DOI: 10.37317/biul-2015-0020

ANDRZEJ KASPERSKI [1]
RENATA KASPERSKA [2]
[1] Wydział Nauk Biologicznych, Katedra Biotechnologii, Uniwersytet Zielonogórski, ul. Szafrana 1
65-516 Zielona Góra
[2] Instytut Inżynierii Bezpieczeństwa i Nauk o Pracy, Uniwersytet Zielonogórski, ul. Szafrana 4
65-516 Zielona Góra

Zastosowanie n-wymiarowej macierzy kropkowej do analizy zmienności genetycznej roślin*

Application of n-dimensional dot-matrix to analysis of plant genetic diversity

Celem pracy jest przedstawienie możliwości wykorzystania nowej metody do analizy zmienności genetycznej organizmów, w tym organizmów o znaczeniu przemysłowym. Zaproponowana metoda działa na dwóch poziomach — na poziomie porównań aminokwasów oraz na poziomie porównań ich kodonów, tj. na poziomie kodu genetycznego. W pracy przedstawiono analizę inhibitorów proteinaz z nasion dyniowatych oraz analizę cytochromu c wybranych roślin uprawnych. Ponadto, za pomocą nowej metody dwupoziomowej analizy zmienności genetycznej otrzymano wyniki, które wykorzystano do interpretacji drzew filogenetycznych zbudowanych dla analizowanych sekwencji.

Słowa kluczowe: n-wymiarowa macierz kropkowa, statystyczna ocena przyrównań, transwersja, tranzycja, zmienność genetyczna

The aim of this study is to present the possibilities of using a new method for genetic variability analysis of organisms, including organisms of industrial use. The proposed method works at two levels — at the amino-acid comparison level and at the amino-acid codon comparison level, i.e. at the genetic code level. The paper presents analysis of proteinase inhibitors from squash seeds and analysis of cytochrome c of selected crops. Moreover, the results obtained using the new method of genetic variability two-level analysis, have been used to interpret the phylogenetic trees constructed for the analyzed sequences.

Key words: n-dimensional dot-matrix, statistical evaluation of alignments, transversion, transition, genetic variation

* Praca przedstawiona na konferencji IHAR — PIB, Zakopane, 3 lutego 2015 roku

Redaktor prowadzący: Wiesław Mądry

WSTĘP

Przyczyną zmienności genetycznej organizmów są mutacje powodujące zmiany w sekwencjach DNA. Powstające zróżnicowanie genetyczne organizmów, w tym roślin uprawnych, jest istotne m.in. dla odporności danego gatunku na wirusy i bakterie. Ponadto, stopień zmienności genetycznej organizmów informuje o ich przeszłości. Zrozumienie zmienności genetycznej organizmów oraz ich metabolizmu stanowi podstawę optymalnego zaprojektowania procesów biotechnologicznych (Kasperski i in., 2015). Ze względu na złożoność obliczeń, np. podczas oceny interakcji genotypu z biotycznymi i antybiotycznymi czynnikami środowiska w trakcie trwania okresu wegetacji roślin, w komputerowych analizach zmienności często wykorzystywane są metody sztucznej inteligencji, w tym sztuczne sieci neuronowe (Janaszek i in., 2011). Badanie zmienności genetycznej organizmów może zostać wykonane poprzez budowanie drzew filogenetycznych, w których długości gałęzi pokazują rozmiar zmian genetycznych między poszczególnymi taksonami. Analiza drzewa filogenetycznego pozwala poznać pokrewieństwo pomiędzy organizmami oraz lepiej zrozumieć mechanizmy narastania zmienności genetycznej. Do wygenerowania drzew filogenetycznych w komputerowych implementacjach wykorzystywane są metody łączenia sąsiadów (ang. Neighbor Joining, NJ), maksymalnej wiarygodności (ang. Maximum Likelihood, ML), maksymalnej oszczędności (ang. Maximum Parsimony, MP), wnioskowania Bayesowskiego (ang. Bayesian Inference, BI) (Hall, 2008). Wygenerowanie drzewa filogenetycznego jedną spośród metod NJ, ML, MP, BI poprzedzone jest zawsze wzajemnym dopasowaniem sekwencji. Wzajemne dopasowanie sekwencji może zostać wykonane z wykorzystaniem różnych algorytmów oraz ich komputerowych implementacji, np. ClustalW, MUSCLE, ProbCons, T-Coffee. Wybór algorytmu decyduje o uzyskanym dopasowaniu oraz w konsekwencji o długościach gałęzi i topologii drzewa filogenetycznego. Dopasowanie sekwencji zależy także od ustawionych wartości parametrów, w wyniku czego możemy uzyskać wiele dopasowań dla różnych algorytmów i dla różnych ustawień parametrów. Im większa zmienność sekwencji, a w szczególności im większa ilość insercji i delecji, tym większa ilość możliwych dopasowań. Ilość możliwych drzew filogenetycznych zależy w istotny sposób od ilości sekwencji, np. dla 4 taksonów ilość możliwych drzew ukorzenionych jest równa 15, a dla 50 taksonów ilość możliwych drzew ukorzenionych jest większa od ilości atomów we wszechświecie. Teoretycznie wybór najlepszego drzewa wymaga przeanalizowania każdego z możliwych drzew, co w praktyce jest zadaniem niemożliwym do wykonania dla większej ilości sekwencji. W konsekwencji, w praktycznych rozwiązaniach stosowane są różne heurystyki. Reasumując, ze względu na złożoność zagadnienia dokładna rekonstrukcja zmienności genetycznej organizmów poprzez próby określenia rzeczywistych drzew filogenetycznych najczęściej nie jest możliwa. Powoduje to konieczność poszukiwania nowych metod, które pozwolą na wiarygodniejsze określenie zmienności genetycznej organizmów. Zaproponowana w tej pracy metoda bazuje na metodzie macierzy kropkowej. Jedną z zalet metody macierzy kropkowej jest możliwość znalezienia wszystkich możliwych podobieństw reszt przyrównywanych sekwencji. Metoda ta posiada także wady (Xiong, 2006): brak

wiarygodności statystycznej oceny przyrównań, możliwość wykonywania przyrównań sekwencji tylko parami oraz konieczność „ręcznego" łączenia sąsiednich fragmentów identyczności. W niniejszej pracy, poprzez implementację nowej metody pokazano, w jaki sposób można wyeliminować te ograniczenia i uzyskać w pełni funkcjonalną n-wymiarową metodę macierzy kropkowej, która posiada statystyczną ocenę uzyskanych fragmentów identyczności oraz umożliwia automatyczne łączenie sąsiednich fragmentów identyczności.

MATERIAŁ I METODY

Przedstawione w pracy wyniki analizy zostały wykonane dla sekwencji inhibitorów proteinaz z nasion dyniowatych oraz sekwencji cytochromu c wybranych roślin uprawnych. Sekwencje zostały pobrane z wykorzystaniem serwisów internetowych Protein Blast i NCBI oraz informacji zawartych w literaturze (m.in. Leluk, 2000 a) i są dostępne pod adresem http://www.uz.zgora.pl/~akaspers/pliki/Zakopane2015/sekwencje.txt.

Wzajemne dopasowanie sekwencji zostało wykonane z wykorzystaniem programu ClustalW (Dereeper i in., 2008; Thompson i in., 1994). Drzewa filogenetyczne zbudowano metodą maksymalnej wiarygodności (ang. Maximum Likelihood, ML), wykorzystując program MEGA6 (Tamura i in., 2013). Inne, ważniejsze metody tworzenia drzew filogenetycznych (w tym metoda ML) zostały szczegółowo opisane np. w książce „Łatwe drzewa filogenetyczne" (Hall, 2008). Do szacowania wiarygodności drzew filogenetycznych wykorzystano metodę samopróbkowania (ang. bootstrap) dla ilości replikacji równej 500.

Algorytm nowej metody dwupoziomowej analizy zmienności genetycznej organizmów został zaimplementowany w języku C# w programie dotPicker działającym pod systemem operacyjnym Windows. Program dotPicker jest dostępny pod adresem http://www.uz.zgora.pl/~akaspers/others.html. Algorytm zaproponowanej metody jest rozszerzeniem metody macierzy kropkowej (ang. dot-matrix method) o możliwość m.in. wyszukiwania podobieństw dla więcej niż dwóch sekwencji. Czyszczenie macierzy kropkowej zostało wykonane dla wielkości okna (ang. window size) równej 100 i progu identyczności (ang. identity threshold) równego 10 (Kasperski i Kasperska, 2014). Algorytm n-wymiarowej macierzy kropkowej jest powtarzany dla każdej z sekwencji ustawionej jako pierwsza i pozostałych sekwencji ustawionych w kolejności alfabetycznej (Kasperski i Kasperska, 2012). Współrzędne uzyskanych fragmentów identyczności po każdym kroku algorytmu są zapamiętywane, co umożliwia automatyczne wykrycie i połączenie sąsiednich fragmentów identyczności.

Nowa metoda wykorzystuje podejście semihomologiczne (ang. semihomology approach), które polega na wykrywaniu mutacji jednopunktowych (tj. tranzycji i transwersji) w kodonach porównywanych aminokwasów (Leluk, 1998, 2000 a, 2000 b; Leluk i in., 2001, 2003). W podejściu semihomologicznym wykorzystywany jest trójwymiarowy diagram relacji genetycznych występujących pomiędzy aminokwasami (Leluk i in., 2003). Podejście to umożliwia określenie możliwych mechanizmów zmienności i przebiegu zmienności ewolucyjnej (Kuśka i in., 2005).

Kolejne etapy działania algorytmu nowej metody są następujące:
a) w pierwszym kroku algorytmu pierwsza sekwencja (ze zbioru rozpatrywanych sekwencji) jest odkładana na osi Y macierzy, a druga sekwencja (ze zbioru rozpatrywanych sekwencji) jest odkładana na osi X. Uruchamiana jest standardowa metoda macierzy kropkowej z dodatkowym wykrywaniem pozycji semihomologicznych (tj. tranzycji i transwersji), w wyniku której otrzymujemy zbiór fragmentów identyczności;
b) w drugim kroku trzecia sekwencja (ze zbioru rozpatrywanych sekwencji) jest odkładana na osi X macierzy. Standardowa metoda macierzy kropkowej z wykrywaniem pozycji semihomologicznych jest powtarzana dla każdego otrzymanego w pierwszym kroku fragmentu identyczności odkładanego kolejno na osi Y. W rezultacie otrzymujemy zmodyfikowany zbiór fragmentów identyczności;
c) w ostatnim kroku ostatnia sekwencja (ze zbioru rozpatrywanych sekwencji) jest odkładana na osi X macierzy. Standardowa metoda macierzy kropkowej z wykrywaniem pozycji semihomologicznych jest powtarzana dla każdego otrzymanego w przedostatnim kroku fragmentu identyczności odkładanego kolejno na osi Y. W rezultacie otrzymujemy zmodyfikowany zbiór fragmentów identyczności będący wynikowym zbiorem fragmentów identyczności dla danej sekwencji ustawionej jako pierwsza;
d) z wynikowego zbioru fragmentów identyczności (tj. zbioru fragmentów identyczności określonego w punkcie c) wybierany jest najlepszy fragment identyczności (zgodnie ze stosowaną w tej pracy oceną fragmentu identyczności);
e) powyższy algorytm jest powtarzany dla każdej sekwencji (ze zbioru rozpatrywanych sekwencji) odkładanej jako pierwsza na osi Y macierzy.

Stosowana w tej pracy ocena fragmentu identyczności polega na określeniu ilości pozycji dla otrzymanego fragmentu identyczności:
a) pozycji konserwatywnych typu:
"R" — w danej pozycji występują te same aminokwasy,
"+" — występuje co najmniej 80% porównań tych samych aminokwasów,
b) pozycji semikonserwatywnych typu:
"#" — w danej pozycji występują tylko porównania tych samych aminokwasów i tranzycje,
"@"— występują tylko porównania tych samych aminokwasów, tranzycie i transwersje,
"$" — występują tylko porównania tych samych aminokwasów i transwersje,
"*" — występuje co najmniej 80% porównań tych samych aminokwasów, tranzycje i transwersji w danej pozycji.
"-" — oznacza pozostałe pozycje z dwupunktowymi lub trójpunktowymi mutacjami.

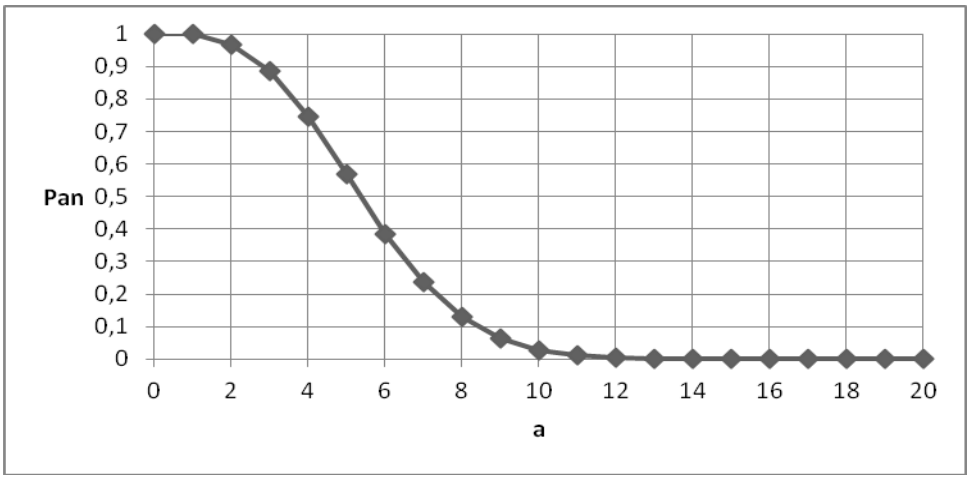
Najlepsza ocena fragmentu identyczności oznacza kolejno największą liczbę pozycji konserwatywnych (tj. największą sumę pozycji "R" i "+") oraz największą liczbę pozycji semikonserwatywnych (tj. największą sumę pozycji "#", "@", "$" i "*").

Statystyczną ocenę przyrównań uzyskano poprzez obliczanie współczynnika P_{an} dla każdego otrzymanego fragmentu identyczności. Współczynnik P_{an} obliczany był według wzoru (Kasperski i Kasperska, 2012):

$$P_{an} = \frac{\sum_{k=a}^{n} \binom{n}{k} x^k (x(x-1))^{n-k}}{x^{2n}}$$

gdzie:
x = 20 dla sekwencji aminokwasowych,
n — długość fragmentu identyczności,
a — ilość identycznych pozycji.

Wartość współczynnika P_{an} przyjmuje wartości z przedziału [0, 1] i informuje o losowości danego przyrównania. Wartość P_{an} = 1 oznacza, że przyrównanie jest całkowicie losowe, tzn., że otrzymany fragment identyczności jest całkowicie losowy. Wartość P_{an} = 0 oznacza, że przyrównanie jest całkowicie nielosowe, tzn., że otrzymany fragment identyczności jest całkowicie nielosowy. Wartość współczynnika P_{an} może zostać obliczona np. z wykorzystaniem programu dotPicker po wybraniu opcji Tools -> Analyser. Na rysunku 1 została przedstawiona przykładowa zależność współczynnika P_{an} od ilości identycznych pozycji (tj. pozycji "a") dla fragmentu identyczności o długości n = 100. Zgodnie z rysunkiem 1 wzrost ilości identycznych pozycji powoduje zmniejszanie się współczynnika P_{an}, co świadczy o zmniejszaniu się losowości przyrównania wraz ze wzrostem ilości identycznych pozycji w danym przyrównaniu.



Rys. 1. Przykładowa zależność współczynnika P_{an} od ilości identycznych pozycji
Fig. 1. The exemplary dependence of the P_{an} factor on the number of identical positions

OMÓWIENIE WYNIKÓW I DYSKUSJA

W pracy zostaną zaprezentowane wyniki analizy inhibitorów proteinaz z nasion dyniowatych oraz cytochromu c wybranych roślin uprawnych.

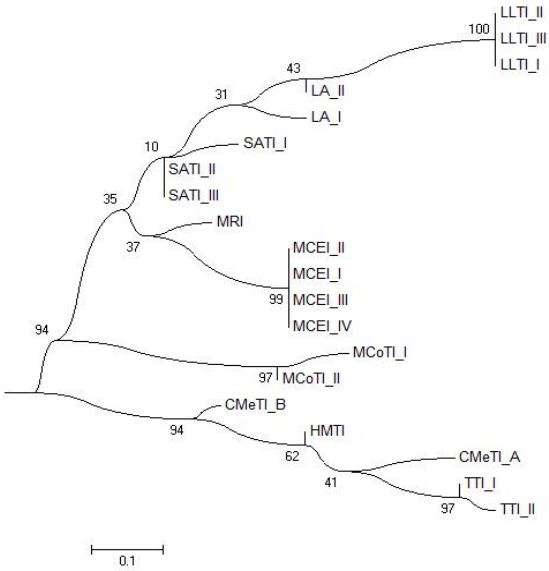
Badanie zmienności proteinaz z nasion dyniowatych

Analizę zmienności genetycznej proteinaz z nasion dyniowatych rozpoczęto od obliczenia logarytmu wiarygodności drzew zbudowanych metodą ML (maksymalnej wiarygodności, ang. Maximum Likelihood) dla tych sekwencji i różnych modeli substytucji aminokwasów (tab. 1). Metoda ML jest metodą statystyczną, umożliwiającą określenie drzewa, którego wiarygodność jest największa (Higgs i Attwood, 2005). Najbardziej wiarygodne drzewo (o największej wartości logarytmu wiarygodności) otrzymano dla modelu WAG (tab. 1).

Tabela 1

Logarytmy wiarygodności drzew ML dla proteinaz z nasion dyniowatych
Logarithms of ML tree likelihoods for proteinase inhibitors from squash seeds

Model substytucji aminokwasów Amino-acid substitution model	Logarytm wiarygodności drzewa Logarithm of tree likelihood
Dayhoffa	-320,29
JTT	-321,59
LG	-328,65
Poissona	-344,29
WAG	-320,14



Rys. 2. Drzewo ML dla inhibitorów proteinaz z nasion dyniowatych, otrzymane dla modelu WAG substytucji aminokwasów
Fig. 2. ML tree for proteinase inhibitors from squash seeds, obtained for the amino-acid substitution WAG model

Drzewo to zostało przedstawione na rysunku 2. W drzewie ML wiarygodności kladów, zawierających sekwencje położone najbliżej korzenia (tj. sekwencje SATI_II i SATI_III), wynoszą odpowiednio: 94%, 35% i 10% (rys. 2). Nie ma reguły, która określałaby minimalną wartość określoną metodą samopróbkowania, aby uznać węzeł za wiarygodny.

W praktyce przyjmuje się, że taką minimalną wartością wiarygodności jest 70% (Higgs i Attwood, 2005). Z tego powodu określenie na podstawie rysunku 2, które sekwencje w rzeczywistości leżą najbliżej korzenia, wymaga dodatkowego sprawdzenia. Uzyskane wyniki sprawdzono wykonując dwupoziomową analizę zmienności genetycznej organizmów. Ustawiając każdą z sekwencji jako pierwszą otrzymano fragmenty identyczności oraz ich oceny, które zostały przedstawione na rysunku 3.

sekwencja ustawiona jako pierwsza	**fragment identyczności**	**ocena fragmentu identyczności [R/+/#/@/$/*/-]**
SATI_II	CP@I-+-C**+-DC@-@C@C---G*CG	[11/2/0/4/0/3/7]
SATI_III	CP@I-+-C**+-DC@-@C@C---G*CG	[11/2/0/4/0/3/7]
SATI_I	CP@I-+-C**+-DC@-@C-C-*-G*CG	[11/2/0/3/0/4/7]
LA_II	CP@I-+-C*-+-DC@*@C@C---G-CG	[11/2/0/4/0/2/8]
LA_I	CP@I-+-C--+-DC**@C@C---G*CG	[11/2/0/3/0/3/8]
CMeTI_B	CP@I-+*C*-+-DC@--C-C--*G*CG	[11/2/0/2/0/4/8]
HMTI	CP@I-+*C*-+-DC@--C-C--*G*CG	[11/2/0/2/0/4/8]
MRI	CP@I-+-C**+-DC@--C-C---G*CG	[11/2/0/2/0/3/9]
MCEI_I	CP*I-+-C**+-DC@--C@C---G-CG	[11/2/0/2/0/3/9]
MCEI_II	CP*I-+-C**+-DC@--C@C---G-CG	[11/2/0/2/0/3/9]
MCEI_III	CP*I-+-C**+-DC@--C@C---G-CG	[11/2/0/2/0/3/9]
MCEI_IV	CP*I-+-C**+-DC@--C@C---G-CG	[11/2/0/2/0/3/9]
CMeTI_A	CP@I-+-C*-+-DC*--C-C--*G*CG	[11/2/0/1/0/4/9]
LLTI_I	CP@I-+-C*-+-DC@--C-C---G-CG	[11/2/0/2/0/1/11]
LLTI_II	CP@I-+-C*-+-DC@--C-C---G-CG	[11/2/0/2/0/1/11]
LLTI_III	CP@I-+-C*-+-DC@--C-C---G-CG	[11/2/0/2/0/1/11]
MCoTI_II	CP-I-$*C@*+-DC**-C@C-**G*CG	[11/1/0/2/1/7/5]
MCoTI_I	CP-I---C@*+-DC**-C@C-**G*CG	[11/1/0/2/0/6/7]
TTI_I	CP@I-+-C*-#-DC@--C-C--*G*CG	[11/1/1/2/0/3/9]
TTI_II	CP@I-+-C*-#-DC@--C-C--*G*CG	[11/1/1/2/0/3/9]

Rys. 3. Analiza inhibitorów proteinaz z nasion dyniowatych dla każdej sekwencji ustawionej jako pierwsza
Fig. 3. Analysis of proteinase inhibitors from squash seeds for each sequence put as a first one

Najlepszą ocenę otrzymanych fragmentów identyczności uzyskano dla sekwencji SATI_II i SATI_III ustawionych jako pierwsze. Najlepsza ocena może świadczyć o tym, że sekwencje SATI_II i SATI_III najmniej zmutowały podczas ewolucji w stosunku do pozostałych sekwencji i dlatego ich umieszczenie najbliżej korzenia drzewa jest uzasadnione. Z rysunku 3 wynika także, że największą różnicę w stosunku do pozostałych sekwencji wykazują sekwencje TTI_I i TTI_II, tzn. te sekwencje powinny być umiejscowione w dużej odległości od korzenia drzewa, co jest zgodne z wynikami przedstawionymi na rysunku 2.

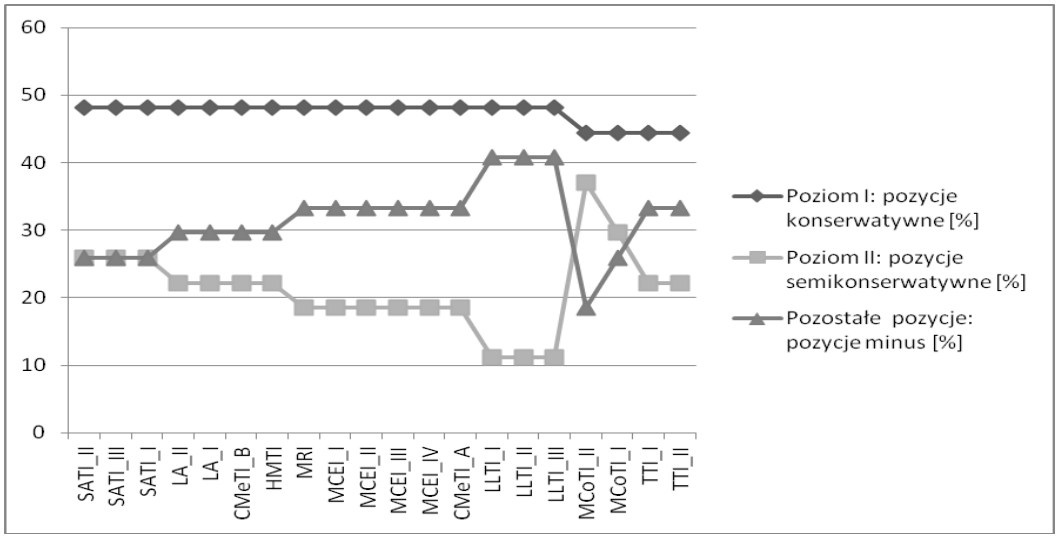
Na rysunku 4 przestawiono przykładową analizę inhibitorów proteinaz z nasion dyniowatych dla sekwencji SATI_II ustawionej jako pierwsza. Wynikowy fragment identyczności CP@I-+-C**+-DC@-@C@C---G*CG jest „zawieszony” w 20-wymiarowej przestrzeni sekwencji. W pozycji 10 uzyskano "@", ponieważ w tej pozycji istnieje 13 homologii (porównania R z R), 4 transwersje (porównania R z L), dwie tranzycje (porównania R z K). W pozycji 13 uzyskano "+", ponieważ w tej pozycji istnieje ponad 80% homologii

(porównań M z M). W pozycji 16 uzyskano "*", ponieważ w tej pozycji istnieje 15 homologii (porównania K z K), jedna transwersja (porównanie K z Q), dwie tranzycje (porównania K z R), jedna mutacja dwupunktowa (porównanie K z S).

```
                         1234567890123456789012345678901234 5
SATI_II                  ----GRICPRILMECKRDSDCLAECICQ-SGYCG-
CMeTI_A                  -----GICPRILMPCKTDDDCMLDCRCLSNGYCG-
CMeTI_B                  -----VGCPRILMKCKTDRDCLTGCTCKRNGYCG-
HMTI                     -----VGCPRILMKCKTDDDCLLGCKCLSNGYCG-
LA_I                     ------ICPRILMECSHDSDCFGECICLSSGYCG-
LA_II                    -----IRCPRIYMECKHDSDCLGECICLESGFCG-
LLTI_I                   ----ERRCPRIYMECKHDSDCLADCVCLEHGICGG
LLTI_II                  -----RRCPRIYMECKHDSDCLADCVCLEHGICG-
LLTI_III                 ----ERRCPRIYMECKHDSDCLADCVCLEHGICG-
MCEI_I                   -----RICPLIWMECKRDSDCLAQCIC-VDGHCG-
MCEI_II                  -----RICPLIWMECKRDSDCLAQCIC-VDGHCG-
MCEI_III                 ---EERICPLIWMECKRDSDCLAQCIC-VDGHCG-
MCEI_IV                  --EEERICPLIWMECKRDSDCLAQCIC-VDGHCG-
MCoTI_I                  SGSDGGVCPKILQRCRRDSDCPGACICRGNGYCG-
MCoTI_II                 SGSDGGVCPKILKKCRRDSDCPGACICRGNGYCG-
MRI                      -----GICPRILMECKRDSDCLAQCVCKRQGYCG-
SATI_I                   -----RVCPRILMRCKRDSDCLAECTCQGSGYCG-
SATI_III                 --ERGRICPRILMECKRDSDCLAECICQ-SGYCG-
TTI_I                    -------CPRILMPCKVNDDCLRGCKCLSNGYCG-
TTI_II                   -------CPRILMPCQVNDDCLRGCKCLSNGYCG-
uzyskany fragment              CP@I-+-C**+-DC@-@C@C---G*CG
identyczności
```

Rys. 4. Przykładowa analiza inhibitorów proteinaz z nasion dyniowatych dla sekwencji SATI_II ustawionej jako pierwsza
Fig. 4. Exemplary analysis of proteinase inhibitors from squash seeds for the SATI_II sequence put as a first one

Dynamika zmienności proteinaz z nasion dyniowatych została przedstawiona na rysunku 5. Stosunkowo nieduża (poniżej 50%) ilość pozycji konserwatywnych (tj. pozycji typu "R" i "+") oraz duży zakres zmian (od ok. 10% do ok. 40%) pozycji semikonserwatywnych (tj. pozycji typu "#", "@", "$" i "*") i pozycji Minus ("-") świadczą o dużej zmienności proteinaz z nasion dyniowatych. Jest to wynik zgodny z wynikiem uzyskanym dla drzewa filogenetycznego (rys. 2), dla którego otrzymane dystanse ewolucyjne są również duże (skala drzewa filogenetycznego równa 0,1).



Rys. 5. Dynamika zmienności inhibitorów proteinaz z nasion dyniowatych rozpatrywana na dwóch poziomach

Poziom I: pozycje konserwatywne [%] = 100*("R"+"+")/ilość pozycji
Poziom II: pozycje semikonserwatywne [%] = 100*("#"+"@"+"$"+"*")/ilość pozycji
Pozycje minus [%] = 100*("-")/ ilość pozycji

Fig. 5. Variability dynamics of proteinase inhibitors from squash seeds considered at two levels

Level I: conservative positions [%] = 100*("R"+"+")/ number of positions
Level II: semiconservative positions [%] = 100*("#"+"@"+"$"+"*")/ number of positions
Minus positions [%] = 100*("-")/ number of positions

		[R/+/#/@/$/*/-]	P_{an}
SATI_III	GRICPRILMECKRDSDCLAECICQ-SGYCG	[29/0/0/0/0/0/1]	6.67E-37
SATI_I	R#CPRILM-CKRDSDCLAEC#CQ-SGYCG	[25/0/2/0/0/0/2]	7.40E-29
MRI	ICPRILMECKRDSDCLA$C#C$--GYCG	[23/0/1/0/2/0/2]	1.18E-25
LA_I	ICPRILMEC-#DSDC$$ECIC$-SGYCG	[22/0/1/0/3/0/2]	8.65E-24
MCEI_I	RICPIMECKRDSDCLA$CIC---G#CG	[22/0/1/0/3/0/3]	3.36E-23
MCEI_II	RICPIMECKRDSDCLA$CIC---G#CG	[22/0/1/0/3/0/3]	3.36E-23
MCEI_III	RICPIMECKRDSDCLA$CIC---G#CG	[22/0/1/0/3/0/3]	3.36E-23
MCEI_IV	RICPIMECKRDSDCLA$CIC---G#CG	[22/0/1/0/3/0/3]	3.36E-23
LA_II	CPRI-MECK#DSDCL$ECIC$-SG$CG	[21/0/1/0/3/0/2]	1.38E-22
LLTI_I	R$CPRI-MECK#DSDCLA$C#C$--G-CG	[20/0/2/0/3/0/4]	7.84E-20
LLTI_II	R$CPRI-MECK#DSDCLA$C#C$--G-CG	[20/0/2/0/3/0/4]	7.84E-20
LLTI_III	R$CPRI-MECK#DSDCLA$C#C$--G-CG	[20/0/2/0/3/0/4]	7.84E-20
MCoTI_II	G##CP#IL$#C#RDSDC#$$CIC#-#GYCG	[18/0/8/0/3/0/1]	2.32E-16
MCoTI_I	G##CP#IL--C#RDSDC#$$CIC#-#GYCG	[18/0/7/0/2/0/3]	2.32E-16
CMeTI_B	CPRILM#CKDDCL##C#C$-#GYCG	[18/0/5/0/3/0/1]~	1.52E-17
HMTI	CPRILM#CK$D-DCL-#C$C$-#GYCG	[18/0/3/0/3/0/3]	1.52E-17
CMeTI_A	ICPRILM-CK$D-DC$-CC$-#GYCG	[18/0/1/0/5/0/4]	3.98E-17
TTI_I	CPRILM-CK-#-DCL-#CC-#GYCG	[17/0/3/0/2/0/5]	5.21E-16
TTI_II	CPRILM-C$-#-DCL-#C$C$-#GYCG	[16/0/3/0/3/0/5]	1.54E-14

Rys. 6. Analiza inhibitorów proteinaz z nasion dyniowatych dla porównania każdej sekwencji z sekwencją SATI_II. Współczynnik P_{an} określa statystyczną ocenę uzyskanych przyrównań

Fig. 6. Analysis of proteinase inhibitors from squash seeds for comparison of each sequence with the SATI_II sequence. The P_{an} factor determines the statistical evaluation of the alignments

Dodatkowe sprawdzenie wykonano dla jednej z sekwencji uznawanej za najbliżej położoną od korzenia, tj. porównując sekwencję SATI_II po kolei ze wszystkimi pozostałymi sekwencjami. Na rysunku 6 przedstawiono wyniki porównania, w kolejności od wyniku mającego ocenę najlepszą do wyniku o najgorszej ocenie. Najmniejszy dystans od sekwencji SATI_II, przy zastosowaniu hipotezy jednopunktowej mutacji, uzyskano dla sekwencji SATI_III, SATI_I i MRI, co jest zgodne również z rysunkiem 2. Ponadto, wyniki przedstawione na rysunku 6 pokazują dużą zgodność zaproponowanego sposobu oceny fragmentów identyczności (wartości podane w nawiasie kwadratowym) z wartością współczynnika P_{an}, który określa ich statystyczną ocenę. Zwiększenie zmienności poszczególnych sekwencji w stosunku do sekwencji SATI_II powoduje zwiększenie współczynnika P_{an} (wyjątek stanowi pozycja oznaczona za pomocą znaku "~").

Badanie zmienności wybranych roślin uprawnych na podstawie analizy cytochromu c

Analizę zmienności genetycznej wybranych roślin uprawnych rozpoczęto od obliczenia logarytmu wiarygodności drzew zbudowanych metodą ML dla cytochromu c i różnych modeli substytucji aminokwasów (tab. 2).

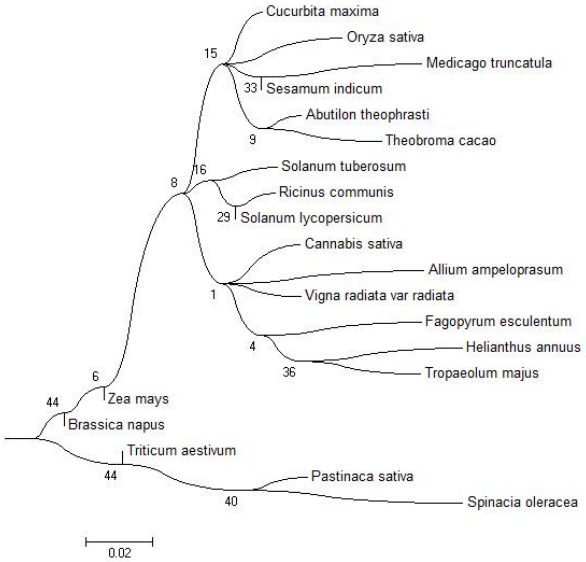
Tabela 2

Logarytmy wiarygodności drzew ML dla cytochromu c wybranych roślin uprawnych
Logarithms of ML tree likelihoods for cytochrome c of selected crops

Model substytucji aminokwasów Amino-acid substitution model	Logarytm wiarygodności drzewa Logarithm of tree likelihood
Dayhoffa	-637,77
JTT	-647,19
LG	-658,64
Poissona	-703,36
WAG	-648,84

Najbardziej wiarygodne drzewo (o największej wartości logarytmu wiarygodności) otrzymano dla modelu Dayhoffa (tab. 2). Drzewo to zostało przedstawione na rysunku 7. Z tabeli 2 i rysunku 7 wynika, że zarówno logarytm wiarygodności otrzymanego drzewa, jak i wiarygodność poszczególnych kladów są stosunkowo małe. Małą wiarygodność kladów (tj. poniżej 50%) dla cytochromu c wybranych roślin uprawnych uzyskiwano zarówno dla modelu Dayhoffa, jak i innych modeli substytucji aminokwasów (tj. dla modeli JTT, LG, Poissona, WAG) oraz dla większych ilości powtórzeń równej 2000 w metodzie samopróbkowania. Mała wiarygodność kladów powoduje konieczność dodatkowego sprawdzenia uzyskanych wyników. Sprawdzenie wyników wykonano wykonując dwupoziomową analizę zmienności genetycznej organizmów.

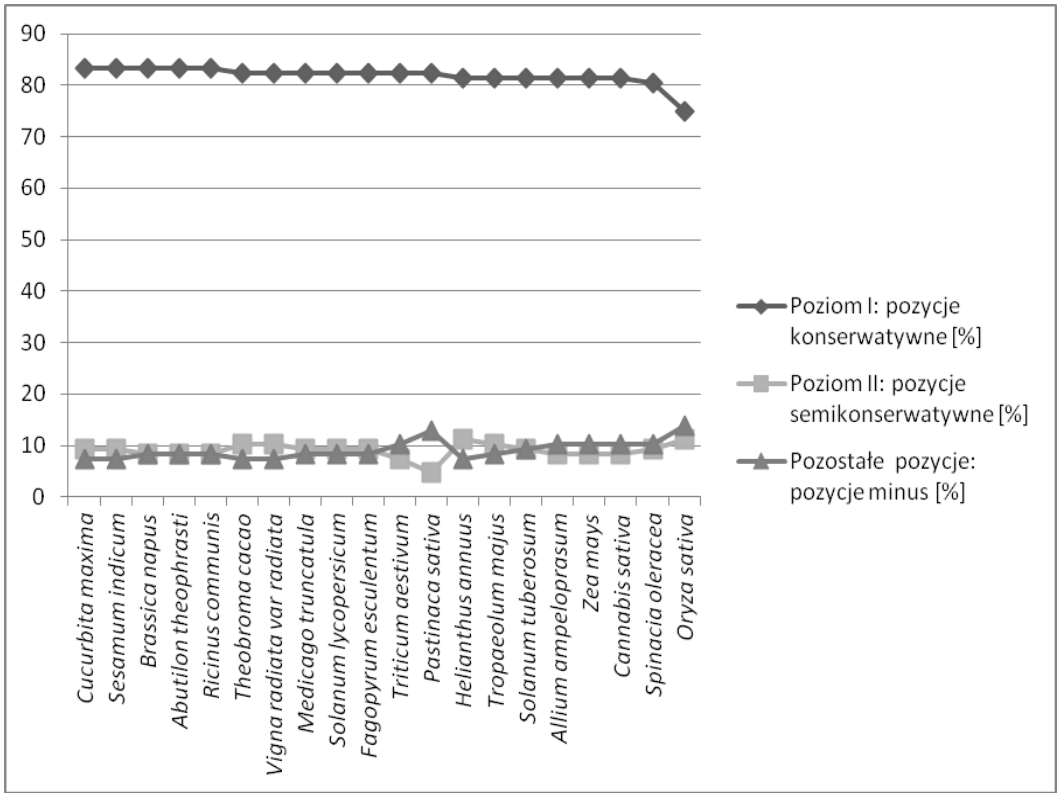
Ustawiając każdą z sekwencji jako pierwszą otrzymano fragmenty identyczności o ocenach przedstawionych na rysunku 8. Rysunek 8 pokazuje, że podobieństwo pomiędzy sekwencjami jest duże, tzn. ilość pozycji konserwatywnych ("R", "+") jest znacznie większa od ilości pozycji "-", dla każdej sekwencji ustawionej jako pierwsza. Można również zauważyć, że ilość pozycji "$", jest większa od ilości pozycji "#". Ponadto, ilość pozycji "*" jest większa zarówno od ilości pozycji "$" (dla *Pastinaca sativa* równa), jak i pozycji "@" (dla *Oryza sativa* równa).



Rys. 7. Drzewo ML dla cytochromu c wybranych roślin uprawnych, otrzymane dla modelu Dayhoffa substytucji aminokwasów
Fig. 7. ML tree for cytochrome c of selected crops, obtained for the amino-acid substitution Dayhoff's model

	[R/ +/#/@/$/*/-]
Cucurbita maxima	[71/19/0/4/1/5/8]
Sesamum indicum	[71/19/0/2/2/6/8]
Brassica napus	[71/19/0/3/2/4/9]
Abutilon theophrasti	[71/19/0/3/2/4/9]
Ricinus communis	[71/19/0/2/3/4/9]
Theobroma cacao	[71/18/1/3/2/5/8]
Vigna radiata var radiata	[71/18/0/2/3/6/8]
Medicago truncatula	[71/18/1/1/2/6/9]
Solanum lycopersicum	[71/18/0/2/3/5/9]
Fagopyrum esculentum	[71/18/0/2/3/5/9]
Triticum aestivum	[71/18/0/2/2/4/11]
Pastinaca sativa	[71/18/0/1/2/2/14]
Helianthus annuus	[71/17/0/2/4/6/8]
Tropaeolum majus	[71/17/0/2/3/6/9]
Solanum tuberosum	[71/17/0/2/3/5/10]
Allium ampeloprasum	[71/17/1/1/3/4/11]
Zea mays	[71/17/0/3/2/4/11]
Cannabis sativa	[71/17/0/2/3/4/11]
Spinacia oleracea	[71/16/2/1/3/4/11]
Oryza sativa	[71/10/1/4/3/4/15]

Rys. 8. Analiza cytochromu c wybranych roślin uprawnych dla każdej sekwencji ustawionej jako pierwsza
Fig. 8. Analysis of cytochrome c of selected crops for each sequence put as a first one



Rys. 9. Dynamika zmienności wybranych roślin uprawnych na podstawie analizy cytochromu c rozpatrywana na dwóch poziomach

Poziom I: pozycje konserwatywne [%] = 100*("R"+"+")/ilość pozycji
Poziom II: pozycje semikonserwatywne [%] = 100*("#"+"@"+"$"+"*")/ilość pozycji
Pozycje minus [%] = 100*("-")/ ilość pozycji

Fig. 9. Variability dynamics of selected crops based on the analysis of cytochrome c considered at two levels

Level I: conservative positions [%] = 100*("R"+"+")/number of positions
Level II: semiconservative positions [%] = 100*("#"+"@"+"$"+"*")/ number of positions
Minus positions [%] = 100*("-")/ number of positions

Dynamika zmienności wybranych roślin uprawnych, określona na podstawie analizy cytochromu c, została przedstawiona na rysunku 9. Duża ilość (powyżej 70%) pozycji konserwatywnych (tj. pozycji "R" i "+") oraz stosunkowo mała (poniżej 15%) ilość pozycji semikonserwatywnych (tj. pozycji typu "#", "@", "$" i "*") i pozycji Minus ("-") mogą świadczyć o małej zmienności cytochromu c wybranych roślin uprawnych. Jest to wynik zgodny z wynikiem otrzymanym dla drzewa filogenetycznego (rys. 7), dla którego otrzymane dystanse ewolucyjne są również małe (skala drzewa filogenetycznego równa 0,02).

Dodatkowe sprawdzenie wykonano porównując sekwencję *Bassica napus* (którą można uznać za leżącą najbliżej korzenia na podstawie rysunków 7 i 8) po kolei ze wszystkimi

sekwencjami. Uporządkowane wyniki porównania, od oceny najlepszej do najgorszej, przedstawiono na rysunku 10. Dla tej analizy ilość transwersji (tj. pozycji "$") przy porównaniu każdej sekwencji z sekwencją *Brassica napus* jest zazwyczaj większa (z trzema wyjątkami dla *Pastinaca sativa*, *Cucurbita Maxima* i *Oryza sativa*) od ilości tranzycji (tj. pozycji "#"). Najmniejszy dystans od sekwencji *Bassica napus* przy zastosowaniu hipotezy jednopunktowej mutacji uzyskano dla sekwencji *Cucurbita maxima*, co jest zgodne z wynikiem przedstawionym na rysunku 8. Rysunki 8 i 10 potwierdzają także dużą odległość sekwencji *Oryza sativa* i *Spinacia oleracea* od korzenia drzewa. Ponadto, wyniki przedstawione na rysunku 10 pokazują dużą zgodność zaproponowanego sposobu oceny fragmentów identyczności (wartości podane w nawiasie kwadratowym) z wartością współczynnika P_{an}, który określa ich statystyczną ocenę. Zwiększenie zmienności poszczególnych sekwencji w stosunku do sekwencji *Brassica napus* powoduje zwiększenie współczynnika P_{an} (z jednym wyjątkiem, który oznaczono za pomocą znaku "~").

	[R/+/#/@/$/*/-]	**P_{an}**
Cucurbita maxima	[108/0/1/0/1/0/2]	1.57E-134
Vigna radiata var radiata	[106/0/1/0/3/0/2]	2.18E-129
Zea mays	[103/0/1/0/4/0/4]	3.44E-122
Solanum lycopersicum	[102/0/3/0/5/0/2]	6.74E-120
Solanum tuberosum	[102/0/3/0/4/0/3]	6.74E-120
Triticum aestivum	[102/0/2/0/3/0/4]~	6.33E-121
Ricinus communis	[101/0/1/0/4/0/6]	1.19E-117
Theobroma cacao	[100/0/1/0/8/0/1]	2.23E-117
Tropaeolum majus	[100/0/2/0/5/0/5]	1.90E-115
Sesamum indicum	[100/0/1/0/5/0/6]	1.90E-115
Medicago truncatula	[99/0/2/0/7/0/2]	3.86E-115
Abutilon theophrasti	[99/0/1/0/6/0/6]	2.78E-113
Helianthus annuus	[98/0/2/0/8/0/2]	6.06E-113
Fagopyrum esculentum	[98/0/1/0/7/0/5]	4.91E-112
Cannabis sativa	[98/0/2/0/3/0/9]	3.74E-111
Pastinaca sativa	[97/0/3/0/2/0/10]	4.64E-109
Allium ampeloprasum	[96/0/1/0/6/0/9]	5.35E-107
Spinacia oleracea	[95/0/3/0/5/0/8]	9.17E-106
Oryza sativa	[92/0/3/0/3/0/13]	9.00E-100

Rys. 10. Analiza cytochromu c wybranych roślin uprawnych dla porównania każdej sekwencji z sekwencją *Brassica napus*. Współczynnik P_{an} określa statystyczną ocenę uzyskanych przyrównań
Fig. 10. Analysis of cytochrome c of selected crops for comparison of each sequence with the *Brassica napus* sequence. The P_{an} factor determines the statistical evaluation of the alignments

PODSUMOWANIE I WNIOSKI

W pracy przedstawiono analizę zmienności genetycznej wybranych organizmów za pomocą nowej metody. W przedstawionej metodzie porównywanie sekwencji następuje na poziomie aminokwasowym i w przypadku, gdy porównywane aminokwasy są różne

następuje automatyczna zmiana poziomu i porównanie danych aminokwasów jest wykonywane na poziomie ich kodonów. Implementacja nowej metody wskazuje na możliwości usunięcia ograniczeń standardowej metody macierzy kropkowej poprzez: a) wprowadzenie sposobu oceny otrzymanych fragmentów identyczności, który jest zgodny ze statystyczną oceną przyrównań, określoną przez współczynnik P_{an}, b) zaproponowanie algorytmu n-wymiarowej macierzy kropkowej, który rozszerza możliwości standardowej metody macierzy kropkowej o możliwość przyrównania dowolnej ilości sekwencji oraz uwzględnienie podejścia semihomologicznego, c) zapamiętywanie w czasie obliczeń współrzędnych odnalezionych fragmentów identyczności, co umożliwia automatyczne wykrycie i łączenie sąsiednich fragmentów identyczności. Metoda poprawnie działa zarówno dla dużych zbiorów sekwencji, jak i dla zbiorów sekwencji o niewysokim stopniu podobieństwa. Dla sekwencji o niewysokim stopniu podobieństwa ilość fragmentów identyczności, otrzymywanych po każdym z etapów działania algorytmu, może szybko maleć. W przypadku, gdy ilość fragmentów identyczności zmaleje do zera, konieczna może być m.in. zmiana wartości parametrów, tj. zwiększenie wielkości okna i/lub zmniejszenie progu identyczności w celu zwiększenia ilości otrzymywanych fragmentów identyczności. W przypadku dużych zbiorów sekwencji (tj. powyżej 50) konieczne może być również podzielenie wejściowego zbioru sekwencji na mniejsze podzbiory i rozpatrywanie każdego podzbioru osobno. W przypadku zbiorów sekwencji o dużym stopniu podobieństwa ilość otrzymywanych fragmentów identyczności po każdym z etapów działania algorytmu może być bardzo duża (np. kilkaset), przez co czas obliczeń może być długi. W takim przypadku konieczne może być m.in. zmniejszenie wielkości okna i/lub zwiększenie progu identyczności w celu redukcji ilości otrzymywanych fragmentów identyczności. Zastosowanie dwupoziomowej analizy zmienności genetycznej organizmów ułatwia zrozumienie i interpretację wyników, szczególnie interpretację zmian na poszczególnych pozycjach oraz ocenę otrzymanych fragmentów identyczności w kontekście jednopunktowej mutacji. Zaproponowany sposób określenia charakterystyki zmian (rys. 5 i 9) pozwolił na ocenę dynamiki zmian inhibitorów proteinaz z nasion dyniowatych (duża zmienność) oraz sekwencji cytochromu c wybranych roślin uprawnych (mała zmienność). Dynamiki zmian określone za pomocą przedstawionej metody dwupoziomowej analizy zmienności, są zgodne z wynikami uzyskanymi dla drzew filogenetycznych otrzymanych metodą maksymalnej wiarygodności (rys. 2 i 7).

LITERATURA

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J. F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36: 465 — 469.

Hall B. 2008. Łatwe drzewa filogenetyczne. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

Higgs P. G., Attwood T. K. 2005. Bioinformatics and Molecular Evolution. Blackwell Publishing Company, Oxford.

Janaszek M., Mańkowski D. R., Kozdój J. 2011. Sieci neuronowe typu MLP w prognozowaniu plonu jęczmienia jarego. Biul. IHAR 259: 93 — 112.

Kasperski A., Kasperska R. 2012. A novel method of sequence similarity evaluation in n-dimensional sequence space. Curr. Bioinformatics 7: 295 — 303.

Kasperski A., Kasperska R. 2014. Identification of protein family representatives. Curr. Bioinformatics 9: 414 — 425.
Kasperski A., Sun K., Tian Y. 2016. New approach to control of the dissolved oxygen concentration in a biomass-driven self-cycling biochemical process. Chem. Eng. Commun. 203 (1): 75 — 93.
Kuśka J., Leluk J., Lesyng B. 2005. Zmienność mutacyjna w homologicznych rodzinach kinaz. Bio-Algorithms and Med-Systems 1 (1/2): 125 — 128.
Leluk J. 1998. A new algorithm for analysis of the homology in protein primary structure. Comput. Chem. 22 (1): 123 — 131.
Leluk J. 2000 a. Serine proteinase inhibitor family in squash seeds: mutational variability mechanism and correlation. Cell. Biol. Mol. Lett. 5: 91 — 106.
Leluk J. 2000 b. Regularities in mutational variability in selected protein families and the Markovian model of amino acid replacement. Comput. Chem. 24(1): 659 — 672.
Leluk J., Hanus-Lorenz B., Sikorski A. F. 2001. Application of genetic semihomology algorithm to theoretical studies on various protein families. Acta Biochim. Pol. 48: 21 — 33.
Leluk J., Konieczny L., Roterman I. 2003. Search for structural similarity in proteins. Bioinformatics 19: 117 — 124.
Tamura K., Stecher G., Peterson D., Filipski A. i Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30: 2725 — 2729.
Thompson J. D., Higgins D. G., Gibson T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673 — 4680.
Xiong J. 2006. Essential Bioinformatics, Cambridge University Press.