной области; 7) построение тепловой карты похожести по полученным расстояниям. Первым разработанным дескриптором была гистограмма распределения цвета на изображении. Для этого RGB каналы изображения объединялись в один путем слияния значений, то есть значение R (red) канала занимает первых 8 бит числа, значение G (green) канала занимает следующие 8 бит, а значение B (blue) канала – последние 8 бит. После чего происходит построение гистограммы с заданным количеством значений; для эффективности был выбран размер гистограммы в 256 элементов. Далее был разработан улучшенный алгоритм вычисления дескриптора с помощью гистограммы – адаптивная

цветовая гистограмма, которая представляла из себя гистограмму распределения цветов из 256-цветовой палитры и состояла из количества элементов, равного размеру палитры. Последним разработанным дескриптором была матрица совместной встречаемости цветов. Эта матрица была построена так же и использовала палитру цветов. В (*i*, *j*) элемент матрицы заносили количество пикселей цвета, стоящего в палитре на *i*-м месте, которые граничат с пикселем цвета, стоящего в палитре на *j*-м месте. Все дескрипторы нормализовались на разрешение области, на которой считается дескриптор. Использование поиска похожих участков и дескрипторов позволило получить карту похожести (рис. 4) при выбранной области, что при выбранном опухолевом участке позволяет найти похожие на него опухолевые участки в пределах одного изображения.



a *b*

Рис. 4. Пример карты похожести: *a* – срез, окрашенный гематоксилином и эозином; *b* – цветовая карта схожести (выделены опухолевые клетки)

Fig. 4. An example of a similarity map: *a* – a whole-slide image stained with hematoxylin and eosin; *b* – a color similarity map (allocated neoplastic cells)

На данном этапе разработки автоматизированной программы предварительной диагностики рака молочной железы полученный алгоритм может быть использован для обучения врачей-интернов с целью изучения гистологических форм инвазивного протокового рака молочной железы, а также тканевой и клеточной компоновки опухоли в гистологических срезах.

Заключение

В ходе выполнения работы были рассмотрены проблемы, которые возникают при обработке и анализе полнослайдовых гистологических изображений инвазивного протокового рака молочной железы. Учитывая рассмотренные трудности, предложено два алгоритма, решающие некоторые из них. Алгоритм для нахождения цвета стандартных химических маркеров (гематоксилин и эозин), использующихся при окраске ткани, позволяет проводить нормализацию цветового пространства полнослайдового гистологического изображения. Алгоритм сегментации ткани на полнослайдовом изображении позволяет уменьшить область для обработки и сократить время, затрачиваемое на вычисления. В конце приведен алгоритм, который находит опухолевые области полуавтоматическим способом, используя поиск похожих для построения тепловой карты похожести выбранного региона на остальные регионы изображения.

Список литературы / References

1. Dimitriou N., Arandjelovic O., Caie P.D. Deep Learning for Whole Slide Image Analysis: An Overview. *Frontiers of Medicine.* October 19, 2019;1-11. https://www.researchgate.net/publication/336671999.
2. DICOM Whole Slide Imaging. – URL:http://dicom.nema.org/Dicom/DICOMWSI/ (дата обращения 21.10.2020).
3. Macenko M., Niethammer M., Marron J. S., Borland D., Woosley J. T., Guan X., Schmitt C., Thomas N. E. A method for normalizing histology slides for quantitative analysis. *In IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009;1107-1110. DOI: 10.1109/ISBI.2009.5193250.
4. Bankhead P., Loughrey M. B., Fernández J.A., Dombrowski Y., McArt D.G., Dunne P.D., McQuaid S., Gray R.T., Murray L.J., Coleman H.G., James J.A., Salto-Tellez M., Hamilton P.W. QuPath: Open source software for digital pathology image analysis. *Scientific Reports.* 2017;7(1):168-78. DOI:10.1038/s41598-017-17204-5.

Вклад авторов

Семеник И.А. и Деревянко М.А. осуществили подготовку гистологических образцов ткани рака молочной железы, сканирование гистологических препаратов.

Рябцева С.Н. и Москаленко Р.А. провели морфологическую верификацию гистологических срезов, оценили результаты точности выделения зон в полнослайдовых изображениях после использования алгоритма.

Малышев В.Д., Довбыш А.С. и Савченко Т.Р. участвовали в разработке алгоритмов.

Рябцева С.Н. и Малышев В.Д. подготовили рукопись статьи.

Ковалев В.А. и Романюк А.Н. осуществили общее руководство работой и подготовку рукописи статьи.

Authors' contribution

Siamionik I.A. and Derevyanko M.A. prepared the histological tissue samples of breast cancer and scanned the histological slides.

Rjabceva S.N. and Moskalenko R.A. performed morphological verification of histological slides, checked the results of indicators in whole-slide tissue image after using the algorithm.

Malyshev V.D., Dovbysh A.S. and Savchenko T.R. participated in the algorithm development.

Rjabceva S.N. and Malyshev V.D. prepared the manuscript of the article.

Kovalev V.A. and Romaniuk A.N. provided general guidance and preparation of the manuscript.

Сведения об авторах

Рябцева С.Н., к.м.н., заведующий лабораторией «Центр электронной и световой микроскопии» Института физиологии Национальной академии наук Беларуси.

Ковалев В.А., к.т.н., заведующий лабораторией анализа биомедицинских изображений Объединенного института проблем информатики Национальной академии наук Беларуси.

Малышев В.Д., инженер-программист лаборатории анализа биомедицинских изображений Объединенного института проблем информатики Национальной академии наук Беларуси.

Семеник И.А., к.б.н., старший научный сотрудник лаборатории «Центр электронной и световой микроскопии» Института физиологии Национальной академии наук Беларуси.

Деревянко М.А., к.б.н., старший научный сотрудник лаборатории «Центр электронной и световой микроскопии» Института физиологии Национальной академии наук Беларуси.

Москаленко Р.А., д.м.н.. доцент кафедры патологической анатомии Сумского государственного университета.

Довбыш А.С., д.т.н., профессор, заведующий кафедрой компьютерных наук Сумского государственного университета.

Савченко Т.Р., студент кафедры компьютерных наук Сумского государственного университета.

Романюк А.Н., д.м.н., профессор, заведующий кафедрой патологической анатомии Сумского государственного университета.

Information about the authors

Rjabceva S.N., PhD, the Head of Laboratory "Center of Electron and Light Microscopy" of Institute of Physiology of the National Academy of Sciences of Belarus.

Kovalev V.A., PhD, the Head of Laboratory "Biomedical Image Analysis" of United Institute of Informatics Problems of the National Academy of Sciences of Belarus.

Malyshev V.D., Software Engineer of Laboratory "Biomedical Image Analysis" of United Institute of Informatics Problems of the National Academy of Sciences of Belarus.

Siamionik I.A., PhD, Senior Researcher of Laboratory "Center of Electron and Light Microscopy" of Institute of Physiology of the National Academy of Sciences of Belarus.

Derevyanko M.A., PhD, Senior Researcher of Laboratory "Center of Electron and Light Microscopy" of Institute of Physiology of the National Academy of Sciences of Belarus.

Moskalenko R.A., D.Sci, Associate Professor of the Pathological Anatomy Department of Sumy State University.

Dovbysh A.S., D.Sci, Professor Professor of the Computer Sciences Department of Sumy State University.

Savchenko T.R., Student of the Computer Sciences Department of Sumy State University.

Romaniuk A.N., D.Sci, Professor of the Pathological Anatomy Department of Sumy State University.

Адрес для корреспонденции

220072, Республика Беларусь,
г. Минск, ул. Академическая, 28,
Институт физиологии Национальной академии наук Беларуси
тел. +375-29-315-65-21;
Рябцева Светлана Николаевна

Address for correspondence

220072, Republic of Belarus
Minsk, Academicheskaya str., 28,
Institute of Physiology of the National Academy of Sciences of Belarus
tel. +375-29-315-65-21;
Rjabceva Svetlana Nikolaevna