DOI: 10.37317/biul-2011-0014

DARIUSZ R. MAŃKOWSKI [1]
ZBIGNIEW LAUDAŃSKI [2]
MONIKA JANASZEK [3]
[1] Pracownia Ekonomiki Nasiennictwa i Hodowli Roślin
Zakład Nasiennictwa i Nasionoznawstwa
Instytut Hodowli i Aklimatyzacji Roślin — Państwowy Instytut Badawczy w Radzikowie
[3] Katedra Ekonometrii i Statystyki
Wydział Zastosowań Informatyki i Matematyki
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
[3] Katedra Podstaw Inżynierii
Wydział Inżynierii Produkcji
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Przydatność wybranych miar podobieństwa dla danych binarnych do analiz wielocechowych w badaniach molekularnych

The application of chosen similarity measures for binary data in multivariate analysis in molecular experiments

W pracy przedstawiono możliwości wykorzystania ośmiu miar podobieństwa genetycznego w analizie danych binarnych, będących matematycznym obrazem żeli elektroforetycznych uzyskiwanych w badaniach molekularnych. Scharakteryzowano miary zgodności (Gowera), Jaccarda, Nei'a i Li (Dice'a), Hamanna, Ochiai, współczynnik Y Yule'a, współczynnik Q Yule'a oraz zero-jedynkowy odpowiednik współczynnika korelacji Pearsona (phi 4-point correlation). Na przykładzie analizy porównawczej 14 odmian marchwi jadalnej (*Daucus carota* L.) przedstawiono wykorzystanie tych miar w analizach wielocechowych — analizie skupień metodą UPGMA oraz analizie głównych współrzędnych PCoA. Przedstawiono i omówiono wyniki przeprowadzonych analiz oraz opisano różnice pomiędzy nimi. Porównano istniejące w literaturze miary podobieństwa dla danych molekularnych pod względem zgodności wyników uzyskiwanych z analiz statystycznych.

Słowa kluczowe: dane binarne, analizy molekularne, miary podobieństwa, PCoA, analiza skupień, marchew jadalna

The article presents the possibility of using eight measures of genetic similarity in analysis of binary data which are a mathematical image of the electrophoresis gels obtained in molecular studies. We characterized similarity measures: simple matching (Gower), Jaccard, Nei and Li (Dice), Hamann, Ochiai, Yule Y coefficient, Yule Q coefficient and zero-one equivalent of the Pearson correlation coefficient (phi 4-point correlation). Then, the example of a comparative analysis of 14 varieties of carrots (*Daucus carota* L.) presents the use of these measures in a multivariate analysis — UPGMA cluster analysis and principal coordinates analysis PCoA. The results of the analysis and the differences

between them were presented and discussed. The similarity measures for the molecular data existing in the literature were compared in terms of results compliance obtained from statistical analyses.

Key words: binary data, molecular analysis, similarity measures, PCoA, cluster analysis, carrot

WSTĘP

Ocena zróżnicowania genetycznego obiektów, odmian, populacji czy też gatunków jest bardzo często przedmiotem badań z zakresu biologii molekularnej, genetyki, hodowli roślin i bioinformatyki. Do lat 70. XX wieku ocena zróżnicowania pomiędzy obiektami opierała się w głównej mierze na markerach morfologicznych, fizjologicznych i cytologicznych, izoenzymach oraz na analizach porównawczych z doświadczeń ścisłych, ocenie heterozji i analizie zmienności w krzyżowaniach (Rief i in., 2005).

Obecnie najpowszechniej stosowane i uznane za najbardziej dokładne są metody oparte na markerach molekularnych. Wykorzystuje się je w badaniu genetycznych relacji pomiędzy różnymi genotypami w bankach genów oraz w procesach hodowlanych. W tych badaniach bardzo ważny staje się dobór właściwej miary podobieństwa (p) lub zróżnicowania ($z = 1 - p$) genotypów. Miary spotykane w literaturze posiadają różne właściwości matematyczne i cechują się różną przydatnością do prowadzonych badań i analiz (Rief i in., 2005). Odnaleźć można szereg prac charakteryzujących matematyczne własności miar podobieństwa/zróżnicowania, np. Goodman (1972), Gower (1985), Gower i Legendre (1986), Takezaki i Nei (1996), Rief i wsp. (2005), Siatkowski i wsp. (2010).

Dane pochodzące z analiz molekularnych są najczęściej matematycznym zapisem obrazów żeli elektroforetycznych, a ściślej binarnym (zero-jedynkowym) zapisem pasm uzyskiwanych w wykonywanych analizach molekularnych. Fakt wystąpienia prążka o danej masie na elektroforogramie jest zapisywany jako 1, brak prążka o tej masie jako 0. W ten sposób matematyczny obraz wszystkich pasm dla analizowanej puli genotypów ma postać zero-jedynkowej macierzy. Dane te mają charakter skategoryzowany, to znaczy, że rzeczywista różnica pomiędzy wystąpieniem i brakiem danego prążka nie może być zapisana matematycznie jako różnica (1–0). Z tego powodu do analizy takich danych nie powinno się wykorzystywać klasycznych miar odległości pomiędzy obiektami, takich jak odległość Euklidesa, Czebyszewa, itp.

Celem pracy było porównanie istniejących w literaturze miar podobieństwa dla danych molekularnych pod względem zgodności wyników uzyskiwanych ze statystycznych analiz wielocechowych, takich jak analiza skupień i analiza głównych współrzędnych.

MATERIAŁ I METODY

Materiał badawczy

Materiał do badań stanowiło 14 odmian marchwi jadalnej (*Daucus carrota* L.). Odmiany te były powszechnie uprawiane w Polsce na cele przemysłowe (przetwórstwo spożywcze). Badane obiekty pochodziły z kolekcji odmian uprawianych na polu pokazowym w Rolniczym Zakładzie Doświadczalnym SGGW w Żelaznej. Listę badanych odmian oraz informacje o nich zamieszczono w tabeli 1.

Tabela 1

Wykaz odmian marchwi wykorzystanych w badaniach
List of carrot cultivars used in experiment

Odmiana Cultivar	Kraj pochodzenia Country of origin	Hodowca Breeder
Canada (F1)	NL	Bejo Zaden B.V.
Finezja	PL	SPÓJNIA Hodowla i Nasiennictwo Ogrodnicze Sp. z o.o.
Florida (F1)	NL	Bejo Zaden B.V.
Kathmandu (F1)	NL	Bejo Zaden B.V.
Kazan (F1)	NL	Bejo Zaden B.V.
Laguna (F1)	NL	Nunhems B.V.
Macon (F1)	NL	Rijk Zwaan Zaadteelt en Zaadhandel B.V.
Maxima (F1)	DE	Agri-Saaten GmbH
Mazurska	PL	—
Prodigy (F1)	US	Seminis Vegetable Seeds, Inc.
Recoleta (F1)	US	Seminis Vegetable Seeds, Inc.
Sirkana (F1)	FR	Nunza B.V.
Sugarsnax (F1)	NL	Nunhems B.V.
Trafford (F1)	NL	Rijk Zwaan Zaadteelt en Zaadhandel B.V.

Analizy molekularne

Analizę zmienności genetycznej badanych odmian marchwi przeprowadzono z zastosowaniem semispecyficznego PCR (Rafalski, 2004). Technika z zastosowaniem łańcuchowej reakcji polimeryzacji jest wykorzystywana do amplifikacji sekwencji DNA z użyciem pary oligonukleotydowych starterów, z których każdy jest komplementarny do jednego końca docelowej sekwencji DNA. Startery są wydłużane ku sobie za pomocą termostabilnej polimerazy DNA, którą wyizolowuje się z bakterii termofilnych. Cykl, następujących po sobie reakcji, obejmuje denaturację, przyłączenie startera i polimeryzację. Mieszanina reakcyjna zawiera matrycę, startery, bufory, enzymy oraz dNTP i Mg^{2+}. W pierwszym cyklu cząsteczka DNA rozdziela się na 2 nici (ulega denaturacji) w wyniku ogrzewania do 95°C. Przyłączenie starterów jest możliwe po obniżeniu temperatury do około 55°C. Po ich przyłączeniu podnosi się temperaturę mieszaniny reakcyjnej do 72°C, aby umożliwić optymalną polimeryzację.

W pierwszym etapie polimeryzacji każda nić docelowa DNA jest kopiowana od miejsca przyłączenia startera na różną długość, co trwa aż do rozpoczęcia drugiego cyklu, w którym mieszanina reakcyjna jest ponownie ogrzewana do 95°C, aby zdenaturować nowo powstałe cząsteczki DNA. W kolejnym etapie drugi starter może wiązać się do powstałej nowej nici DNA i podczas polimeryzacji kopiować tę nić aż do miejsca, w którym znajdował się początek pierwszego startera. W ten sposób w końcowym etapie drugiego cyklu występują już nowo syntetyzowane cząsteczki o właściwej długości. W następnych cyklach liczba takich cząsteczek wzrasta. Jeżeli reakcja PCR zachodzi z wydajnością 100%, jedna docelowa cząsteczka po *n* cyklach ulega amplifikacji z krotnością równą 2^n.

Preparatykę DNA przeprowadzono zgodnie z metodyką opisaną przez Davisa i wsp. (1986). DNA otrzymano z próbek naci badanych odmian marchwi jadalnej, a jego stężenie oznaczono fluorymetrycznie, zgodnie z instrukcją fluorymetru TKO 100 (Hoefer Scientific Instruments, San Francisco, USA). Powielanie fragmentów DNA przeprowadzono w termocyklerze UNO II (Biometra, Göttingen, Niemcy), w mieszaninie zawierającej w

objętości 20µl: 10 lub 15,2 ng DNA roślinnego, 1×bufor, 2,5 mM $MgCl_2$, 1,2 µM startera, 200 µM dNTP i jedną jednostkę termostabilnej polimerazy. Produkty amplifikacji rozdzielono elektroforetycznie w 1,5% żelu agarozowym w buforze TBE i wizualizowano w świetle ultrafioletowym przy użyciu bromku etydyny. Fotografie rozdziałów elektroforetycznych analizowano w programie „Fragment NT" (Molecular Dynamice, Sunnyvale, CA, USA), przy użyciu którego utworzono macierz wyników, zapisaną w systemie binarnym. Obecność fragmentu DNA o określonej masie (wyrażonej w parach zasad) oznaczano jako 1, a brak tego fragmentu jako 0. Kolumny macierzy odpowiadały fragmentom DNA, a wiersze macierzy stanowiły odmiany marchwi. Do dalszych analiz wykorzystano tylko pasma polimorficzne (Janaszek, 2008).

Miary podobieństwa i zróżnicowania

W literaturze można odnaleźć wiele miar podobieństwa lub zróżnicowania (nazwanych tak dla odróżnienia od miar bliskości lub odległości, wykorzystywanych w analizach polegających na grupowaniu wielocechowym obiektów opisanych za pomocą zmiennych ilościowych) używanych do analizy danych o częstości występowania genów bądź fragmentów DNA (Nei, 1978; Reif i in., 2005). W dalszej części pracy zaprezentowano pewną grupę miar podobieństwa, które mogą być stosowane w analizach danych zero-jedynkowych pochodzących z badań molekularnych.

Miary podobieństwa dla danych binarnych wyznacza się w oparciu o tabelę krzyżową rozkładu częstości występowania zer i jedynek w obrębie dwóch porównywanych genotypów (tab. 2).

Tabela 2

Tabela krzyżowa rozkładu częstości występowania zer i jedynek w obrębie porównywanych genotypów X oraz Y

Cross-table of zeros and ones frequencies within compared genotypes X and Y

Częstotliwość Frequency		Genotyp X Genotype X	
		1	0
Genotyp Y Genotype Y	1	*a*	*b*
	0	*c*	*d*

W oparciu o wyznaczone z tabeli krzyżowej (tab. 2) częstości występowania par: 1–1 (*a*); 1–0 (*b*); 0–1 (*c*); 0–0 (*d*), u porównywanych genotypów, opracowano wiele miar ich podobieństwa (Reif i in., 2005; Laudański i Mańkowski, 2007; Siatkowski i in., 2010). Poniżej opisano wybrane miary podobieństwa, omawiane w niniejszej pracy:

— Miara zgodności — Gowera (Gower, 1971; Sneath i Sokal, 1973; Backhaus i in., 2000)

$$p(X;Y) = \frac{a+d}{a+b+c+d} \tag{1}$$

— Miara Jaccarda (Jaccard, 1908)

$$p(X;Y) = \frac{a}{a+b+c} \tag{2}$$

— Miara Neia i Li — Dice (Dice, 1945; Nei i Li, 1979)

$$p(X;Y) = \frac{2a}{2a+b+c} = \frac{2a}{(a+b)+(a+c)} \tag{3}$$

— Miara Hamanna (Hamann, 1961)

$$p(X;Y) = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{4}$$

— Współczynnik Y Yule'a (Lienert i von Eye, 1986)

$$p(X;Y) = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \tag{5}$$

— Współczynnik Q Yule'a (Lienert i von Eye, 1986)

$$p(X;Y) = \frac{ad-bc}{ad+bc} \tag{6}$$

— Miara Ochiai (Ochiai, 1957)

$$p(X;Y) = \sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}} \tag{7}$$

— Phi 4-point correlation — zerojedynkowy odpowiednik współczynnika korelacji Pearsona (Guilford, 1936)

$$p(X;Y) = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{8}$$

Analiza skupień (CA)

Analiza skupień obejmuje metody ilościowe, które umożliwiają porównanie i uporządkowanie obiektów tworzących pewien zbiór. Celem analizy skupień jest przede wszystkim podział badanych obiektów na grupy reprezentujące zbiorowości obiektów nie różniących się znacząco między sobą. Uzyskany w ten sposób podział, oprócz odkrycia nieznanej struktury zjawiska, pozwala na wyodrębnienie zasadniczych właściwości uzyskanych skupień (Laudański i Mańkowski, 2007). W praktyce, w badaniach taksonomicznych i molekularnych stosuje się techniki hierarchicznej analizy skupień. W przypadku tej techniki skupienia tworzą drzewa binarne. Najczęściej stosuje się tu metody aglomeracyjne tworzenia skupień. Polegają one na sukcesywnym łączeniu skupień (początkowo zakłada się, że każdy badany obiekt stanowi oddzielne skupienie) będących w jak najmniejszej odległości od siebie.

W przypadku klasycznej analizy skupień do oceny bliskości obiektów wykorzystuje się miary odległości pomiędzy tymi obiektami. Dla danych binarnych, pochodzących z badań molekularnych, wykorzystuje się natomiast miary zróżnicowania obiektów, czyli przeciwieństwo miar podobieństwa. Zróżnicowanie wyznacza się wg wzoru:

$$z(X;Y) = 1 - p(X;Y). \tag{9}$$

W przypadku badań molekularnych najczęściej stosowanym w hierarchicznej analizie skupień sposobem obliczania odległości pomiędzy skupieniami jest metoda UPGMA — średniej odległości między skupieniami (Liu i in., 2000; Guthridge i in., 2001; Díaz-Perales i in., 2002; Manimekalai i Nagarajan, 2006; Huang i in., 2007; Moncada i in., 2007). Odległość pomiędzy dwoma skupieniami, A oraz B, traktowana jest tu jako średnia arytmetyczna odległości między wszystkimi parami obiektów należących do skupień A oraz B, tzn.:

$$d(A;B) = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} z\left(O_{Ai};O_{Bj}\right)}{n_A \cdot n_B}. \tag{10}$$

Wadą metody UPGMA jest jej skłonność do łączenia skupień o niskich wariancjach i tworzenia klas o wariancjach podobnych (Sokal and Michener, 1958; Sneath i Sokal, 1973).

W wyniku hierarchicznej analizy skupień uzyskuje się wykres przypominający drzewo, zwany dendrogramem. Z takiego dendrogramu można odczytać kolejność tworzenia się skupień. Im połączenie pomiędzy obiektami lub skupieniami jest bliższe początku wykresu tym obiekty tworzące dane skupienie są mniej zróżnicowane między sobą.

W praktyce liczbę powstałych skupień, na które dzieli się badaną zbiorowość, określa się uznaniowo na podstawie obserwacji struktury dendrogramu. Taka ocena, z racji subiektywności, budzi znaczne kontrowersje. W przypadku analiz jednocechowych, znane są i powszechnie stosowane metody grupowania obiektów na znacząco różniące się od siebie i te nie wykazujące takiego zróżnicowania (np. procedury porównań wielokrotnych w analizie wariancji). Propozycje metod wskazywania właściwej liczby skupień w analizach wielocechowych (Mc Queen, 1966; Lance i Williams, 1967; Karoński, 1971; Caliński i Harabasz, 1974; Harabasz i Karoński, 1977; Chudzik i Karoński, 1979; Sarle, 1983; Sieczko, 2003; Kaczmarek i in., 2008) nie znalazły powszechnego zastosowania, a ocena nadal ma charakter uznaniowy.

W niniejszej pracy zwyczajową ocenę dendrogramu oraz subiektywną metodę określenia liczby skupień uzupełniono o metodę doboru optymalnej liczby skupień dokonywaną na podstawie takich kryteriów, jak współczynnik R^2 i semi-cząstkowy R^2 (Timm, 2002).

W analizie funkcji regresji współczynnik determinacji — R^2 jest miarą całkowitej wariancji zmiennej zależnej opisanej przez zmienne niezależne. W analizie wariancji, uwzględniającej analizę trendów, R^2 jest definiowane jako stosunek sumy kwadratów odchyleń dla analizowanego modelu do łącznej sumy kwadratów odchyleń i jest miarą całkowitej wariacji zmiennej zależnej opisanej przez analizowany model liniowy. W analizie skupień współczynnik R^2 jest miarą całkowitej wariancji cech wszystkich analizowanych obiektów, wynikającą z podziału zbiorowości na określoną liczbę skupień. Dla n skupień całkowita suma kwadratów odchyleń wynosi:

$$T = \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2, \tag{11}$$

a suma kwadratów odchyleń wewnątrz *k*-tego skupienia wynosi:

$$SSE_k = \sum\nolimits_i \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2. \tag{12}$$

Wówczas, współczynnik R^2 dla *k* skupień jest definiowany jako:

$$R_k^2 = \frac{T - \sum_k SSE_k}{T}. \tag{13}$$

Dla liczy skupień $k = n$ każda $SSE_k = 0$, więc $R^2 = 1$. Wraz ze zmniejszaniem się liczby skupień od *n* do 1 łączna wartość sum kwadratów odchyleń wewnątrz skupień rośnie, co z kolei powoduje zmniejszenie się wartości R^2, aż do $R^2 = 0$ przy jednym skupieniu zawierającym wszystkie badane obiekty.

Alternatywnie dla połączenia skupień A i B w skupienie Z można wyznaczyć semi-cząstkowy współczynnik R^2 (semipartial R^2) jako:

$$SR^2 = R_k^2 - R_{k-1}^2. \tag{14}$$

Statystyka SR^2 wyraża stosunek $SSE_Z - (SSE_A + SSE_B)$ gdzie skupienia A oraz B zostały połączone w skupienie Z, do łącznej sumy kwadratów odchyleń *T,* wyrażonej wzorem (11). Większe wartości SR^2 odpowiadają łączeniu bardziej odległych skupień.

Analiza głównych współrzędnych (PCoA)

Punktem wyjścia do analizy głównych współrzędnych jest analiza składowych głównych (PCA — Principal Component Analysis). Jest to wielocechowa analiza danych, zaproponowana przez Carla Pearsona na początku XX-ego wieku (Timm, 2002). Znaną i używaną do dziś formę tej analizy zaproponowali Hotelling (1933) i Rao (1964). Jest to metoda rzutowania danych, w przestrzeni zredukowanej, poprzez kombinacje liniowe oryginalnych zmiennych (wzajemnie nieskorelowanych składowych głównych), które zachowują maksimum oryginalnej wariancji danych. Celem jest wyrażenie wielowymiarowych obserwacji przy użyciu małej liczby współrzędnych. Celem PCA jest taki obrót układu współrzędnych, aby maksymalizować w pierwszej kolejności wariancję pierwszej współrzędnej, następnie wariancję drugiej współrzędnej, itd. Tak przekształcone wartości współrzędnych nazywane są ładunkami składowych głównych. W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe składowe główne. Źródłem danych do analizy PCA może być macierz kowariancji pomiędzy obiektami wyznaczana na podstawie informacji wielocechowych. Najczęściej jednak wykorzystuje się do obliczeń macierz korelacji, będącą standaryzowaną macierzą kowariancji.

Opisaną powyżej właściwość analizy PCA do redukcji wymiarów charakteryzujących analizowane obiekty wykorzystał Gower (1966). Zastępując macierz korelacji macierzą podobieństwa można przeprowadzić analizę składowych głównych, w wyniku której

uzyska się możliwość rzutowania obiektów w nowej dwu- (2D) lub trójwymiarowej (3D) przestrzeni współrzędnych, reprezentujących maksimum zmienności zawartej w źródłowym zbiorze danych (Kenkel, 2006). Składowe główne wyznaczone w wyniku tej analizy nazywane są współrzędnymi głównymi. A tak przeprowadzona analiza nazywana jest analizą głównych współrzędnych (PCoA — Principal Coordinate Analysis). Niektórzy badacze (Krzanowski, 2004; Kenkel, 2006) zaliczają PCoA do grupy analiz określonych mianem skalowania wielowymiarowego (multidimensional scaling). Celem skalowania wielowymiarowego jest bowiem wyrażenie wielowymiarowych obserwacji przy użyciu małej liczby współrzędnych tak, aby możliwie najlepiej zachować relacje między nimi.

Wynikiem analizy PCoA jest wykres przedstawiający rozmieszczenie badanych obiektów w przestrzeni dwóch lub trzech współrzędnych głównych. Na podstawie takiego przestrzennego rozmieszczenia obiektów można opisać strukturę populacji z jakiej pochodzą badane obiekty oraz określić stopień ich zróżnicowania (Sneath i Sokal, 1973; Kenkel, 2006).

Przetwarzanie danych oraz wyznaczenie wartości miar podobieństwa przeprowadzono w arkuszu kalkulacyjnym Ms Excel 2007. Analizy statystyczne wykonano w Systemie SAS® w wersji 9.2 (SAS Institute Inc., 2009).

OMÓWIENIE WYNIKÓW

Analiza skupień metodą UPGMA

W publikowanych pracach badawczych, dotyczących oceny zróżnicowania genetycznego różnych populacji roślinnych, analiza skupień jest analizą wykorzystywaną stosunkowo najczęściej. W pracach tych stosowana jest aglomeracja metodą średniego wiązania, a ocena powstających skupień dokonywana jest uznaniowo, na podstawie analizy uzyskanych dendrogramów (Liu i in., 2000; Guthridge i in., 2001; Díaz-Perales i in., 2002; Manimekalai i Nagarajan, 2006; Huang i in., 2007; Moncada i in., 2007).

Aby ocenić przydatność omawianych miar podobieństwa, w pierwszym etapie badań przeprowadzono hierarchiczną analizę skupień metodą UPGMA. W jej wyniku sporządzono dendrogramy charakteryzujące badane odmiany (rys. 1). Analiza wizualna umożliwiła wskazanie, zależnie od zastosowanej miary podobieństwa, od trzech do sześciu skupień badanych odmian marchwi jadalnej:

— sześć grup, do których należały odmiany: 1) Maxima; 2) Finezja, Prodigy, Recoleta; 3) Kazan, Laguna; 4) Florida, Canada, Kathmandu, Sirkana, Trafford, Macon; 5) Mazurska; 6) Sugarsnax, uzyskano stosując miary Gowera (zgodności), Hamanna, zerojedynkowy odpowiednik współczynnika korelacji Pearsona oraz współczynnik Y Yuley'a;
— pięć grup, do których należały odmiany: 1) Maxima; 2) Finezja, Prodigy, Recoleta; 3) Kazan, Laguna; 4) Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; 5) Sugarsnax, uzyskano stosując miary Jaccarda, Nei'a i Li (Dice'a) oraz Ochiai;
— trzy wyraźne skupienia: 1) Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; 2) Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; 3) Sugarsnax, uzyskano natomiast stosując współczynnik Q Yule'a. Dodatkowo zastosowanie tego

współczynnika pozwoliło na wyraźniejsze zilustrowanie struktury zmienności pomiędzy badanymi odmianami marchwi.

Równocześnie wyznaczono wartości współczynników R^2 i SR^2 dla poszczególnych poziomów agregacji w analizie skupień (tab. 3).

Tabela 3

Wyniki aglomeracji badanych odmian marchwi jadalnej uzyskane w analizie skupień metodą średniego wiązania

The results of the agglomeration of the studied carrot cultivars obtained by UPGMA cluster analysis

Liczba skupień Number of clusters	Miara podobieństa Similarity measure															
	zgodności (Gowera) simple matching (Gower)		Jaccarda Jaccard		Nei'a i Li (Dice'a) Nei and Li (Dice)		Hammana Hamman		współczynnik Y Yule'a Yule Y coefficient		współczynnik Q Yule'a Yule Q coefficient		Ochiai Ochiai		zero-jedynkowy odpowiednik współczynnik korelacji Pearsona Phi 4-point correlation	
	R^2	SR^2	R^2	SR^2	R^2	SR^2	R^2	SR^2	R^2	SR^2	R^2	SR^2	R^2	SR^2	R^2	SR^2
14	1,000	—	1,000	—	1,000	—	1,000	—	1,000	—	1,000	—	1,000	—	1,000	—
13	0,963	0,037	0,959	0,041	0,967	0,033	0,963	0,037	0,990	0,010	0,965	0,035	0,967	0,033	0,964	0,036
12	0,918	0,045	0,908	0,051	0,923	0,043	0,918	0,045	0,971	0,019	0,920	0,045	0,924	0,043	0,919	0,045
11	0,871	0,047	0,853	0,055	0,875	0,049	0,871	0,047	0,950	0,021	0,873	0,047	0,875	0,049	0,872	0,047
10	0,817	0,054	0,788	0,065	0,816	0,059	0,817	0,054	0,921	0,029	0,818	0,054	0,816	0,059	0,817	0,054
9	0,762	0,055	0,725	0,062	0,759	0,057	0,762	0,055	0,890	0,031	0,763	0,056	0,759	0,057	0,762	0,056
8	0,701	0,061	0,662	0,063	0,702	0,058	0,701	0,061	0,852	0,037	0,702	0,061	0,701	0,058	0,701	0,061
7	0,612	0,089	0,600	0,063	0,645	0,057	0,612	0,089	0,811	0,041	0,612	0,090	0,644	0,057	0,612	0,089
6	0,548	0,063	0,531	0,069	0,581	0,064	0,548	0,063	0,731	0,080	0,549	0,063	0,581	0,063	0,549	0,063
5	0,471	0,077	0,446	0,085	0,497	0,084	0,471	0,077	0,667	0,064	0,473	0,076	0,497	0,084	0,473	0,076
4	0,392	0,079	0,357	0,089	0,407	0,090	0,392	0,079	0,599	0,068	0,395	0,079	0,406	0,091	0,383	0,089
3	0,300	0,092	0,252	0,105	0,296	0,110	0,300	0,092	0,510	0,089	0,305	0,090	0,296	0,110	0,265	0,119
2	**0,172**	**0,128**	**0,153**	**0,099**	**0,191**	**0,105**	**0,172**	**0,128**	**0,347**	**0,163**	**0,176**	**0,129**	**0,192**	**0,104**	**0,175**	**0,090**
1	0,000	0,172	0,000	0,153	0,000	0,191	0,000	0,172	0,000	0,347	0,000	0,176	0,000	0,192	0,000	0,175

Niezależnie od stosowanej miary podobieństwa, w przypadku każdej analizy obserwowano dynamiczny, niemal jednostajny spadek poziomu wartości współczynnika R^2. W praktyce, optymalnej liczbie skupień, odpowiada ten poziom aglomeracji, w którym następuje znaczące obniżenie wartości współczynnika R^2, czyli znaczący wzrost współczynnika SR^2. W tym wypadku taki skok obserwowano dość późno bo na etapie 2 skupień, gdy wartość współczynnika R^2 obniżała się do poziomu 0,2–0,3. W wyniku analizy stwierdzono, że badane odmiany marchwi jadalnej były silnie zróżnicowane, co znacznie utrudniało wskazanie optymalnej liczby skupień. Badane odmiany marchwi jadalnej podzielono na dwa skupienia niezależnie od użytej miary podobieństwa: 1) Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; 2) Sugarsnax.

Porównanie podziałów, uzyskanych na podstawie oceny wizualnej — uznaniowej oraz statystycznych współczynników doboru optymalnej liczby skupień, pozwoliło na stwierdzenie, iż zdarza się, że oceny wykonywane wizualnie prowadzą do zbyt szczegółowych podziałów, które nie znajdują uzasadnienia w analizie zmienności

wewnątrz i pomiędzy uzyskanymi skupieniami. Innym możliwym wynikiem jest podział zbyt ogólny, który nie oddaje w całości zróżnicowania obiektów występującego w zbiorze danych. Prezentowana analiza jest przykładem tej pierwszej możliwości. Dodatkowo wskazuje, że dostępne kryteria statystyczne, służące do wskazywania optymalnej liczby skupień nie muszą dawać jednoznacznych wyników, a posługiwanie się nimi może okazać się skomplikowane.

Analiza głównych współrzędnych PCoA

Analizę głównych współrzędnych w literaturze, dotyczącej badań zróżnicowania genetycznego u roślin, spotyka się zdecydowanie rzadziej niż omawianą wcześniej analizę skupień. Dodatkowo zaobserwować można znaczne problemy z terminologią związaną z tą metodą, w szczególności w pracach publikowanych w czasopismach polskojęzycznych. Z racji pewnych podobieństw analiza ta jest często mylona z analizą składowych głównych lub analizą czynnikową.

Tabela 4

Wartości własne i stopień wyjaśnienia zmienności przez wyznaczone współrzędne główne w analizie PCoA

Eigenvalues and the degree of explanation of variation by the principal coordinates analysis PCoA

Współrzędne główne Principal coordinates	Zgodności (Gowera) Simple matching (Gower)			Miara Jaccarda Jaccard coefficient			Nei'a i Li (Dice'a) Nei and Li (Dice)			Miara Hamanna Hamann measure		
	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien-ności cumula-tive % of variation	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien-ności cumula-tive % of variation	wartość własna eigenval ue	% zmien-ności % of variance	skumulo wany % zmien-ności cumula-tive % of variation	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien– ności cumula-tive % of variation
1	2,3497	67,62	67,62	1,7324	49,86	49,86	2,2708	65,35	65,35	1,2557	36,14	36,14
2	0,1419	4,08	71,70	0,1956	5,63	55,48	0,1561	4,49	69,84	0,2679	7,71	43,85
3	0,1184	3,41	75,11	0,1771	5,10	60,58	0,1336	3,84	73,69	0,2303	6,63	50,48
4	0,1086	3,13	78,24	0,1576	4,54	65,12	0,1127	3,24	76,93	0,2171	6,25	56,72
5	0,1054	3,03	81,27	0,1532	4,41	69,53	0,1091	3,14	80,07	0,2066	5,95	62,67
6	0,0886	2,55	83,82	0,1391	4,00	73,53	0,0960	2,76	82,83	0,1767	5,09	67,75
7	0,0860	2,48	86,30	0,1298	3,74	77,26	0,0876	2,52	85,35	0,1715	4,94	72,69
8	0,0807	2,32	88,32	0,1273	3,66	80,93	0,0855	2,46	87,81	0,1613	4,64	77,33
9	0,0763	2,20	90,81	0,1229	3,54	84,46	0,0814	2,34	90,15	0,1517	4,37	81,70
10	0,0717	2,06	92,88	0,1189	3,42	87,89	0,0781	2,25	92,40	0,1420	4,09	85,78
11	0,0701	2,02	94,90	0,1146	3,30	91,18	0,0738	2,12	94,53	0,1399	4,03	89,81
12	0,0658	1,89	96,79	0,1122	3,23	94,41	0,0718	2,07	96,59	0,1315	3,78	93,60
13	0,0570	1,64	98,43	0,0996	2,87	97,28	0,0611	1,76	98,35	0,1138	3,28	96,87
14	0,0545	1,57	100,00	0,0945	2,72	100,00	0,0573	1,65	100,00	0,1087	3,13	100,00

c. d. Tabela 4

Współrzędne główne Principal coordinates	Współczynnik Y Yuley'a Yuley's Y coefficient			Współczynnik Q Yuley'a Yuley's Q coefficient			Miara Ochiai Ochiai measure			Zero-jedynkowy odpowiednik współczynnika korelacji Pearsona Phi 4-point correlation		
	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien-ności cumula–tive % of variation	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien-ności cumula–tive % of variation	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmienno ści cumula–tive % of variation	wartość własna eigen-value	% zmien-ności % of variance	skumulo wany % zmien-ności cumula–tive % of variation
1	1,2594	36,24	36,24	2,0490	58,97	58,97	2,2734	65,43	65,43	1,2526	36,05	36,05
2	0,2668	7,68	43,92	0,2694	7,75	66,72	0,1558	4,48	69,91	0,2664	7,67	43,71
3	0,2335	6,72	50,64	0,2050	5,90	72,62	0,1335	3,84	73,75	0,2338	6,73	50,44
4	0,2188	6,30	56,94	0,1720	4,95	77,57	0,1127	3,24	77,00	0,2183	6,28	56,72
5	0,2073	5,97	62,90	0,1598	4,60	82,17	0,1091	3,14	80,14	0,2075	5,97	62,69
6	0,1744	5,02	67,92	0,1111	3,20	85,36	0,0959	2,76	82,90	0,1751	5,04	67,73
7	0,1732	4,98	72,91	0,1080	3,11	88,47	0,0873	2,51	85,41	0,1737	5,00	72,73
8	0,1607	4,62	77,53	0,0918	2,64	91,11	0,0855	2,46	87,87	0,1612	4,64	77,37
9	0,1476	4,25	81,78	0,0735	2,12	93,23	0,0802	2,31	90,18	0,1491	4,29	81,66
10	0,1422	4,09	85,87	0,0664	1,91	95,14	0,0781	2,25	92,43	0,1428	4,11	85,77
11	0,1396	4,02	89,89	0,0609	1,75	96,89	0,0738	2,12	94,55	0,1404	4,04	89,81
12	0,1307	3,76	93,65	0,0513	1,48	98,36	0,0714	2,05	96,60	0,1317	3,79	93,60
13	0,1139	3,28	96,93	0,0309	0,89	99,25	0,0610	1,75	98,36	0,1143	3,29	96,89
14	0,1067	3,07	100,00	0,0260	0,75	100,00	0,0571	1,64	100,00	0,1082	3,11	100,00

Zastosowanie analizy PCoA, z racji jej matematycznej konstrukcji i zdolności do redukcji wymiarów przestrzeni danych, wiąże się z utratą części informacji dotyczących charakterystyki badanych obiektów (Sneath i Sokal, 1973; Kenkel, 2006). Teoretycznie wiązać się to może z pogorszeniem jakości klasyfikacji badanych genotypów. W praktyce jednak prowadzi najczęściej do 'odszumienia' (usunięcia zbędnych informacji zacierających rzeczywisty obraz danych) struktury wielocechowej zmienności badanych obiektów, co z kolei powinno owocować bardziej klarowną klasyfikacją przestrzenną.

Aby ocenić przydatność omawianych miar podobieństwa przeprowadzono analizę głównych współrzędnych. Wyniki analizy głównych współrzędnych dla wszystkich miar podobieństwa (tab. 4) wykazały, że największy udział w wytłumaczeniu obserwowanej zmienności — od 36% do 68% — ma pierwsza współrzędna główna, druga współrzędna główna tłumaczyła od 4% do 8% obserwowanej zmienności pomiędzy badanymi odmianami marchwi jadalnej.

Na podstawie wykresów rozmieszczenia analizowanych obiektów w układzie dwóch pierwszych współrzędnych głównych (rys. 2) stwierdzono, że badane odmiany marchwi jadalnej dzielą się na 2 do 3 wyraźnych grup. Dla miary podobieństwa Hamanna, współczynnika Y Yule'a oraz zero-jedynkowego współczynnika korelacji uzyskano wyraźny podział na trzy grupy odmian, dla miary Ochiai i współczynnika Q Yuleya podział na trzy grupy nie jest już tak wyraźny, natomiast dla miar Gowera (zgodności), Jaccarda, Nei'a i Li (Dice'a) uzyskano podział na dwie grupy odmian.

W tabeli 5 zestawiono wyniki klasyfikacji badanych odmian marchwi jadalnej przeprowadzonej za pomocą analizy skupień (wariant oceny wizualnej oraz z wykorzystaniem współczynników R^2 i SR^2) oraz analizy głównych współrzędnych PCoA.

Tabela 5

Podsumowanie wyników klasyfikacji badanych odmian marchwi jadalnej z zastosowaniem wybranych współczynników podobieństwa dla danych binarnych
Summary of results of the classification of the tested cultivars carrots using the selected coefficients of similarity for binary data

Miara podobieństwa Similarity measure	Analiza skupień — Cluster analysis		Analiza głównych współrzędnych Principal coordinate analysis
	wizualne określenie liczby skupień visual determination of the number of clusters	określenie liczby skupień w oparciu o współczynniki R^2 i SR^2 determination the number of clusters based on the R^2 and SR^2 coefficients	
1	2	3	4
Zgodności (Gowera) Simple matching (Gower)	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon; **5** — Mazurska; **6** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax
Jaccarda Jaccard	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **5** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax
Nei'a i Li (Dice'a) Nei and Li (Dice)	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **5** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax
Hammana Hamman	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon; **5** — Mazurska; **6** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **2** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **3** — Sugarsnax
Współczynnik Y Yule'a Yule Y coefficient	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon; **5** — Mazurska; **6** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **2** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **3** — Sugarsnax
Współczynnik Q Yule'a Yule Q coefficient	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **2** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **3** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1a** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **1b** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax

c.d. Tabela 5

1	2	3	4
Ochiai Ochiai	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **5** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1a** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **1b** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax
Zerojedynkowy odpowiednik współczynnik korelacji Pearsona Phi 4-point correlation	**1** — Maxima; **2** — Finezja, Prodigy, Recoleta; **3** — Kazan, Laguna; **4** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon; **5** — Mazurska; **6** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna, Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **2** — Sugarsnax	**1** — Maxima, Finezja, Prodigy, Recoleta, Kazan, Laguna; **2** — Florida, Canada, Kathmandu, Sirkana, Trafford, Macon, Mazurska; **3** — Sugarsnax

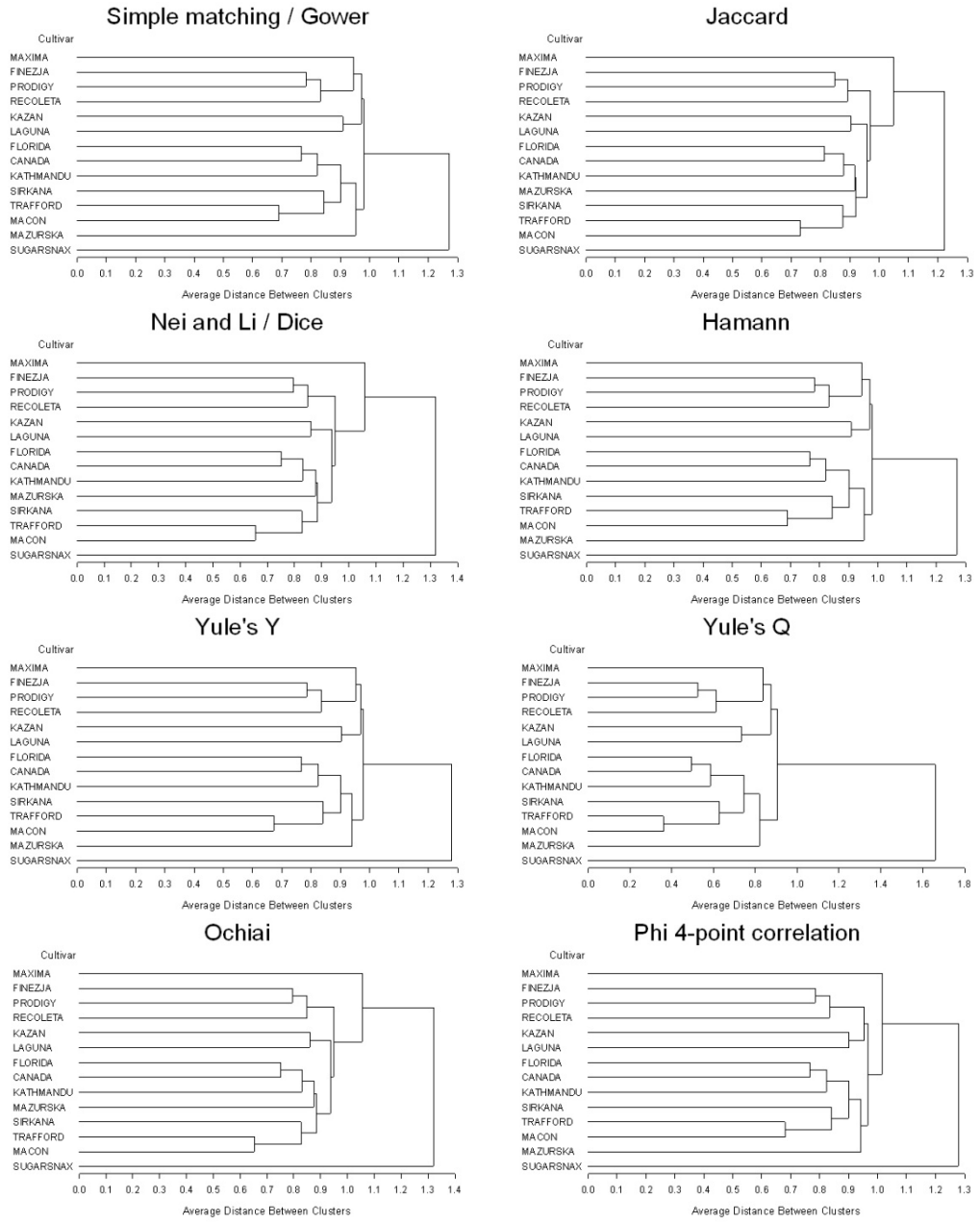
Na podstawie przedstawionego zestawienia stwierdzono, że zgodne klasyfikacje uzyskano w przypadku analizy skupień z wykorzystaniem współczynników R^2 i SR^2 do wskazania właściwej liczby skupień oraz analizy PCoA przy wykorzystaniu miar podobieństwa: Gowera (zgodności), Jaccarda, Nei'a i Li (Dice'a). Ponadto zbliżone wyniki uzyskano przy zastosowaniu współczynnika Q Yule'a oraz mierze podobieństwa Ochiai . Ocena wizualna w analizie skupień wskazywała na istnienie większej liczby skupień niż wymienione wcześniej metody. Wyjątek stanowiło wykorzystanie do oceny podobieństwa pomiędzy badanymi odmianami współczynnika Q Yule'a. W tym przypadku obydwa warianty grupowania z zastosowaniem analizy skupień oraz wyniki analizy głównych współrzędnych były bardzo podobne.

DYSKUSJA I PODSUMOWANIE

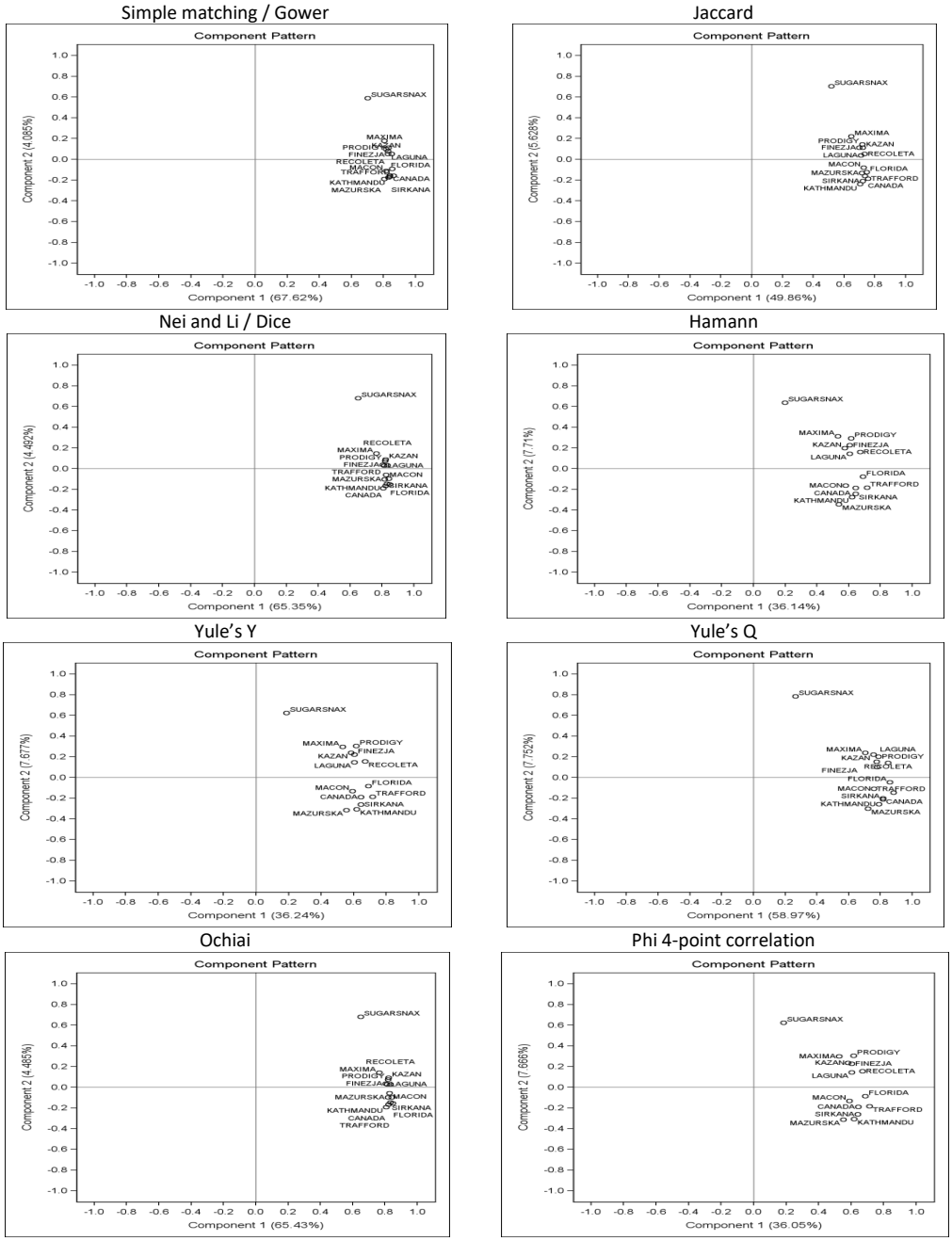
Omawiane miary podobieństwa były już w przeszłości opisywane i oceniane. Jednak w praktyce najczęściej spotyka się tylko trzy spośród nich (miarę Jaccarda, Nei'a i Li — Dice'a oraz miarę Gowera — zgodności).

W prezentowanym przykładzie badano 14 odmian przemysłowych marchwi jadalnej. Dane pochodziły z analizy semispecyficznego PCR i były zapisane w formie macierzy zero-jedynkowej. Następnie wyznaczono macierze podobieństwa badanych obiektów. Do tego celu wykorzystano 8 miar bliskości stworzonych dla danych binarnych (dychotomicznych). Następnie macierze te wykorzystano do dwóch analiz statystycznych. Po przekształceniu z macierzy podobieństwa do macierzy zróżnicowania wykonano analizę skupień metodą UPGMA. Natomiast na macierzach podobieństwa przeprowadzono analizę głównych współrzędnych bez rotacji.

Prezentowane w pracy osiem miar podobieństwa dla danych binarnych można podzielić, ze względu na sposób ich wyznaczania, na trzy grypy. Pierwszą stanowią miary, w obliczaniu których uwzględnia się jedynie pary prążków 1:1 (prążki występują w obydwu porównywanych genotypach) jako pary zgodne oraz 1:0 i 0:1 (prążek występuje tylko w jednym z porównywanych genotypów) jako pary niezgodne.



Rys. 1. Dendrogramy analizy skupień metodą UPGMA wykonane w oparciu o osiem analizowanych miar podobieństwa dla danych z analizy semispecyficznego PCR czternastu odmian marchwi jadalnej
Fig. 1. Dendrograms of cluster analysis performed by UPGMA method based on eight measures of similarity of the analysed data from semi-specific PCR analysis of fourteen varieties of carrot



Rys. 2. Wykresy obrazujące rozmieszczenie badanych odmian marchwi jadalnej w układzie dwóch pierwszych współrzędnych głównych dla ośmiu analizowanych miar podobieństwa

Fig. 2. Charts showing the distribution of tested varieties of carrots in system of the first two principal coordinates for the eight analyzed measures of similarity

Do tych miar zaliczyć można miarę Jaccarda, miarę Neia i Li (Dice'a), oraz miarę Ochiai. Miary te nie uwzględniają faktu, że brak wystąpienia danego prążka w obydwu porównywanych genotypach może świadczyć o ich podobieństwie. Druga grupa miar uwzględnia fakt występowania par 0:0 (brak prążka u obydwu porównywanych genotypów). W przypadku tej grupy miar efekt wystąpienia obszarów zgodnych (1:1 oraz 0:0) jest jednak pomniejszany przez efekt występowania obszarów niezgodnych (1:0 oraz 0:1). Do tych miar zaliczyć można miarę Hammana, współczynniki Q oraz Y Yule'a oraz zero-jedynkowy odpowiednik współczynnika korelacji. Do trzeciej grupy miar podobieństwa zaliczyć można miarę zgodności (Gowera). Miara ta wyraża udział fragmentów zgodnych (1:1 oraz 0:0) w zbiorze porównywanych pasm.

Różny sposób wyznaczania wartości opisywanych miar podobieństwa wiąże się ściśle z rodzajem uzyskiwanego wyniku. Dobór właściwej miary powinien być więc uzasadniony celem badań i rodzajem prowadzonych analiz molekularnych.

Zastosowanie analizy skupień do porównywania badanych genotypów, w analizach danych pochodzących z badań molekularnych, analizy skupień pozwala na opisanie struktury podobieństwa tych genotypów. Ma to szczególne znaczenie w badaniach filogenetycznych czy też badaniach związanych z oceną tożsamości genotypów. Niestety znaną wadą analizy skupień jest brak wyraźnej odpowiedzi jak powinien wyglądać właściwy podział na grupy genotypów podobnych. Ta metoda eksploracji danych nie daje klarownej odpowiedzi. W większości przypadków badacze stosują różne kryteria uznaniowe. W związku z tą 'ułomnością' analizy skupień cały czas trwają prace nad opracowaniem jasnego i dokładnego kryterium pozwalającego na wyciągnięcie wniosków co do faktycznego podziału badanych genotypów na podobne i różne. Jedną z takich propozycji jest zastosowanie opisanych w niniejszej pracy współczynników R^2 i semi-cząstkowego R^2. Niestety w przypadku znaczącego zróżnicowania pomiędzy badanymi genotypami (jak ma to miejsce w opisywanym w pracy przykładzie) współczynniki te wskazują na podziały bardzo ogólne i w bardzo niskim stopniu oddające rzeczywiste zróżnicowanie badanych genotypów (tab. 3).

Analiza głównych współrzędnych PCoA wydaje się być ciekawym sposobem przestrzennego opisu (2D i 3D) zróżnicowania genotypów w badaniach molekularnych. Niestety, podobnie jak inne analizy wielocechowe polegające na redukcji wymiarów, użycie jej wiąże się z utratą części informacji (obserwowanej zmienności). Wybierając tylko dwie lub trzy współrzędne główne decydujemy się na pewną redukcję informacji. Oczywiście stosowana w analizie PCoA metoda analizy składowych głównych Hotellinga wskazuje w pierwszej kolejności te składowe (współrzędne) główne, które niosą najwięcej informacji, jednak odcięcie dalszych, mniej istotnych składowych (współrzędnych) prowadzić może nie tylko do usunięcia szumu (zakłóceń losowych), lecz również pewnej części informacji. Bardzo wyraźnie widać to w omawianym w niniejszej pracy przykładzie. Niezależnie od stosowanej miary podobieństwa, wyznaczone dwie pierwsze główne współrzędne tłumaczyły jedynie od 42% do 72% obserwowanej zmienności. Pozostała część tej zmienności nie była brana pod uwagę przy rozmieszczaniu badanych obiektów w przestrzeni ograniczonej przez te dwie główne współrzędne.

Zastosowanie analizy skupień połączonej z uznaniowym podziałem badanych odmian na skupienia pozwoliło na wskazanie od 3 do 6 skupień odmian marchwi jadalnej zależnie od metody wyznaczania miary podobieństwa. Jednak już przy tym podejściu można było zaobserwować znaczące, ale i równomierne zróżnicowanie badanych odmian, które objawiało się stosunkowo dalekim położeniem połączeń pomiędzy skupieniami względem początku układu współrzędnych. Spośród badanych odmian jedynie odmiana Sugarsnax wyraźnie oddzielała się od pozostałych. W grupie porównywanych miar podobieństwa, przy opisywanym podejściu oraz przy tak zróżnicowanej puli porównywanych genotypów najciekawsze wyniki uzyskano stosując współczynnik Q Yule'a. Pozwolił on na uzyskanie wyraźnego i bardzo czytelnego obrazu (rys. 1) struktury podobieństwa badanych odmian marchwi jadalnej.

Wykorzystanie do identyfikacji skupień współczynnika R^2 i semi-cząstkowego R^2 pozwoliło na podział badanych odmian marchwi jadalnej na dwa skupienia niezależnie od zastosowanej miary podobieństwa. Jedynie 9–20% całkowitej wariancji cech wszystkich analizowanych odmian, wynikało z uzyskanego podziału badanej zbiorowości genotypów marchwi jadalnej na dwa skupienia. Uzyskane podziały wskazywały na to, że najbardziej różna od pozostałych była odmiana Sugarsnax, co skutkowało oddzieleniem jej od pozostałych badanych odmianami utworzeniem odrębnego skupienia (tab. 5).

Zastosowanie analizy głównych współrzędnych i rzutowanie przestrzeni wielowymiarowej w której opisane były badane odmiany marchwi, na przestrzeń dwuwymiarową ograniczoną przez pierwsze dwie główne współrzędne pozwoliło na pogrupowanie badanych odmian na dwie do trzech grup (skupień) zależnie od zastosowanej miary podobieństwa. W przypadku miary zgodności (Gowera), miary Jaccarda oraz miary Nei'a i Li (Dice'a) uzyskano wyraźny podział na dwa skupienia (rys. 2, tab. 5). Podział ten pokrywał się w pełni z podziałem uzyskanym z zastosowaniem analizy skupień z uwzględnieniem współczynnika R^2 do wyznaczania liczby skupień. W przypadku miary Ochiai oraz współczynnika Q Yule'a uzyskano podział na skupienia bardzo podobny do podziału uzyskanego z zastosowaniem analizy skupień z uwzględnieniem współczynnika R^2 do wyznaczania liczby skupień. W przypadku tych miar podobieństwa można było jednak zauważyć niezbyt wyraźny podział dużego skupienia zawierającego wszystkie badane odmiany za wyjątkiem odmiany Sugersnax, na dwa mniejsze skupienia (oznaczone jako 1 a oraz 1 b w tab. 5). W przypadku miary Hammana, współczynnika Y Yule'a oraz zero-jedynkowego odpowiednika współczynnika korelacji stwierdzono, że badane odmiany podzieliły się na trzy wyraźne skupienia. We wszystkich przypadkach ostatnie, najbardziej oddalone od pozostałych odmian skupienie stanowiła odmiana Sugarsnax. Jak można zauważyć (tab. 4) podział na większe liczby skupień uzyskano dla tych miar podobieństwa, dla których dwie pierwsze główne współrzędne opisywały mniejszy zasób obserwowanej zmienności (ok. 43%) w stosunku do pozostałych badanych miar podobieństwa (55–72%).

Obserwowane rozbieżności w uzyskiwanych wynikach można wytłumaczyć całym szeregiem zjawisk. Jako pierwsze należy wymienić — różne metody wyznaczania miar podobieństwa, specyfikę analizy skupień oraz specyfikę analizy głównych współrzędnych. Ponadto znaczącym czynnikiem był różny stopień obserwowanej zmienności,

wykorzystanej do podziału na skupienia. Dodatkowym elementem znacznie utrudniającym ocenę była specyficzna pula badanych odmian, które cechowały się znacznym zróżnicowaniem, jednak na dość podobnym poziomie pomiędzy poszczególnymi parami porównywanych genotypów. W puli odmian marchwi jadalnej poddanej badaniom jedynie odmiana Sugarsnax cechowała się znaczącym oddaleniem od pozostałych odmian.

LITERATURA

Backhaus K., Erichson B., Plinke W., Weiber R. 2000. Multivariaten Analysemethoden. Eine anwendungsorientierte Einführung. Springer, Berlin.

Caliński T., Harabasz J. S. 1974. A dendrite method for cluster analysis. Communications in Statistics, vol. 3: 1 — 27.

Chudzik H., Karoński M. 1979. Skupianie obserwacji metodą k-średnich. Roczniki AR w Poznaniu, Algorytmy Biomedyczne i Statystyczne, 78: 133 — 152.

Davis L. G., Dibner M. D., Battey J. F. 1986. Basic methods in molecular biology. Elsevier Sci. Publ., New York: 42 — 43.

Díaz-Perales A., Linacero R., Vázquez A. M. 2002. Analysis of genetic relationships among 22 European barley varieties based on two PCR markers. Euphytica, 129: 53 — 60.

Dice L. R. 1945. Measures of the amount of ecologic association between species. Ecology, 26: 297 — 302.

Duda R. O., Hart P. E. 1973. Pattern Classification and Scene Analysis. New York: John Wiley & Sons.

Goodman M. M. 1972. Distance analysis in biology. Syst. Zool.: 174 — 186.

Gower J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53: 325 — 338.

Gower J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics, 27: 857 — 874.

Gower J. C. 1985. Measures of similarity, dissimilarity and distances. In: Klotz S. *et al.* (ed.), Encyclopedia of statistical sciences. Vol. 5. Wiley & Sons, New York, USA.

Gower J. C., Legendre P. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classification, 3: 5 — 48.

Guilford J. 1936. Psychometric Methods. New York: McGraw–Hill Book Company, Inc.

Guthridge K. M., Dupal M. P., Kölliker R., Jones E. S., Smith K. F., Forster J. W. 2001. AFLP analysis of genetic diversity within and between populations of perennial ryegrass (*Lolium perenne* L.). Euphytica, 122: 191 — 201.

Hamann U. 1961: Merkmalsbestand und Verwandtschafsbeziehungen der farinosae. Ein Beitrag zum System der Monokotyledonen. Willdenowia, 2: 639 — 768.

Harabasz J. S., Karoński M. 1977. Dendrytowa metoda analizy skupień. Roczniki AR w Poznaniu, Algorytmy Biomedyczne i Statystyczne, 57: 135 — 148.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24: 417 — 441, 498 — 520.

Huang X.-Q., Wolf M., Ganal M. W., Orford S., Koebner R. M. D., Röder M. S. 2007. Did modern plant breeding lead to genetic erosion in European winter wheat varieties? Crop Sci., 47: 343 — 349.

Jaccard P. 1908. Nouvelles recherches sur la distribution florae. Bull. Soc. Vaud. Sci. Nat., 44: 223 — 270.

Janaszek M. 2008. Identyfikacja cech korzeni marchwi jadalnej z wykorzystaniem komputerowej analizy obrazów. SGGW, Warszawa, rozprawa doktorska.

Kaczmarek Z., Czajka S., Adamska E. 2008. Propozycja metody grupowania obiektów jedno i wielocechowych z zastosowaniem odległości Mahalanobisa i analizy skupień. Biuletyn IHAR, Nr 249: 9 — 18.

Karoński M. 1971. Algorytm grupowania populacji w rozkładach metodą krok po kroku. Roczniki AR w Poznaniu, Algorytmy Biomedyczne i Statystyczne, 4: 30 — 33.

Kenkel N. C. 2006. On selecting an appropriate multivariate analysis. Canadian Journal of Plant Science, 86: 663 — 676.

Krzanowski W. J. 2004. Biplots for multifactorial analysis of distance. Biometrics, 60: 517 — 524.

Lance G. M., Williams W. T. 1967. A general theory of classificatory sorting strategies. Hierarchical Systems, Computer Journal, 9: 373 — 380.
Laudański Z., Mańkowski D. R. 2007. Planowanie i wnioskowanie statystyczne w badaniach rolniczych. IHAR Radzików.
Lienert G. A., von Eye A. 1986. Yule-Coefficients for Second- and Higher-Order Associations. Biometrical Journal, 28: 539 — 545.
Liu F., von Bothmer R., Salomon B. 2000. Genetic diversity in European accessions of the barley core collection as detected by isozyme electrophoresis. Genetic Resources and Crop Evolution, 47: 571 — 581.
Manimekalai R., Nagarajan P. 2006. Interrelationships among coconut (*Cocos nucifera* L.) accessions using RAPD technique. Genetic Resources and Crop Evolution, 53: 1137 — 1144.
Mc Queen J. B. 1966. Some methods for classification and analysis of multivariate observations. Proc. Fifth Barkeley Symposium on Mathematical Statistics and Probability Theory. Barkeley University of California Press, vol.1: 281 — 287.
Moncada K. M., Ehlke N. J., Muehlbauer G. J., Sheaffer C. C., Wyse D. L., DeHaan L. R., 2007. Genetic variation in three native plant species across the State of Minnesota. Crop Sci., 47: 2379 — 2389.
Nei M. 1978. The theory of genetic distance and evolution of human races. Jpn. J. Hum. Genet., 23: 341 — 369.
Nei M., Li W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA, 76: 5269 — 5273.
Ochiai A. 1957. Zoographic studies on the soleoid fishes found in Japan and its neighboring regions. Bull. Japan Soc. Sci. Fish., 22: 526 — 530.
Rafalski A. 2004.Semi-specyficzny PCR w badaniach genetyczno-hodowlanych roślin. Monografie i Rozprawy Naukowe IHAR, Nr 23.
Rao C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyã, A26: 329 — 358.
Reif J. C., Melcinger A. E., Frisch M. 2005. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Science, 45: 1 — 7.
Sarle W. S. 1983. Cubic Clustering Criterion. SAS Technical Report A-108, Cary, NC: SAS Institute Inc.
SAS Institute Inc. 2009. SAS/STAT 9.2 user's guide. Second edition. SAS Institute Inc., Cary, NC, USA.
Siatkowski I., Goszczurna T., Szabelska A., Zyprych J. 2010. Coefficients of dissimilarity and similarity with application. Colloquium Biometricum, 40: 13 — 23.
Sieczko L. 2003. Kryteria wstępnego przecięcia dendrogramu w hierarchicznej analizie skupień. Colloquium Biometryczne, 33: 249 — 258.
Sneath P. H. A., Sokal R. R. 1973. Numerical taxonomy. Freeman, San Francisco.
Sokal R. R., Michener C. D. 1958. A statistical method for evaluating systemic relationships. University of Kansas Science Bulletin, 38: 1409 — 1438.
Takezaki N., Nei M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics, 144: 389 — 399.
Timm N. H. 2002. Applied multivariate analysis. New York, USA: Springer-Verlag Inc.