Statystyka i filogenetyka

/ 30
Liczba punktów (wypełnia KGOB)

PESEL											Imię i nazwisko	Grupa				Nr
												Czerwona	Niebieska	Zielona	Żółta	

Zaznacz znakiem X swoją grupę

Czas: 90 min.
Łączna liczba punktów do zdobycia: 30

Odpowiedzi zapisz w miejscu na to przeznaczonym przy każdym z zadań używając długopisu lub pióra z **czarnym atramentem**.

Zadanie 1

Twoim celem jest oszacowanie relacji pokrewieństwa między pięcioma gatunkami motyli (A-E), których ogólne sylwetki oraz powiększenia głowy zostały przedstawione na planszy **P1**. Zadanie należy wykonać w czterech krokach:

1. Zakodowanie dziecięciu cech jakościowych w postaci surowej macierzy cech.
2. Obliczenie na podstawie surowej macierzy cech macierzy odległości między gatunkami.
3. Zastosowanie algorytmu UPGMA w celu zgrupowania gatunków w klastry.
4. Przedstawienie oszacowanych relacji pokrewieństwa w postaci drzewa filogenetycznego.

Szczegółowe instrukcje, jak wykonać zadanie, znajdują się przy każdym z kroków. Przykład zastosowania algorytmu UPGMA znajduje się w załączniku **Z1**.

KROK 1: Kodowanie cech

Dla każdego z gatunków należy określić stan każdej z 10 cech opisanych poniżej.

1. Przyoczka (*ocelli*)
   0 – przyoczek brak
   1 – przyoczka obecne
2. Wzór na odwłoku
   0 – brak
   1 – obecny
3. Wzór na skrzydłach I-pary
   0 – brak
   1 – obecny
4. Wzór na skrzydłach II-pary
   0 – brak
   1 – obecny
5. Kolor skrzydeł
   0 – skrzydła szare
   1 – skrzydła jaskrawo zabarwione
6. Wielkość oczu
   0 – oczy małe (zdecydowanie mniejsze od głowy)
   1 – oczy duże (+/- wielkości głowy)
7. Budowa czułków
   0 – czułki szczoteczkowate (gęsto owłosione)
   1 – czułki nitkowate z buławką
8. Długość czułków
   0 – czułki krótkie (<1 cm)
   1 – czułki długie (>1 cm)
9. Obecność ostróg (*caudae*) na skrzydłach II-pary
   0 – ostróg brak
   1 – ostrogi obecne
10. Długość trąbki (*proboscis*)
    0 – trąbka krótka (<1 cm)
    1 – trąbka długa (>1 cm)

1.1 Kodowanie cech morfologicznych (5 pkt)

(puste pola należy uzupełnić za pomocą cyfr „0” lub „1”)

	Cecha									
Gatunek	**1. Przyoczka**	**2. Wzór na odwłoku**	**3. Wzór skrzyd. I pary**	**4. Wzór skrzyd. II pary**	**5. Kolor skrzydeł**	**6. Wielkość oczu**	**7. Budowa czułków**	**8. Długość czułków**	**9. Obecność ostróg**	**10. Długość trąbki**
A										
B										
C										
D										
E										

KROK 2: Obliczenie macierzy odległości

Na podstawie zakodowanych cech pięciu gatunków motyli oblicz odległości między poszczególnymi gatunkami. Jako miarę odległości między dwoma gatunkami przyjmij liczbę cech, dla których występują różnice w kodowaniu.

1.2 Macierz odległości między gatunkami (5 pkt)

(puste pola należy uzupełnić odpowiednimi wartościami z zakresu 0–10)

	A	**B**	**C**	**D**	**E**
A					
B					
C					
D					
E					

KROK 3: Zastosowanie algorytmu UPGMA

Na podstawie obliczonej macierzy odległości wykonaj grupowanie gatunków w klastry za pomocą algorytmu UPGMA. Elementarny przykład zastosowania metody przypominający zasadę prowadzenia obliczeń znajduje się w załączniku **Z1**.

1.3.1 Utworzenie pierwszego klastra (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	
Wiek nowego klastra:	

1.3.2. Nowa macierz odległości między klastrami (1 pkt)

(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

1.3.3 Utworzenie drugiego klastra (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	
Wiek nowego klastra:	

1.3.4. Nowa macierz odległości między klastrami (1 pkt)

(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

1.3.5 Utworzenie trzeciego klastra (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	
Wiek nowego klastra:	

1.3.6 Nowa macierz odległości między klastrami (1 pkt)
(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

1.3.7. Określenie długości drzewa (odległości od liści do korzenia drzewa) (1pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	ABCDE
Wiek drzewa:	

KROK 4: Rysowanie drzewa filogenetycznego

Na podstawie wyników algorytmu UPGMA narysuj w poniższej ramce drzewo filogenetyczne przedstawiające relacje pokrewieństwa między pięcioma gatunkami motyli (A-E). Zachowaj proporcje pomiędzy długościami gałęzi. Przy każdej z gałęzi zapisz jej długość.

1.4 Diagram przedstawiający drzewo filogenetyczne (3 pkt)

Zadanie 2

Twoim celem jest sprawdzenie, czy dwa okazy roślin *Chaerophyllum bulbosum* (baldaszkowate – Apiaceae) różnią się średnią wielkością pyłku. Zadanie należy wykonać w dwóch krokach.

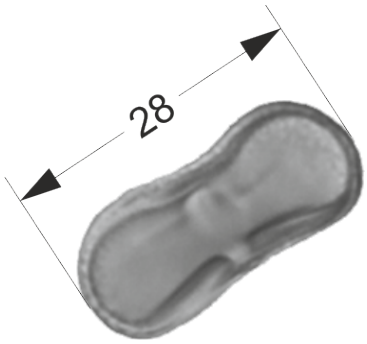
1. Wykonywanie pomiarów długości pyłku na podstawie mikrofotografii pobranych prób.
2. Testowanie hipotezy o równości średnich za pomocą testu t studenta dla prób niezależnych.

Szczegółowe instrukcje, jak wykonać zadanie znajdują się przy każdym z kroków.

KROK 1: Pomiary pyłku

Próby pyłku pobrane z okazów A i B przedstawiono w postaci mikrofotografii na planszy **P2**. W obydwu przypadkach zastosowano powiększenie 1000 × – jeden milimetr na planszy odpowiada w rzeczywistości jednemu mikrometrowi.

Zmierz za pomocą linijki odległość biegunową (wymiar pyłku wzdłuż jego dłuższej osi) każdego z numerowanych ziaren pyłku z dokładnością do 1 µm, a wyniki zapisz w poniższej tabeli.



Przykład pomiaru odległości biegunowej pyłku (brak zachowanej skali – rysunek poglądowy).

2.1 Pomiary odległości biegunowej (5 pkt)

(puste pola należy uzupełnić wynikami pomiarów w µm)

	Ziarno pyłku									
Okaz	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
A										
B										

KROK 2: Analiza statystyczna

Na podstawie wykonanych pomiarów przetestuj hipotezę o równości średniej odległości biegunowej u dwóch zbadanych okazów.

2.2.1 Oblicz statystykę testu wg poniższego wzoru. (2 pkt)

$$t_0 = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$\bar{x}_i$ średnia z i-tej próby
n_i liczebność i-tej próby
s_i^2 wariancja i-tej próby

Wzór na odchylenie standardowe z próby:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

$t_0 =$ .....

2.2.2 Oblicz liczbę stopni swobody wg poniższego wzoru. (1 pkt)

$$df = n_1 + n_2 - 2$$

$df =$ .....

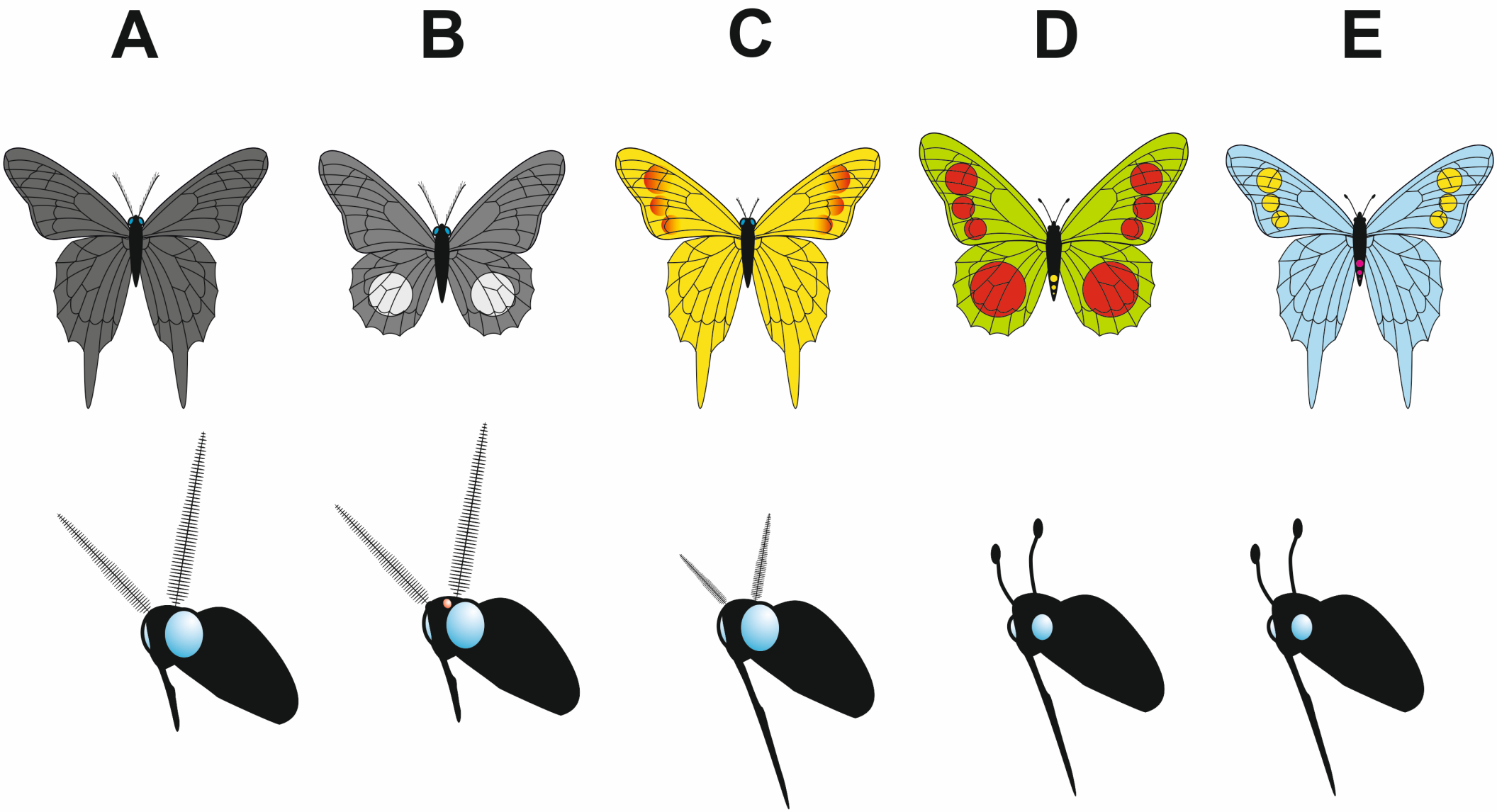
2.2.3 Korzystając z tabeli zawierającej wartości krytyczne statystyki testowej (załącznik Z2) znajdź przedział w jakim znajduje się p-wartość. (1 pkt)

..... $< p <$ .....

2.2.4 Zdecyduj, czy odrzucamy hipotezę zerową na poziomie istotności alfa = 0,05. Odpowiedź uzasadnij. (1 pkt)

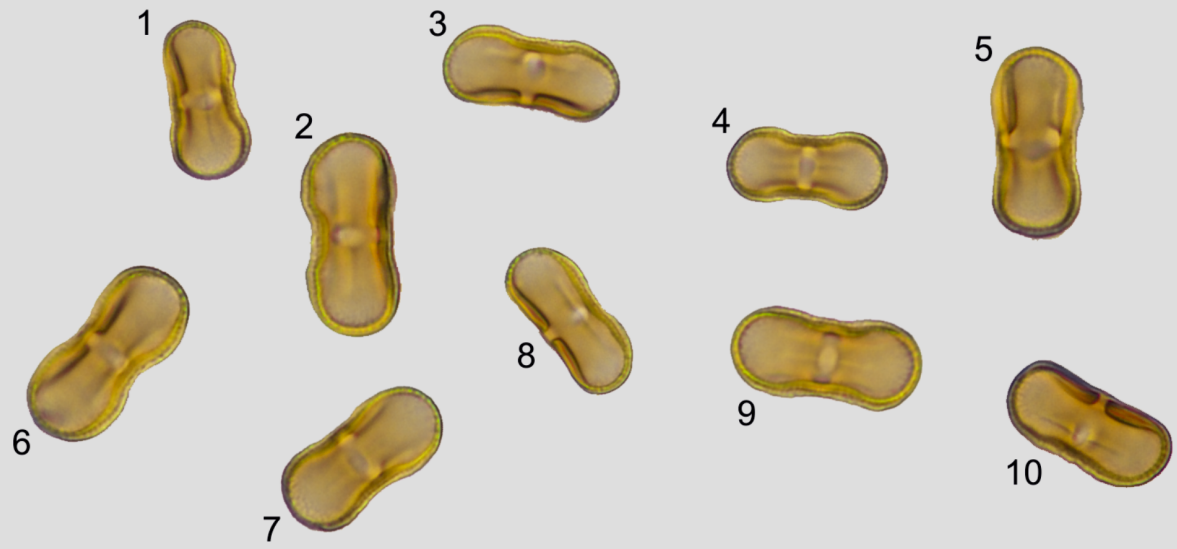
.....
.....
.....

Plansza P1: Sylwetki pięciu gatunków motyli oraz powiększenie głowy

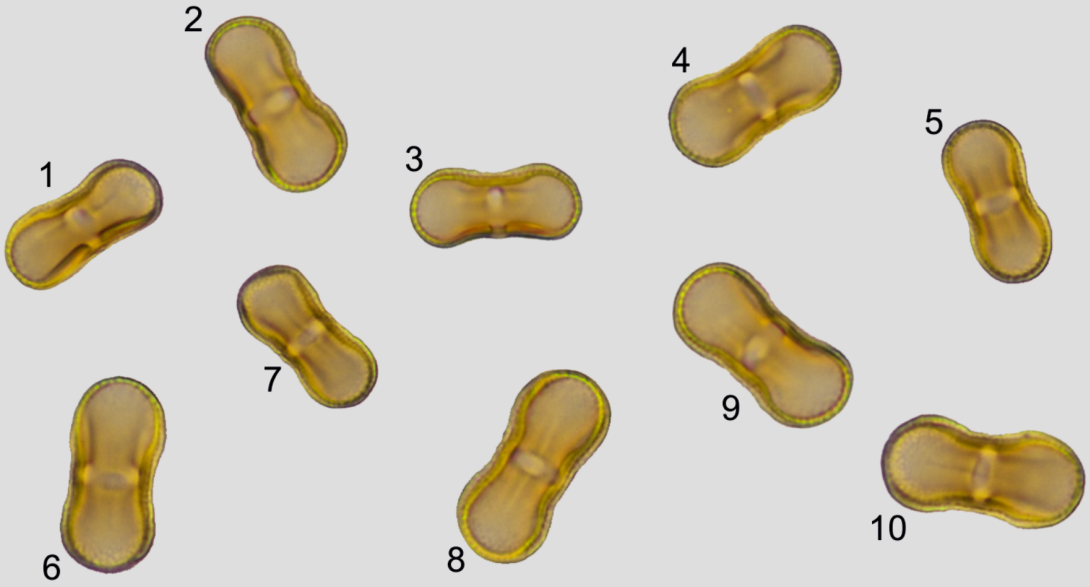


Plansza P2: Próbki pyłku dwóch okazów roślin

Okaz A



Okaz B



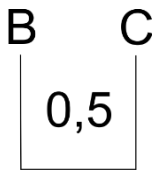
Załącznik Z1

Przykład zastosowania algorytmu UPGMA

1. Wyjściowa macierz odległości:

	A	B	C	D
A				
B	4			
C	5	1		
D	2	2	3	

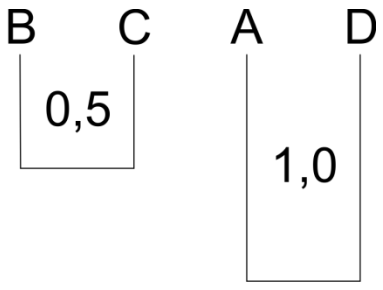
2. Najbliżej siebie są taksony B i C, które łączymy gałęzią o długości 1, którą łamiemy dokładnie w środku jej długości (powstaje klaster BC o wieku równym 0,5):



3. Następnie na podstawie wyjściowej macierzy przeliczamy odległości między taksonami A, D i klastrem BC: (2 + 3) / 2 oraz (4 + 5) / 2

	A	BC	D
A			
B C	4,5		
D	2	2,5	

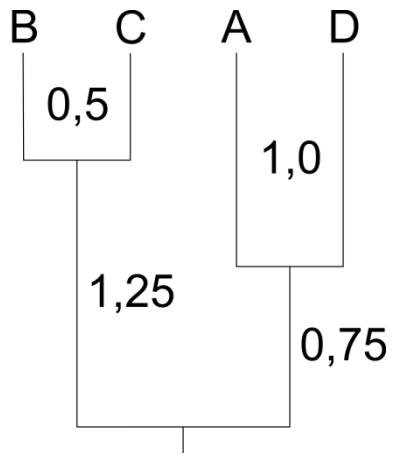
4. Teraz najbliżej siebie są taksony A i D więc łączymy je gałęzią o długości 2, którą znów łamiemy w połowie długości (powstaje klaster AD o wieku równym 1):



5. Następnie znów na podstawie wyjściowej macierzy musimy przeliczyć odległości miedzy klastrami BC i AD: (2 + 3 + 4 + 5) / 4. W tym celu uśredniamy indywidualne odległości między B–A, C–D, C–A i C–D, a stanowiące wszystkie możliwe relacje między gatunkami należącymi do różnych klastrów.

	AD	BC
AD		
BC	3,5	

6. Łączymy klaster AD i BC gałęzią o długości 3,5 złamaną w połowie, co daje 1,75 licząc od liści do korzenia (powstaje kompletne drzewo o długości równej 1,75):



Załącznik Z2

Wartości krytyczne statystyki testu t-studenta.

Liczba stopni swobody	p-wartość w teście dwustronnym							
	0,2	0,1	0,05	0,04	0,02	0,01	0,002	0,001
1	3,07768	6,31375	12,7062	15,8945	31,8205	63,6568	318,306	636,627
2	1,88562	2,91999	4,30265	4,84873	6,96456	9,92484	22,3272	31,5990
3	1,63774	2,35336	3,18245	3,48191	4,54070	5,84091	10,2145	12,9240
4	1,53321	2,13185	2,77644	2,99853	3,74695	4,60409	7,17318	8,61031
5	1,47588	2,01505	2,57058	2,75651	3,36493	4,03214	5,89344	6,86884
6	1,43976	1,94318	2,44691	2,61224	3,14267	3,70743	5,20763	5,95880
7	1,41492	1,89458	2,36462	2,51675	2,99795	3,49948	4,78528	5,40787
8	1,39682	1,85955	2,30600	2,44898	2,89646	3,35539	4,50079	5,04130
9	1,38303	1,83311	2,26216	2,39844	2,82144	3,24984	4,29681	4,78092
10	1,37218	1,81246	2,22814	2,35931	2,76377	3,16927	4,14370	4,58691
11	1,36343	1,79588	2,20099	2,32814	2,71808	3,10581	4,02470	4,43697
12	1,35622	1,78229	2,17881	2,30272	2,68100	3,05454	3,92963	4,31779
13	1,35017	1,77093	2,16037	2,28160	2,65031	3,01228	3,85198	4,22083
14	1,34503	1,76131	2,14479	2,26378	2,62449	2,97684	3,78739	4,14045
15	1,34061	1,75305	2,13145	2,24854	2,60248	2,94671	3,73283	4,07276
16	1,33676	1,74588	2,11991	2,23536	2,58349	2,92078	3,68615	4,01500
17	1,33338	1,73961	2,10982	2,22385	2,56693	2,89823	3,64576	3,96512
18	1,33039	1,73406	2,10092	2,21370	2,55238	2,87844	3,61048	3,92164
19	1,32773	1,72913	2,09302	2,20470	2,53948	2,86094	3,57940	3,88341
20	1,32534	1,72472	2,08596	2,19666	2,52798	2,84534	3,55181	3,84952
21	1,32319	1,72074	2,07961	2,18943	2,51765	2,83136	3,52715	3,81927
22	1,32124	1,71714	2,07387	2,18289	2,50832	2,81876	3,50499	3,79214
23	1,31946	1,71387	2,06866	2,17696	2,49987	2,80734	3,48496	3,76762
24	1,31784	1,71088	2,06390	2,17154	2,49216	2,79694	3,46678	3,74539
25	1,31635	1,70814	2,05954	2,16659	2,48511	2,78744	3,45019	3,72514
26	1,31497	1,70562	2,05553	2,16203	2,47863	2,77871	3,43500	3,70660
27	1,31370	1,70329	2,05183	2,15783	2,47266	2,77068	3,42103	3,68959
28	1,31253	1,70113	2,04841	2,15393	2,46714	2,76326	3,40816	3,67391
29	1,31143	1,69913	2,04523	2,15033	2,46202	2,75639	3,39624	3,65941
30	1,31041	1,69726	2,04227	2,14697	2,45726	2,75000	3,38519	3,64596
31	1,30946	1,69552	2,03951	2,14383	2,45282	2,74404	3,37490	3,63345
32	1,30857	1,69389	2,03693	2,14090	2,44868	2,73848	3,36531	3,62180
33	1,30774	1,69236	2,03452	2,13816	2,44479	2,73328	3,35634	3,61091
34	1,30695	1,69092	2,03224	2,13558	2,44115	2,72840	3,34793	3,60072
35	1,30621	1,68957	2,03011	2,13316	2,43772	2,72381	3,34004	3,59115
36	1,30551	1,68830	2,02809	2,13087	2,43449	2,71948	3,33262	3,58215
37	1,30485	1,68709	2,02619	2,12871	2,43145	2,71541	3,32563	3,57367
38	1,30423	1,68595	2,02439	2,12667	2,42857	2,71156	3,31903	3,56568
39	1,30364	1,68488	2,02269	2,12474	2,42584	2,70791	3,31279	3,55811
40	1,30308	1,68385	2,02108	2,12291	2,42326	2,70446	3,30688	3,55096

Zasady oceniania rozwiązań zadań

Zadanie 1

1.1 Kodowanie cech (5 pkt)

- 0,5 pkt – za poprawne zakodowanie każdej z cech dla wszystkich pięciu gatunków.

Prawidłowe rozwiązanie:

	Cecha									
Gatunek	1. Przyoczka	2. Wzór na odwłoku	3. Wzór skrzyd. I pary	4. Wzór skrzyd. II pary	5. Kolor skrzydeł	6. Wielkość oczu	7. Budowa czułków	8. Długość czułków	9. Obecność ostróg	10. Długość trąbki
A	0	0	0	0	0	1	0	1	1	0
B	1	0	0	1	0	1	0	1	0	0
C	0	0	1	0	1	1	0	0	1	1
D	0	1	1	1	1	0	1	0	0	1
E	0	1	1	0	1	0	1	0	1	1

UWAGA: Cechy są binarne, a więc zmiana kodowania 0 na 1 i 1 na 0 nie ma więc wpływu na dalsze rozwiązanie zadania (algorytm UPGMA). Innymi słowy kodowanie jest arbitralne, ale od uczestnika było wymagane, żeby podążał za schematem przedstawionym we wstępie do zadania.

2.2 Obliczenie macierzy odległości (5 pkt)

- 0,5 pkt – za poprawne obliczenie każdej z dziesięciu odległości między gatunkami.

UWAGA: Błędy w kodowaniu cech popełnione w poprzednim kroku nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność obliczeń na podstawie kodowania cech wykonanego przez uczestnika, pod warunkiem że w tabeli nie ma zostawionych pustych komórek.

Jeżeli jednak w kroku 1 nie uzyskano rozwiązania lub jest ono niekompletne, to za krok 2 jest przyznawane bezwzględnie 0 pkt, a rozwiązanie zadania nie jest dalej oceniane.

Prawidłowe rozwiązanie:

	A	**B**	**C**	**D**	**E**
A					
B	3				
C	4	7			
D	9	8	5		
E	7	10	3	2	

Krok 3: Zastosowanie algorytmu UPGMA (7 pkt)

- 1 pkt – za poprawne określenie składu gatunkowego i wieku każdego z nowych klastrów przyznawany jest 1 pkt.
- 1 pkt – za poprawne określenie długości całego drzewa przyznawany jest 1 pkt.
- 1 pkt – za poprawne obliczenie każdej kolejnej macierzy odległości.

UWAGA: Błędy rachunkowe popełnione w trakcie stosowania algorytmu obniżają punktację jedynie za dany podpunkt, w którym popełniono błąd. Następnie egzaminator przelicza rozwiązanie dalszych podpunktów z uwzględnieniem popełnionego błędu.

Błędy popełnione w poprzednim kroku podczas obliczania macierzy odległości nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność obliczeń na podstawie macierzy odległości otrzymanej przez uczestnika, pod warunkiem, że w tabeli nie ma zostawionych pustych komórek (rozwiązanie jest kompletne).

Jeżeli jednak w kroku 2 nie uzyskano rozwiązania lub jest ono niekompletne, to za krok 3 jest przyznawane bezwzględnie 0 pkt, a rozwiązanie zadania nie jest dalej oceniane.

Prawidłowe rozwiązanie:

1.3.1 Utworzenie pierwszego klastra i macierz odległości między klastrami (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	D, E
Wiek nowego klastra:	1

1.3.2 Nowa macierz odległości między klastrami (1 pkt)

(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

	A	B	C	D+E
A				
B	3			
C	4	7		
D+E	8	9	4	

1.3.3 Utworzenie drugiego klastra i macierz odległości między klastrami (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	A, B
Wiek nowego klastra:	1,5

1.3.4 Nowa macierz odległości między klastrami (1 pkt)

(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

	A+B	C	D+E
A+B			
C	5,5		
D+E	8,5	4	

1.3.5 Utworzenie trzeciego klastra i macierz odległości między klastrami (1 pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	D, E, C
Wiek nowego klastra:	2

1.3.6 Nowa macierz odległości między klastrami (1 pkt)

(w nagłówki kolumn i wierszy należy wpisać oznaczenia literowe wszystkich gatunków wchodzących w skład klastra)

	A+B	(D+E)+C
A+B		
(D+E)+C	7,5	

1.3.7 Określenie długości drzewa (odległości od liści do korzenia drzewa) (1pkt)

Oznaczenia literowe gatunków wchodzących w skład nowego klastra:	A, B, C, D, E
Wiek drzewa:	3,75

1.4 Rysowanie drzewa filogenetycznego (3 pkt)

Podczas oceniania są sprawdzane następujące kryteria w przedstawionej kolejności:

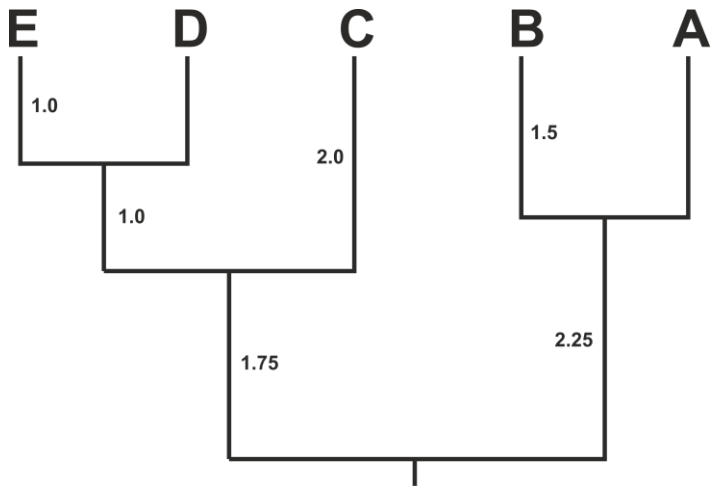
- Zachowanie prawidłowej topologii,
- Zachowanie równej odległości od korzenia drzewa do każdego z gatunków,
- Zachowanie właściwych długości gałęzi.

3 pkt – za spełnienie łączenie trzech kryteriów.
2 pkt – za spełnienie łącznie dwóch pierwszych kryteriów.
1 pkt – za spełnienie jedynie pierwszego kryterium.

UWAGA: Błędy popełnione w poprzednim kroku podczas stosowania algorytmu UPGMA nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność kreślenia drzewa na podstawie uzyskanych przez ucznia klastrów gatunków i ich wieków.

Jeżeli jednak w kroku 3 nie uzyskano rozwiązania lub jest ono niekompletne, to za krok 4 jest przyznawane bezwzględnie 0 pkt

Prawidłowe rozwiązanie:



Zadanie 2

2.1 Pomiary pyłku (5 pkt)

- 0,25 pkt – za prawidłowe zmierzenie każdego z ziaren pyłku z tolerancją +/- jeden mikrometr.

Prawidłowe rozwiązanie:

	Ziarno pyłku									
Okaz	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
A	20	26	22	20	24	25	23	21	24	22
B	22	24	22	24	21	25	21	27	26	26

2.2.1 Obliczenie statystyki testu (2 pkt)

- 2 pkt – za obliczenie statystyki testu dokładnością do 4 miejsc po przecinku.
- 1 pkt – za obliczenie statystyki testu z dokładnością do 3 miejsc po przecinku.

Prawidłowe rozwiązanie: $t_0 = 1{,}1545$

UWAGA: Błędy popełnione w poprzednim podpunkcie podczas pomiarów pyłku nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność obliczeń na podstawie pomiarów wykonanych przez uczestnika pod warunkiem, że pomiary pyłku zostały wykonane dla każdego ziarna.

2.2.2 Obliczenie liczby stopni swobody (1 pkt)

- 1 pkt – za prawidłowe obliczenie wartości parametru.

Prawidłowe rozwiązanie: $df = 18$

2.2.3 Wyznaczenie przedziału dla p-wartości (1 pkt)

- 1 pkt – za prawidłowe wskazanie przedziału zgodnie z rozdzielczością tabeli podanej w załączniku Z2.

UWAGA: Błędy popełnione w poprzednim podpunkcie podczas wyznaczania wartości statystyki testu nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność podejmowania decyzji na podstawie wartości statystyki testu obliczonej przez uczestnika pod warunkiem, że w ogóle została ona wyznaczona.

Prawidłowe rozwiązanie: $0{,}2 < p < 1$

Uwaga merytoryczna: p-wartość jest rodzajem prawdopodobieństwa, a więc musi się zawierać w przedziale od 0 do 1.

2.2.4 Decyzja o odrzuceniu hipotezy zerowej (1 pkt)

- 1 pkt – za podjęcie prawidłowej decyzji wraz z uzasadnieniem odnoszącym się do właściwego porównania otrzymanej p-wartości z zadanym poziomem istotności.

UWAGA: Błędy popełnione w poprzednim podpunkcie podczas wyznaczania przedziału dla p-wartości nie mają wpływu na ocenę rozwiązania zadania. Sprawdzana jest poprawność podejmowania decyzji na podstawie przedziału wyznaczonego przez uczestnika pod warunkiem, że przedział jest skonstruowany poprawnie pod kątem formalnym.

Przykładowe rozwiązanie:

Nie ma podstaw do odrzucenia hipotezy zerowej ponieważ p-wartość jest większa od założonego poziomu istotności.