AUTOMATYKA • 2008 • Tom 12 • Zeszyt 3

Jarosław Gocławski*, Joanna Sekulska-Nalewajko*, Patryk Anioł**

Metoda automatycznego wyznaczania indeksu mitotycznego populacji komórek cebuli z wykorzystaniem drzewa decyzyjnego

1. Wprowadzenie

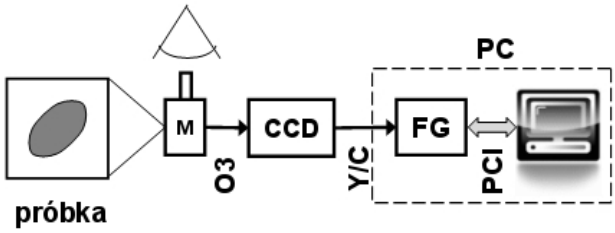
Mitoza jest procesem przemian w jądrze komórkowym i cytoplazmie zmierzającym do wytworzenia komórek potomnych, będących dokładnymi kopiami komórek somatycznych organizmów. Najważniejsze przemiany w trakcie mitozy zachodzą w jądrze komórkowym, którego materiał ulega organizacji w chromosomy, rozdzielane następnie między komórki potomne. Za pomocą mikroskopu można zaobserwować zmianę liczby składników jądra komórkowego, ich kształtu oraz położenia. Obserwacje występowania mitozy i jej etapów umożliwiają pomiar aktywności mitotycznej w populacji komórek. Wiedza na temat intensywności podziałów komórek jest wykorzystywana do oceny wpływu różnych czynników na stan organizmów i ich zdolność do przeżycia w warunkach skażeń chemicznych lub promieniowania jonizującego [1, 2, 3]. Elementarnym wskaźnikiem aktywności mitotycznej populacji komórek jest indeks mitotyczny związany z udziałem procentowym komórek danej populacji podlegających mitozie [4].

Istnieje kilka rozwiązań komercyjnych umożliwiających automatyczny pomiar indeksu mitotycznego [5], jednakże wewnętrzne algorytmy ich działania nie są udostępnione. Wobec różnych sposobów przygotowania preparatów, w zależności od badanych tkanek, należy stosować różne metody automatycznej obróbki obrazu w szczególności w zakresie segmentacji [6]. Autorzy zaproponowali określone rozwiązania segmentacji, ekstrakcji cech oraz klasyfikacji dostosowane do preparatów tkanki merystematycznej cebuli (*Allium cepa*), powszechnie wykorzystywanych w badaniach naukowych w Katedrze Cytologii i Cytochemii Uniwersytetu Łódzkiego.

2. Stanowisko badawcze

System akwizycji obrazów składa się z mikroskopu optycznego OLYMPUS CH20 z halogenowym źródłem światła, pracującego w jasnym polu, kamery CCD i komputera osobistego PC z kartą akwizycji FG podłączoną do magistrali PCI (rys. 1).

* Katedra Informatyki Stosowanej, Politechnika Łódzka
** Katedra Cytologii i Cytochemii Roślin, Uniwersytet Łódzki



Rys. 1. Schemat blokowy stanowiska pomiarowego. M – mikroskop optyczny, O3 –trinokular, CCD – kamera CCD z interfejsem Y/C, FG – karta akwizycji obrazu, PC – komputer z oprogramowaniem

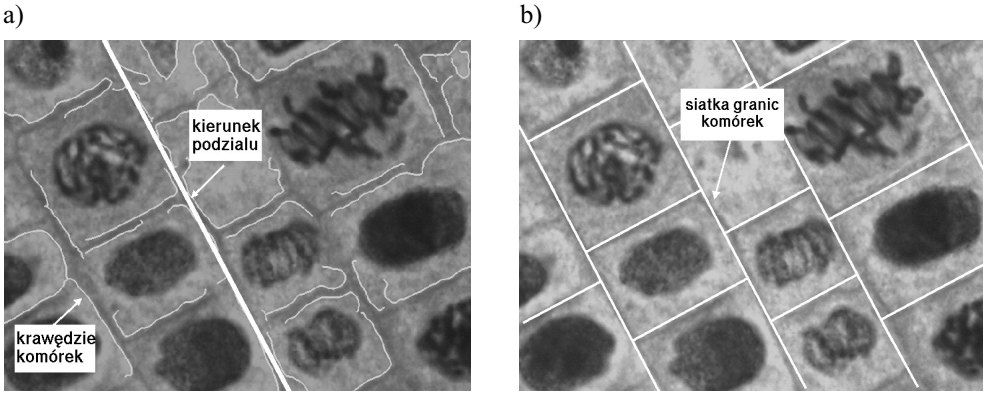
Światło z mikroskopu przechodzące przez cienką warstwę badanego preparatu i układ optyczny z obiektywem 100× tworzy obraz RGB na matrycy CCD kamery. Kamera podłączona jako trzeci okular mikroskopu wyprowadza obraz PAL (768×576) w formie sygnału analogowego Y/C do karty grabbera w komputerze.

Preparaty stanowią przekroje merystemu korzeniowego cebuli. Stolik mikroskopowy w chwili obecnej pozycjonuje się manualnie w kierunkach *X*, *Y*, *Z* w celu wprowadzenia w pole widzenia i zogniskowania kolejnych obiektów jąder komórkowych. Obrazy odwzorowane podczas akwizycji są zapisywane do plików bitmapy w celu dalszej obróbki. Segmentacja komórek, jąder komórkowych lub ich składowych podczas mitozy oraz rozpoznawanie udziału jąder interfazowych w badanej populacji zostały przeprowadzone za pomocą zestawu procedur opracowanego dla środowiska MATLAB R14.

3. Segmentacja obrazów z preparatów *Allium cepa*

Badane preparaty, wybarwione metodą Feulgena, zapewniają tłumienie światła przechodzącego przez nie, proporcjonalnie do zawartości DNA w danym miejscu próbki. Wobec tego, jądra komórkowe lub ich oderwane fragmenty z chromosomami zawierającymi DNA są widoczne na obrazach preparatów jako obszary ciemniejsze od otaczającego je tła. Odpowiedni dobór barwnika umożliwia także uwidocznienie większości ścian komórek jako ciemnych krawędzi, ale skontrastowanych słabiej w porównaniu ze skupieniami chromosomów jąder. Komórki pochodzące z centralnych części przekrojów korzeniowych mają kształt prawie prostokątny i dzielą się podczas mitozy, z zachowaniem określonego kierunku w obrębie pola jednego obrazu przy stosowanym powiększeniu mikroskopu (rys. 2).

Autorzy opracowali wcześniej metodę wyznaczania kierunku podziału komórek w obrazie oraz wyodrębniania poszczególnych komórek [7]. Wykorzystuje ona wykrywanie krawędzi komórek [8] (rys. 2) za pomocą detektora Canny'ego [9] przy równoczesnym maskowaniu krawędzi elementów jąder. W drugim kroku tej metody przeprowadza się segmentację skupisk chromosomów tworzących jądra lub ich fragmenty. Podstawę jej stanowi lokalne progowanie pod maską krawędzi elementów jądra w obszarze ROI każdej komórki.



Rys. 2. Przykładowy obraz preparatu *Allium*: a) z wykrytymi krawędziami poszczególnych komórek i wyznaczonym kierunkiem podziału komórkowego; b) z wykrytymi automatycznie granicami komórek

W ramach segmentacji przeprowadza się także redukcję artefaktów wynikających z dwóch przyczyn:

1) zetknięcia się jądra ze ścianą komórki,
2) nakładania się obrazu badanego jądra z cieniami innych jąder spoza płaszczyzny ostrości.

4. Algorytmy uczenia faz mitozy oraz wyznaczania indeksu mitotycznego

Indeks mitotyczny opisuje poziom (stopień) mitozy w badanej populacji komórek. Definiuje się go jako [4, 6]:

$$\mathrm{MI} = \frac{1}{|S|}\sum_{i}\delta\left(C_i\left(S_i\right) \neq I\right) \tag{1}$$

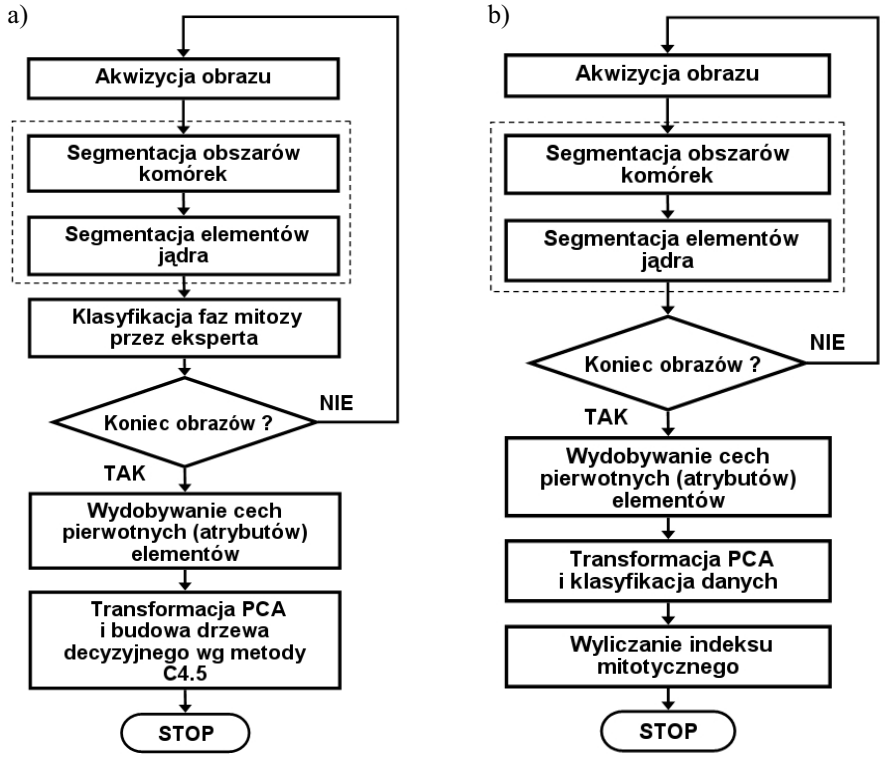
gdzie:

$\delta(w_i)$ – operator konwersji wartości *false* abo *true* wyrażenia logicznego w_i, na odpowiednią wartość liczbową 0 albo 1,
$C_i(S_i)$ – klasa przypisana i-temu elementowi $S_i \in S$ zbioru danych opisujących jądra komórek,
I – klasa oznaczająca interfazę,
$|S|$ – liczność zbioru S.

Aby wyznaczyć indeks MI powinno się policzyć wszystkie komórki zawierające jądra i rozpoznać oraz zliczyć te, które są w stadium mitozy tj. poza interfazą. Ponieważ klasy mitozy są znane i dobrze rozróżnialne wizualnie, trzeba zdefiniować charakterystyczne dla

nich cechy (atrybuty). Dane opisujące poszczególne jądra komórek to zestawy wartości tych atrybutów. Należy nauczyć automatyczny system rozpoznający interpretowania tych wartości.

Na rysunkach 3a i 3b przedstawiono główne moduły proponowanych algorytmów uczenia i klasyfikacji wraz ze wstępnym przetwarzaniem danych obrazowych. Omawiane w rozdziale 3 segmentacje zostały objęte wspólnym blokiem oznaczonym linią przerywaną. Podczas uczenia faz mitozy ekspert zaznacza dowolny element jądra komórki wymaskowanego w procesie segmentacji i wybiera z listy przypisywaną mu fazę (jedną z siedmiu). Na obecnym etapie prac nie uwzględnia się komórek z jądrami nierozpoznawalnymi, tylko usuwa je z listy danych.



Rys. 3. Przebiegi algorytmów: a) uczenia faz mitozy; b) estymacji indeksu mitotycznego MI

Po automatycznej ekstrakcji atrybutów wykonywana jest ich transformacja metodą PCA [10, 11] i ich redukcja do czterech składowych głównych. Algorytm uczący buduje binarne drzewo decyzyjne na podstawie transformowanych cech (wtórnych). Drzewo to jest wykorzystywane w algorytmie klasyfikującym (rys. 3b) do rozpoznania stanów interfazy lub mitozy (bez rozróżniania jej poszczególnych stadiów). Ostatecznie wyliczany jest indeks mitotyczny według równania (1).

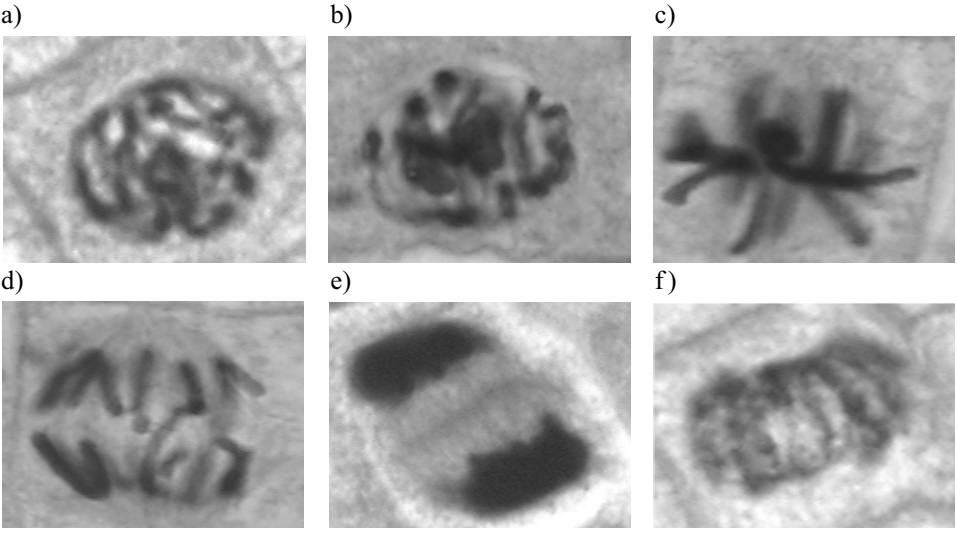
Drzewa decyzyjne uważa się za odpowiednie narzędzia do klasyfikacji różnych typów danych pozyskiwanych z preparatów biologicznych komórek, okrzemek i różnych innych mikroorganizmów [12]. Umożliwiają one osiąganie względnie małych błędów klasyfikacji, zwłaszcza przy niewielkich ilościach danych i dużej liczbie atrybutów, w porównaniu z metodami sieci neuronowych lub klasyfikatorami statystycznymi. Przy rozwiązywaniu problemu identyfikacji stadiów mitozy podejście wykorzystujące drzewa decyzyjne naśladuje sposób rozumowania ekspertów wykonujących to samo zadanie.

5. Ekstrakcja cech jąder komórkowych

Aby rozróżnić fazy mitozy i interfazę, autorzy wybrali określone cechy topologiczne, geometryczne i teksturalne, jakimi można scharakteryzować jądro komórkowe lub jego poszczególne elementy – skupienia chromosomów. Są to również cechy brane pod uwagę przez specjalistów podejmujących decyzję o klasyfikacji faz jąder komórkowych. Jako podstawę do wydobywania cech wzięto obrazy luminancji preparatu komórkowego i binarne maski całych jąder lub ich elementów po segmentacji, przypisane poszczególnym komórkom.

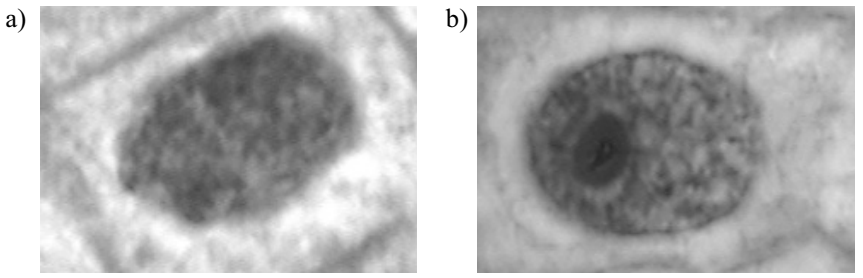
Zgodnie ze stanem wiedzy biologów przyjęto model zakładający sześć klas jąder komórkowych [13] odpowiadających poszczególnym fazom mitozy: wczesnej i późnej profazie $\{P_1, P_2\}$, metafazie $\{M\}$, anafazie $\{A\}$ oraz wczesnej i późnej telofazie $\{T_1, T_2\}$ (rys. 4).

$$C = \{I, P_1, P_2, M, A, T_1, T_2\} \tag{2}$$



Rys. 4. Przykładowe obrazy jąder *Allium* w różnych stadiach mitozy: a) wczesna profaza P_1; b) późna profaza P_2; c) metafaza M; d) anafaza A; e) wczesna telofaza T_1; f) późna telofaza T_2

Siódmą klasę reprezentują jądra interfazowe {I} występujące pomiędzy cyklami mitozy (rys. 5). Klasy te wymieniono w równaniu (2).



Rys. 5. Przykładowe obrazy jąder *Allium* w stadium interfazy: a) bez jąderka; b) z widocznym pojedynczym jąderkiem

Na podstawie obrazów źródłowych jąder i ich binarnych masek łatwo zauważyć, że jądra lub ich elementy w interfazie, wczesnej profazie i wczesnej lub późnej telofazie mają stosunkowo słabo rozwinięte brzegi oraz posiadają kształty zbliżone do elipsy w odróżnieniu od pozostałych rodzajów faz. Ponadto jądra w stadium interfazy oraz wczesnej telofazy nie zawierają wewnątrz istotnych „prześwitów" wynikających z rozsuwania się w procesie podziału mitotycznego ciemno wybarwionych chromosomów. Obserwowane wahania jasności w ich wnętrzu wynikają z lokalnych wahań koncentracji chromatyny. Najsilniejsze wahania tego typu dotyczą jąderek – ultraelementów odpowiedzialnych za syntezę RNA. Występują one w postaci od jednego do trzech obszarów o kształcie elipsoidalnym, strukturalnie jednorodnych i wyraźnie ciemniejszych od otaczającego je materiału jądra (rys. 5b). Ponieważ mają one wpływ na charakterystykę zmian jasności wewnątrz jądra powinny być rozróżniane poprzez dodatkową segmentację. Poza interfazą, jąderka można zaobserwować także w stadium wczesnej profazy i późnej telofazy. Tylko interfaza, wczesna profaza i późna telofaza charakteryzują się istnieniem jednego obiektu jądra. W pozostałych fazach obserwuje się wiele elementów zdezintegrowanego jądra – oddzielnych skupisk chromosomów ostatecznie koncentrujących się na dwóch przeciwnych końcach komórki w celu utworzenia jąder potomnych.

Ilość obserwowanych elementów jądra – skupisk chromosomów stanowi wobec tego jedną z cech wyróżniających klasy odpowiadające wzmiankowanym fazom.

Na podstawie powyższych spostrzeżeń wytypowano niżej wymienione podstawowe cechy inwariantne względem skali, translacji i rotacji obrazu stosowane w zagadnieniach przetwarzania obrazów [12, 14].

– Liczba elementów NE:
  Ilość skupień materiału jądrowego w obrębie komórki. Odpowiada liczbie Eulera wobec eliminacji „dziur" w obiektach. Interfaza, wczesna profaza i późna telofaza z założenia odpowiadają jednoelementowym skupieniom. Na etapie segmentacji usuwane są artefakty o zbyt małej powierzchni i zbyt dużej jasności wewnętrznej.

– Eliptyczność EL [12]:
Wskaźnik oceniający podobieństwo kształtu elementów jądra do elipsy. Zastosowano go w odniesieniu do największego elementu jądra w każdej komórce.

$$\mathrm{EL} = \begin{cases} I_1 \cdot 16\pi^2, & \text{jeżeli} \quad I_1 < 1/\left(16\pi^2\right) \\ 1/\left(I_1 \cdot 16\pi^2\right), & \text{przeciwnie} \end{cases}, \quad I_1 = \frac{\mu_{20}\,\mu_{02} - \mu_{11}^2}{\mu_{00}^4} \tag{3}$$

gdzie: $\mu_{pq} - p, q \in [0,1]$ – momenty centralne 2 rzędu dla maski binarnej elementu. Wskaźnik eliptyczności mieści się w zakresie [0,1] i osiąga wartość 1 dla doskonałej elipsy.

– Zwartość CP [8, 14]:
Stanowi miarę odległości kształtu elementu jądra w badanej komórce od kształtu koła. Dla koła CP = 1, w innych przypadkach jest odpowiednio większe.

$$\mathrm{CP} = \frac{L^2}{4\pi \cdot P} \tag{4}$$

gdzie:
L – obwód maski binarnej elementu jądra,
P – pole powierzchni maski elementu.

– Odchylenie od elipsy ED:
Wskaźnik zaproponowany przez autorów, zbliżony do EL. Ocenia on średnią odległość konturu badanego obiektu od elipsy najlepiej do niego dopasowanej w sensie średniokwadratowym [15]. Wyliczana odległość nie odpowiada mierze euklidesowej, ale jest z nią zgodna. Wskaźnik ten uzyskuje się ze wzoru:

$$\mathrm{ED} = \frac{mean\left|f_1 + f_2 - 2a\right|}{a+b} \tag{5}$$

gdzie:
a, b – duży i mały promień elipsy dopasowanej do elementu jądra
f_1, f_2 – ogniskowe dopasowanej elipsy.

Rozważane cechy teksturalne odnoszą się do wszystkich elementów jądra wewnątrz komórki. Ich definicje oparte są o dane uzyskane z macierzy współwystępowania poziomów jasności GLCM (*Gray Level Cooccurence Matrix*) [16, 17] dla obrazu luminancji o zredukowanej liczbie poziomów jasności. Ponadto korzysta się z macierzy kodowania bieżącej długości GLRL (*Gray Level Run Length Matrix*) [18]. Macierze GLCM i GLRL dotyczą jedynie obszarów wymaskowanych elementów jądra. Jest to realizowane poprzez przesunięcie przyjętego zakresu jasności elementów powyżej poziomu odniesienia 0 zarezerwowanego dla maski podłoża. Następnie pierwszy wiersz i kolumna macierzy GLCM oraz pierwszy wiersz GLRL, wyrażające kontakt z tą maską, są eliminowane z dalszych rozważań. Wykorzystuje się przy tym funkcje MATLAB *graycomatrix* z pakietu *Image Processing Toolbox* oraz *grayrlmatrix* z pakietu uzupełniającego [19].

Po uzgodnieniu z ekspertami przyjęto, że obserwowane zmiany jasności wewnątrz elementów jądra, poza pewnymi wyjątkami, nie wykazują zachowań kierunkowych. Wobec tego uśredniono macierz GLRM dla kierunków: 0^o, 45^o, 90^o, 135^o, oraz macierz GLRL dla kierunku poziomego i pionowego. Przy podanych założeniach wzięto pod uwagę następujące wskaźniki tekstury [20]:

– Kontrast CT:

$$\mathrm{CT} = \sum_{i,j} |i-j|^2 P_D(i,j) \tag{6}$$

gdzie:

P_D – uśredniona kierunkowo macierz GLCM dla odległości D = 10 pikseli,
i, j – indeksy macierzy P_D z zakresu [1, 64].

– Jednorodność H (*Homogeneity*):

$$\mathrm{H} = \sum_{i,j} \frac{P_D(i,j)}{1+|i-j|} \tag{7}$$

gdzie:

P_D – uśredniona kierunkowo macierz GLCM dla D = 10 pikseli,
i, j – indeksy macierzy P_D z zakresu [1, 64].

– Nierównomierność bieżącej długości RLN (*Run Length Nonuniformity*):

$$\mathrm{RLN} = \frac{1}{N_R} \sum_{j=a}^{b} \left(\sum_{i=1}^{M} P_R(i,j) \right)^2 \tag{8}$$

gdzie:

P_R – uśredniona kierunkowo macierz bieżących długości,
$[a, b]$ – badany zakres bieżących długości P_R (a = 5, b = 10),
N_R – suma wartości elementów macierzy P_R w podanych zakresach,
M – ilość poziomów jasności dla analizy (M = 8).

6. Budowa drzewa decyzyjnego

Autorzy napisali procedurę budowy drzewa decyzyjnego opartą na algorytmie C4.5 [21, 22, 23], niezależnie od dostępnej w środowisku MATLAB (*Statistics Toolbox*) [10] funkcji *treefit* dysponującej kilkoma innymi kryteriami dyskryminacji w węzłach. Jest ona dostosowana do konkretnego zadania rozpoznawania stanu mitozy i umożliwia kontrolę warunków wytwarzania liści, włączając w to odpowiednie przycinanie drzewa w trakcie jego budowy. Została napisana w „M" kodzie z intencją późniejszego jej przetłumaczenia na język C. Zgodnie z algorytmem C4.5, progi dyskryminacji w węzłach są wyznaczane na podstawie tzw. współczynnika wzmocnienia informacyjnego IGR (*Information Gain Ratio*).

W określonym węźle drzewa wylicza się, dla każdego z atrybutów danych A i każdego z możliwych położeń progu dyskryminacji T w jego dziedzinie S_A, współczynniki wzmocnienia informacyjnego IGR [12, 21]. Następnie wybiera się próg dyskryminacji T atrybutu A o największym współczynniku IGR_M (równ. (9)).

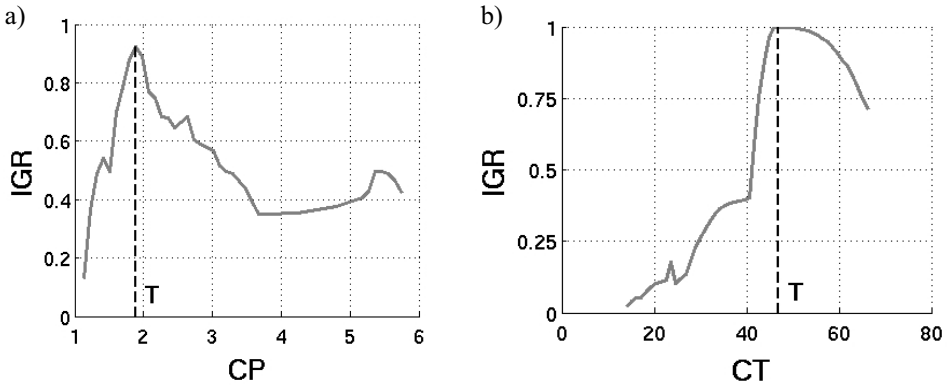
$$IRG_M(T_M, A) = \max_T \frac{I(S_A) - I_T(T, S_A)}{I_S(T, S_A)} \tag{9}$$

gdzie:

$I(S_A)$ – entropia klas w zbiorze S_A danych atrybutu A przed podziałem,

$I_T(T, S_A)$ – suma ważona entropii klas w zbiorach S_{A1}, S_{A2} po podziale progiem T,

$I_S(T, S_A)$ – entropia podziału danych progiem T.



Rys. 6. Przykładowe przebiegi *IGR* w węzłach drzewa decyzyjnego z rysunku 7: a) zwartości CP na poziomie 1; b) kontrastu CT na poziomie 4

Rysunki 6a i 6b przedstawiają przykładowe przebiegi współczynników IGR dla drzewa z rysunku 7.

Poniżej pokazano przebieg zastosowanego algorytmu budowy drzewa. W poniższym opisie $|S|$ oznacza liczność zbioru danych S, $k_I = 1\%$ – progowy błąd rozpoznawania interfazy, $k_S = 1\%$ – próg przycinania drzewa wyrażony poprzez różnicę względną błędów klasyfikacji węzła N i jego potomków N_1, N_2. Symbole I oraz $\sim I$ opisują przynależność danej opisanej zbiorem atrybutów odpowiednio do klasy interfazy albo mitozy.

Drzewo (S)

Krok 1. Jeśli jest spełniony warunek (10) to: oznacz węzeł jako liść, przypisz mu klasę dominującej fazy I albo $\sim I$, wyjdź z procedury.

$$|I| < k_I |S| \quad \text{lub} \quad |I| \geq (1 - k_I)|S| \tag{10}$$

Krok 2. Oblicz dla każdego atrybutu A wartości maksymalne IGR_M w zbiorze danych S_A i odpowiadające im progi dyskryminacji T_M.

$$IGR_M\left(T_M, A\right) = \max_T \left(IGR\left(T, A\right)\right) \tag{11}$$

Krok 3. Wybierz atrybut A z największą wartością IGR_{MAX} i odpowiadający mu próg dyskryminacji T_{MAX}.

$$IGR_{MAX}\left(T_{MAX}\right) = \max_A \left(IGR_M\left(T_M, A\right)\right) \tag{12}$$

Krok 4. Podziel zbiór danych S na podzbiory $S_1 = \{S: S_A = T_{MAX}\}$, $S_2 = \{S: S_A > T_{MAX}\}$ za pomocą progu T_{MAX} wybranego atrybutu A.

Krok 5. Jeżeli $S_1 = \varnothing$ albo $S_2 = \varnothing$, to: oznacz węzeł jako liść, przypisz mu klasę dominującej fazy I albo $\sim I$, wróć z procedury.

Krok 6. Jeżeli jest spełniony warunek przycinania (13), to: oznacz węzeł jako liść, przypisz mu klasę dominującej fazy I albo $\sim I$, wróć z procedury. E_S, E_{S1}, E_{S2} – ilości błędnie sklasyfikowanych danych w zbiorze S i jego potomkach.

$$\frac{E_S - E_{S_1} - E_{S_2}}{|S|} < k_S \tag{13}$$

Krok 7. Zapamiętaj bieżący węzeł drzewa N do celów klasyfikacji. N_1, N_2 oznaczają węzły potomne odpowiadające zbiorom S_1, S_2.

$$N(T_{MAX}, N_1, N_2) \tag{14}$$

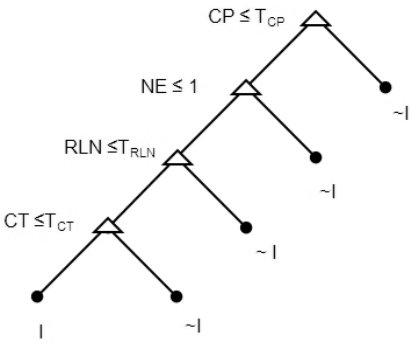
Krok 8. Wywołaj rekurencyjnie:

$$\begin{aligned} &Drzewo(S_1) \\ &Drzewo(S_2) \end{aligned} \tag{15}$$

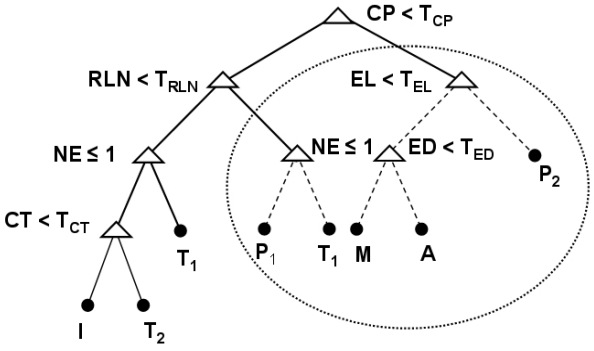
Krok 9. Wróć.

Algorytm proponowany przez autorów pozwala na natychmiastowe uzyskanie drzewa przyciętego metodą REP (*Reduced Error Pruning*) [24] poprzez sprawdzanie warunku (13) rozrostu drzewa.

Dla porównania, na rysunku 8 przedstawiono drzewo utworzone na zbiorze testowym S metodą *treefit* programu MATLAB, przyjmując jako kryterium dyskryminacji metodę dewiancji statystycznej. Liczby poziomów niezbędnych do oddzielenia interfazy i struktury obydwu uzyskanych drzew są podobne.



Rys. 7. Proponowane drzewo decyzyjne dla zbioru testowego *S* do klasyfikacji stadiów interfazy *I* oraz mitozy *~I*



Rys. 8. Drzewo decyzyjne uzyskane metodą *treefit* do klasyfikacji stadiów interfazy *I* oraz mitozy *~I* z wykorzystaniem reguły podziału przez dewiancję statystyczną (*statistical deviance*)

7. Klasyfikacja i wyznaczanie indeksu MI

Błąd klasyfikacji dla proponowanego drzewa klasyfikacji został oceniony przy użyciu metody 10-krotnej walidacji skrośnej [10, 12].

Cały dostępny zbiór danych *S* o liczności $|S|$ opisany przez siedem atrybutów $A = \{\text{NE, EL, CP, ED, CT, H, RLN}\}$ podzielono losowo na $k = 10$ wzajemnie rozłącznych podzbiorów $\{S_1, S_2, \ldots, S_k\}$, o zbliżonych rozmiarach. Klasyfikator był trenowany i testowany 10 razy; za *i*-tym razem ($1 = i = 10$) trenowany na zbiorze $S \backslash S_i$ i testowany na S_i. Do podziału losowego danych na grupy wykorzystano funkcję MATLAB-a [25]:

$$indeksy = crossvalind('Kfold', [S], k) \tag{16}$$

gdzie:

$[S]$ – zbiór danych w formie tablicy z kolumnami atrybutów (cech),
$k = 10$ – liczba grup danych,
indeksy – wektor indeksów grup danych.

Za miarę błędu klasyfikacji przyjęto liczbę błędnych klasyfikacji odniesioną do liczby próbek w zbiorze danych [26]:

$$e = \frac{1}{|S|}\sum_{i=1}^{|S|}\delta\left(C_i\left(S\right) \neq C_i'\ \left(S\right)\right) \tag{17}$$

gdzie:

$\delta(w_i)$ – operator konwersji wartości *false* abo *true* wyrażenia logicznego w_i, na odpowiednią wartość liczbową 0 albo 1,
$|S|$ – liczność zbioru S klasyfikowanych komórek,
C_i, C_i' – wartości klas rzeczywistej oraz rozpoznanej w dziedzinie $\{I, \sim I\}$ przypisane i-tej danej zbioru S.

Estymowany indeks mitotyczny MI′:

$$\mathrm{MI}' = \frac{1}{|S|}\sum_{i}\delta\left(C_i'\ \left(S\right) \neq I\right) \tag{18}$$

gdzie:

$\delta(w_i)$ – operator konwersji wartości *false* abo *true* wyrażenia logicznego w_i, na odpowiednią wartość liczbową 0 albo 1,
$C_i'(S)$ – wynik klasyfikacji przypisany i-temu elementowi zbioru S,
I – klasa oznaczająca interfazę,
$|S|$ – liczność zbioru S klasyfikowanych komórek.

Jeżeli postawić hipotezę [27] o przynależności danej komórki *Allium* do stadium interfazy, to łatwo zauważyć, że błąd klasyfikacji (równ. (17)) jest sumą błędów pierwszego i drugiego rodzaju dla tej hipotezy. Z równań (1), (17), (18) wynika, że przyjmowany standardowo błąd pomiaru indeksu mitotycznego e_{MI} (równ. (19))

$$e_{MI} = \left|\mathrm{MI}' - \mathrm{MI}\right| \tag{19}$$

jest różnicą obu typów błędów. Należy więc oczekiwać, że będzie on niższy niż e, zwłaszcza dla populacji I oraz $\sim I$ z podobnym prawdopodobieństwem a'priori.

8. Wyniki pomiarów i wnioski

Przeprowadzono wyznaczanie indeksu mitotycznego dla populacji komórek *Allium* o liczności $|S| = 172$ próbki. Wzięto pod uwagę prezentowany wyżej algorytm C4.5 i proponowany zestaw siedmiu cech pierwotnych $A = \{\mathrm{NE}, \mathrm{EL}, \mathrm{CP}, \mathrm{ED}, \mathrm{CT}, \mathrm{H}, \mathrm{RLN}\}$. Uzyskane wyniki przedstawiono w tabeli 1.

Tabela 1
Wyniki dwóch serii 10 prób 10-krotnych walidacji skrośnych z użyciem procedury na bazie C4.5

Nr walidacji	Cechy pierwotne			PCA, 4 komponenty główne		
	Błąd klasyfikacji [%]	Liczba węzłów drzewa	Błąd estymatora MI′ [%]	Błąd klasyfikacji [%]	Liczba węzłów drzewa	Błąd estymatora MI′ [%]
1	8,14	2÷5	0	4,65	2÷4	0
2	6,98	3÷6	0	4,65	3÷5	0
3	6,98	2÷6	0	5,81	3÷4	0
4	6,98	3÷ 7	0	5,81	2÷7	0
5	5,81	2÷6	0	4,65	3÷6	0
6	8,14	3÷5	0	5,81	1÷4	0
7	8,14	3÷6	0	5,81	3÷4	0
8	6,98	3÷5	0	4,65	2÷4	0
9	6,98	2÷6	0	5,81	2÷5	0
10	4,65	3÷6	0	4,65	3÷5	0
Średnio	6,98	–	0	5,23	–	0

MI=0.5116

Do rozpatrywanych danych zastosowano także metodę PCA [11]. Wyznaczono składowe główne dla wektora cech pierwotnych jako wektory własne ich macierzy kowariancji. Dokonano projekcji danych na przestrzeń wektorów własnych. Wykorzystano funkcję *princomp* z pakietu *StatisticsToolbox* [10] środowiska MATLAB.

Wyniki klasyfikacji dla czterech głównych cech (o największych wartościach własnych) zobrazowano w tabeli 2.

Dla 10 prób 10-krotnych walidacji uzyskano średni błąd klasyfikacji o 1,75% niższy niż w przypadku pełnego zestawu cech pierwotnych. Podobny poziom błędów klasyfikacji otrzymano dla drzewa z wykorzystaniem reguły dewiancji statystycznej do wyznaczania progów decyzyjnych w węzłach (tab. 2).

Błąd estymacji indeksu mitotycznego e_{MI} z przyczyn wyjaśnionych w poprzednim rozdziale jest niewykrywalny, tzn. poniżej 0,6% dla populacji $|S| = 172$ komórek.

Wyniki pomiarów wskazują na możliwość skutecznego pomiaru indeksu mitotycznego populacji komórek z wykorzystaniem proponowanych atrybutów. Czas wykonywania algorytmu C4.5, w porównaniu z metodami wbudowanymi MATLAB-a, jest dłuższy (tab. 3) z uwagi na wykorzystywanie przy jego budowie języka ”M” środowiska MATLAB-a i brak optymalizacji zapisów macierzowych. Po jej zastosowaniu lub zakodowaniu proce-

dur w języku C czas ten ulegnie istotnemu skróceniu. Jeżeli bierze się pod uwagę całość prezentowanej metody (rys. 3a, 3b), to dużo większy nakład czasu przypada na wykonywanie procedur segmentacji komórek, elementów jąder i ekstrakcji cech.

Tabela 2
Wyniki serii 10 prób 10-krotnych walidacji skrośnych przy użyciu funkcji *treefit*, podziału na bazie dewiancji statystycznej i redukcji do 4 komponentów głównych metodą PCA

Nr walidacji	Błąd klasyfikacji [%]	Błąd estymatora MI´ [%]
1	6,98	0
2	3,49	0
3	4,65	0
4	6,98	0
5	6,98	0
6	6,98	0
7	4,65	0
8	5,81	0
9	5,81	0
10	6,98	0
Średnio	5,93	0

MI=0.5116

Tabela 3
Porównanie średnich czasów wykonywania serii 10 prób 10-krotnych walidacji skrośnych dla różnych algorytmów uczenia/ klasyfikacji

Metoda	Średni czas uczenia [s]	Średni czas klasyfikacji [s]
Metoda C4.5 bez PCA	4,93	0,05
Metoda C4.5 z PCA(4)	4,01	0,05
Dewiancja statystyczna bez PCA	0,71	0,12
Dewiancja statystyczna z PCA (4)	0,53	0,12

Dewiancja statystyczna – funkcja *treefit*, Pentium 4, 3 GHz

Aby skutecznie wydobywać cechy teksturalne interfazy eliminuje się obszary jąderek z rozważanych powierzchni. W naszych badaniach, eliminując jąderka, posłużyliśmy się informacją *a priori* o przynależności jądra do klasy interfazy. W warunkach praktycznych brak tej informacji należy zastąpić pomocniczą klasyfikacją z wykorzystaniem drzewa zawierającego atrybuty inne niż teksturalne. Będzie to powiększać błędy klasyfikacji stanów mitozy. Do pełnej automatyzacji pomiarów MI należy zastosować automatyczny napęd stolika w kierunkach *X*, *Y*, *Z*. Aby uzyskać ostre krawędzie jąder komórkowych przy powiększeniu obiektywu 100×, trzeba kilkakrotnie zmieniać płaszczyznę ostrości w tym samym polu widzenia.

W przyszłości autorzy zamierzają zoptymalizować przebieg algorytmu rozpoznawania, aby poprawić jego wydajność. Należałoby także zastosować kartę akwizycji obrazu współpracującą ze środowiskiem MATLAB i zapewniającą przesyłanie obrazu bezpośrednio do jego przestrzeni pamięci. Ponadto dołączone będą inne cechy elementów jąder umożliwiające skuteczną klasyfikację poszczególnych stadiów mitozy.

Literatura

[1] Kovalchuk O., Kovalchuk I., Arkchipov A., Telyuk P., Hohn B., Kovalchuk L., *The Allium cepa chromosome aberration test reliably measures genotoxicity of soils of inhabited areas in the Ukraine contaminated by the Chernobyl accident*. Mutation Research, 415, 1998, 47–57.

[2] Johnson K.L., Nath. J., Pluth J.M., Tucker J. D., *The distribution of chromosome damage*, *non-reciprocal translocations and clonal aberrations in lymphocytes from Chernobyl clean-up workers*. Mutation Research, 439, 1999, 77–85.

[3] Staykova T.A., Ivanova E.N., Velcheva I.G., *Cytogenetic effect of heavy-metal and cyanide in contaminated waters from the region of southwest Bulgaria*. Journal of Cell and Molecular Biology, 4, 2005, 41–46.

[4] Santiago F.I., Cannen R.E.L., *Mitotic index and chromosomal changes in Allium cepa as affected by an organophosphate pesticide*, *malathion*. Philippine Journal of Science, Vol. 128 (1), 1999, 49–54.

[5] *Cellscan* (*22.04.2008*). http://www.imstarsa.com/ productsservices/ ondemandplatforms/cellscan/.

[6] Sundblad L., Geladi P., Dunberg A., Sundberg B., *The use of image analysis and automation for measuring mitotic index in apical conifer meristems*. Journal of Experimental Botany, 49 (327), 1749–1756.

[7] Goclawski J., Sekulska-Nalewajko J., Aniol P., *The segmentation of meristematic Allium cell images and extraction of nuclei features for the purpose of mitotic index evaluation*. Proceedings of the IVth International Conference of Young Scientists MEMSTECH' 2008, Lviv-Polyana 2008, 123–128.

[8] Tadeusiewicz R., Korohoda P.: *Komputerowa analiza i przetwarzanie obrazów.* Kraków, Wydawnictwo Fundacji Postępu i Telekomunikacji 1997.

[9] Canny J., *A Computational Approach to Edge Detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 6, 1986, 679–698.

[10] The Mathworks, *Statistics Toolbox User's Guide*. www.maths.lth.se/matstat/staff/ krys/ Program/ stats_tb.pdf.

[11] Jolliffe I.T., *Principal Component Analysis*. 2nd Edition, Springer, 2002.

[12] Du Buf H., Bayer M., *Automatic Diatom Identification*. World Scientific Publishing 2002.

[13] Woźny A., Michejda J., Ratajczak L., *Podstawy biologii komórki roślinnej*. Poznań, Wydawnictwo UAM 2001.

[14] Malina W., Smiatacz M., *Metody cyfrowego przetwarzania obrazów*. Warszawa, Akademicka Oficyna Wydawnicza EXIT 2005.
[15] Fitzgibbon A., Pilu M., Fisher B., *Direct Least Squares Fitting of Ellipses*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(5), May 1999, 476–480.
[16] Haralick R.M., Shanmugam K., Dinstein I., *Textural features for Image Classification*. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 6, 1973, 610–621.
[17] Albregtsen F., *Statistical Texture Measures Computed from Gray Level Coocurence matrices*. Image Processing Laboratory Department of Informatics University of Oslo, Nov. 1995.
[18] Tang X., *Texture Information in Run-Length Matrices.* IEEE Transactions on Image Processing, Vol. 7, No. 11, 1998, 1602–1609.
[19] Wei X., *Gray Level Run Length Matrix Toolbox v1.0 Software*. Beijing Aeronautical Technology Research Center, 2007
[20] The Mathworks, *Image Processing Toolbox User's Guide*. http://www.mathworks. com/access/ helpdesk/ help/ toolbox/images/.
[21] Ruggieri S., *Efficient C4.5*. IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 2, 2002, 438–444.
[22] Quinlan J.R., *C4.5: Programs for Machine Learning.* San Mateo, California, Morgan Kaufman 1993.
[23] Quinlan J.R., *Improved Use of Continuous Attributes in C4.5.* Journal of Artifcial Intelligence Research, 4, 1996, 77–90.
[24] Esposito F., Malerba D., Semeraro G., Kay J., *A comparative analysis of methods for pruning decision trees*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, Issue 5, 1997, 476–491.
[25] The Mathworks, *Bioinformatics Toolbox*. http://www.mathworks.com/access/ helpdesk/ help/toolbox/bioinfo/.
[26] Stąpor K., *Automatyczna klasyfikacja obiektów*. Warszawa, Akademicka Oficyna Wydawnicza EXIT 2005.
[27] Pawłowski Z., *Statystyka matematyczna.* Warszawa, PWN 1980.