Gdański Uniwersytet Medyczny

Agata Czerniecka

Nowa metoda obliczeniowa porównywania sekwencji białek

Rozprawa doktorska

Promotor: dr hab. Dorota Bielińska-Wąż

Zakład Informatyki Radiologicznej i Statystyki

Gdańskiego Uniwersytetu Medycznego

Gdańsk, 2017

Podziękowania

Dr hab. Dorocie Bielińskiej-Wąż
za pomoc, motywację i niewyczerpalne pokłady cierpliwości

mojemu mężowi Rafałowi i mojej rodzinie za wspracie i wiarę we mnie

Streszczenie

Metody porównywania sekwencji białek alternatywne w stosunku do powszechnie stosowanych metod opartych na dopasowywaniu sekwencji należą do nowej generacji metod w ramach nauk bioinformatycznych. Okazuje się, że za pomocą nowych metod można ujawnić różne aspekty podobieństwa, które standardowo nie są rozpatrywane. W obecnej pracy sformułowano nową metodę porównywania sekwencji białek, która uwzględnia rozkład poszczególnych aminokwasów biogennych w sekwencji analogicznie jak uwzględnia się rozkład mas bryły sztywnej w dynamice (A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16-23. (2016 Impact Factor 2.801)).

W celu uzyskania nowych wielkości numerycznych (tak zwanych deskryptorów) opisujących sekwencje białkowe zostały zastosowane metody, które do tej pory nie były stosowane do opisu tego typu obiektów. Sekwencja aminokwasów została przedstawiona jako zestaw punktów materialnych w 20-wymiarowej przestrzeni. Dystrybucje punktów w przestrzeni uzyskano za pomocą metody przesunięć. W ten sposób został otrzymany abstrakcyjny 20-wymiarowy dynamiczny graf. Jako nowe deskryptory sekwencji zostały zaproponowane unormowane główne momenty bezwładności tych grafów.

Do ilustracji metody przeprowadzono analizę podobieństwa sekwencji dehydrogenazy NADH podjednostki 5 (ND5) dla dziewięciu różnych gatunków i podjednostki 6 (ND6) dla ośmiu różnych gatunków. Wykonano również dodatkowe obliczenia w celu porównania 20–wymiarowej dynamicznej reprezentacji sekwencji białkowych z innymi alternatywnymi metodami. Jako dane posłużyły następujące sekwencje: helikazy bakulowirusów dla jedenastu gatunków, ND5 dla dwudziestu dwóch gatunków i wirusa grypy typu A dla dwudziestu ośmiu gatunków.

Za pomocą nowej metody wyznaczono macierz podobieństwa i jej reprezentację za pomocą drzewa filogenetycznego. Zaletą zaproponowanej metody jest brak ograniczeń w długości sekwencji. 20-wymiarowa dynamiczna reprezentacja sekwencji białek jest wygodnym i niezawodnym narzędziem do rozwiązywania wielu problemów w medycynie i biologii, w których niezbędna jest analiza podobieństwa sekwencji białek. W szczególności metoda ta umożliwia tworzenie drzew filogenetycznych.

Abstract

Alignment-free methods constitute a fast developing branch of bioinformatics. It appears that by using new methods we are able to reveal different aspects of similarity, which normally would not be considered. These methods are not demanding computationally, and often offer both numerical and graphical tools for sequence comparisons. In the present work, a new method of comparison of protein sequences which takes into account the distribution of individual amino acids in the sequence has been introduced (A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16-23. (2016 Impact Factor 2.801)). This idea is analogous to the considering the mass distribution of the rigid body in the dynamics.

The sequence of amino acids is represented by a set of point masses in a 20D space. The distribution of points in the space is obtained by applying the method of „A Walk" in the 20D space. In this way, an abstract 20-dimensional dynamic graph may be obtained. 20D moments of inertia are proposed as new descriptors of protein sequences.

The calculations have been performed for a standard set of data used for introducing new alignment-free methods of comparison of protein sequences: the NADH dehydrogenase subunit 5 (ND5) protein sequences of nine species and for subunit 6 (ND6) of eight species. In order to compare 20D-dynamic Representation of Protein Sequences with alternative methods of other authors additional calculations were made. Computations have been performed for the following samples: helicase protein sequences of twelve baculoviruses, ND5 of twenty two species and neuraminidase proteins of influenza A virus of twenty eight species.

Using this new method, the similarity matrix is determined along with its representation by means of the phylogenetic tree. The advantage of the proposed method is the removal of limitations on lengths of the sequences. 20D-dynamic Representation of Protein Sequences is a convenient and reliable tool for sequence comparison. In particular, this method allows to create phylogenetic trees.

Spis treści

I. CEL

Celem rozprawy doktorskiej jest rozbudowanie klasy metod porównywania sekwencji białek. W szczególności, została opracowana nowa metoda obliczeniowa nazwana przez nas „20-wymiarowa dynamiczna reprezentacja sekwencji białek". Istotnym elementem rozprawy jest krytyczny przegląd literatury dotyczącej metod porównywania sekwencji białek, alternatywnych do tych które są oparte na dopasowaniu sekwencji.

II. TEZA

Niestandardowe metody porównywania sekwencji białek są wydajniejsze pod względem obliczeniowym i są w stanie wykazać więcej aspektów podobieństwa. Omawiane metody obliczeniowe są również wolne od niejednoznaczności, a ich wyniki mogą posłużyć do dalszych obliczeń, mających na celu ustalenie pokrewieństwa ewolucyjnego. Ponadto niniejsze metody mogą być wykorzystane w naukach biomedycznych, do rozwiązywania problemów, które wymagają badania podobieństwa sekwencji białek.

III. OSIĄGNIĘCIE CELU

Stworzenie nowej metody porównywania sekwencji białek, która nie tylko w sposób numeryczny, ale także graficzny pozwoli na przedstawienie badanych sekwencji. Przeprowadzenie analizy stanu wiedzy, związanego z metodami porównywania lub dopasowania sekwencji białek.

Nowa metoda została opublikowana w artykule: A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16-23. (2016 Impact Factor 2.801).

Metoda ta została przedstawiona na konferencji krajowej i międzynarodowej:

I Konferencja Doktorantów Pomorza BioMed Session, Gdańsk, 12.12.2015r., poster: *„20-wymiarowa Dynamiczna Reprezentacja Sekwencji Białek"*.

BIOMATH 2016 - International Conference on Mathematical Methods and Models in Biosciences, Blagoevgard, Bułgaria, 19-25.06.2016r., poster: *„20D-dynamic representation of protein sequences"*.

IV. UDOWODNIENIE TEZY I ROZWIĄZANIE PROBLEMU

Została opracowana nowa metoda, która została nazwana 20-wymiarową dynamiczną reprezentacją sekwencji białek. W metodzie tej, sekwencja białek jest reprezentowana przez zbiór punktów materialnych w przestrzeni 20-wymiarowej. Jako nowe charakterystyki liczbowe, zwane w tej dziedzinie deskryptorami, zostały zaproponowane 20-wymiarowe momenty bezwładności. Materiałem do badań są sekwencje białkowe, ogólnodostępne w bazie danych Protein Data Bank. Obliczenia zostały wykonane za pomocą metod numerycznych w programie MATLAB, a sama wizualizacja za pomocą programu Gnuplot.

WSTĘP

Chcąc odpowiedzieć sobie na pytanie jak bardzo podobne są do siebie interesujące nas gatunki bądź białka musimy znać ilościowy model ewolucji. Do wyznaczenia takiego modelu ewolucyjnego mogą nam posłużyć metody standardowe, czyli oparte na dopasowywaniu sekwencji. Pierwszym takim modelem była macierz PAM (1). Z macierzy podstawień PAM otrzymuje się macierze punktacji, które służą do wyznaczania dopasowań sekwencji białek. Okazuje się jednak, że metody standardowe są wymagające obliczeniowo, dlatego warto znaleźć alternatywne metody. Przełomem w tej dziedzinie okazują się być metody niestandardowe (*z ang. alignment-free methods)* polegające na wyznaczeniu w sposób numeryczny macierzy podobieństwa.

Metody niestandardowe miały możliwość rozwijania się dzięki połączeniu takich dziedzin jak informatyka, matematyka, fizyka oraz chemia. Stanowią one szybko rozwijającą się gałąź bioinformatyki. Te metody często nie wymagają skomplikowanych obliczeń, oferując za to zarówno numeryczne jak również graficzne narzędzia do porównywania sekwencji (tak zwane graficzne reprezentacje sekwencji biologicznych), które niewątpliwie są ich dodatkowym atutem.

Metody niestandardowe możemy podzielić na dwie grupy: Graficzne Reprezentacje i Metody Numeryczne, np. przesunięcie w przestrzeni wielowymiarowej. W przypadku przestrzeni wielowymiarowej jesteśmy ograniczeni tylko do metod numerycznych, chyba że przedstawimy rzuty tego 20-wymiarowego grafu w formie 2D lub 3D, wtedy możemy mówić również o reprezentacji graficznej. Początkowo metody niestandardowe były sformułowane tylko dla DNA (2,3), a od pewnego czasu sprawdzają się również w przypadku sekwencji białek (4,5). Niewątpliwie największym problemem matematycznym dla opisywanego zagadnienia jest wielowymiarowość z uwagi na 20 aminokwasów biogennych, które budują białka. W przypadku DNA czy RNA rozważamy tylko 4 zasady azotowe.

Jedną z głównych wad metod mających na celu porównać ze sobą białka jest zjawisko degeneracji, czyli fakt, iż różne sekwencje białkowe mogą być przedstawione przez ten sam wykres. Zjawisko degeneracji jest szczególnie zauważalne, w momencie przekształcenia skomplikowanych struktur w prostszy dwuwymiarowy wykres, który jest po prostu czytelniejszy dla ludzkiego oka. Powodem wystąpienia tego zjawiska jest przesunięcie w przestrzeni 2D, które może się odbywać w obie strony wzdłuż tego samego śladu. W celu minimalizacji wystąpienia tego zjawiska, powstało wiele metod (6,7). Sposobem zapobiegania tego problemu podczas przesunięcia w przestrzeni 2D było wprowadzenie dwuwymiarowej reprezentacji sekwencji DNA (8–11). W tym

podejściu sekwencja jest przedstawiana w postaci zbioru punktów. Jeżeli przesunięcie jest wykonywane kilka razy wzdłuż tej samej ścieżki, wówczas zwiększa się masę jego poszczególnych punktów. Dzięki temu uzyskano metodę całkowicie wolną od degeneracji. Inspiracją do numerycznego opisu wykresu był jeden z zestawów deskryptorów dla dwuwymiarowych wykresów reprezentacji sekwencji białek – dwuwymiarowych momentów bezwładności (8).

Dwuwymiarowa reprezentacja sekwencji łańcucha DNA również została przedstawiona w postaci trójwymiarowej. W konsekwencji powstały dwie metody reprezentacji sekwencji DNA w 3D: pierwsza zaproponowana przez Aram i Iranmanesh (12), a druga przez P. Wąż oraz D. Bielińska-Wąż (13). Trójwymiarowe wykresy reprezentujące sekwencje DNA mogą być określane przy użyciu wartości trójwymiarowych momentów bezwładności (13,14).

Niestandardowa koncepcja tworzenia wykresów na podstawie momentów bezwładności została zapoczątkowana w celu wyznaczenia dwuwymiarowych wykresów oraz została ostatnio zastosowana przez Y. Yao i innych (15,16). Autorzy charakteryzują wykresy reprezentujące sekwencje białka, przy wykorzystaniu dwuwymiarowych momentów bezwładności (15) oraz momentów bezwładności 3D (16). Trójwymiarowe momenty bezwładności zostały również zastosowane jako charakterystyki dla wykresów sekwencji białek przez Hou i innych (17).

W postaci molekularnych deskryptorów zostały również wprowadzone jednowymiarowe momenty bezwładności. W tym przypadku zostało wykorzystane dynamiczne podejście do numerycznej charakterystyki molekularnych widm w podczerwieni (18,19).

W niniejszej rozprawie doktorskiej, zostały przedstawione 20-wymiarowe momenty bezwładności jako nowe deskryptory sekwencji białkowych. Obecnie metody niestandardowe należą do szybko rozwijającej się dziedziny bioinformatyki (4,5,15,16,20–39) (przegląd metod: (26)). Zaproponowana metoda dostarcza nowe i niestandardowe narzędzie do porównywania oraz filogenetycznej analizy sekwencji białek.

Potencjalne zastosowanie nowej metody może być bardzo szerokie. Przykładowo metodę opisywaną w niniejszej rozprawie doktorskiej można zastosować do poszukiwania predykcyjnych modeli, na podstawie których można wywnioskować aktywność biologiczną nowej sekwencji, jak w badaniach białek QSAR (19,40–42). Wiele interesujących zastosowań różnych rodzajów deskryptorów sekwencji biologicznych do badań QSAR, do reprezentacji innych źródeł informacji, tj. widm masowych w surowicy krwi w proteomice klinicznej, molekularne dynamiczne trajektorie, itd., można znaleźć w (43–51). W związku z tym nowe deskryptory, które zostały zaproponowane w niniejszej rozprawie doktorskiej wydają się być niezwykle ważne.

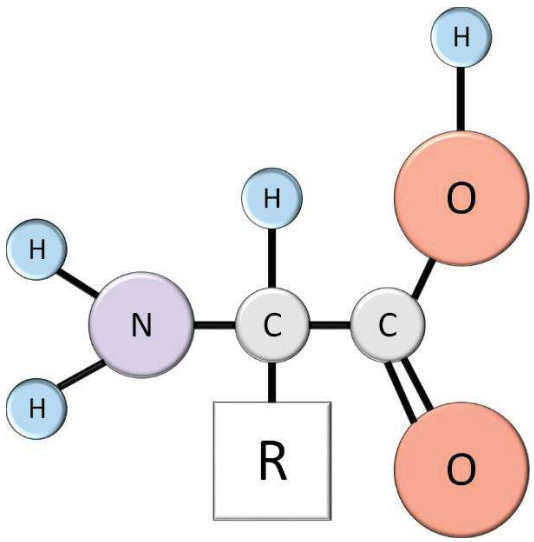
1. CZĘŚĆ TEORETYCZNA

1.1. Co to są białka?

Białka zaliczają się do grupy biopolimerów składających się z liniowego łańcucha aminokwasowego (52). Stanowią główną część budulcową enzymów, pełniących między innymi rolę katalizatorów w układach biologicznych (53). Dodatkowo pełnią również inne role, takie jak: mechaniczne, strukturalne czy sygnalizujące.

Na budowę białka mają wpływ właściwości chemiczne aminokwasów oraz ich sekwencja w łańcuchu białkowym, który z tego powodu przybiera strukturę trójwymiarową. To właśnie jego struktura decyduje o właściwości cząsteczki białkowej. Sekwencja, zgodnie z którą uporządkowane są aminokwasy, zapisana jest w materiale genetycznym komórki w postaci genu (54).

Wszystkie aminokwasy składają się z jednego atomu wodoru, grupy karboksylowej i aminowej oraz reszty aminokwasowej dołączonej do centralnego atomu węgla (*rysunek nr 1*). To właśnie reszty aminokwasowe determinują charakter aminokwasu: jego reaktywność, rozpuszczalność czy polarność, przy zachowaniu nienaruszonej struktury (55).



Rysunek 1. Struktura aminokwasu (NH3 – grupa aminowa*, - COOH – grupa karboksylowa, R – reszta aminokwasowa)*

Struktura białka, ze względu na jej przestrzenność, może być opisana na czterech poziomach (56):

- pierwszorzędowa struktura białka: opisuje kolejność sekwencji aminokwasów,
- drugorzędowa struktura białka: opisuje ułożenie łańcuchów aminokwasowych w przestrzeni, m.in. alfa – helisa, beta – harmonijka,
- trzeciorzędowa struktura białka: opisuje sposób ułożenia elementów wchodzących w skład struktury drugorzędowej,
- czwartorzędowa struktura białka: kompleksy podjednostek – opisuje wzajemne położenie łańcuchów polipeptydowych oraz struktur z grupy prostetycznej.

W niniejszej pracy szczególna uwaga poświęcona będzie pierwszorzędowej strukturze białka, skupiając się na analizie sekwencji aminokwasów.

1.2. Drzewa filogenetyczne

W tej pracy doktorskiej również wykorzystywane będą drzewa filogenetyczne, które głównie służyć będą do ukazania pokrewieństwa pomiędzy populacjami. Jednak dodatkowo ten sposób przedstawiania zależności może być także wykorzystany do wyrażenia w jaki sposób narastała zmienność w danej populacji lub dla jednostki.

W skład drzewa filogenetycznego wchodzą:

- liście – opisują daną jednostkę taksonomiczną (np. populacje),
- węzły – opisują rozgałęzienie się jednostek taksonomicznych,
- gałęzie – opisują związki pomiędzy porównywanymi jednostkami taksonomicznymi, ich długości oznaczają liczbę zmian, które miały miejsce w danej linii ewolucyjnej,
- korzeń – wspólny przodek dla wszystkich jednostek taksonomicznych.

Drzewa filogenetyczne można podzielić na ukorzenione oraz nieukorzenione.

Drzewa ukorzenione cechuje wspólny przodek – czyli korzeń, dając możliwość wyrażenia kierunku przebiegu ewolucji. Natomiast drzewa nieukorzenione nie będą wskazywały pochodzenia danych jednostek taksonomicznych, a jedynie zależności pomiędzy tymi jednostkami (57).

Drzewo filogenetyczne posłużyło do porównania naszej metody z innymi metodami. Do jego wyznaczenia zastosowano metodę odległościową UPGMA, która jest zaimplementowana w pakiecie Mega 6.0 (58). Dla metody UPGMA danymi wejściowymi jest nasza macierz podobieństwa, która stanowi dane do ustalenia topologii drzewa i długości drzewa.

1.3. Porównywanie sekwencji białkowych

Chcąc porównywać ze sobą sekwencje białkowe musimy posiadać ilościowy model ewolucji, który posłuży nam do wyznaczania odległości ewolucyjnych pomiędzy gatunkami, a także tworzenia drzew filogenetycznych. Pierwszym modelem powszechnie używanym w bioinformatyce do opisywania ewolucji sekwencji DNA był model Jukensa-Cantora (1969) (59). Umożliwił obliczenie odległości ewolucyjnej d (odległość Jukesa-Cantora) opierającej się na stosunku liczby podstawień do długości sekwencji, oznaczonym literą D we wzorze nr 1. Model ten jest oparty na wskaźniku tempa podstawień i przedstawiany za pomocą macierzy o wymiarach 4×4, co związane jest z ilością nukleotydów (A, G, C, T).

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right). \quad (1)$$

W celu badania ewolucji sekwencji białek stworzono nową macierz o znacznie większych wymiarach, a mianowicie 20×20, ze względu na to, że każde białko jest przedstawiane przy pomocy dwudziestu biogennych aminokwasów. W rezultacie, powstała pierwsza macierz tempa podstawień dla sekwencji białek, tzw. macierz PAM (*z ang. Point Accepted Mutations*), określająca utrwalone mutacje punktowe. Macierz ta ma wymiary 20×20 i uwzględnia każdy możliwy typ podstawienia aminokwasowego i nadaje się najlepiej do podobnych białek (60).
Alternatywą do macierzy PAM, która nie wymaga konstrukcji modelu ewolucyjnego, jest macierz BLOSUM (*z ang. BLOcks SUbstitution Matrix*).

1.4. Graficzne metody porównywania białek

GRANCH (*z ang. Graphical representation and numerical characterizaion*) to graficzna i numeryczna reprezentacja nukleotydów i białek (61).

Graficzna reprezentacja białek jest równie pożądana jak graficzna reprezentacja DNA. Jednak badania nad reprezentacją białek ukazały się znacznie później, głównie ze względu na ich 20-wymiarowość. Wielowymiarowość jest wyzwaniem zarówno rozpatrując wizualizacje białek, jak również wartości liczbowe opisujące te białka, tzw. deskryptory. Wartości te mogą posłużyć do dalszych obliczeń np. wyznaczania drzew filogenetycznych. Reprezentacje białek, można przedstawić na trzy różne sposoby: przy pomocy modyfikacji istniejących już metod graficznej reprezentacji DNA, wyboru jednej spośród $20!$ alternatywnych uporządkowań aminokwasów, albo również niegraficznej reprezentacji białek, bazującej na ominięciu selekcji spośród $20!$ możliwości dopasowań dwudziestu aminokwasów.

Poniżej zostanie przedstawionych pięć wybranych typów graficznej reprezentacji białek.

1.4.1. Magic Square

Magic Square jest metodą, która została stworzona przez Jeffrey'a (1990 rok), w której zastosowano grę w chaos (CGS – Chaos Game Reprezentation) do wyrysowania sekwencji DNA (62,63). Jest to metoda 2D, posiadająca dwie główne zalety. Pierwszą z nich jest możliwość przedstawienia graficznego dowolnej sekwencji DNA/białka na pojedynczej kartce papieru. Drugą zaletą jest brak utraty informacji o strukturze białka.

Graficzna reprezentacja tej metody polega na poruszaniu się po kwadracie, utworzonym przy pomocy czterech nukleotydów o następujących współrzędnych:

- (1,-1) – tymina (T),
- (-1,-1) – adenina (A),
- (-1,1) – cytozyna (C),
- (1,-1) – guanina (G).

Kreślenie graficznego modelu rozpoczyna się w środku układu współrzędnych, w punkcie (0,0). Odpowiednio poruszamy się w kierunku nukleotydu (wierzchołku kwadratu), zgodnie z kolejnością sekwencji. Zawsze pokonujemy odcinek równy połowie drogi, która dzieli nas od danego nukleotydu i w tym miejscu stawiamy kropkę, która będzie początkiem dla kolejnego ruchu. Tak długo poruszamy się w kwadracie, aż wyrysujemy całą sekwencję. W ten sposób powstaje nam obraz składający się z takiej liczby znaczników, ile wnosi długość badanej sekwencji DNA. Natomiast możliwość utworzenia poszczególnych aminokwasów z trójek nukleotydów pozwoliła na

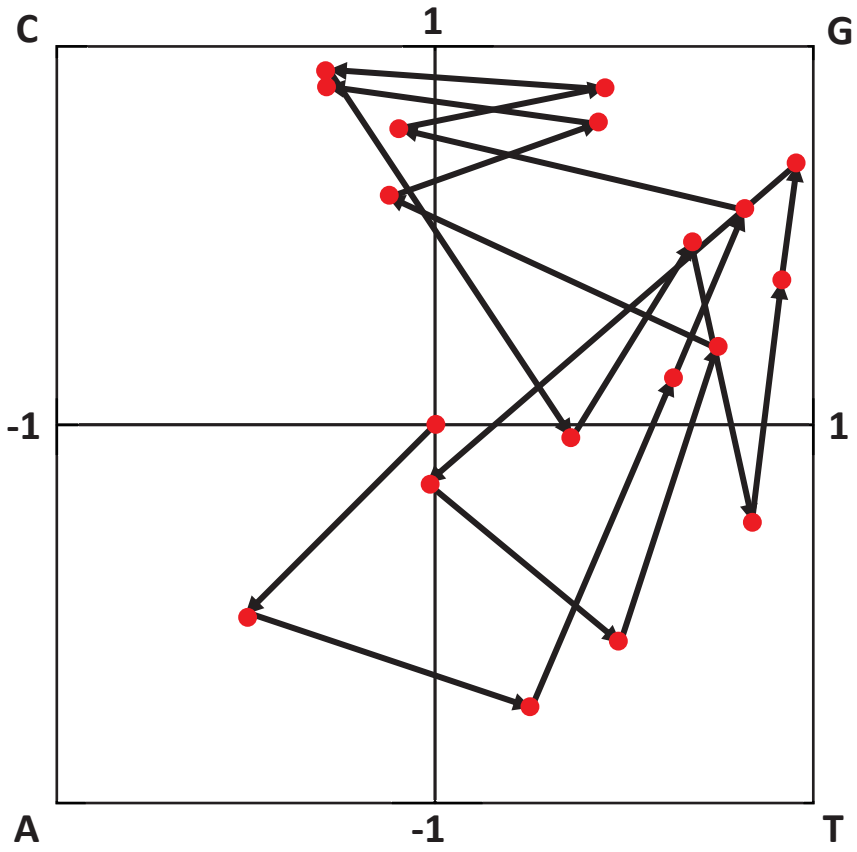
zobrazowanie tą metodą białek. W przypadku aminokwasów sekwencja składająca się z np. 170 aminokwasów będzie miała 510 kropek (61).

Dla zobrazowania metody wyrysowania modelu, można posłużyć się poniższym przykładem Magicznego Kwadratu dla sześciu pierwszych aminokwasów kodujących hormon peptydowy – insulinę: **MALWMR** (ATG GCG CTG TGG ATG CGC), który został przedstawiony na *rysunku nr 2.*

Zgodnie z zasadami metody, wyrysowywanie rozpoczęto od środka układu współrzędnych. Każdy kolejny krok został przedstawiony w postaci strzałek, których groty wskazują na punkty będące właśnie postacią modelu.

Każda strzałka jest zobrazowaniem kolejnej zasady azotowej nukleotydów, kodującej dany aminokwas wchodzący w skład podanego fragmentu łańcucha peptydowego insuliny człowieka.

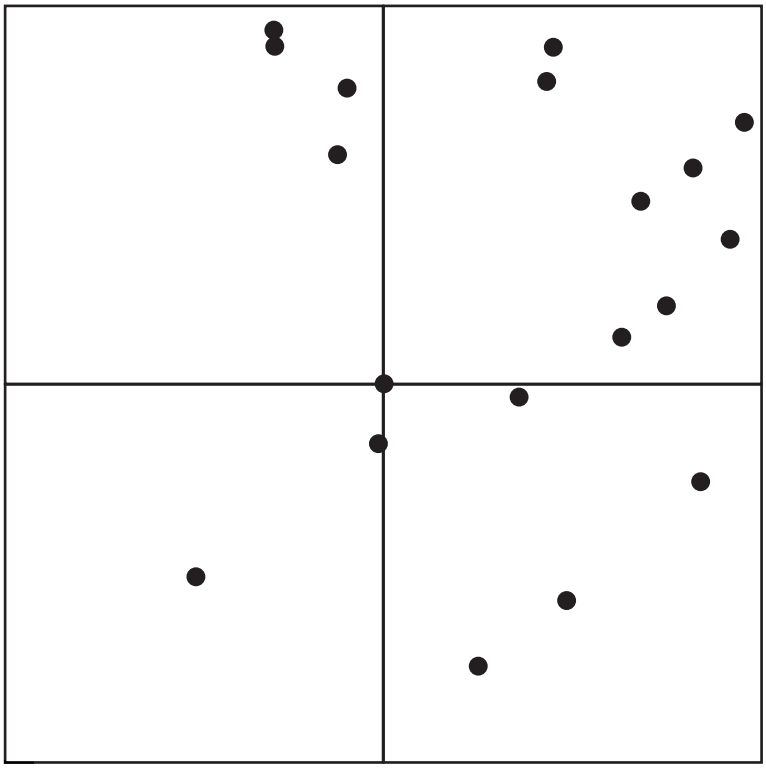
Rezultat został przedstawiony poniżej.



Rysunek 2. Model fragmentu łańcucha insuliny człowieka wyrysowany metodą Magic Square – ukazanie sposobu wykreślania

Na powyższym rysunku został przedstawiony, krok po kroku, sposób, w jaki został stworzony model łańcucha peptydowego metodą Magic Square. Jednak w rzeczywistości na model nie przenosi się strzałek, które jedynie zaciemniają obraz.

Poniżej został przedstawiony finalny model analizowanego fragmentu łańcucha, zadaną metodą Magic Square.



Rysunek 3. Finalny model fragmentu łańcucha insuliny człowieka wyrysowany metodą Magic Square

Po otrzymaniu ostatecznej graficznej reprezentacji sekwencji białkowej można zauważyć, że sposób reprezentacji jest bardzo skondensowany.

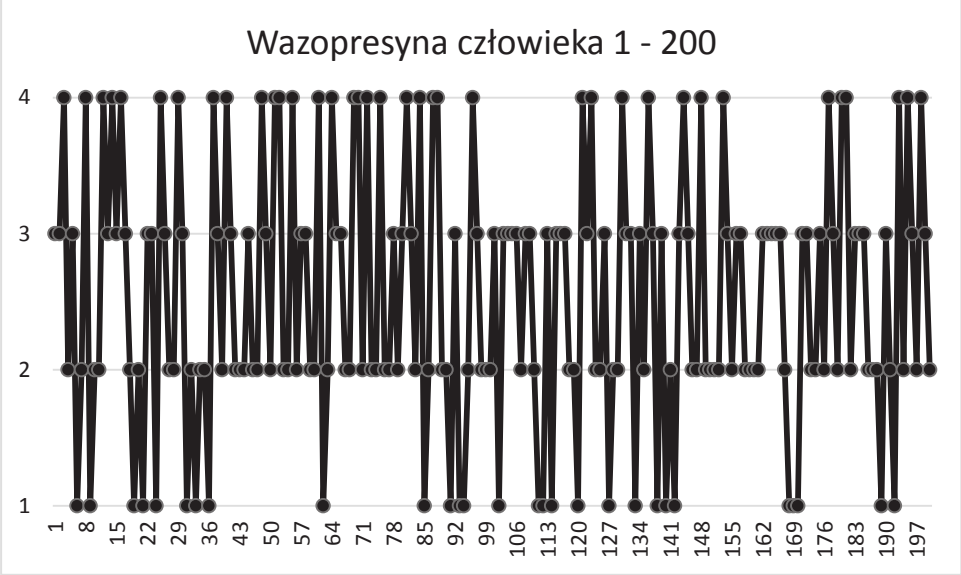
Dobierając odpowiednią skalę, w taki sposób można przedstawić model bardzo długich sekwencji białkowych, bez utraty informacji o strukturze białka.

1.4.2. Virtual Genetic Code

Virtual Genetic Code jest metodą zarówno 2D jak również 3D, która ze względu na wirtualny dobór kodonu dla danego aminokwasu jest obarczona pewnymi błędami, które wynikają z tego dopasowania (61).

Znając trójki nukleotydów kodujące aminokwasy biogenne wchodzące w skład białka możemy za pomocą reprezentacji graficznych stworzonych dla łańcuchów DNA zwizualizować to białko. Natomiast w przypadku, gdy nie znamy trójek nukleotydów budujących danych aminokwas musimy posłużyć się metodą Virtual Genetic Code, która umożliwia przypisanie każdemu aminokwasowi unikalnego kodonu. Oprócz metioniny (ATG) i tryptofanu (TGG) wszystkie pozostałe aminokwasy mogą posiadać więcej niż jedną możliwość zapisu kodonu. Jednak ta metoda polega na przypisaniu każdemu aminokwasowi tylko jednego kodonu, stąd wynikają pewne błędy przybliżeń. Wirtualny kod został tak zaprojektowany, że każda zasada azotowa (A, C, G, T) występuje 15 razy. Reprezentację graficzną danego białka możemy przedstawić za pomocą czterech poziomych osi oznaczonych odpowiednio A-1, C-2, G-3 i T-4.

Na poniższym rysunku został przedstawiony model początku łańcucha białkowego sekwencji wazopresyny człowieka, przy pomocy czterech poziomych osi omawianej metody.



Rysunek 4. Reprezentacja łańcucha peptydowego wazopresyny ludzkiej według metody Virtual Genetic Code z wykorzystaniem czterech poziomych osi

Otrzymany w ten sposób wykres 2D, przedstawiony na *rysunku nr 4.* wydaje się być mało czytelny, a samo porównywanie sekwencji za pomocą takich wykresów wydaje się

być dość czasochłonne, gdyż na pierwszy rzut oka ciężko jest zobaczyć jakiekolwiek zależności pomiędzy sekwencjami. Pewną alternatywą dla tej metody jest przedstawienie dwóch sekwencji łącząc je arytmetycznie, co powoduje, że wykres staje się bardziej przystępny i możemy od razu z wykresu powiedzieć coś na temat porównywanych sekwencji, np. iloma aminokwasami różnią się porównywane białka. Tę formę wizualizacji możemy również przedstawić przy użyciu zapisu zero-jedynkowego (binarnego), gdzie występują tylko dwie osie 0 i 1, a wszystkim nukleotydom, które nie leżą w osi 0 przypisuje się wartość 1. Sekwencję białka możemy również przedstawić w postaci 4 osi poziomych, które będą oznaczone następująco: 5, 10, 15 i 20. Dzięki takiemu podejściu skrócimy trzykrotnie wykres danego białka, gdyż nie będziemy na nim przedstawiać poszczególnych nukleotydów, tylko same aminokwasy. Dodatkowo można w łatwy sposób zweryfikować, np. który aminokwas występuje najczęściej, a który najrzadziej. Również dla tego rozwiązania można przedstawić postać binarną, gdzie zaproponowano, że A = 00, G = 01, C= 10, U (T) = 11.

Istnieją jeszcze inne możliwości wizualizacji tej metody oprócz zapisu binarnego. Jedną z nich można nazwać spiralą (*z ang. „worm" spiral*), której niewątpliwie zaletą jest możliwość przedstawienia sekwencji białka w sposób ścisły. Niestety jest z tym związana utrata czytelności rysunku.

Podczas opisu metody wirtualnego kodu zostały przedstawione alternatywne możliwości wizualizacyjne aminokwasów, a tym samym samych białek.

1.4.3. 8×8 Tables of Codons

8×8 Tables of Codons jest metodą zaproponowaną przez Mian'a Randić'a, która opiera się na modyfikacji metody Jeffrey'a - Magicznego Kwadratu (opisanej w rozdziale 1.4.1) (61,64).

Pierwszym etapem metody jest przygotowanie tabeli z rozpisanymi kodonami, wykorzystując do tego metodę Jeffrey'a. Wymiary tabeli powinny wynosić 8×8 komórek, dzięki czemu uzyskujemy miejsce na wpisanie 64 możliwych kombinacji tripletów zasad azotowych nukleotydów (C – cytozyna, G – guanina, A – adenina, T – tymina).

Powodem stworzenia tabeli w powyższy sposób jest jedna z cech kodu genetycznego – degeneracja. Zgodnie z tą cechą, każdy aminokwas jest kodowany przez więcej niż jeden triplet, np. arginina jest kodowana na sześć różnych sposób, a cysteina na dwa.

Zatem stworzona tabela powinna zawierać 64 możliwe kodony: 61 tripletów kodujących dwadzieścia biogennych aminokwasów oraz 3 kodony nonsensowne – bez informacji genetycznej o żadnym aminokwasie (niosą informację o zakończeniu translacji).

Tabela jest stworzona na wzór metody Magicznego Kwadratu Jeffrey'a i również z jej wykorzystaniem jest wypełniona odpowiednią kolejnością kodonów. Zgodnie z regułami tej metody, na wspólny Magiczny Kwadrat nanosi się graficzną reprezentację wszystkich 64 kodonów. Po dokonaniu niniejszej czynności, należy przeprowadzić podział Magicznego Kwadratu na cztery równe części oraz przepisać kodony do tabeli w takiej kolejności, w jakiej są graficznie reprezentowane na Magicznym Kwadracie.

Dzięki temu uzyskujemy tabelę, przedstawioną poniżej:

Tabela 1. Kolejność kodonów, zapisana według metody Jeffrey'a

CCC	GCC	CGC	GGC	CCG	GCG	CGG	GGG
ACC	TCC	AGC	TGC	ACG	TCG	AGG	TGG
CAC	GAC	CTC	GTC	CAG	GAG	CTG	GTG
AAC	TAC	ATC	TTC	AAG	TAG	ATG	TTG
CCA	GCA	CGA	GGA	CCT	GCT	CGT	GGT
ACA	TCA	AGA	TGA	ACT	TCT	AGT	TGT
CAA	GAA	CTA	GTA	CAT	GAT	CTT	GTT
AAA	TAA	ATA	TTA	AAT	TAT	ATT	TTT

Analizując powyższą tabelę, Mian Randić zauważył, że można wyspecyfikować cztery grupy komórek, reprezentujące kodony kończące się tą samą azotową zasadą nukleotydu. Randić chciał opracować metodę, która pozwoli mu na stworzenie grup kodonów, odzwierciedlających ogólną tabelę tripletów obrazującą związek pomiędzy kodonami a odpowiadającymi im aminokwasami. Zgodnie z tabelą kodonów, dwie pierwsze zasady azotowe każdego tripletu niosą najwięcej informacji na temat rodzaju aminokwasu, który jest przypisany przez dany kodon. Ostatnia zasada azotowa nukleotydu, która znajduje się w kodonie, jest najmniej ważna, gdyż istnieje kilka rodzajów aminokwasów, które można zakodować znając jedynie dwie pierwsze zasady azotowe danego kodonu. Zgodnie z *tabelą nr 1* utworzone zostały grupy na podstawie właśnie najmniej ważnej, ostatniej zasady azotowej nukleotydu, wchodzącego w skład tripletu. Stąd, Randić stworzył modyfikację metody Jeffrey'a, poprzez zmianę zasad wyznaczania graficznej reprezentacji protein przy pomocy Magicznego Kwadratu.

Podstawą zmodyfikowanej metody była zmiana rozmiaru Magicznego Kwadratu na mniejszy, po każdym wykonanym kroku. Początkowy Magiczny Kwadrat jest zapisywany w taki sam sposób jak u Jeffrey'a. Po pierwszym kroku, czyli zaznaczeniu pierwszą strzałką punktu reprezentującego daną zasadę azotowa nukleotydu, wirtualnie tworzony jest nowy Magiczny Kwadrat, którego środek jest wytyczony właśnie przez pierwszy wyrysowany punkt. Następnie zgodnie z zasadą Jeffrey'a wyrysowywana jest kolejna strzałka, wskazująca grotem na kolejną zasadę azotową. Drugi punkt wskazuje środek następnego Magicznego Kwadratu. W taki sposób zgodnie z następnymi sekwencjami, nowe punkty nanoszone są według coraz to mniejszych Magicznych Kwadratów.

Wyznaczając w ten sposób wszystkie 64 kodony oraz umieszczając je na jednym rysunku, Randić osiągnął podobny efekt, jak podczas tworzenia *tabeli nr 1.* Jednak osiągnął coś znacznie więcej, możliwość stworzenia tabeli kodonów, których pierwsza zasada azotowa dzieli triplety na cztery grupy. W podobny sposób, jak miało to miejsce poprzednio, za sprawą ostatniej zasady azotowej nukleotydu. Dodatkowo, po przeanalizowaniu zostało zauważony fakt, że kodony zapisane w nowej tabeli odpowiadają kodonom z *tabeli nr 1* – są zapisane w odwrotny sposób. Przykładowo dla kodonu z *tabeli nr 1*, odpowiadającego drugiej części Magicznego Kwadratu, ACC, w nowej tabeli odpowiada w tym samym miejscu kodon CCA. Oczywiste jest to, że dla tripletów będących palindromami, zapis pomiędzy tabelami się nie zmienił.

Wynik tejże metody, został przedstawiony poniżej, w postaci *tabeli nr 2.*

Tabela 2. Wynik zmodyfikowanej metody Jeffrey'a

CCC	CCG	CGC	CGG	GCC	GCG	GGC	GGG
CCA	CCT	CGA	CGT	GCA	GCT	GGA	GGT
CAC	CAG	CTC	CTG	GAC	GAG	GTC	GTG
CAA	CAT	CTA	CTT	GAA	GAT	GTA	GTT
ACC	ACG	AGC	AGG	TCC	TCG	TGC	TGG
ACA	ACT	AGA	AGT	TCA	TCT	TGA	TGT
AAC	AAG	ATC	ATG	TAC	TAG	TTC	TTG
AAA	AAT	ATA	ATT	TAA	TAT	TTA	TTT

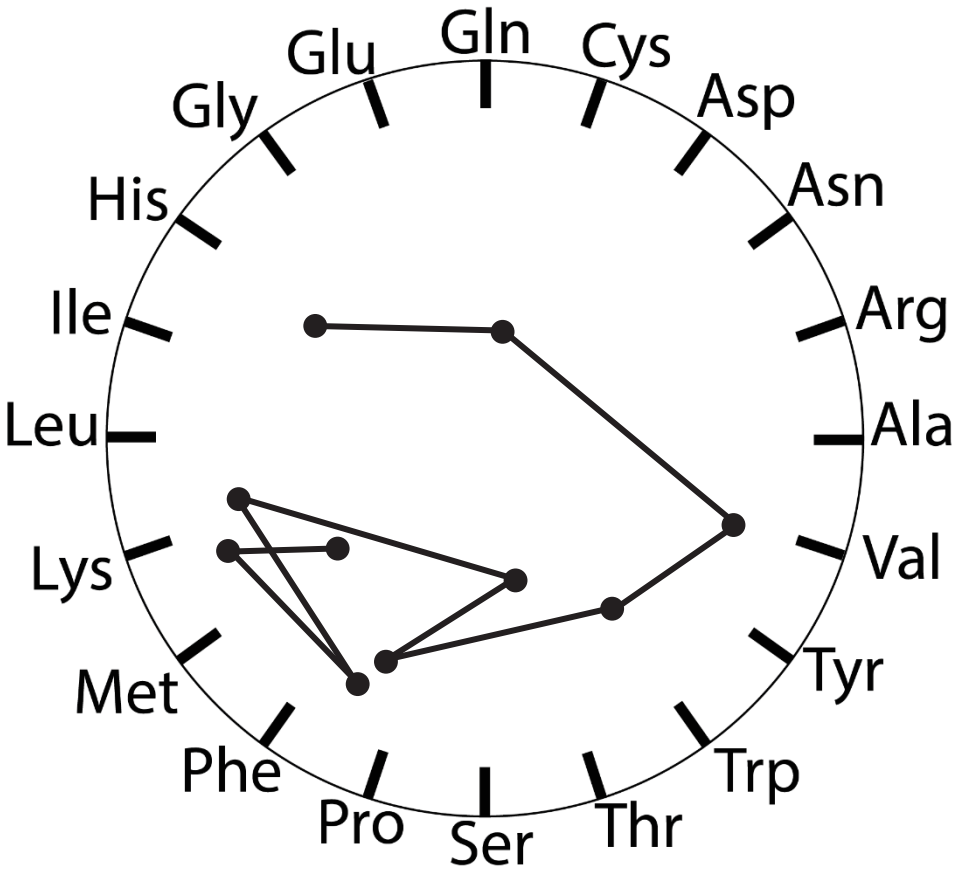
Stosując nową metodę będziemy się poruszać tylko w obrębie jednej ćwiartki, którą wyznaczy nam pierwsza zasada azotowa rozpatrywanego aminokwasu. Wierzchołki odpowiednio oznaczone są A, T, G i C. W związku z tym, że aminokwasy: arginina, lizyna i seryna są kodowane przez sześć różnych kodonów, nie jest możliwe zapisanie ich w obrębie jednej ćwiartki, gdyż rozpoczynają się od innych liter. Kolejnym warunkiem, który spełniono jest fakt, że ten sam aminokwas kodowany przez różne kodony ma mieć inny środek masy. Jak widać w *tabeli nr 1* mamy kilka aminokwasów, które są umieszczone w diagonalny sposób, co wskazuje że mają one te same środki masy. Zatem Randić zaproponował nową tabelę aminokwasów zamieniając miejscami dwie współrzędne Magicznego Kwadratu, a mianowicie G i T, dzięki temu uzyskano różne środki ciężkości. Dodatkowo otrzymano również bliższe sąsiedztwo pomiędzy różnymi trójkami nukleotydów kodujących ten sam aminokwas.

1.4.4. Magic Circle

Magic Circle – jest to nowa metoda zaproponowana przez Randić, Butina i Zupana (21,61). Polega ona na graficznym przedstawieniu aminokwasów za pomocą jednostkowego koła. Na obwodzie koła umieszczane są aminokwasy, odpowiednio co 18°. Kolejność aminokwasów jest alfabetyczna, dlatego zaczynamy od alaniny, umieszczając ją na długości odpowiadającej godzinie 3:00 i kończąc na walinie, poruszając się ruchem przeciwnym do wskazówek ruchu zegara. Niewątpliwą zaletą tej metody tak samo jak w przypadku magicznego kwadratu jest brak utraty danych.

Wyrysowanie interesującego nas białka możemy rozpocząć, gdy znamy jego sekwencję aminokwasową (gdy jest ona podana albo zapisujemy ją za pomocą wirtualnego kodu). Poruszanie się po magicznym kole rozpoczynamy w środku koła i przesuwamy się w kierunku pierwszego aminokwasu, który określa sekwencja aminokwasowa białka. Podobnie jak w metodzie Magicznego Kwadratu Jeffrey'a także w Magicznym Kole pokonywana jest tylko połowa drogi pomiędzy aminokwasami. Po zaznaczeniu pierwszej połowy odległości pomiędzy środkiem a pierwszym aminokwasem przechodzi się do następnego aminokwasu, w którego stronę należy kontynuować dalsze wyrysowywanie trasy. W ten sposób wyznaczanie trasy trwa aż do oznaczenia ostatniego aminokwasu znajdującego się w strukturze białka. W przypadku, gdy liczba aminokwasów jest zbyt duża, wówczas można ograniczyć się tylko do oznaczenia punktowego aminokwasów w obszarze koła, bez wyrysowania krzywej zwanej z języka angielskiego: *zigzag curve*.

Poniżej została przedstawiona graficzna reprezentacja dziesięciu pierwszych aminokwasów glukagonu ludzkiego przedstawiona za pomocą metody Magicznego Koła (Met Lys Ser Ile Tyr Phe Val Ala Gly Leu).



Rysunek 5. Reprezentacja graficzna sekwencji aminokwasów glukagonu człowieka, metodą Magicznego Koła

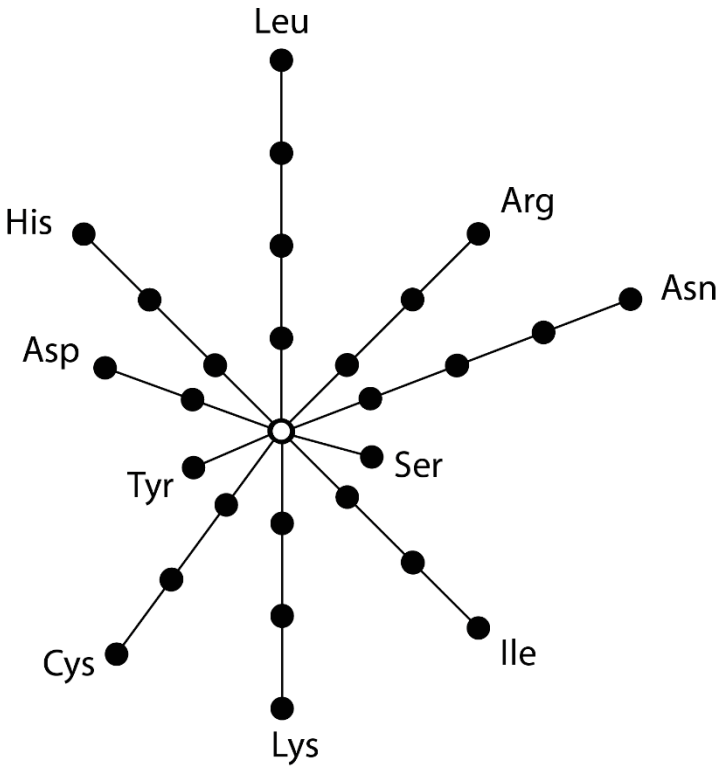
Zdecydowanie metoda Magicznego Koła, jak i Kwadratu ma lepsze zastosowania w celach wizualizacyjnych krótszych sekwencji aminokwasowych, ze względu na brak możliwości wizualnej interpretacji w przypadku długich sekwencji. Porównując krzywe – *zigzag curve* dla różnych gatunków możemy doszukać się pewnego podobieństwa lub jego braku. Lepsze możliwości porównywania sekwencji daje nam wykres przedstawiający różnice pomiędzy współrzędnymi odpowiadających sobie aminokwasów dwóch białek, gdzie oś odciętych przedstawia pary aminokwasów w porównywanych białkach, a oś rzędnych długości pomiędzy ich współrzędnymi (odległość euklidesowa). Główną zaletą takiej reprezentacji białek jest możliwość graficznego przedstawienia podobieństw pomiędzy parą białek, co daje możliwość łatwego zidentyfikowania skali podobieństwa przy jednoczesnym wskazaniu jego miejsca w sekwencji.

1.4.5. Starlike Graphs

Starlike Graphs jest to metoda zaproponowana przez Randić, Zupan i Vikić-Topić, której głównymi zaletami jest brak arbitralnego przyporządkowania aminokwasów do rozpatrywanego obiektu graficznego (61,65). Korzystając z izomorfizmu grafów dla pojedynczego białka możemy zaproponować wiele różnych form tego samego grafu.

Graf przypominający gwiazdę składa się z pojedynczego centralnego punktu (wierzchołek posiadający maksymalny stopień), od którego uchodzą gałęzie o stopniach wierzchołków 2 lub 1 i tylko do tych wierzchołków przypisywane są aminokwasy.

Dobrym przykładem obrazującym sedno całej metody, jest *rysunek nr 6,* gdzie została przedstawiona reprezentacja graficzna dla sekwencji 27 ostatnich aminokwasów łańcucha prolaktyny człowieka.

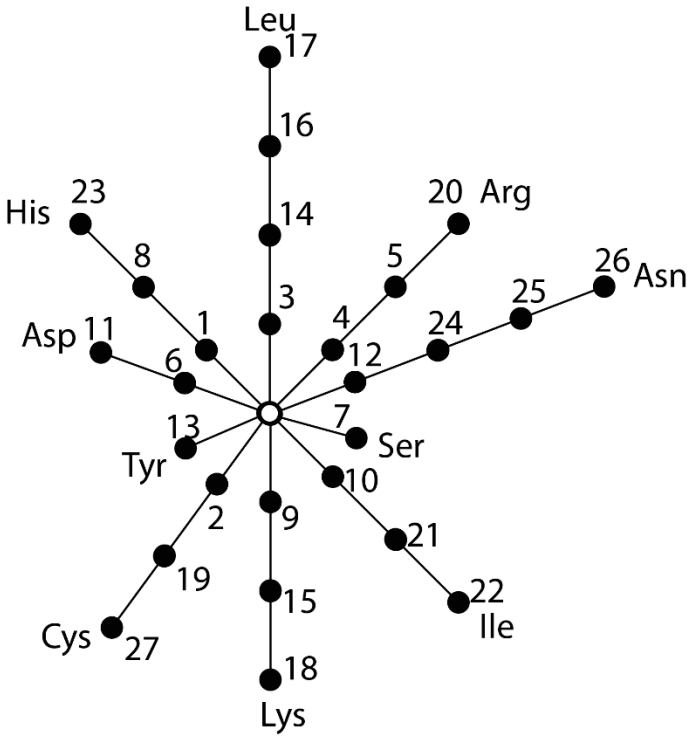


Rysunek 6. Graficzna reprezentacja 27 ostatnich aminokwasów łańcucha prolaktyny człowieka, metodą Starlike Graph

Po przeanalizowaniu powyższego grafu, można zauważyć, że analizowana sekwencja składa się z 10-ciu typów aminokwasów, co zostało przedstawione w postaci 10-ciu różnych wierzchołków grafu. Każda odchodząca gałąź grafu podzielona jest na fragmenty, które oznaczają liczbę aminokwasów, wchodzących w skład sekwencji.

W przypadku grafu przedstawionym na *rysunku nr 6,* można zaobserwować gałęzie o długościach: 1, 2, 3 oraz 4. Pomimo tego, że powyższy graf niesie sobą informację o składzie jakościowym oraz ilościowym analizowanej sekwencji, to nie zdradza struktury logicznej, w jakiej kolejności w sekwencji ułożone są wymienione aminokwasy.

Dlatego została przedstawiona rozszerzona metoda grafu, gdzie opisana została kolejność połączenia aminokwasów, co zostało przedstawione na poniższym rysunku.



Rysunek 7. Graficzna reprezentacja 27 ostatnich aminokwasów łańcucha prolaktyny człowieka, metodą Starlike Graph z numeracją

Opisywane są nie tylko każde gałęzie, ale również oznaczane są odpowiednio wierzchołki. Poszczególne grafy proponowane dla tego samego białka różnią się sąsiadującymi aminokwasami, co w konsekwencji daje różne macierze odległości. Charakteryzujące się tymi samymi wartościami własnymi nazywanymi niezmiennikami macierzy, które nie zależą od numeracji wierzchołków ani graficznej reprezentacji. Ilościowy opis białka w postaci macierzy odległości bazuje na odległościach pomiędzy poszczególnymi wierzchołkami przedstawionymi za pomocą grafu. W ten sposób, dla omawianego przykładu sekwencji prolaktyny powstaje macierz o wymiarach 10×10, na podstawie której można wyznaczyć wartości własne.

Taki numeryczny zapis informacji o danym białku nie wiąże się w żaden sposób z utratą informacji. Maksymalny rozmiar macierzy odległości może wynosić 20×20, co jest podyktowane maksymalną liczbą aminokwasów biogennych, które mogą wystąpić w białku. Istnieje jednak możliwość analizowania sekwencji, w której łączna suma typów aminokwasów, które ją tworzą, jest mniejsza niż 20. Taki przykład został zaprezentowany w omawianej powyżej sekwencji prolaktyny. W takiej sytuacji stosuje się stworzenie „sztucznych" aminokwasów, wypełniających lukę dzielącą nas do uzyskania sumy wszystkich ich rodzajów. Taki zabieg normalizacji usprawnia automatyczne porównywanie sekwencji, które w rzeczywistości różnią się liczbą typów aminokwasów, które budują analizowany łańcuch białkowy.

1.5. Niegraficzne metody porównywania białek – metoda A Walk

A Walk jest to metoda przesunięcia w przestrzeni 20-wymiarowej, gdzie białka traktowane są jako obiekty matematyczne, zapisane w postaci wektorów jednostkowych o 20 składowych (liczbę tę determinuje ilość aminokwasów biogennych) (4,5,61). Można powiedzieć, że jeden aminokwas to pewien krok przesunięcia, który wykonany w całości determinuje białko.

Przesunięcie w 20-wymiarowej przestrzeni polega na przeskanowaniu sekwencji i zapisaniu jej współrzędnych, dokładnie każdej współrzędnej dla poszczególnego aminokwasu. Ilość współrzędnych zależy od długości sekwencji. Przed rozpoczęciem zapisywania wyników reprezentacji łańcucha białkowego przy pomocy metody A Walk, najpierw ustandaryzowuje się strukturę wektorów jednostkowych, reprezentujących dane aminokwasy. Po pierwsze, należy pamiętać, że każdy z wektorów musi być niepowtarzalny, żeby móc odnosić się do danego aminokwasu.

Tabela przedstawiająca dopasowanie wektorów jednostkowych została przedstawiona poniżej.

Tabela 3. Dopasowanie wektorów jednostkowych do reprezentujących przez nie aminokwasów

Nr	Am.	Współrzędne
1	Ala	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
2	Arg	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
3	Asn	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
4	Asp	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
5	Cys	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
6	Glin	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
7	Glu	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
8	Gly	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
9	His	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
10	Ile	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
11	Leu	0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
12	Lys	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
13	Met	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
14	Phe	0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
15	Pro	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16	Ser	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
17	Thr	0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18	Trp	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
19	Tyr	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20	Val	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Jak można zauważyć, każdy z wektorów różni się pozycją jedynki. Kolejność aminokwasów została podana alfabetycznie. Następnie badana sekwencja aminokwasów jest analizowana od pierwszego do ostatniego aminokwasu. Rozpoczynając „kroki po sekwencji” należy stworzyć macierz, która początkowo ma identyczną strukturę jak pierwszy wektor jednostkowy analizowanej sekwencji. Przesuwając się kolejno do następnych aminokwasów, do początkowo stworzonej macierzy, dodawane są kolejne wektory jednostkowe. Dzięki temu po przejściu wszystkich kroków, uzyskujemy oczekiwany efekt metody A Walk. Sposób w jaki tworzona jest analizowana metoda reprezentacji białek, można zobrazować przy pomocy przykładu, sekwencji aminokwasów, np. GIVEQCCASVCSLYQNYCN. Jest to sekwencja łańcucha A insuliny wołu, wybrana specjalnie z powodu tego, że była

pierwszym zsekwencjonowanym białkiem. Rezultat metody, krok po kroku dla powyższego przykładu, został przedstawiony na poniższej tabeli.

Tabela 4. Reprezentacja łańcucha insuliny wołu metodą A Walk, krok po kroku

Nr	Am.	Współrzędne	Częst.	Kw. odl.	Sum.	Pierw. sum.
1	Gly	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0	1	1	1	1.00000
2	Ile	0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0	1	1	2	1.41421
3	Val	1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0	1	1	3	1.73205
4	Glu	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0	1	1	4	2.00000
5	Gln	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0	1	1	5	2.23607
6	Cys	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 0 0	1	1	6	2.44949
7	Cys	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0	2	3	9	3.00000
8	Ala	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 1	1	1	10	3.16228
9	Ser	1 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 1	1	1	11	3.31662
10	Val	2 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 1	2	3	14	3.74166
11	Cys	2 0 0 0 2 0 0 0 0 0 0 1 0 1 1 1 3 0 0 0 1	3	5	19	4.35890
12	Ser	2 0 0 0 3 0 0 0 0 0 0 1 0 1 1 1 3 0 0 0 1	2	3	22	4.69042
13	Leu	2 0 0 0 3 0 0 0 0 1 1 0 1 1 1 3 0 0 0 1	1	1	23	4.79583
14	Tyr	2 1 0 0 3 0 0 0 0 1 1 0 1 1 1 3 0 0 0 1	1	1	24	4.89897
15	Gln	2 1 0 0 3 0 0 0 0 1 1 0 1 1 2 3 0 0 0 1	2	3	27	5.19615
16	Asn	2 1 0 0 3 0 0 0 0 1 1 0 1 1 2 3 0 1 0 1	1	1	28	5.29150
17	Tyr	2 2 0 0 3 0 0 0 0 1 1 0 1 1 2 3 0 1 0 1	2	3	31	5.56776
18	Cys	2 2 0 0 3 0 0 0 0 1 1 0 1 1 2 4 0 1 0 1	4	7	38	6.16441
19	Asn	2 2 0 0 3 0 0 0 0 1 1 0 1 1 2 4 0 2 0 1	2	3	41	6.40312

Oprócz analizowania współrzędnych w 20-wymiarowym wektorze, który po 19-stym kroku obrazuje wynik analizowany metody, można również przedstawić inne parametry. Pierwszym z nich jest częstotliwość występowania danego aminokwasu w całej sekwencji, zapisywana w kolumnie poprzedzającej kolumnę z współrzędnymi. W każdym następnym kroku wartość częstotliwości jest przypisywana dla danego aminokwasu. Następnym parametrem jest kwadrat odległości euklidesowej, którego wartość jest zapisywana również dla poszczególnych aminokwasów, dla danego kroku. Kolejnym parametrem, jest wyrażenie poprzednich parametrów jako suma wartości dla wszystkich kroków, aż do aktualnie analizowanego. Ostatnim parametrem jest pierwiastek sumy, z poprzedniego kroku.

2. MACIERZE SUBSTYTUCJI

2.1. Macierz PAM – Point Accepted Mutations

Model PAM (*z ang. Point Accepted Mutations)* – jest to macierz tempa podstawień aminokwasów, czyli model ewolucji sekwencji białek o wymiarze 20×20. Model PAM jest przydatny do analizy mutacji utrwalonej, czyli takiej mutacji, która pojawiła się w danej populacji oraz uległa utrwaleniu (59).

Pierwsza macierz PAM1 została stworzona przez Margaret Dayhoff (1), dla 71 grup sekwencji aminokwasów, dla których średni współczynnik podobieństwa wynosił 85%. Macierz PAM w swojej nazwie zawiera liczbę, która informuje nas o podobieństwie danych sekwencji. Macierz PAM1 oznacza, że prawdopodobieństwo zamiany danego aminokwasu na inny wynosi 1%, przy czym równoważnie prawdopodobieństwo tego, że dany aminokwas w sekwencji nie zostanie zmieniony wynosi 99%.

Macierz PAM1 tworzy się poprzez rozpisanie dla każdego aminokwasu w sekwencji białkowej częstotliwości jego zmiany w inny aminokwas. Przykładowo dla danej sekwencji aminokwasów, dla dwóch populacji, fenyloalanina występuje 100 razy, z czego w dziewięćdziesięciu miejscach występuje zarówno w jednej jak i drugiej populacji w niezmieniony sposób. Różni się tylko 10 wystąpień, gdzie fenyloalanina (F) została zamieniona 5 razy na tryptofan (W), 4 razy na tyrozynę (Y) i 1 raz na asparaginę (N). Mając takie informacje, możemy obliczyć, że:

$$A_{FW} = \frac{5}{100} = 0.05\,, \tag{2}$$

$$A_{FY} = \frac{4}{100} = 0.04\,, \tag{3}$$

$$A_{FN} = \frac{1}{100} = 0.01\,. \tag{4}$$

Dzięki temu uzyskujemy miarę częstotliwości, z jaką dany aminokwas jest wymieniany na inny, w analizowanych dwóch sekwencjach białkowych. Powyższy przykład przeprowadza się dla wszystkich 20-stu aminokwasów.

Następnie przeprowadza się pomiar częstotliwości występowania aminokwasu na cały zbiór sekwencji białkowych. Dla zobrazowania, jeżeli porównywana sekwencja składa się z 2000 aminokwasów, wówczas korzystając z poprzedniego przykładu dla fenyloalaniny pomiar częstotliwości wynosić będzie:

$$f_F = \frac{100}{2000} = 0.05\,. \tag{5}$$

Następnym krokiem jest oszacowanie szansy na wystąpienie mutacji dla danego aminokwasu. Czyli obliczenie iloczynu wartości częstotliwości występowania aminokwasu na cały zbiór sekwencji białkowych z liczbą wystąpień mutacji dla danego aminokwasu. Dla rozważanego przykładu z fenyloalaniną będzie wynosić to:

$$m_F = f_F \times 10 = 0.5\,. \tag{6}$$

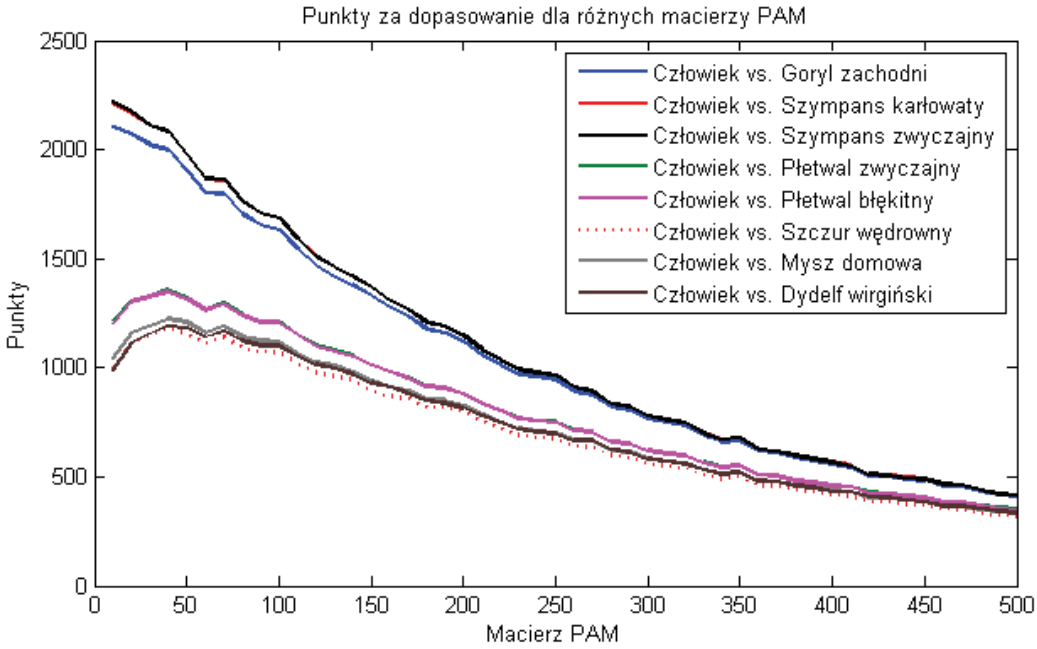
Ostatecznie tworzy się macierz prawdopodobieństw wystąpienia mutacji. Jest to macierz, która zawiera listę wartości, określające prawdopodobieństwo tego, że jeden obiekt zostanie zmutowany w drugi w pewnym określonym i stałym przedziale czasu. Określone w ten sposób wartości tworzą macierz stochastyczną. Dla analizowanego przykładu, wartości komórki macierzy wynosić będą:

$$M_{FW} = m_F \times \frac{A_{FW}}{\sum_1^{20} A_{FW}} = 0.5 \times \frac{0.05}{0.1} = 0.25\,. \tag{7}$$

Dla PAM1 odległość ewolucyjna jest mierzona na podstawie wiedzy, że jedynie 1/100 aminokwasów ulega mutacji. Dlatego w przybliżeniu dla każdej komórki macierzy, dla której wiersz i kolumnę reprezentują te same aminokwasy (przekątna macierzy PAM), wartości dla macierzy PAM1 będą bliskie 0.99.
Mutacje występujące w sekwencji białkowej są od siebie całkowicie niezależne. Dlatego dla białek pochodzących od populacji mniej ze sobą spokrewnionych należy, dla n/100 mutacji, które w nich zaszły pomnożyć macierz PAM1 przez samą siebie n razy. Toteż chcąc przykładowo osiągnąć macierz PAM250, należy pomnożyć macierz PAM1 przez samą siebie 250 razy.

Żeby przetestować działanie metody, posłużono się programem obliczeniowym MATLAB. Stworzono skrypt, w którym wczytano sekwencje aminokwasowe dla łańcuchów białkowych proteiny ND5, dla analizowanych w niniejszej rozprawie doktorskiej gatunków zwierząt, w porównaniu z sekwencją człowieka. Program MATLAB

posiada wbudowaną funkcję, dla której można zliczyć punkty najlepszego dopasowania rodzaju macierzy PAM do dystansu genetycznego analizowanych sekwencji aminokwasowych. Program ten umożliwia użycia macierzy od PAM10 do PAM500 z interwałem co 10. Dlatego na poniższym wykresie widać wartości zliczonych punktów w porównaniu do różnych zastosowanych macierzy PAM, zgodnie z kolejnością: PAM10, PAM20, PAM30, … PAM500.



Rysunek 8. Wykres punktacji różnych macierzy PAM dla analizowanych sekwencji białkowych

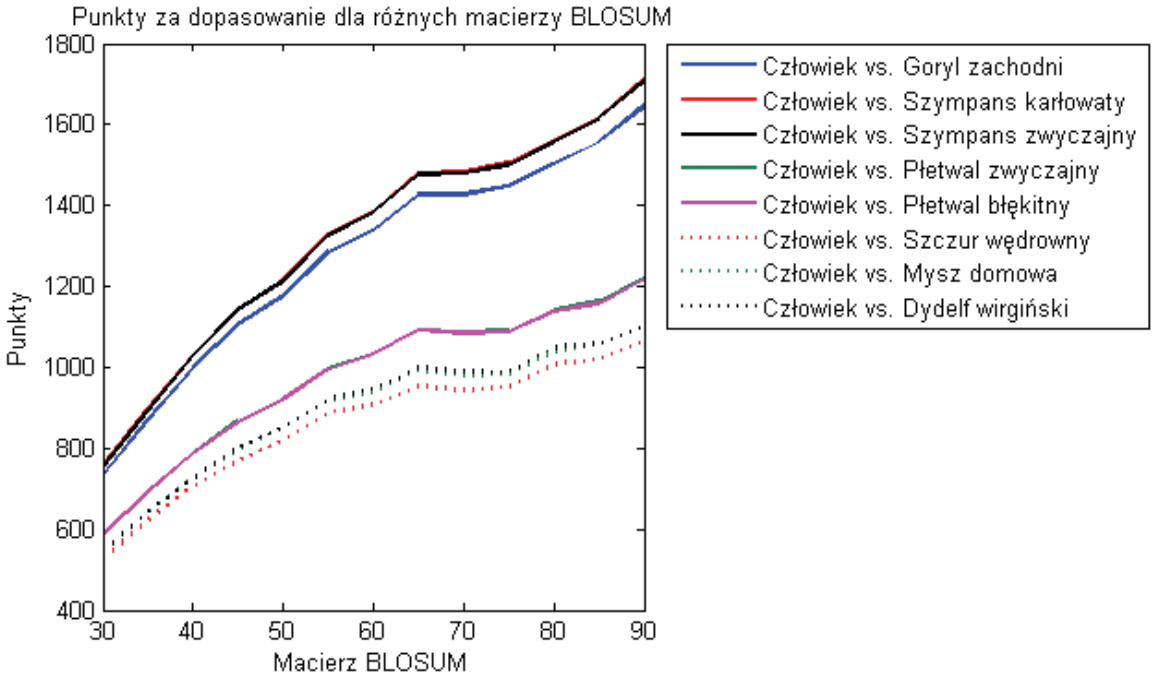
Po przeanalizowaniu powyższego wykresu można zauważyć, że dla sekwencji białka goryla zachodniego, szympansa karłowatego oraz zwyczajnego, wartość maksymalna punktów dla poszczególnych krzywych wykresu odpowiada najmniejszym wartościom PAM. Oznacza to, że odległość ewolucyjna pomiędzy sekwencjami tych gatunków a człowieka jest mniejsza, w porównaniu z resztą analizowanych zwierząt. Można również zauważyć, że wykres dla tych 3 reprezentantów swoich gatunków, jest malejący, a najwyższa wartość znajduje się przed wykresem, najprawdopodobniej oscyluje pomiędzy PAM1 a PAM10. Jest to spowodowane ograniczeniem programu MATLAB, którego funkcja mieści się w zakresie od macierzy PAM10 do macierzy PAM500. Po dokładnym przeanalizowaniu wykresu dla reszty sekwencji najbardziej optymalna była macierz PAM40.

Powyższe wyniki oznaczają, że sekwencja aminokwasów dla białka ND5 badanych gatunków zwierząt w porównaniu z człowiekiem nie jest mocno zróżnicowana, co można stwierdzić na podstawie dopasowania niskich numerów macierzy PAM.

2.2. Macierz BLOSUM – BLOck SUbstitution Matrix

Kolejnym modelem umożliwiającym analizę substytucji aminokwasowych jest macierz BLOSUM (*z ang. BLOck SUbstitution Matrix)* (59). Konstrukcja tej macierzy jest podobna do PAM, jednakże różnią się one między sobą przede wszystkim tym, że w obliczaniu kolejnych macierzy BLOSUM nie korzysta się z ekstrapolacji tylko liczba macierzy BLOSUM określa dokładnie procent identyczności porównywanych sekwencji aminokwasowych. Dlatego im większy element macierzy BLOSUM, tym bardziej porównywane obiekty są bardziej ze sobą spokrewnione.

Poniżej, podobnie jak dla macierzy PAM10, został przedstawiony wykres punktacji różnych wartości macierzy BLOSUM dla analizowanych gatunków zwierząt w porównaniu z sekwencją aminokwasową proteiny ND5 dla człowieka, który również został wygenerowany przy wykorzystaniu programu MATLAB.



Rysunek 9. Wykres punktacji różnych macierzy BLOSUM dla analizowanych sekwencji białkowych

Po przeanalizowaniu powyższego wykresu można zauważyć, że dla wszystkich sekwencji najwyższa punktacja przypadła dla macierzy BLOSUM powyżej 90. Oznacza to, że wszystkie sekwencje aminokwasów proteiny ND5 analizowanych zwierząt są

bardzo blisko spokrewnione z porównywaną sekwencją człowieka, gdyż ich procent identyczności przekracza 90.

Podsumowując można zauważyć jedną z podstawowych różnic pomiędzy użytecznością powyższych metod, na przykładzie analizowanych sekwencji aminokwasowych. Macierz PAM pozwoliła na ogólne rozróżnienie, które z grup zwierząt są najbardziej spokrewnione z człowiekiem, pod względem analizy sekwencji aminokwasów proteiny ND5. Jednak z powodu potrzeby zastosowania funkcji ekstrapolacji bezpośrednio nie otrzymano dokładnej informacji na temat identyczności przeciwstawnych sobie sekwencji. Dzięki zastosowaniu macierzy BLOSUM, można określić, że poziom identyczności dla analizowanych sekwencji zwierząt przekracza 90% w porównaniu z sekwencją człowieka dla proteiny ND5.

3. 20–WYMIAROWA DYNAMICZNA REPREZENTACJA SEKWENCJI BIAŁKOWYCH

3.1. Zarys metody

W niniejszej rozprawie doktorskiej będzie rozszerzana reprezentacja sekwencji łańcuchów DNA do reprezentacji sekwencji białkowych. Jako deskryptory zostały zaproponowane momenty bezwładności (*z ang. moments of inertia*) pewnych 20-wymiarowych struktur nazwanych 20-wymiarowymi dynamicznymi grafami (66). Opisują one sposób rozkładu masy wokół osi obrotu. Im dalej masa jest rozłożona wokół osi obrotu, tym większy jest jej moment bezwładności *I*. Dystrybucje punktów w przestrzeni uzyskano za pomocą metody przesunięć i w ten sposób otrzymano abstrakcyjny 20-wymiarowy dynamiczny graf.

Do utworzenia wykresu używana jest konwencja przesunięcia w 20-wymiarowej przestrzeni. Przesunięcie rozpoczynane jest w początku układu współrzędnych 20D (0, 0, … 0). Każdy krok opisany jest przez wektor jednostkowy, reprezentujący odpowiedni aminokwas w sekwencji. Każdemu aminokwasowi przypisujemy odpowiednią oś w 20-wymiarowym układzie współrzędnych (*tabela nr 5*).

Tabela 5. Przypisanie aminokwasom kolejnych osi w 20-wymiarowym układzie

NR OSI	AMINOKWAS	KOD
1	Alanina	A
2	Cysteina	C
3	Kwas asparaginowy	D
4	Kwas glutaminowy	E
5	Fenyloalanina	F
6	Glicyna	G
7	Histydyna	H
8	Izoleucyna	I
9	Lizyna	K
10	Leucyna	L
11	Metionina	M
12	Asparagina	N
13	Prolina	P
14	Glutamina	Q
15	Arginina	R
16	Seryna	S
17	Treonina	T
18	Walina	V
19	Tryptofan	W
20	Tyrozyna	Y

Na końcu każdego wektora został umiejscowiony punkt materialny o masie: $m = 1$. W konsekwencji otrzymujemy abstrakcyjny, 20-wymiarowy wykres, który składa się z $m = 1$ punktów materialnych w 20-wymiarowej przestrzeni. Rozkład tych punktów pochodzi z sekwencji białkowej, która jest reprezentowana przez ten wykres. Ponieważ masa każdego pojedynczego punktu wynosi jeden to całkowita masa sekwencji obliczana jest wg wzoru:

$$N = \sum_{i=1}^{N} m_i \,, \tag{8}$$

gdzie N jest długością sekwencji (liczbą punktów na wykresie). Współrzędne punktów mas m_i są określone przez przesunięcie w przestrzeni 20D i są oznaczone jako: $x_i^k, k = 1, 2, \dots 20$.

W niniejszej rozprawie doktorskiej, jako nowe deskryptory wykresów 20D zostały zaproponowane momenty bezwładności.

Tensor momentu bezwładności w przestrzeni 20D jest podany przez:

$$\hat{I} = \begin{pmatrix} I_{1\,1} & I_{1\,2} & \dots & I_{1\,k} & \dots & I_{1\,20} \\ I_{2\,1} & I_{2\,2} & \dots & I_{2\,k} & \dots & I_{2\,20} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ I_{j\,1} & I_{j\,2} & \dots & I_{j\,k} & \dots & I_{j\,20} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ I_{20\,1} & I_{20\,2} & \dots & I_{20\,k} & \dots & I_{20\,20} \end{pmatrix}, \tag{9}$$

dla którego główne przekątne określane są wzorem:

$$I_{jj} = \sum_{i=1}^{N} m_i \sum_{k=1}^{20} \left[\hat{x}_i^k \left(1 - \delta_{jk}\right)\right]^2, \tag{10}$$

a pozostałe wartości:

$$I_{jk} = I_{kj} = -\sum_{i=1}^{N} m_i \hat{x}_i^j \, \hat{x}_i^k \,, \tag{11}$$

gdzie:

$$\delta_{ij} = \begin{cases} 1 & i = 1, \\ 0 & i \neq 1 \end{cases} \tag{12}$$

to delta Kroneckera, i $\hat{x}_i^k, = 1,2, \dots 20$ określa współrzędne m_i. Współrzędne zostały zdefiniowane przy użyciu kartezjańskiego systemu współrzędnych, dla którego środek układu wyznaczany jest przez środek masy, np.:

$$\hat{x}_i^k = x_i^k - \bar{x}^k \, , \tag{13}$$

gdzie:

$$\bar{x}^k = \frac{1}{N} \sum_{i=1}^{N} m_i x_i^k \tag{14}$$

to współrzędne środka masy 20 wymiarowego wykresu.
Zagadnienie na wartości własne tensora bezwładności zdefiniowane jest jako:

$$\hat{I}\omega_k = I_k \omega_k, k = 1,2, \dots 20 \, , \tag{15}$$

gdzie: I_k to wartości własne i ω_k to wektory własne. Wartości własne są uzyskiwane poprzez rozwiązanie charakterystycznego równania 20-stego rzędu:

$$det\left(\hat{I} - I\hat{E}\right) = 0 \, , \tag{16}$$

gdzie $\hat{E}$ jest macierzą jednostkową o wymiarze 20×20. Wartości własne $I_1, I_2, \dots I_{20}$ są nazywane głównymi momentami bezwładności.
Jako nowe deskryptory 20-wymiarowego wykresu zostały wybrane znormalizowane główne momenty bezwładności, odpowiednio:

$$r_1 = \sqrt{\frac{I_1}{N}}, \qquad r_2 = \sqrt{\frac{I_2}{N}} \, , \dots \quad r_{20} = \sqrt{\frac{I_{20}}{N}} \, . \tag{17}$$

W celu porównania sekwencji wygodne jest korzystanie z niektórych bezwymiarowych i znormalizowanych miar podobieństwa. W niniejszej rozprawie miarę podobieństwa par sekwencji oznaczonych jako α i β przedstawia się następująco:

$$S(\alpha,\beta) = S(\beta,\alpha) = \left|\frac{D^{\alpha} - D^{\beta}}{D^{\alpha} + D^{\beta}}\right|. \tag{18}$$

Wartości podobieństwa wahają się pomiędzy 0 a 1. Jako deskryptory, które reprezentują sekwencję brane są:

$$D = \sum_{i=1}^{20} r_i\,, \tag{19}$$

gdzie r_i zostały już wcześniej zdefiniowane.

Dla przykładu weźmy model sekwencji aminokwasów: MALWMR. Ponieważ $N = 6$ abstrakcyjny wykres sekwencji składa się z sześciu 20-wymiarowych punktów. Przesunięcie w przestrzeni rozpoczynamy w początku układu współrzędnych. Pierwszym z aminokwasów w sekwencji jest metionina (M), reprezentowana przez wektor jednostkowy: $(0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$.

Pierwsza masa jednostkowa, m_1, zostaje umiejscowiona na końcu powyższego wektora. Następnie m_1 jest punktem wyjścia dla kolejnego kroku, zgodnie z którym następny wektor jednostkowy przedstawia drugi aminokwas w analizowanej sekwencji, itd. 20-wymiarowy wykres składający się z punktów mas $(x_i^1,\ x_i^2, \ldots x_i^{20})$ dla analizowanego modelu sekwencji aminokwasowych przedstawia *tabela nr 6*.

Tabela 6. 20-wymiarowy wykres dla modelu sekwencji aminokwasów MALWMR

$\boldsymbol{m_i}$	$(\boldsymbol{x_i^1},$	$\boldsymbol{x_i^2},$	$\boldsymbol{x_i^3},$	$\boldsymbol{x_i^4},$	$\boldsymbol{x_i^5},$	$\boldsymbol{x_i^6},$	$\boldsymbol{x_i^7},$	$\boldsymbol{x_i^8},$	$\boldsymbol{x_i^9},$	$\boldsymbol{x_i^{10}},$	$\boldsymbol{x_i^{11}},$	$\boldsymbol{x_i^{12}},$	$\boldsymbol{x_i^{13}},$	$\boldsymbol{x_i^{14}},$	$\boldsymbol{x_i^{15}},$	$\boldsymbol{x_i^{16}},$	$\boldsymbol{x_i^{17}},$	$\boldsymbol{x_i^{18}},$	$\boldsymbol{x_i^{19}},$	$\boldsymbol{x_i^{20}})$
$\boldsymbol{m_1}$	(0,	0,	0,	0,	0,	0,	0,	0,	0,	0,	1,	0,	0,	0,	0,	0,	0,	0,	0,	0)
$\boldsymbol{m_2}$	(1,	0,	0,	0,	0,	0,	0,	0,	0,	0,	1,	0,	0,	0,	0,	0,	0,	0,	0,	0)
$\boldsymbol{m_3}$	(1,	0,	0,	0,	0,	0,	0,	0,	0,	1,	1,	0,	0,	0,	0,	0,	0,	0,	0,	0)
$\boldsymbol{m_4}$	(1,	0,	0,	0,	0,	0,	0,	0,	0,	1,	1,	0,	0,	0,	0,	0,	0,	0,	1,	0)
$\boldsymbol{m_5}$	(1,	0,	0,	0,	0,	0,	0,	0,	0,	1,	2,	0,	0,	0,	0,	0,	0,	0,	1,	0)
$\boldsymbol{m_6}$	(1,	0,	0,	0,	0,	0,	0,	0,	0,	1,	2,	0,	0,	0,	1,	0,	0,	0,	1,	0)

W celu wizualizacji wykresów 20D, możemy rzutować je w przestrzeń 2D lub 3D.

3.2. Metoda numeryczna – metoda Jacobiego

W celu wyznaczenia wektorów własnych macierzy wielowymiarowych posłużono się metodami numerycznymi, które umożliwiają znacznie szybsze znalezienie rozwiązania w przypadku problemów wielowymiarowych czy po prostu bardziej skomplikowanych operacji.

W przypadku metody 20-wymiarowej dynamicznej reprezentacji białek znajdowane są wektory własne macierzy 20-wymiarowej. Jedną z metod numerycznych, która okazała się efektywna dla tej macierzy $I_{[20\times20]}$ jest metoda Jacobiego. Metoda ta umożliwia obliczenie wartości własnych i wektorów własnych za pomocą przekształceń ortogonalnych na macierzy a priori $I_{[m\times m]}$, która jest macierzą symetryczną. Do obliczeń użyto tutaj klasycznego algorytmu Jacobiego, gdzie transformacje ortogonalne polegają na zerowaniu w kolejnych krokach iteracji pozadiagonalnej pary elementów o największej co do modułu wartości (para elementów jest rozmieszczona symetrycznie względem przekątnej głównej), aby jak najbardziej zredukować miarę niediagonalności macierzy. W wyniku tych operacji macierz wyjściowa zostaje sprowadzona do postaci (prawie) diagonalnej $A_{[m\times m]}$, gdzie na głównej przekątnej znajdują się wartości własne. Natomiast wektory własne odpowiadające tym wartościom własnym są zapisywane do macierzy $W_{[m\times m]}$.

Zatem naszą pierwotną macierz możemy zapisać zgodnie ze wzorem:

$$I = W \times A \times W^T , \tag{20}$$

gdzie macierz $W_{[m\times m]}$ określona jest za pomocą iloczynu kolejnych transformacji ortogonalnych $Q^{\{i\}}$dla $i = 1,2,\dots,m$ zapisanych w postaci:

$$W = Q^{\{n\}} \times \dots \times Q^{\{n\}} \times \dots \times Q^{\{n\}} \times Q^{\{n\}} , \tag{21}$$

Metoda Jacobiego jest metodą iteracyjną, która korzysta zawsze z wyników poprzedniej iteracji. Iteracje mają na celu wyzerować wszystkie elementy macierzy poza główną przekątną według zasady: wybieramy pary elementów macierzy I rozmieszczonych symetrycznie względem głównej przekątnej i największych co do modułu. Algorytm ten można przedstawić w postaci:

- wyznaczenie elementu wiodącego (największego co do modułu) dla macierzy I,
- obliczenie stałych współczynników dla poszczególnej iteracji,
- wyznaczenie macierzy $I^{\{i+1\}}$,
- wyznaczenie macierzy $W^{\{i+1\}}$,
- sprawdzenie warunku iteracji (jeżeli warunek jest spełniony proces iteracji zostaje zakończony).

Wszystkie wartości własne i odpowiadające im wektory własne zostały obliczone z dokładnością do ε = 0.0001.

Według tego schematu metoda Jacobiego została zaimplementowana w MATLABIE i umożliwiła nam otrzymanie zestawu 20 wartości własnych, które w pracy nazywane są głównymi momentami bezwładności.

3.3. Dane początkowe dla ND5

W celu przedstawienia metody 20 wymiarowej reprezentacji białek przeprowadzono analizę podobieństwa sekwencji dehydrogenazy NADH podjednostki 5 (ND5) dla dziewięciu różnych gatunków. Dane zostały pobrane z biologicznej bazy danych PDB (*z ang. Protein Data Bank*). W *tabeli nr 7* podano numer dostępu do sekwencji aminokwasowej dla danego gatunku, która znajduje się w PDB.

Tabela 7. Dane dotyczące użytych sekwencji ND5

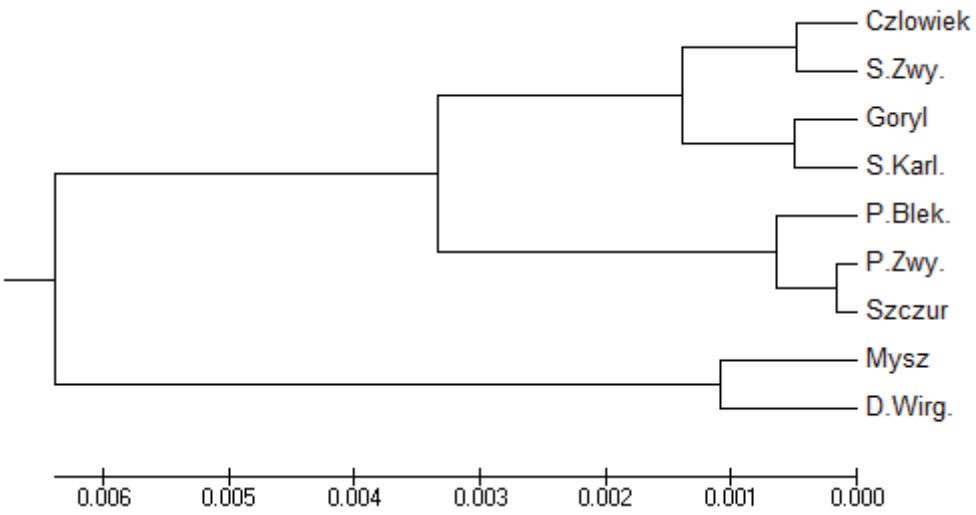
Lp.	Gatunek	Numer dostępu	Dł. sekwencji
1	Człowiek (*Homo sapiens*)	AP_000649	603
2	Goryl zachodni (*Gorilla gorilla*)	NP_008222	603
3	Szympans karłowaty (*Pan paniscus*)	NP_008209	603
4	Szympans zwyczajny (*Pan troglodytes*)	NP_008196	603
5	Płetwal zwyczajny (*Balenoptera physalus*)	NP_006899	606
6	Płetwal błękitny (*Balenoptera musculus*)	NP_007066	606
7	Szczur wędrowny (*Rattus norvegicus*)	AP_004902	610
8	Mysz domowa (*Mus musculus*)	NP_904902	607
9	Dydelf wirgiński (*Didelphis virginiana*)	NP_007105	602

Za pomocą nowej metody wyznaczono macierz podobieństwa i jej reprezentację przy użyciu drzewa filogenetycznego (*rysunek nr 10*). Drzewo filogenetyczne w sposób graficzny oddaje podobieństwo ewolucyjne między gatunkami.

Tabela 8. Macierz podobieństwa dla ND5

Gatunek	Człowiek	Goryl	S.Karł.	S.Zwy.	P.Zwy.	P.Błęk.	Szczur	Mysz	D.Wirg.
Człowiek	0.00000	0.00279	0.00181	0.00094	0.00818	0.00674	0.00787	0.01548	0.01763
Goryl		0.00000	0.00097	0.00372	0.00540	0.00396	0.00508	0.01269	0.01485
S.Karł			0.00000	0.00275	0,00637	0.00493	0.00606	0.01366	0.01582
S.Zwy.				0.00000	0.00912	0.00768	0.00881	0.01642	0.01857
P.Zwy.					0.00000	0.00144	0.00031	0.00729	0.00945
P.Błęk.						0.00000	0.00113	0.00873	0.01089
Szczur							0.00000	0.00761	0.00976
Mysz								0.00000	0.00215
D.Wirg.									0.00000

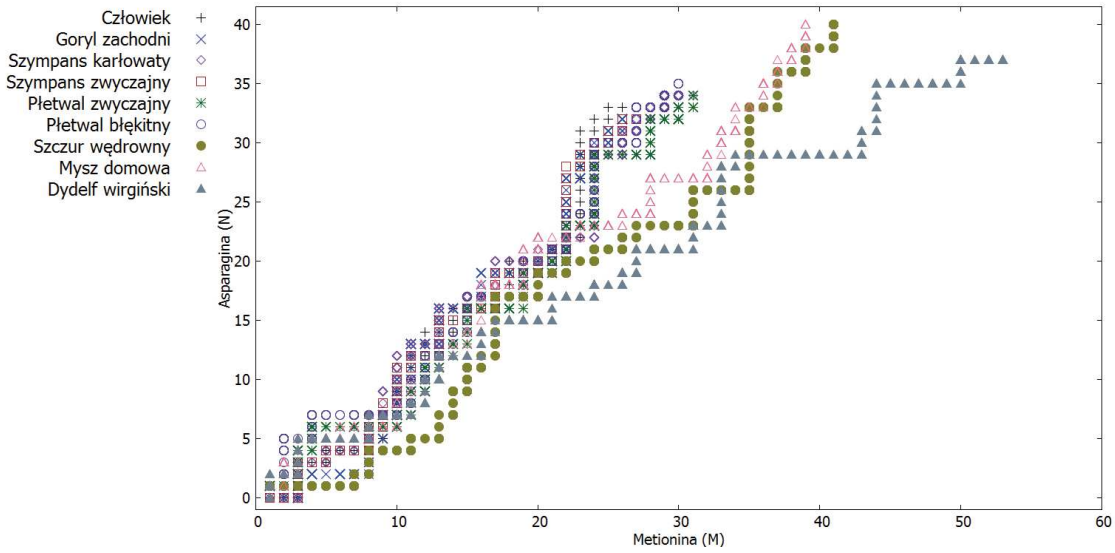
Na podstawie powyższej tabeli (*tabela nr 8*) możemy zauważyć duże podobieństwo pomiędzy sekwencjami człowieka, goryla zachodniego i szympansa zwyczajnego (małe wartości w tabeli). Natomiast sekwencja dydelfa wirgińskiego jest najmniej podobna w porównaniu do innych sekwencji (duże wartości w tabeli) dla rozważanego białka. Warto nadmienić, że inni autorzy również uzyskali podobne wyniki (16).



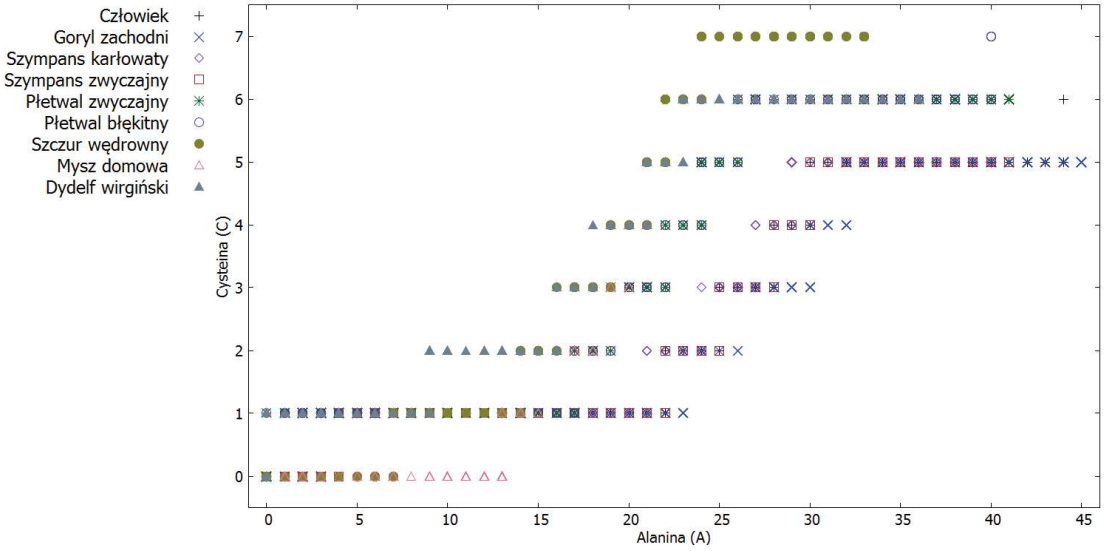
Rysunek 10. Drzewo filogenetyczne dla ND5

Podobieństwo pomiędzy gatunkami zostało także przedstawione za pomocą wykresów 2D (*rysunki 11-30*) i 3D (*rysunki 31-47*). Na podstawie rysunków 2D nr: 11, 18, 21, 22, 25, 26-29 możemy zauważyć, że sekwencja dydelfa wirgińskiego jest odseparowana od pozostałych sekwencji. Natomiast sekwencje człowieka, goryla zachodniego, szympansa zwyczajnego i szympansa karłowatego są do siebie zbliżone. Na tych

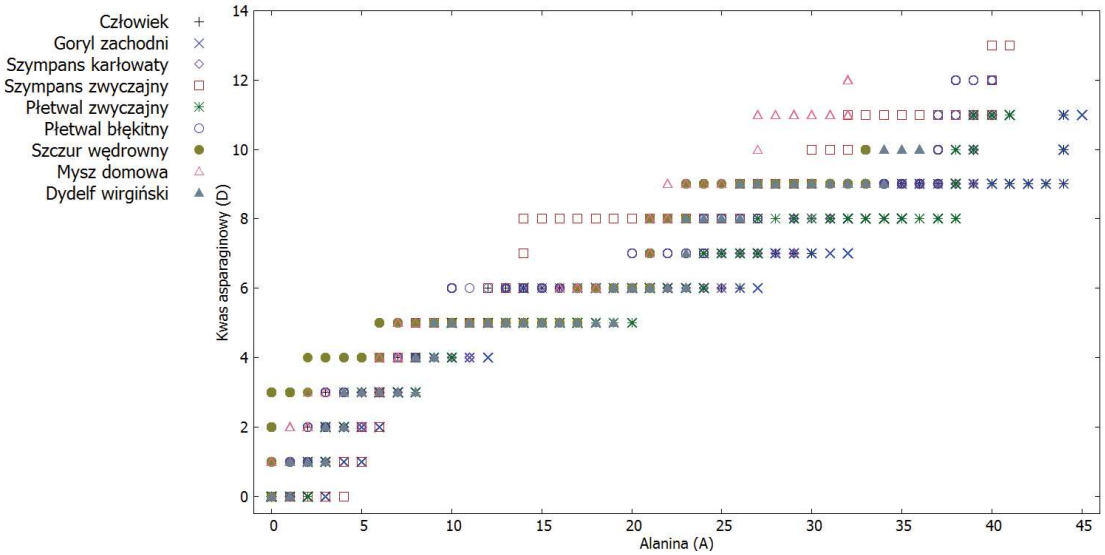
rysunkach również widać podobieństwo pomiędzy dwoma gatunkami: szczurem wędrownym a myszą domową oraz pomiędzy płetwalem zwyczajnym i płetwalem błękitnym. Z kolei rysunki: 12-14, 24 i 30 różnią się tym, że możemy na nich zauważyć podobieństwo pomiędzy sekwencją myszy domowej, szczura wędrownego i dydelfa wirgińskiego. Sekwencje tych 3 gatunków są do siebie zbliżone. Na rysunkach: 15-17 widać podobieństwo pomiędzy człowiekiem, gorylem zachodnim, szympansem karłowatym i szympansem zwyczajnym. Natomiast rysunki: 19, 20 i 23 pokazują podobieństwo pomiędzy człowiekiem i gorylem zachodnim, kolejno pomiędzy szympansem karłowatym i szympansem zwyczajnym, następnie płetwalem zwyczajnym i płetwalem błękitnym oraz pomiędzy szczurem wędrownym i myszą domową. Podobne obserwacje możemy zauważyć dla rysunków 3D. Na rysunkach: 31-36 i 38-47 sekwencje człowieka, goryla zachodniego, szympansa zwyczajnego i szympansa karłowatego są do siebie podobne. Natomiast na podstawie rysunków nr: 32, 35, 41 i 45 zauważalne jest, że sekwencja dydelfa wirgińskiego różni się od pozostałych sekwencji. Jednak w przypadku pozostałych rysunków (z wyjątkiem rysunków nr: 33 i 34, gdzie mysz domowa odstaje od pozostałych sekwencji) sekwencja dydelfa wirgińskiego jest zbliżona do sekwencji szura wędrownego i myszy domowej. Na wszystkich rzutach 3D sekwencje płetwala zwyczajnego i płetwala błękitnego są do siebie podobne.



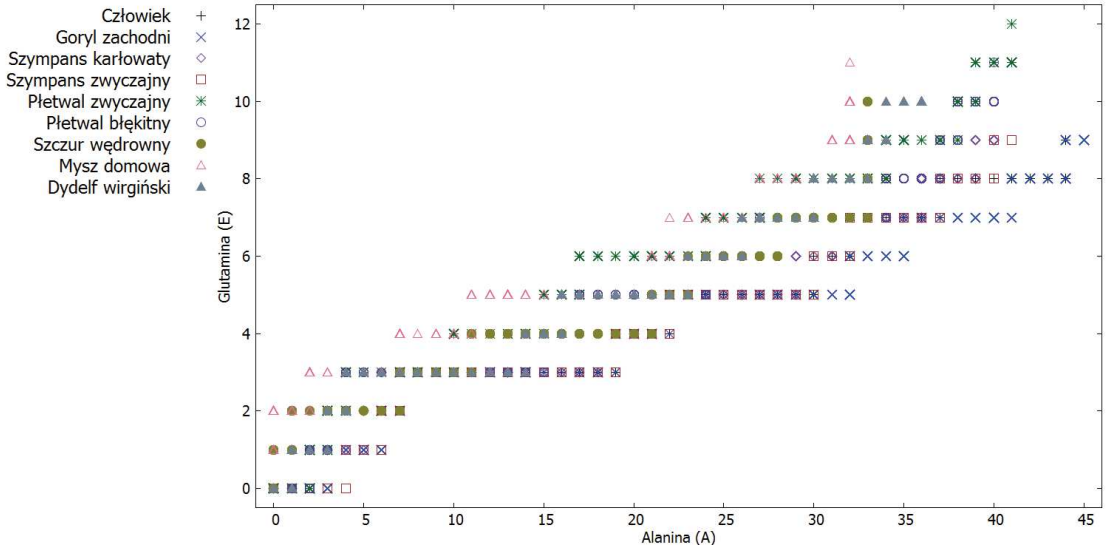
Rysunek 11. Graf 2D - MN dla ND5



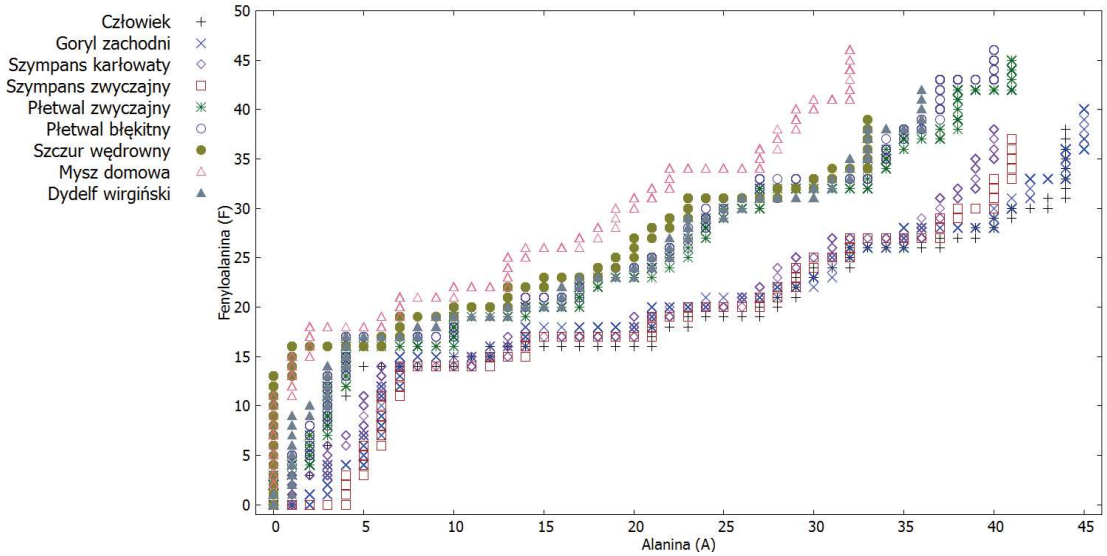
Rysunek 12. Graf 2D - AC dla ND5



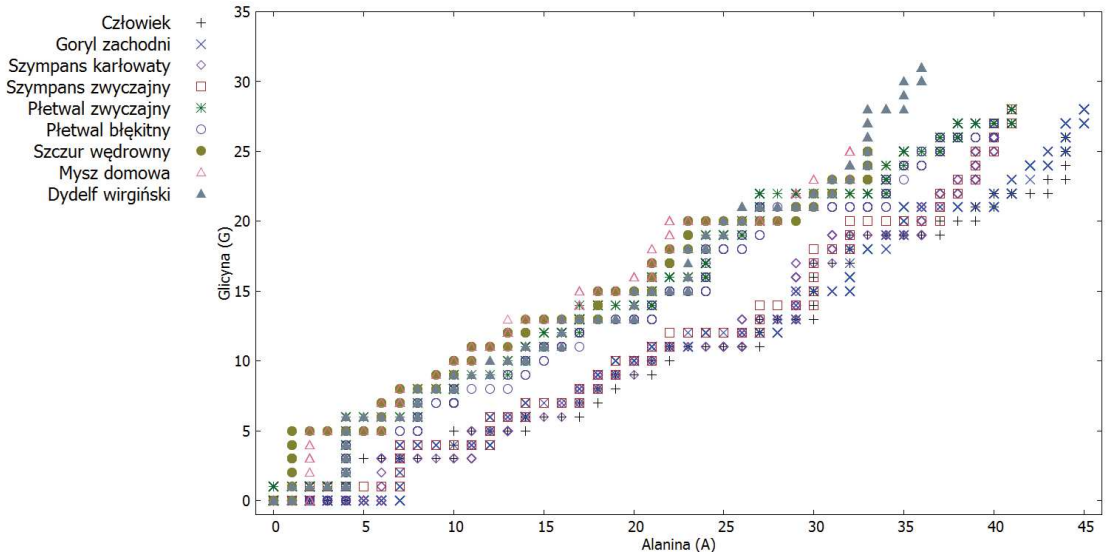
Rysunek 13. Graf 2D - AD dla ND5



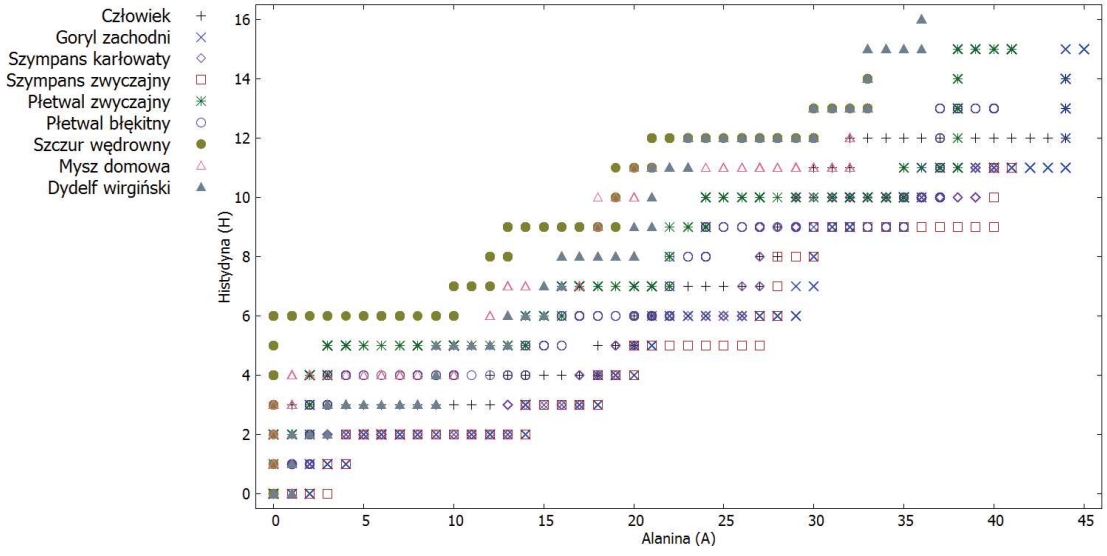
Rysunek 14. Graf 2D - AE dla ND5



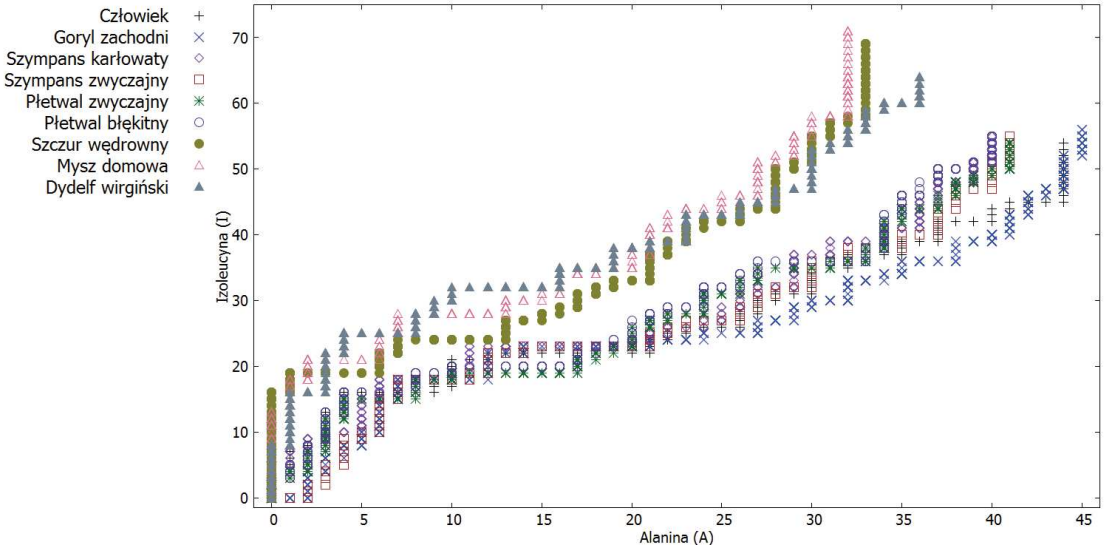
Rysunek 15. Graf 2D - AF dla ND5



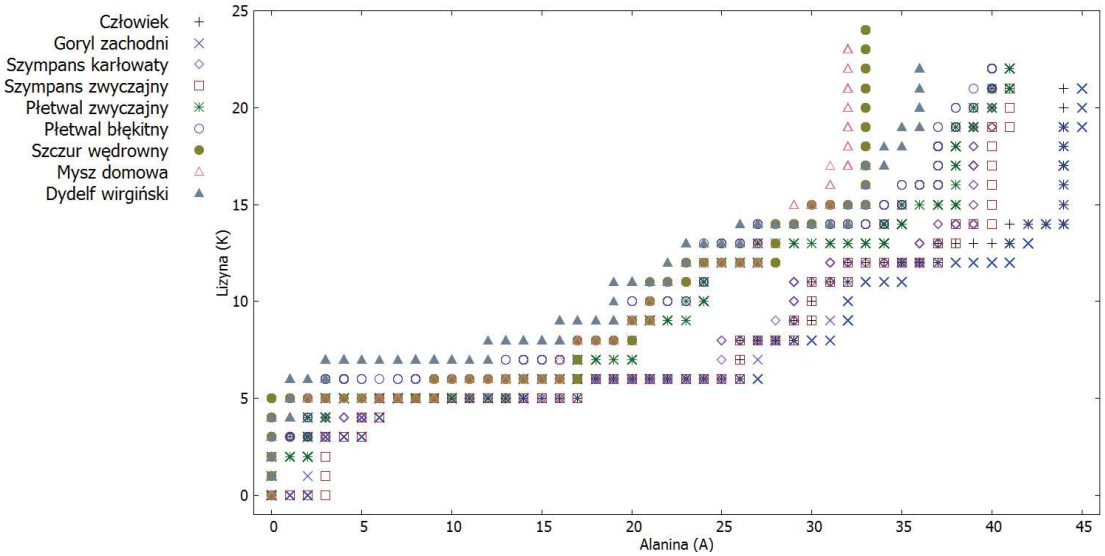
Rysunek 16. Graf 2D - AG dla ND5



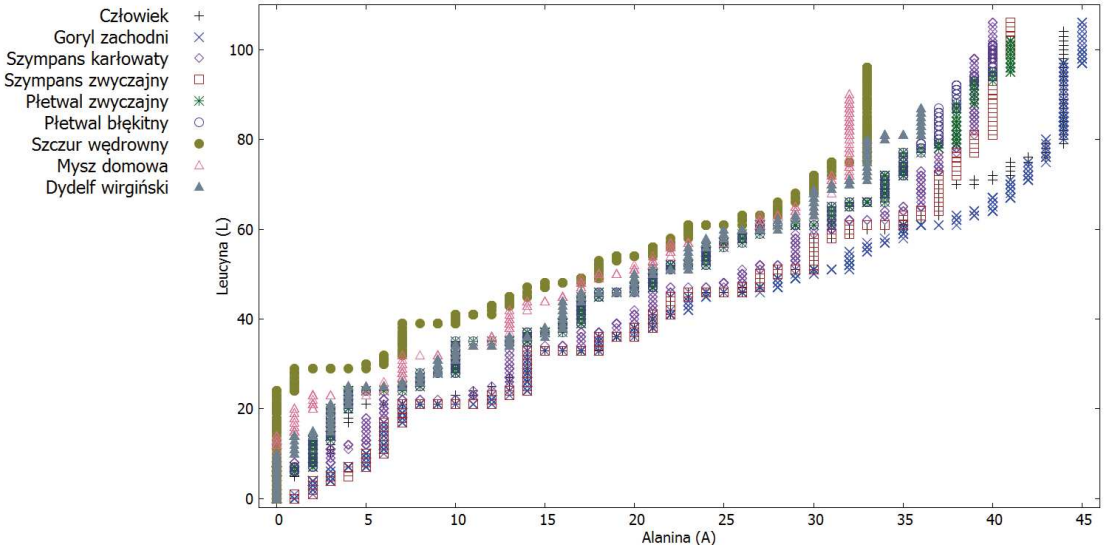
Rysunek 17. Graf 2D - AH dla ND5



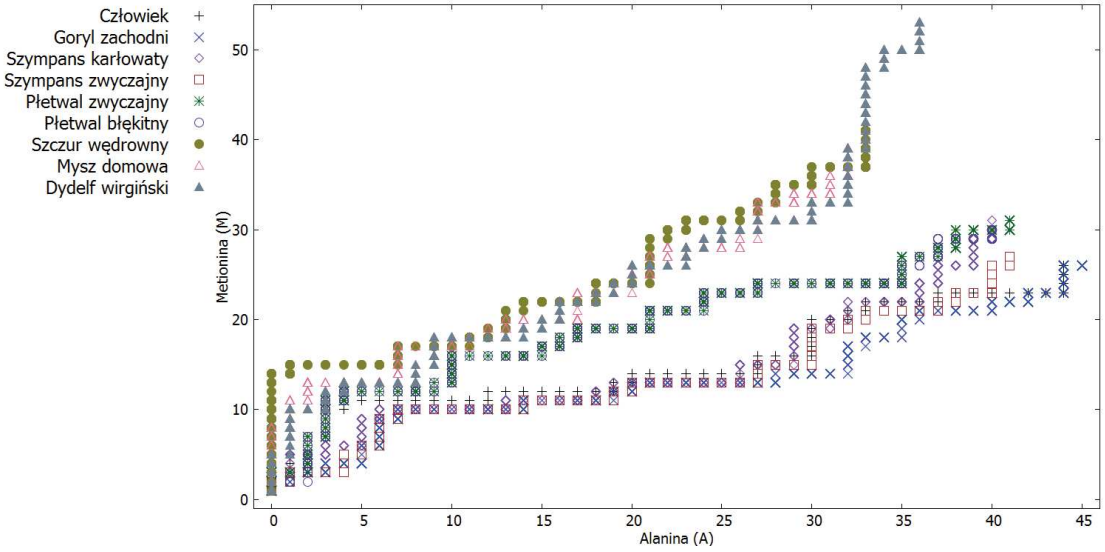
Rysunek 18. Graf 2D - AI dla ND5



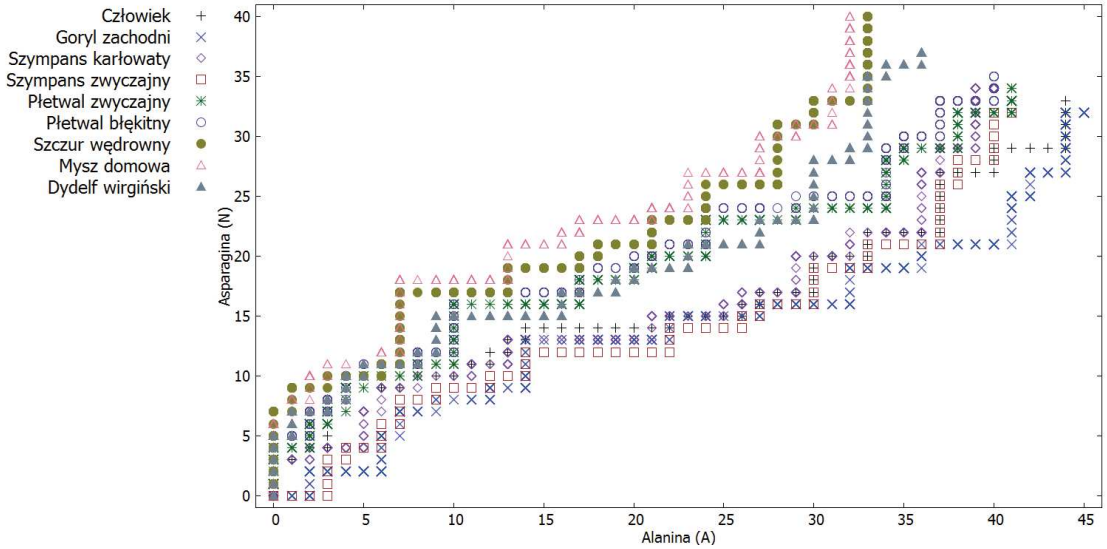
Rysunek 19. Graf 2D - AK dla ND5



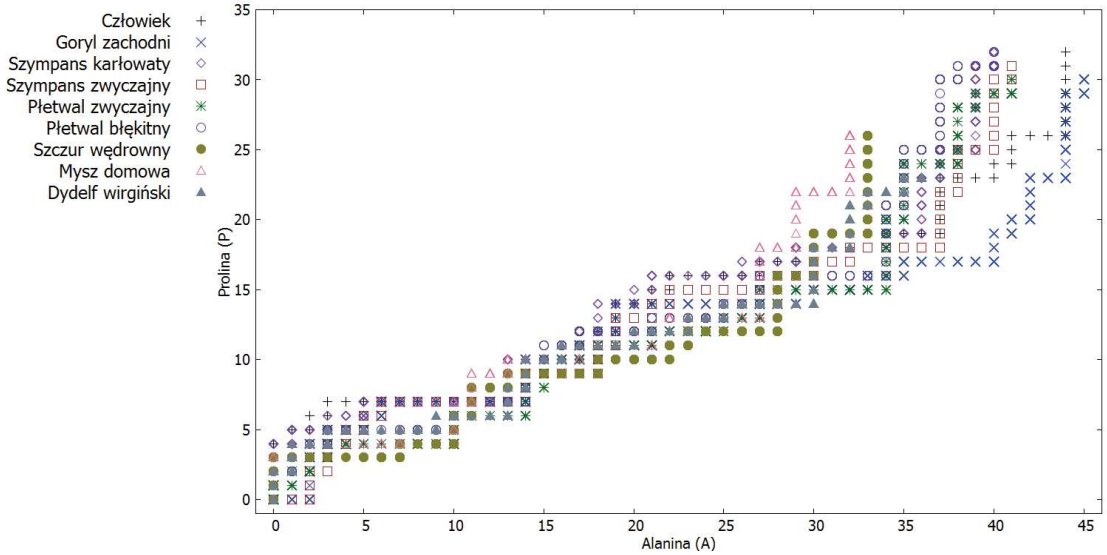
Rysunek 20. Graf 2D - AL dla ND5



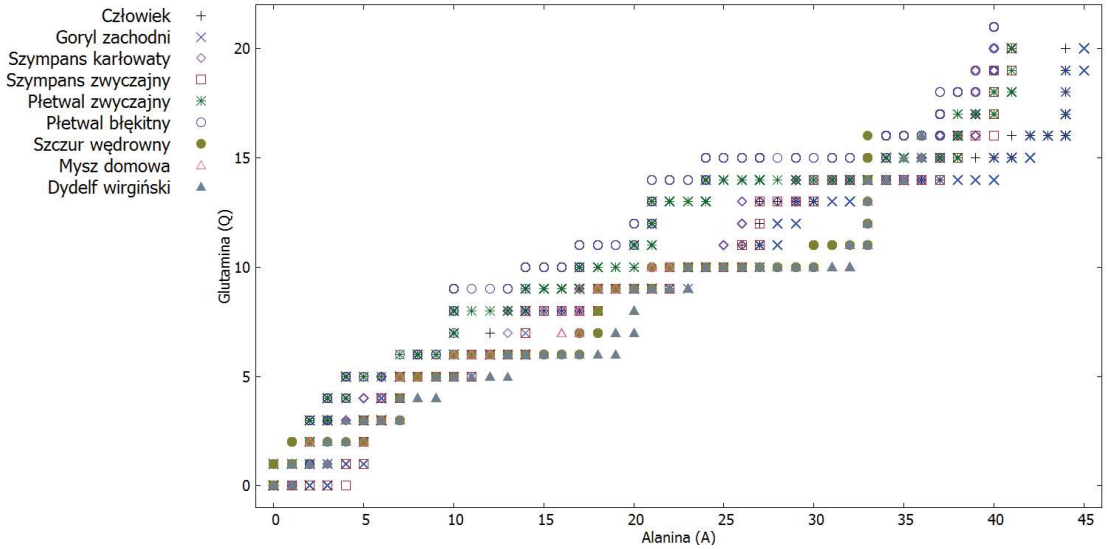
Rysunek 21. Graf 2D - AM dla ND5



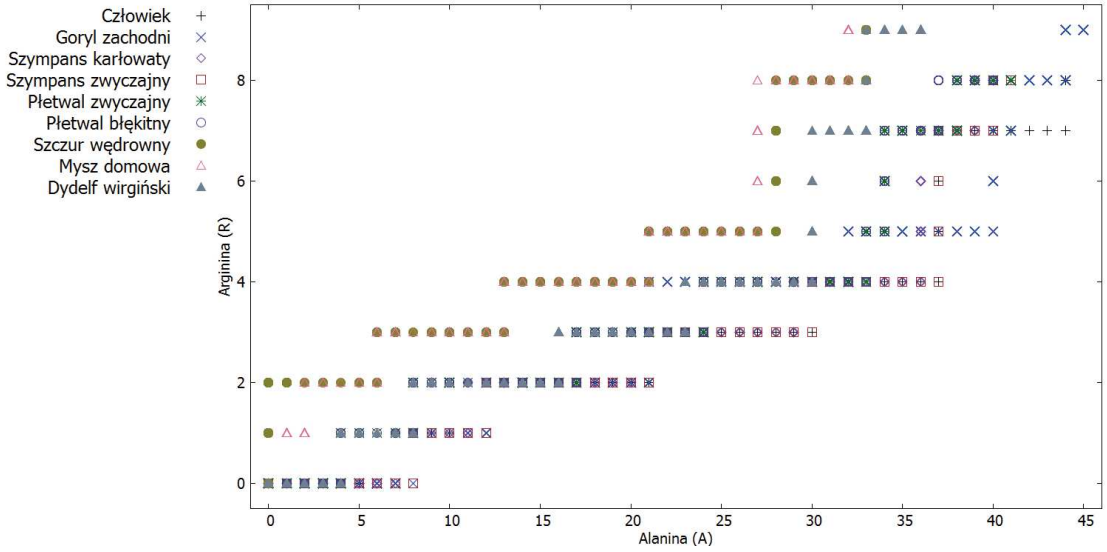
Rysunek 22. Graf 2D - AN dla ND5



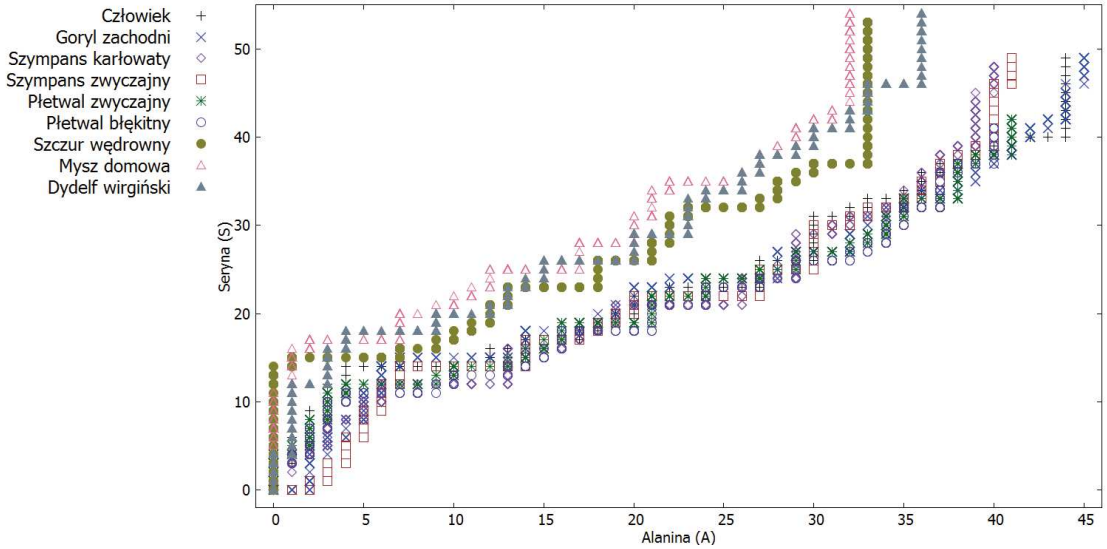
Rysunek 23. Graf 2D - AP dla ND5



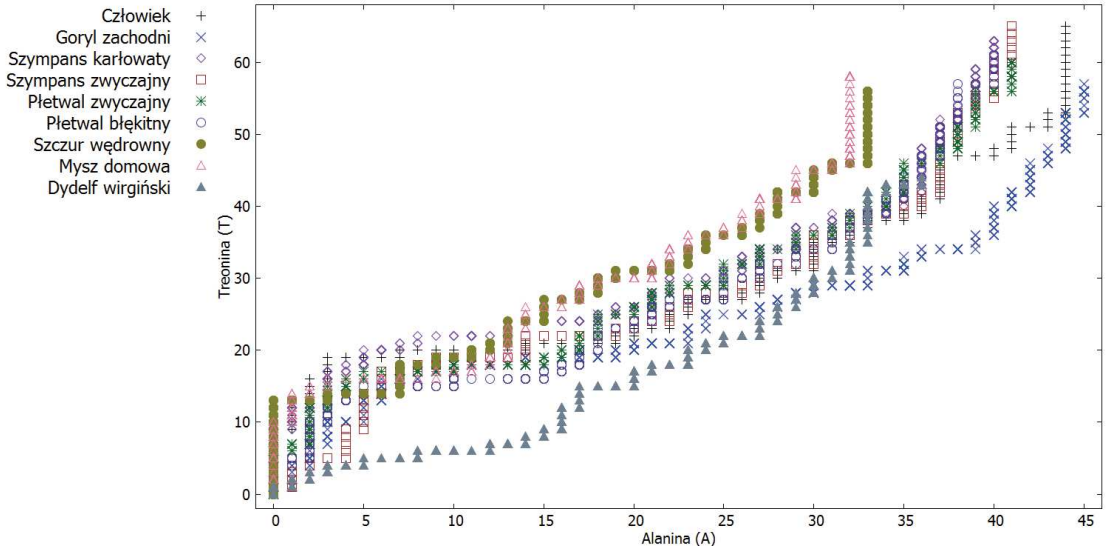
Rysunek 24. Graf 2D - AQ dla ND5



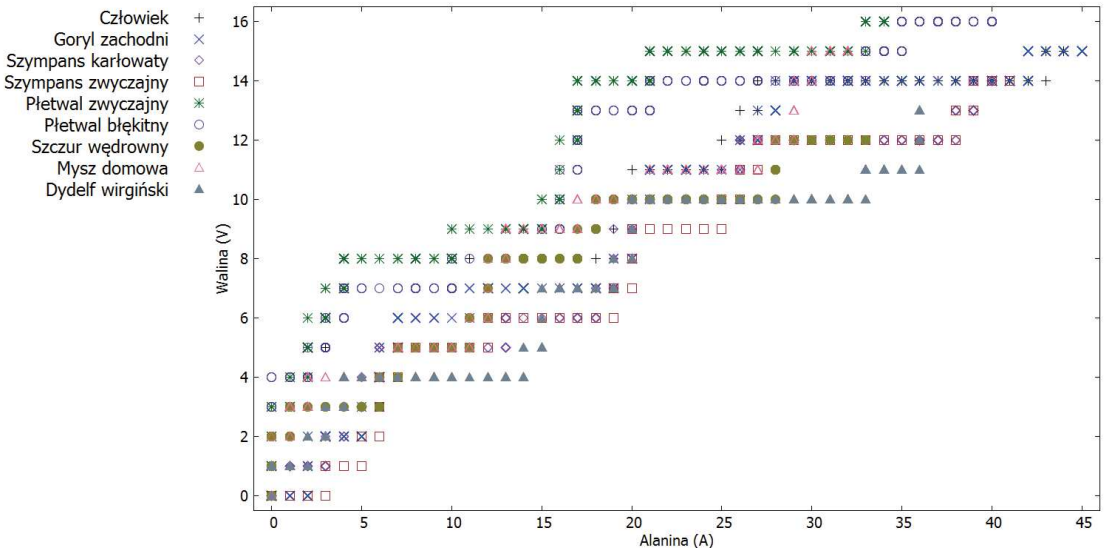
Rysunek 25. Graf 2D - AR dla ND5



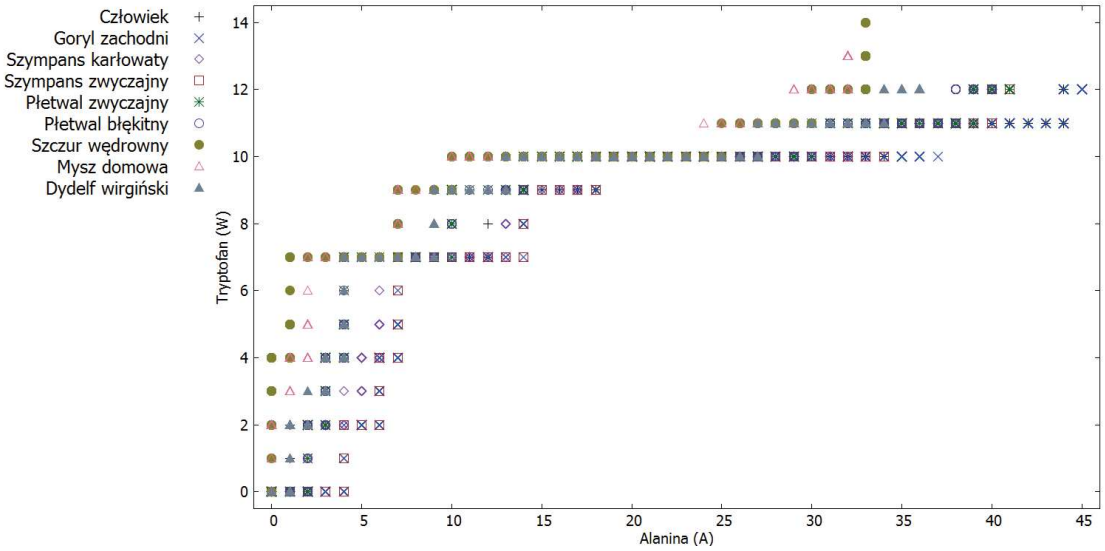
Rysunek 26. Graf 2D - AS dla ND5



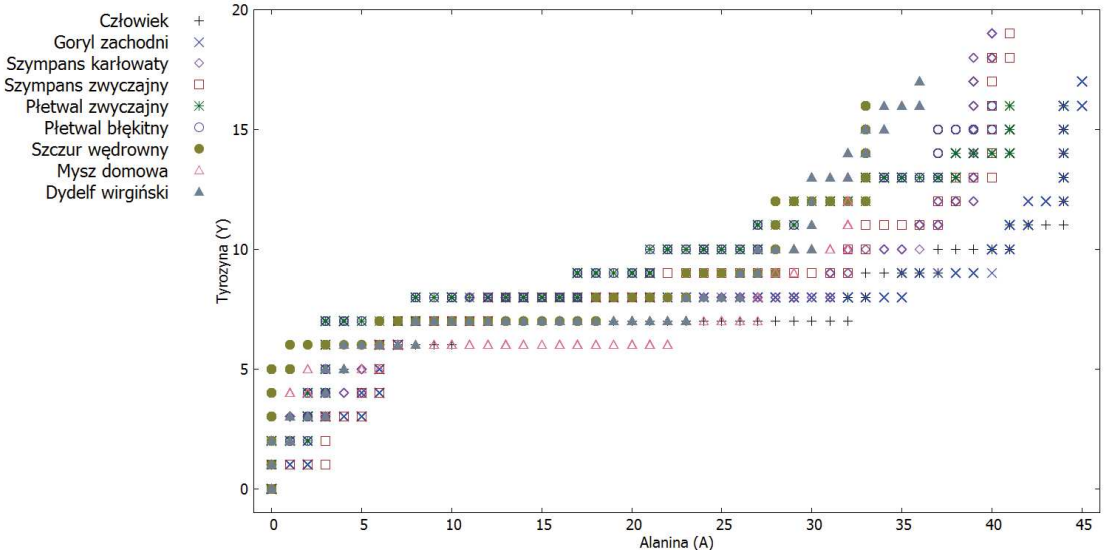
Rysunek 27. Graf 2D - AT dla ND5



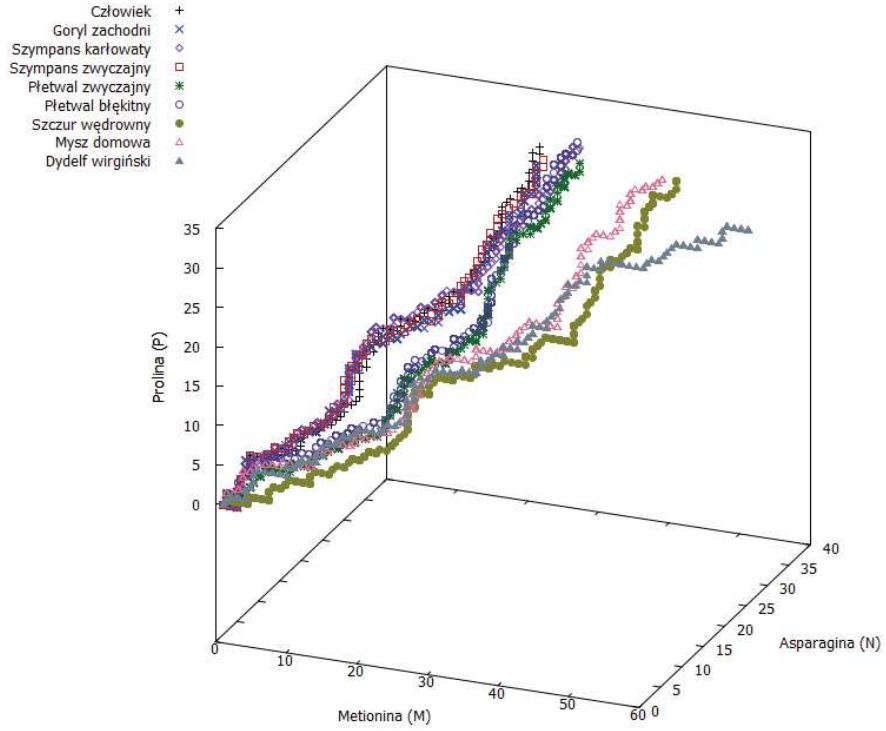
Rysunek 28. Graf 2D - AV dla ND5



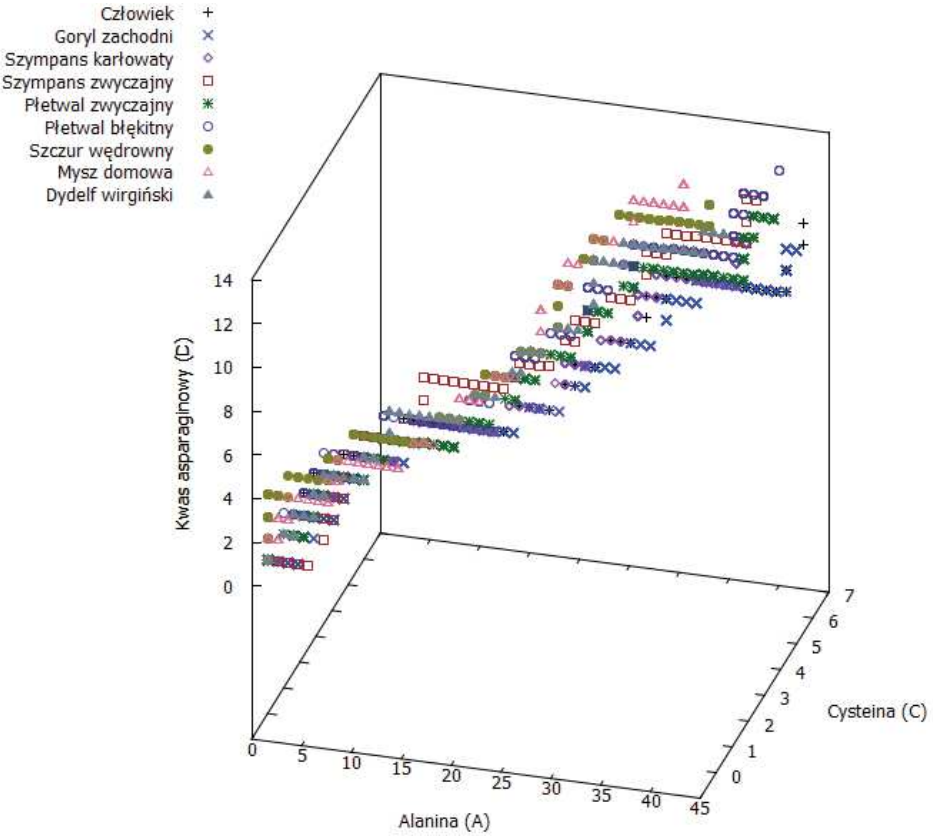
Rysunek 29. Graf 2D - AW dla ND5



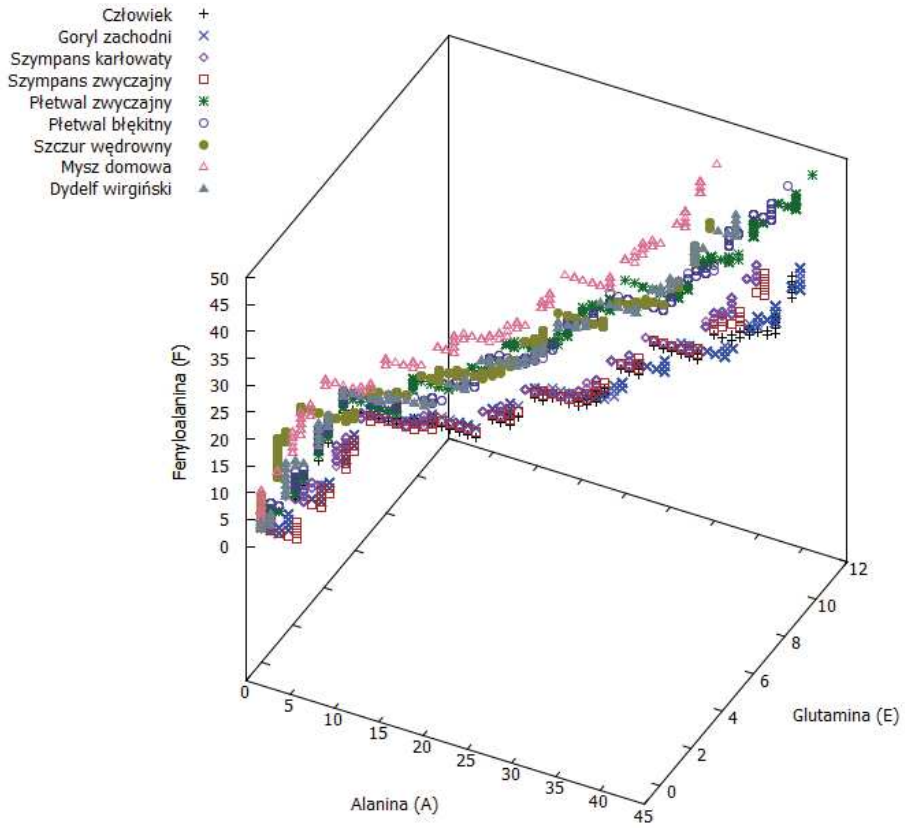
Rysunek 30. Graf 2D - AY dla ND5



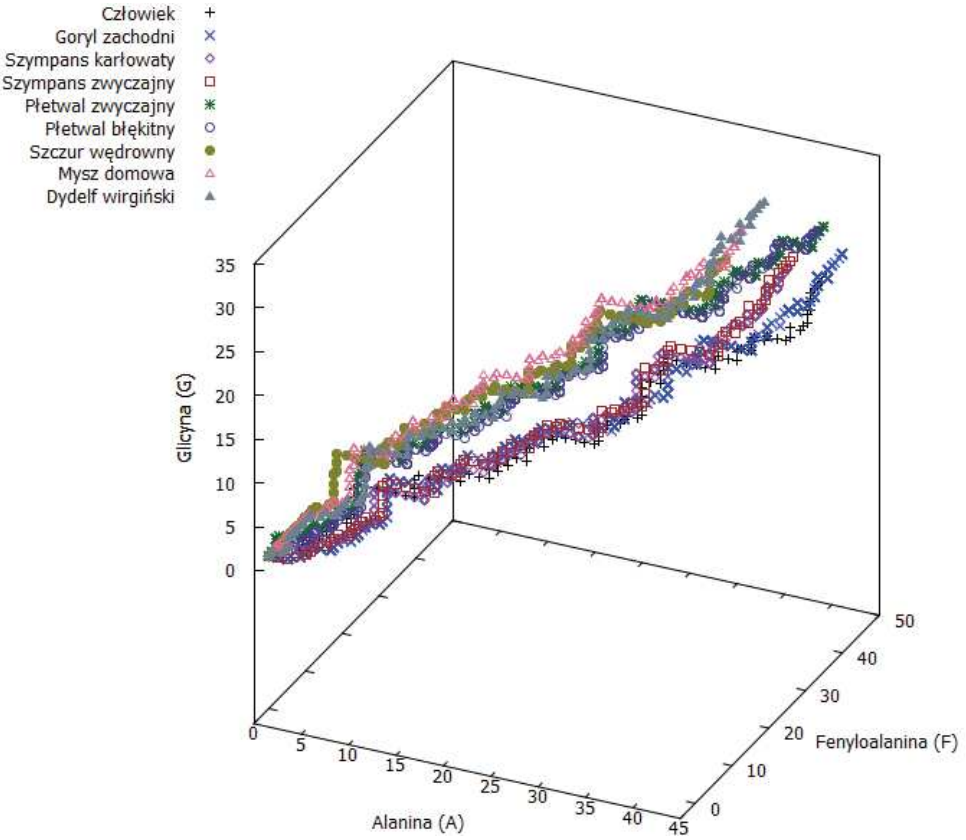
Rysunek 31. Graf 3D - MNP dla ND5



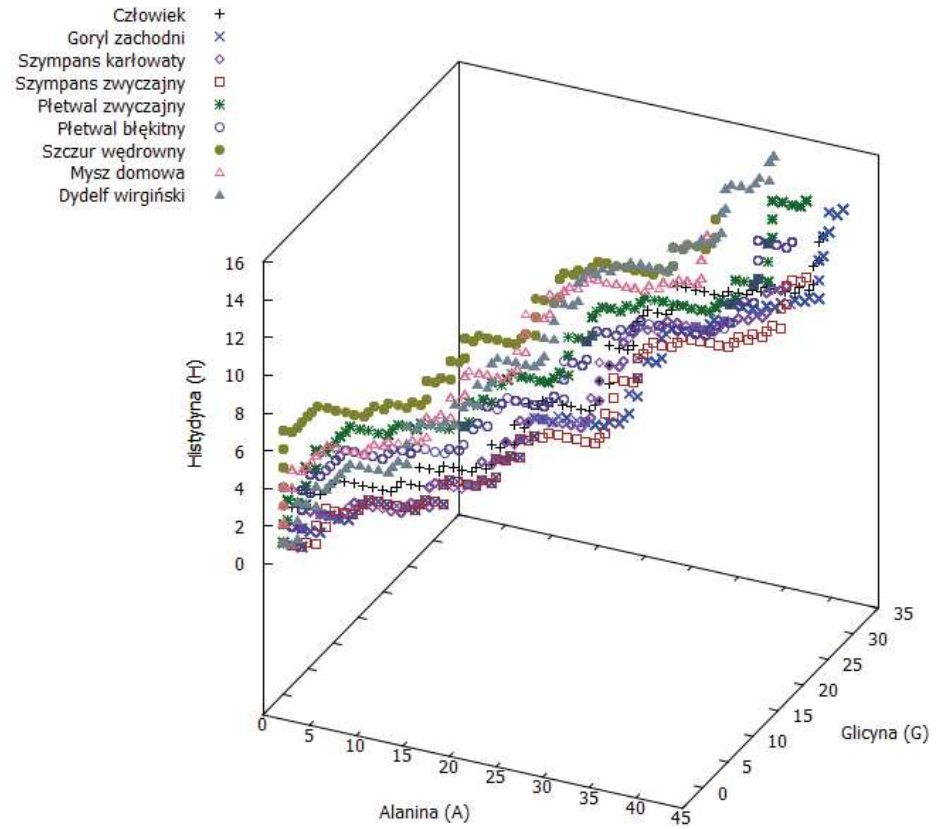
Rysunek 32. Graf 3D - ACD dla ND5



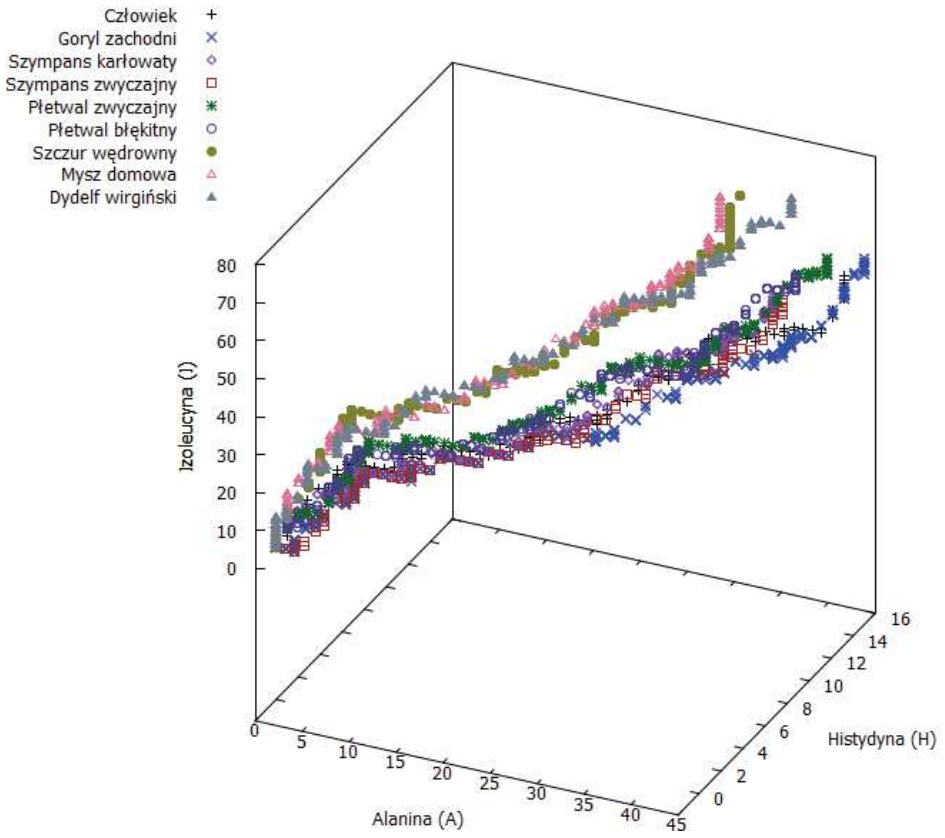
Rysunek 33. Graf 3D - AEF dla ND5



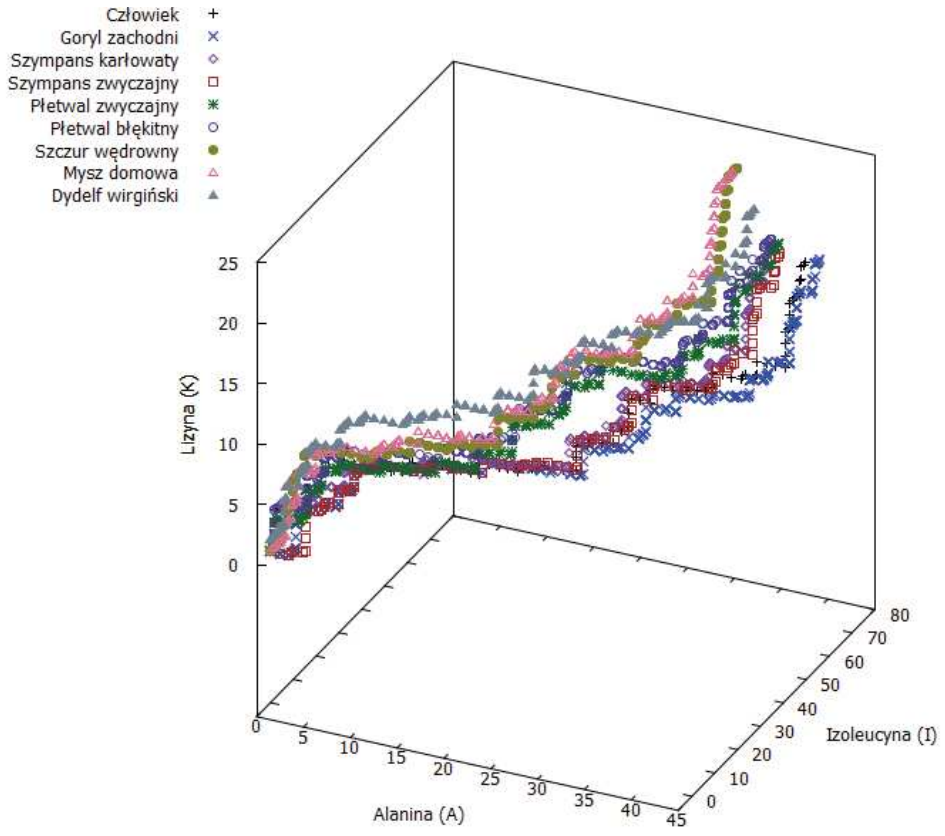
Rysunek 34. Graf 3D - AFG dla ND5



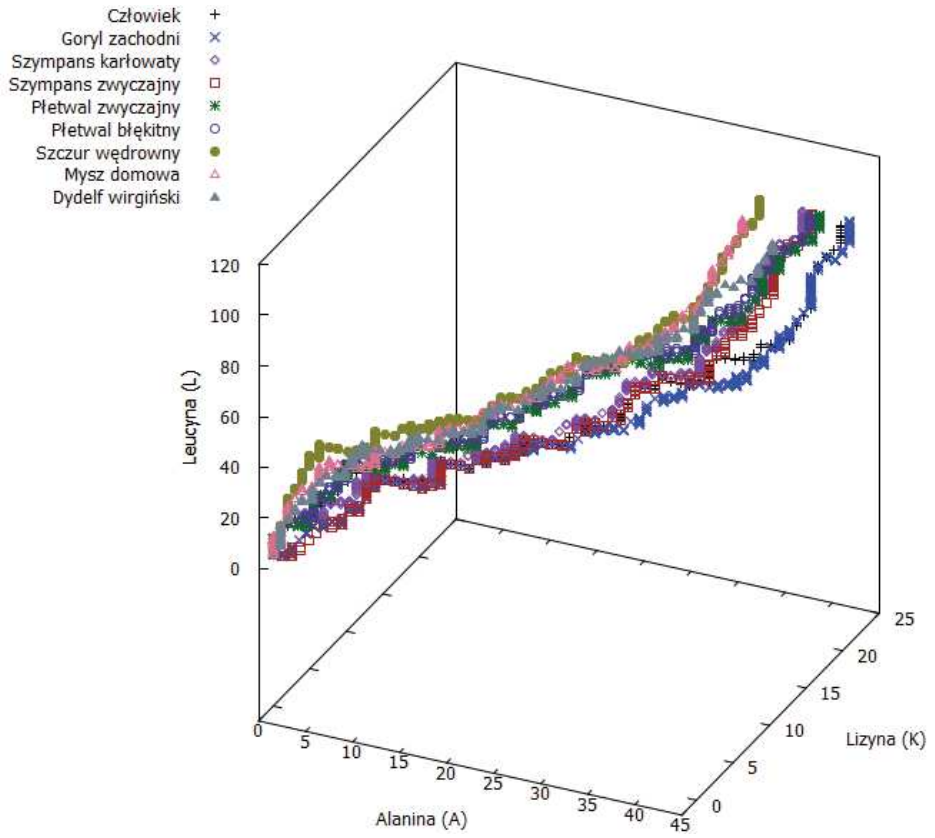
Rysunek 35. Graf 3D - AGH dla ND5



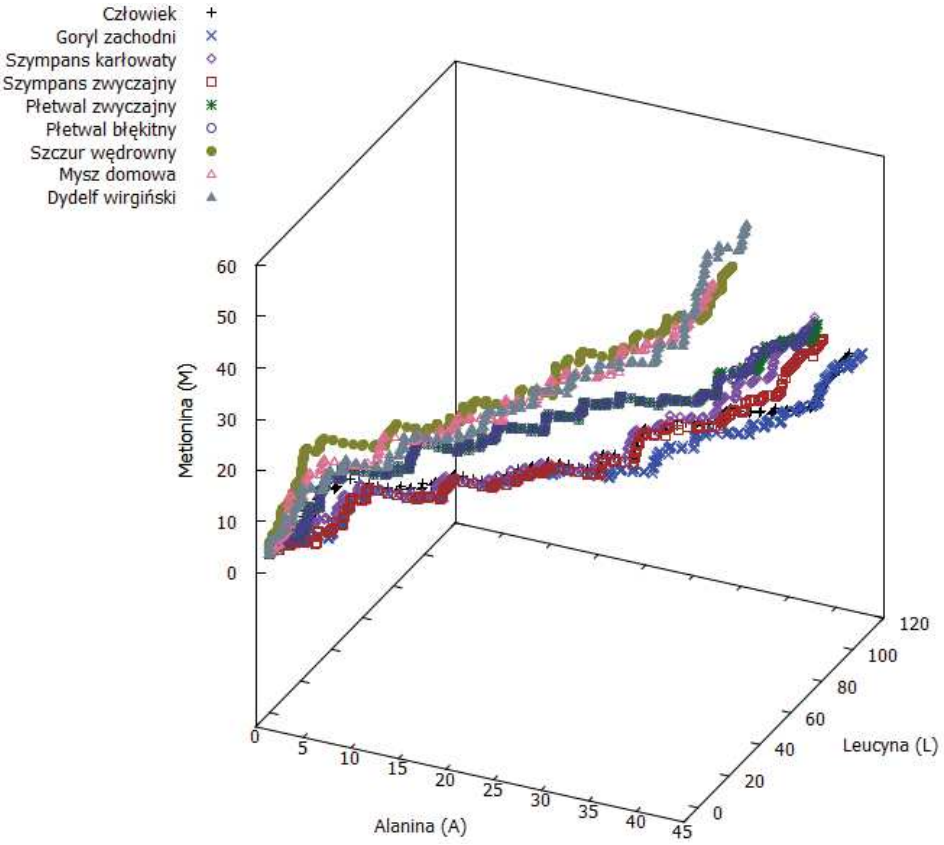
Rysunek 36. Graf 3D - AHI dla ND5



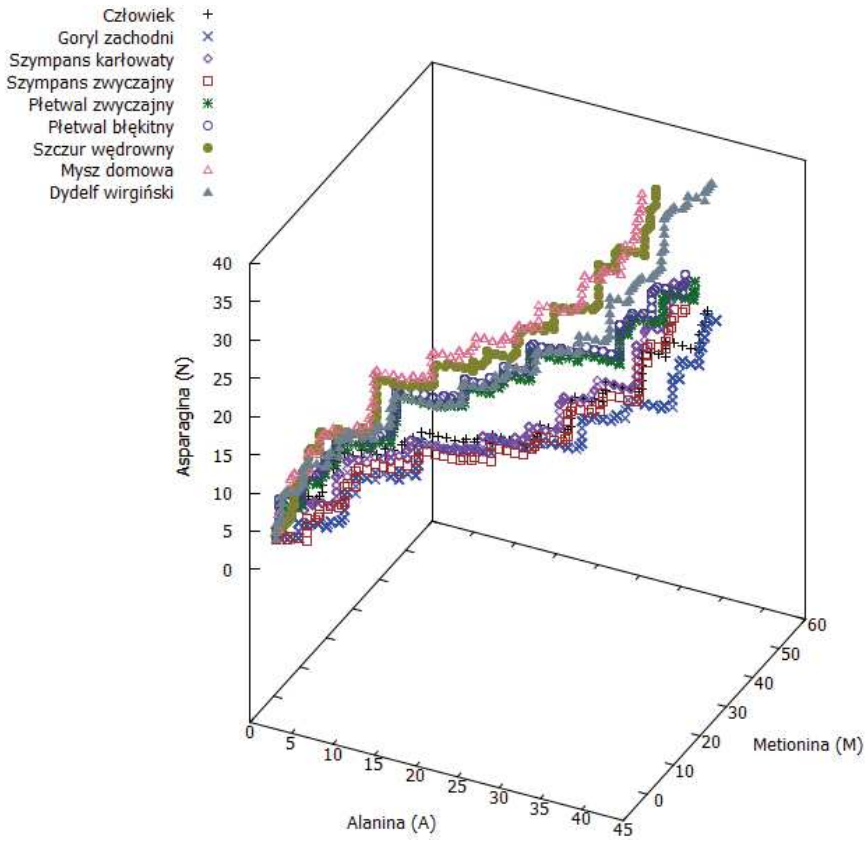
Rysunek 37. Graf 3D - AIK dla ND5



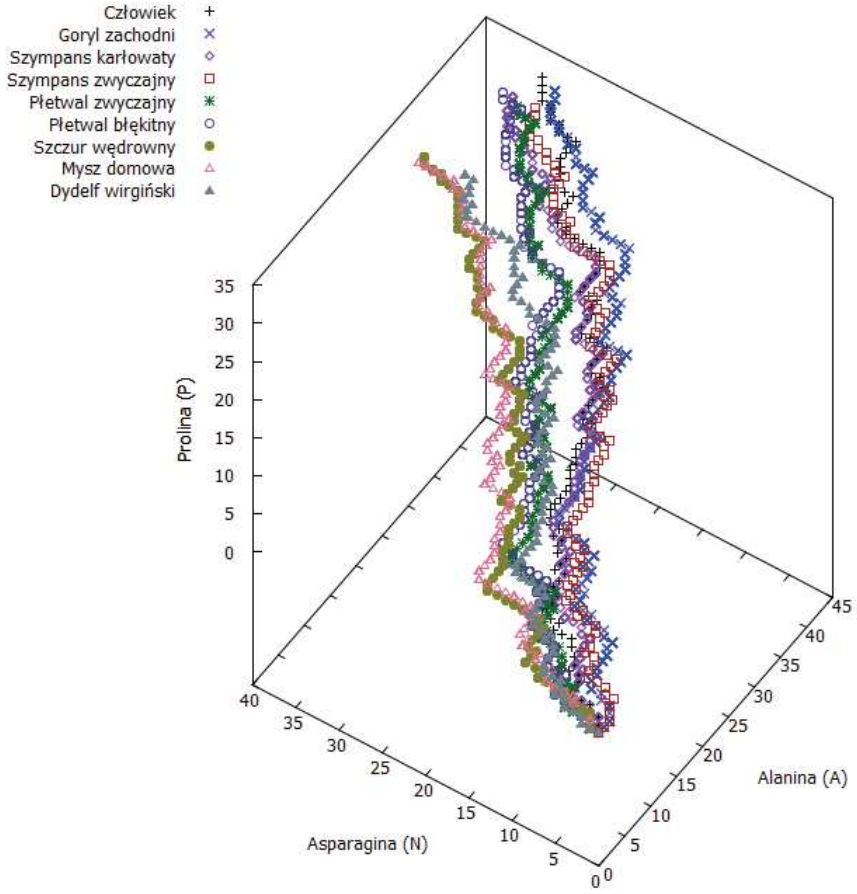
Rysunek 38. Graf 3D - AKL dla ND5



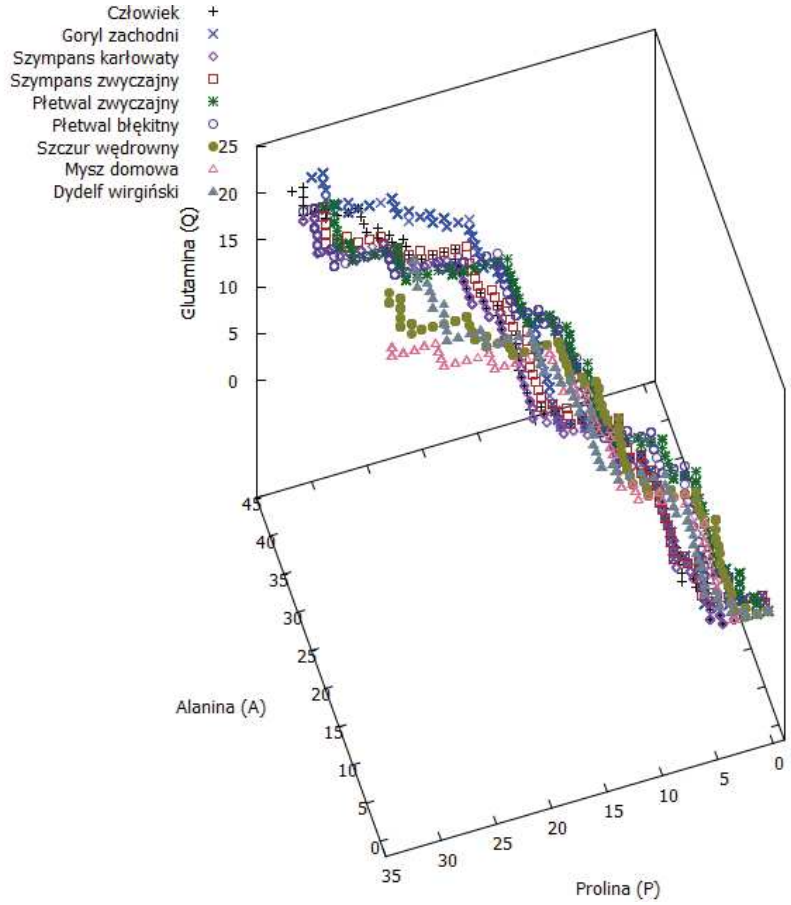
Rysunek 39. Graf 3D - ALM dla ND5



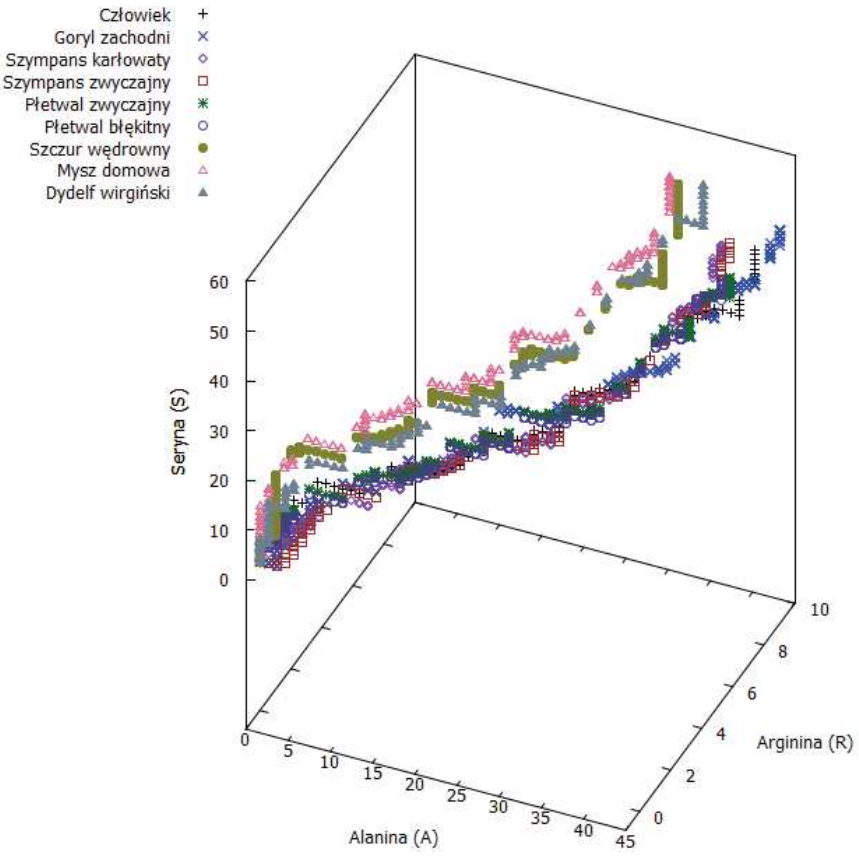
Rysunek 40. Graf 3D - AMN dla ND5



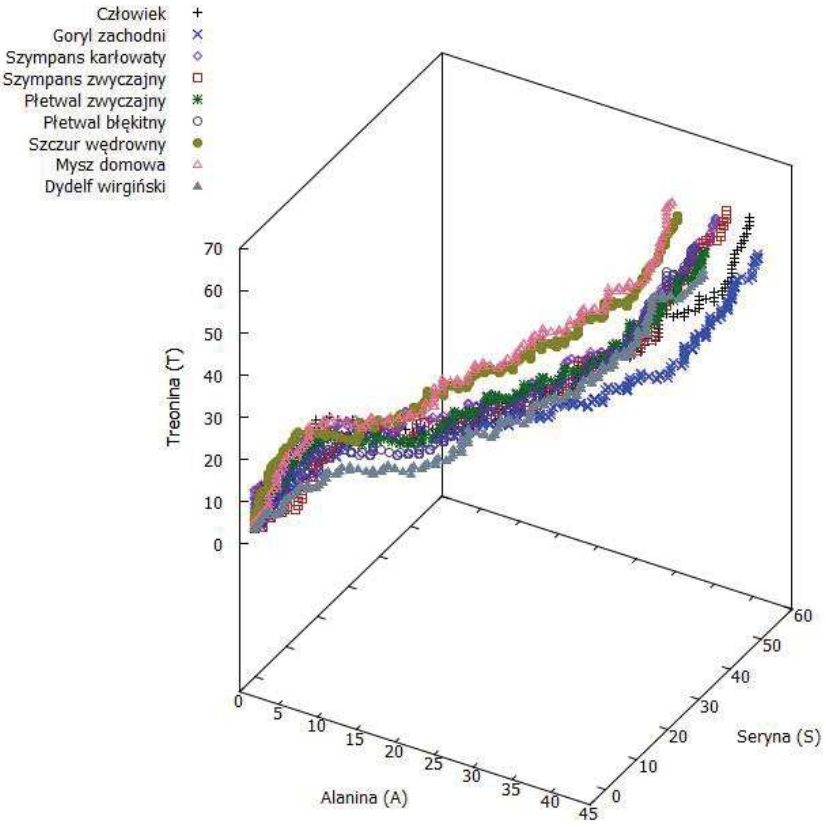
Rysunek 41. Graf 3D - ANP dla ND5



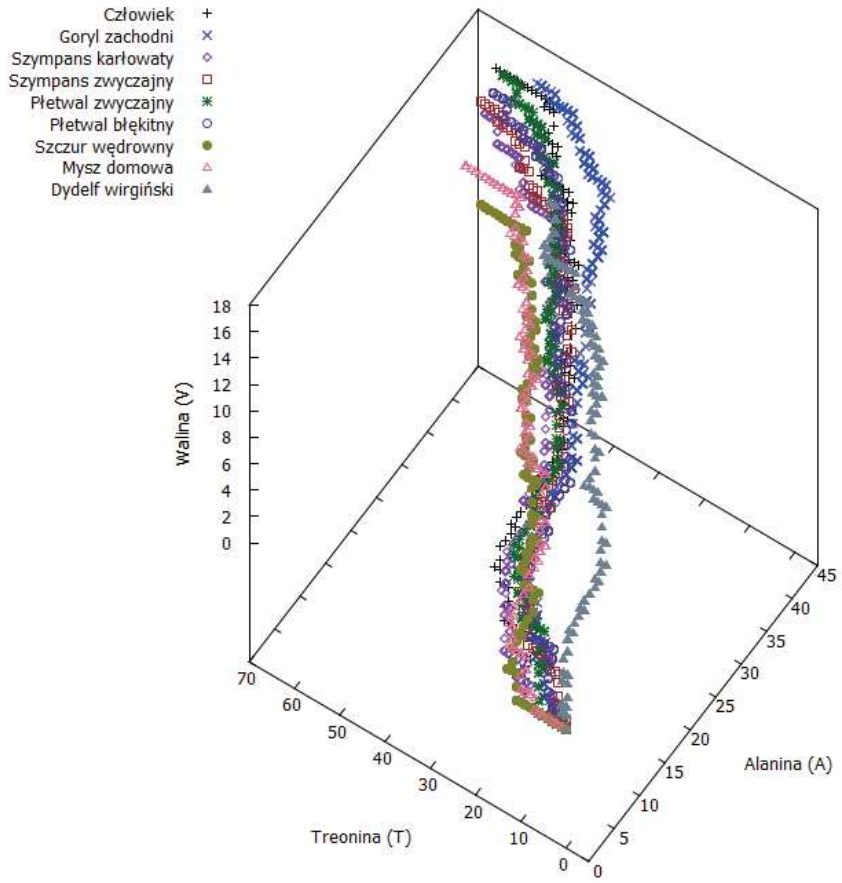
Rysunek 42. Graf 3D - APQ dla ND5



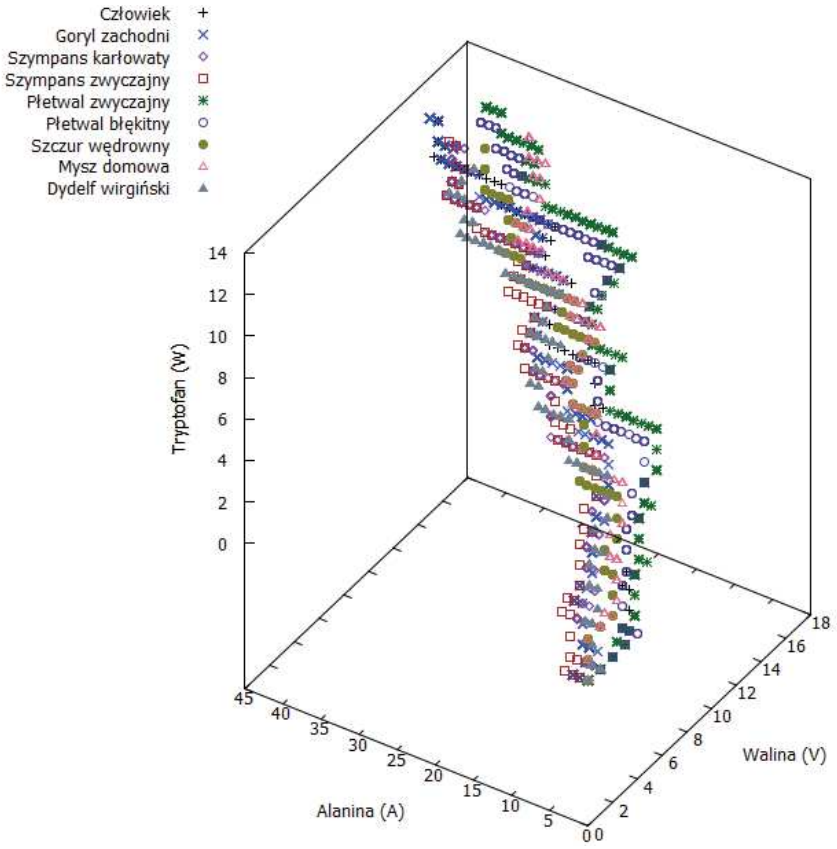
Rysunek 43. Graf 3D - ARS dla ND5



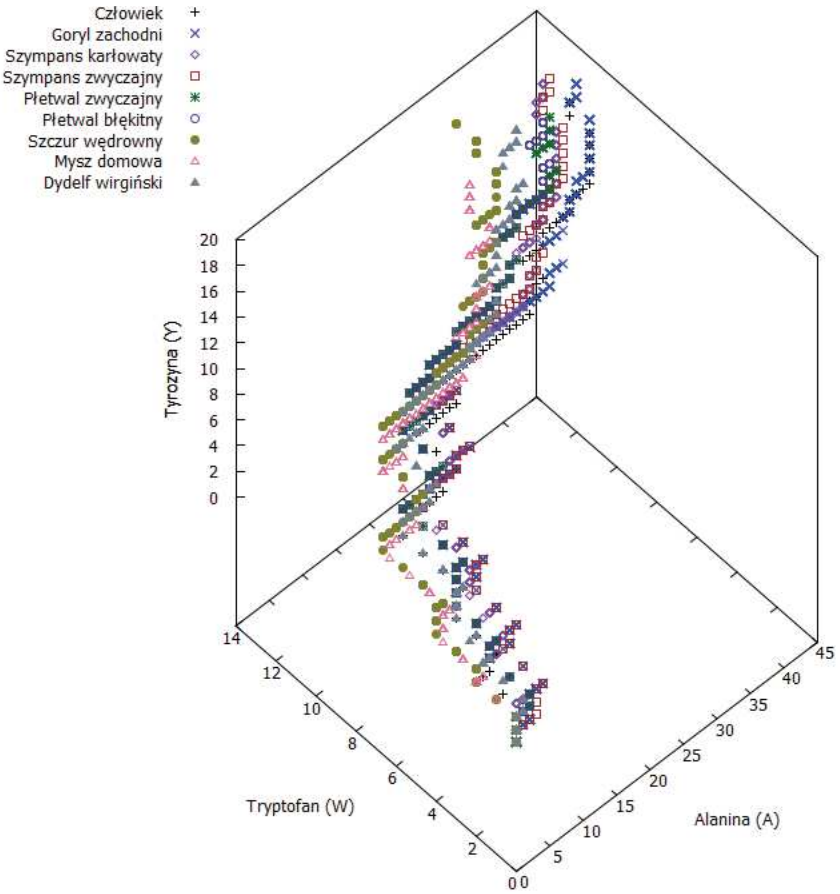
Rysunek 44. Graf 3D - AST dla ND5



Rysunek 45. Graf 3D - ATV dla ND5



Rysunek 46. Graf 3D - AVW dla ND5



Rysunek 47. Graf 3D - AWY dla ND5

3.4. Dane początkowe dla ND6

W celu przedstawienia metody 20 wymiarowej reprezentacji białek przeprowadzono dodatkowe obliczenia dla sekwencji dehydrogenazy NADH podjednostki 6 (ND6) dla ośmiu różnych gatunków. W *tabeli nr 9* podano numer dostępu do sekwencji aminokwasowej dla danego gatunku, która znajduje się w bazie PDB.

Tabela 9. Dane dotyczące użytych sekwencji ND6

Lp.	Gatunek	Numer dostępu	Dł. sekwencji
1	Człowiek (*Homo sapiens*)	AP_000650	174
2	Goryl zachodni (*Gorilla gorilla*)	NP_008223	174
3	Szympans zwyczajny (*Pan troglodytes*)	NP_008197	174
4	Foka pospolita (*Phoca vitulina*)	NP_006939	175
5	Szarytka morska (*Halichoerus grypus*)	NP_006939	175
6	Szczur wędrowny (*Rattus norvegicus*)	AP_004903	172
7	Mysz domowa (*Mus musculus*)	AP_904339	172
8	Kangur górski (*Macropus robustus*)	NP_007405	167

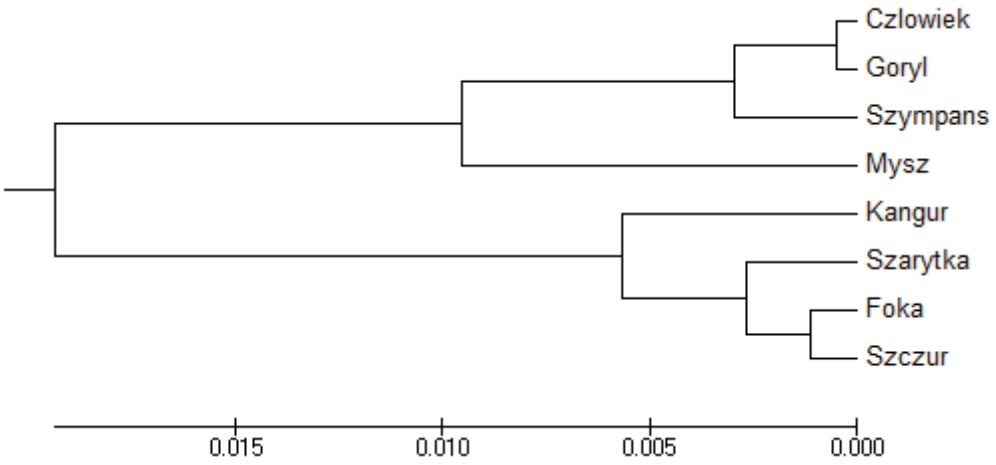
Za pomocą nowej metody wyznaczono macierz podobieństwa i jej reprezentację przy użyciu drzewa filogenetycznego (*rysunek nr 50*).

Tabela 10. Macierz podobieństwa dla ND6

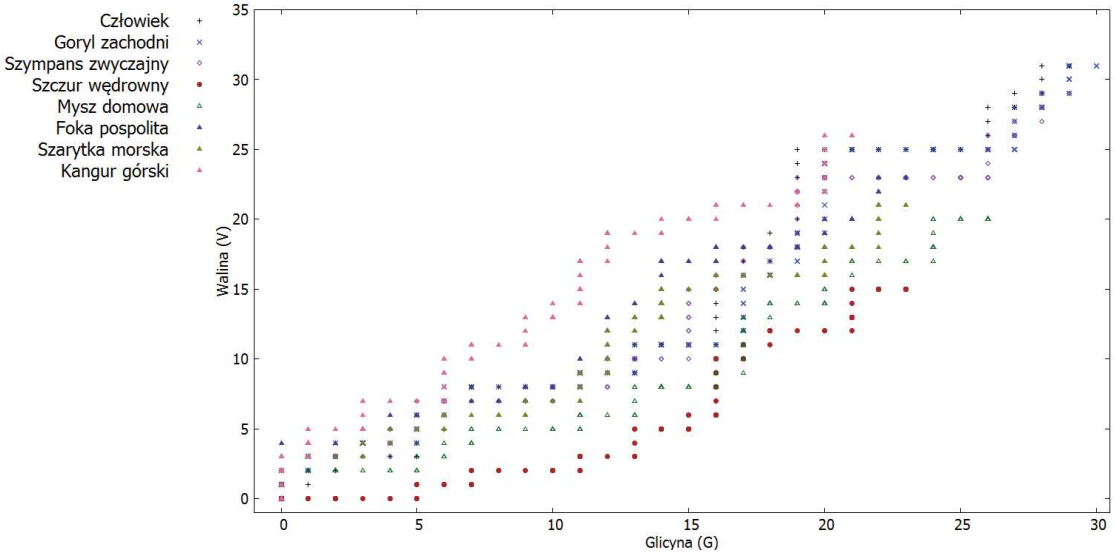
Gatunek	Człowiek	Goryl	Szympans	Foka	Szarytka	Szczur	Mysz	Kangur
Człowiek	0.00000	0.00095	0.00639	0.04249	0.04667	0.04026	0.02156	0.05447
Goryl		0.00000	0.00544	0.04154	0.04572	0.03931	0.02061	0.05352
Szympans			0.00000	0.03611	0.04030	0.03388	0.01517	0.04809
Foka				0.00000	0.00419	0.00224	0.02095	0.01201
Szarytka					0.00000	0.00643	0.02514	0.00781
Szczur						0.00000	0.01871	0.01424
Mysz							0.00000	0.03294
Kangur								0.00000

Na podstawie powyższej tabeli (*tabela nr 10*) możemy zauważyć duże podobieństwo pomiędzy sekwencjami człowieka, goryla zachodniego i szympansa zwyczajnego (małe wartości w tabeli). Natomiast sekwencja kangura górskiego jest najmniej podobna w porównaniu do innych sekwencji (duże wartości w tabeli) dla rozważanego białka. Inni autorzy również uzyskali podobne wyniki (36). Podobieństwo pomiędzy gatunkami zostało także przedstawione za pomocą wykresów 2D (*rysunki 49-63*) i 3D (*rysunki 64-69*). Na wszystkich rysunkach 2D sekwencja kangura górskiego odstaje od pozostałych sekwencji. W przypadku pozostałych sekwencji możemy zauważyć duże podobieństwo pomiędzy człowiekiem, gorylem zachodnim oraz szympansem zwyczajnym, kolejną

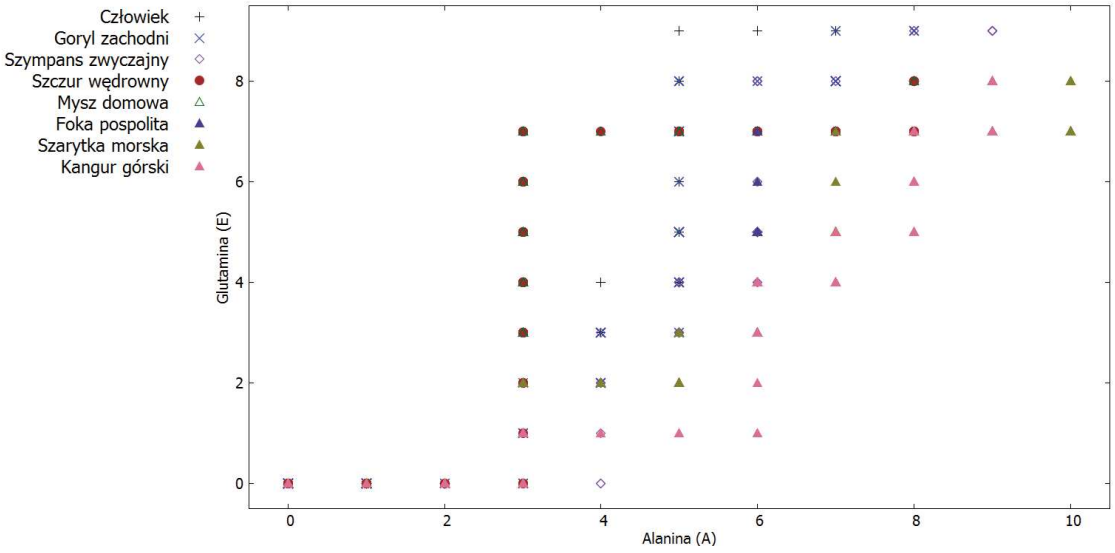
grupą jest szczur wędrowny i mysz domowa (z wyjątkiem rysunków nr 51 i 54) i ostatnią parą wykazującą duże podobieństwo jest foka pospolita i szarytka morska. W przypadku rysunków 3D jest podobnie, sekwencja kangura górskiego w każdym z przypadków jest odseparowana od pozostałych sekwencji. Z kolei pozostałe sekwencji tworzą klastry podobnych sekwencji, pierwszy klaster: człowiek, goryl zachodni i szympans zwyczajny, drugi klaster: szczur wędrowny i mysz domowa i trzeci klaster: foka pospolita i szarytka morska.



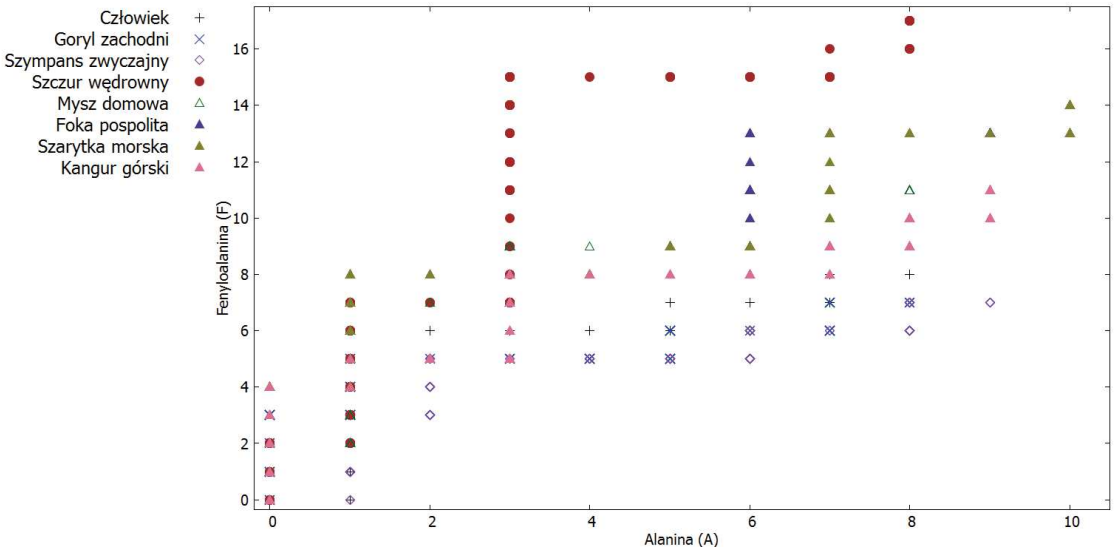
Rysunek 48. Drzewo filogenetyczne dla ND6



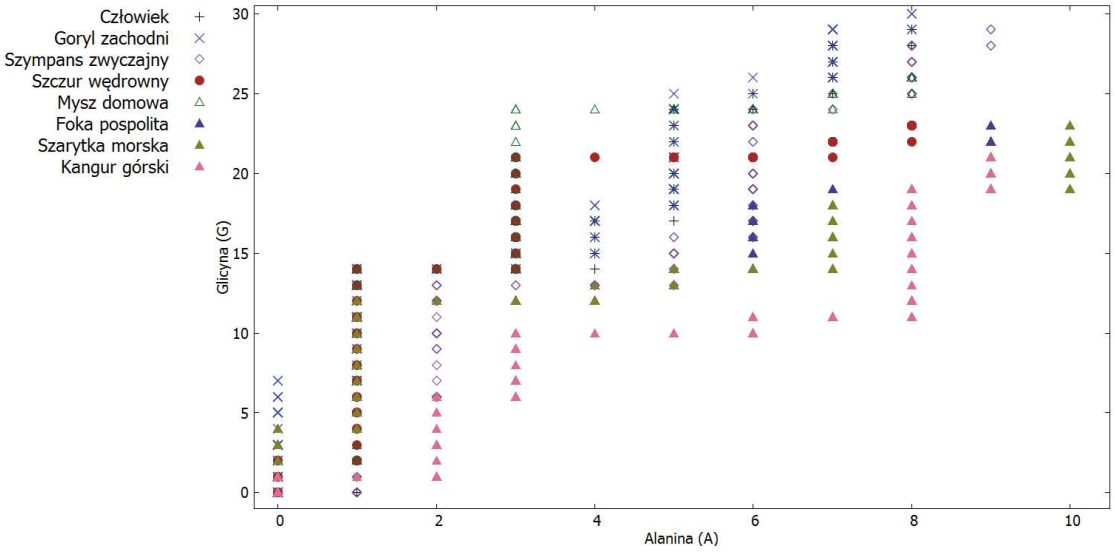
Rysunek 49. Graf 2D - GV dla ND6



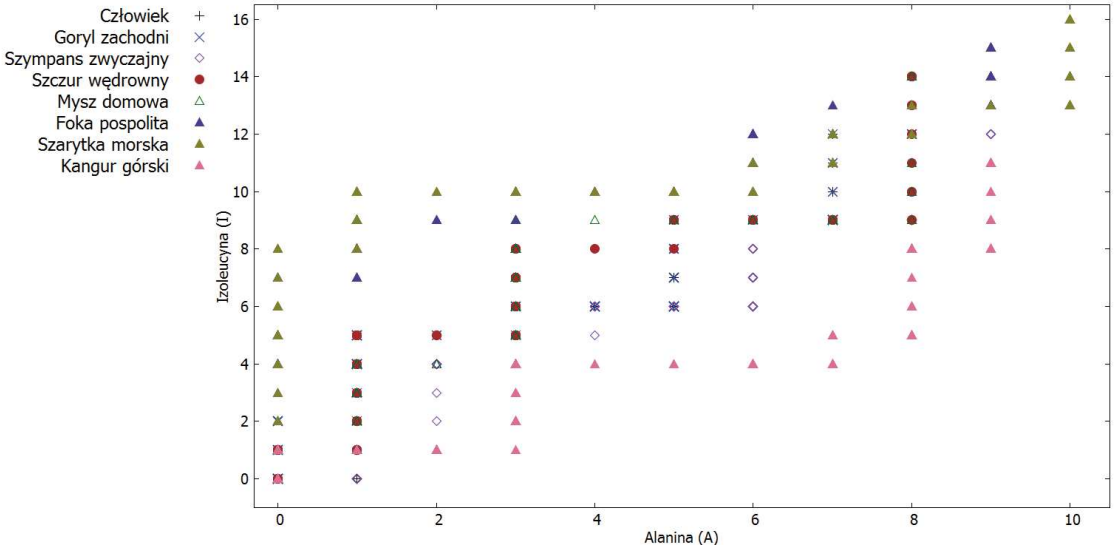
Rysunek 50. Graf 2D - AE dla ND6



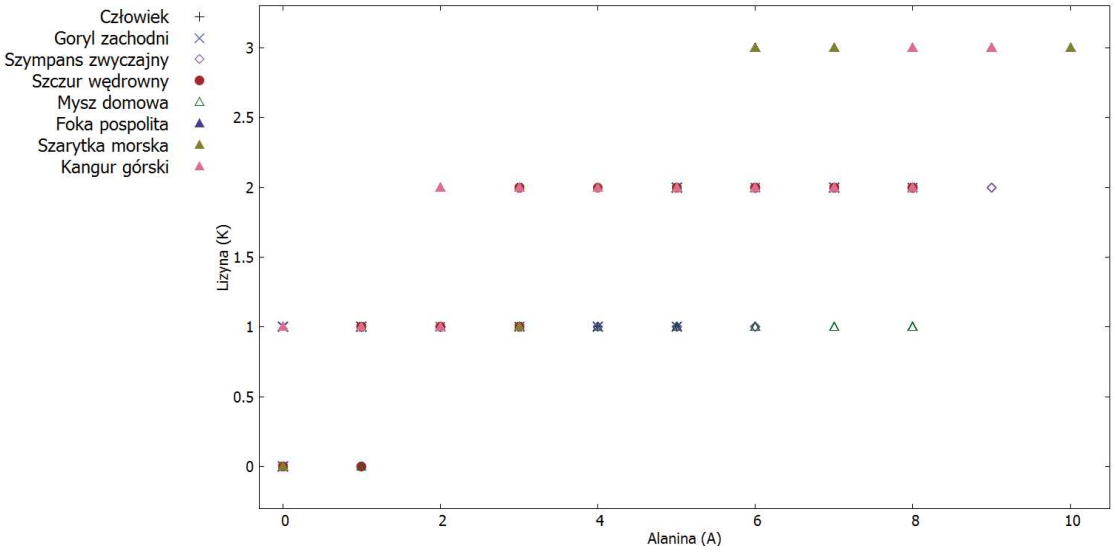
Rysunek 51. Graf 2D - AF dla ND6



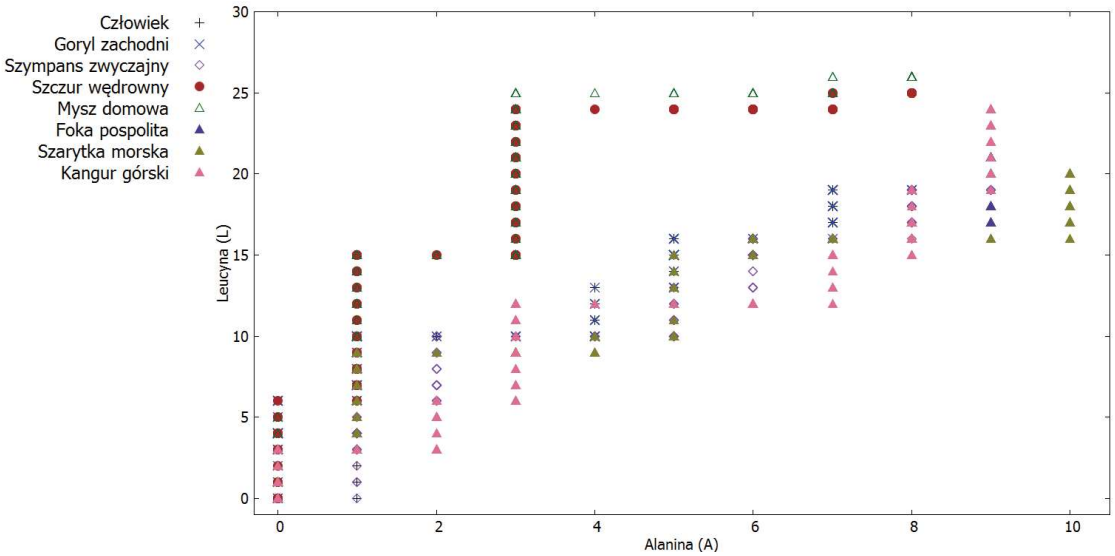
Rysunek 52. Graf 2D - AG dla ND6



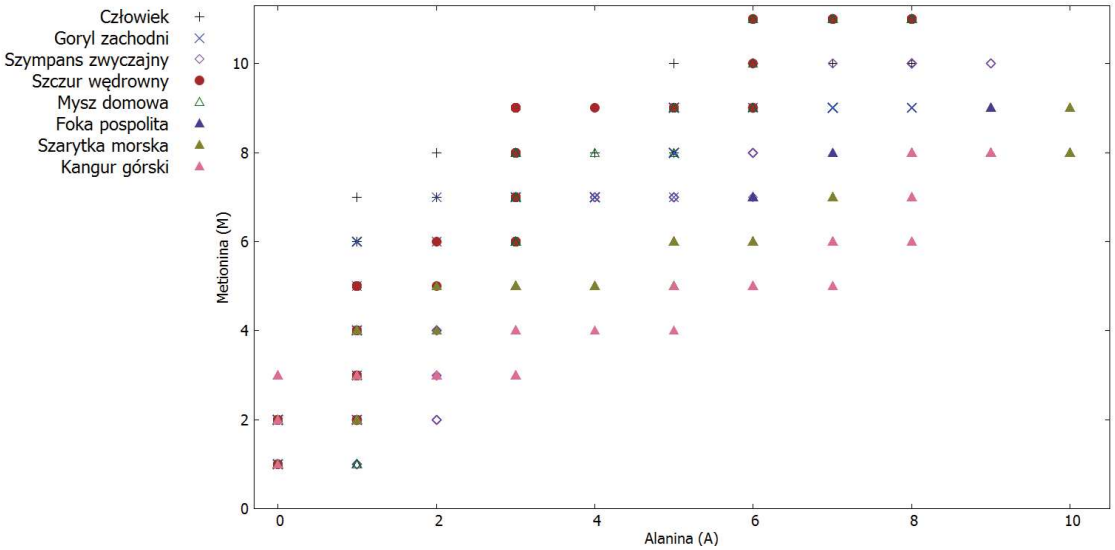
Rysunek 53. Graf 2D - AI dla ND6



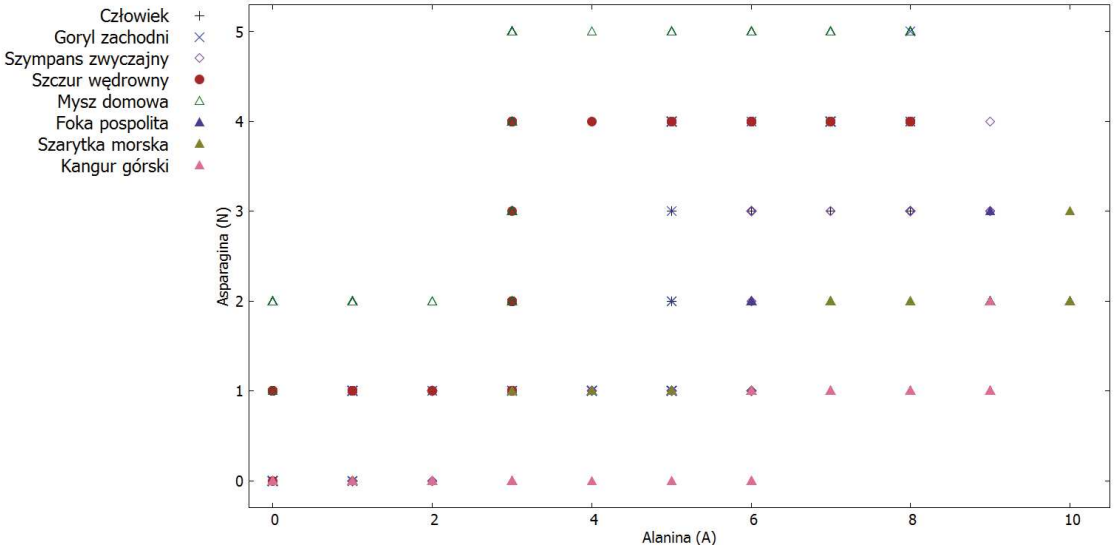
Rysunek 54. Graf 2D - AK dla ND6



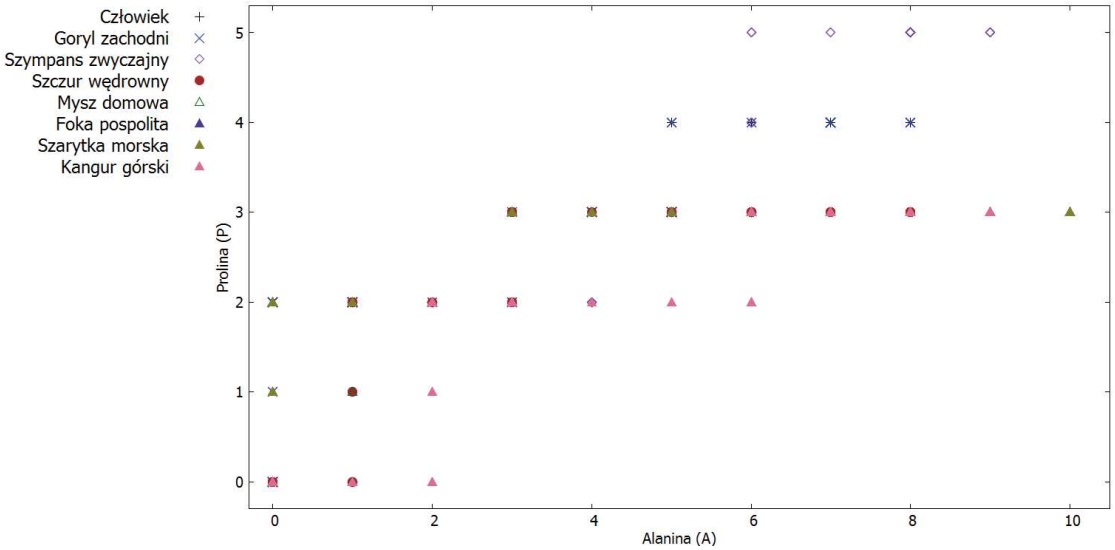
Rysunek 55. Graf 2D - AL dla ND6



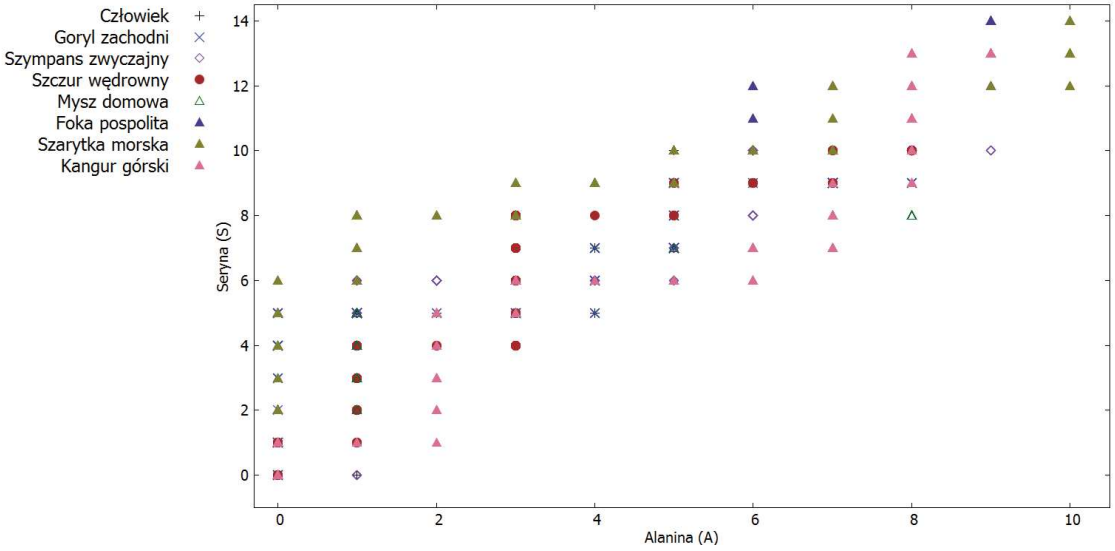
Rysunek 56. Graf 2D - AM dla ND6



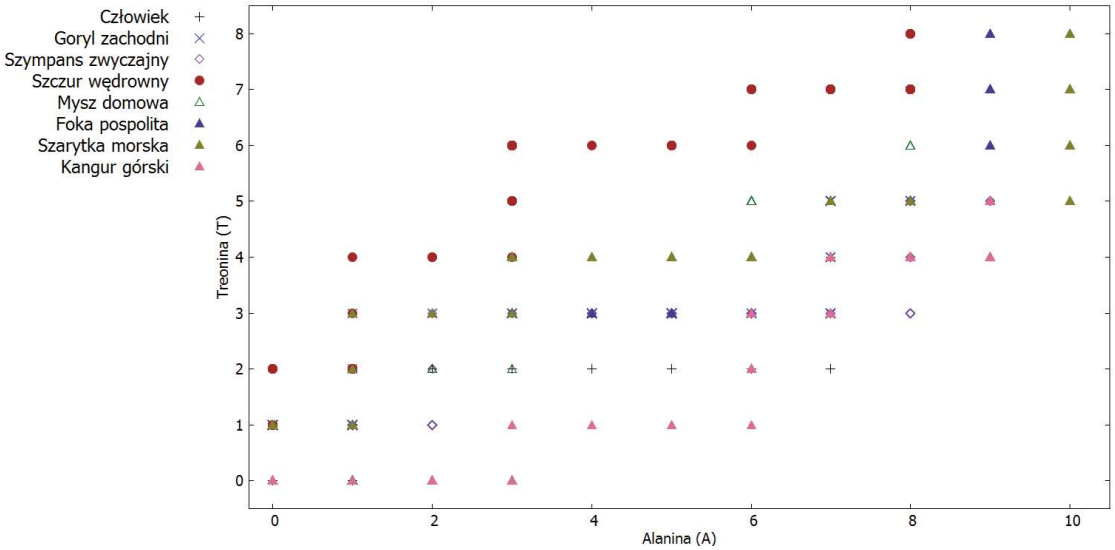
Rysunek 57. Graf 2D - AN dla ND6



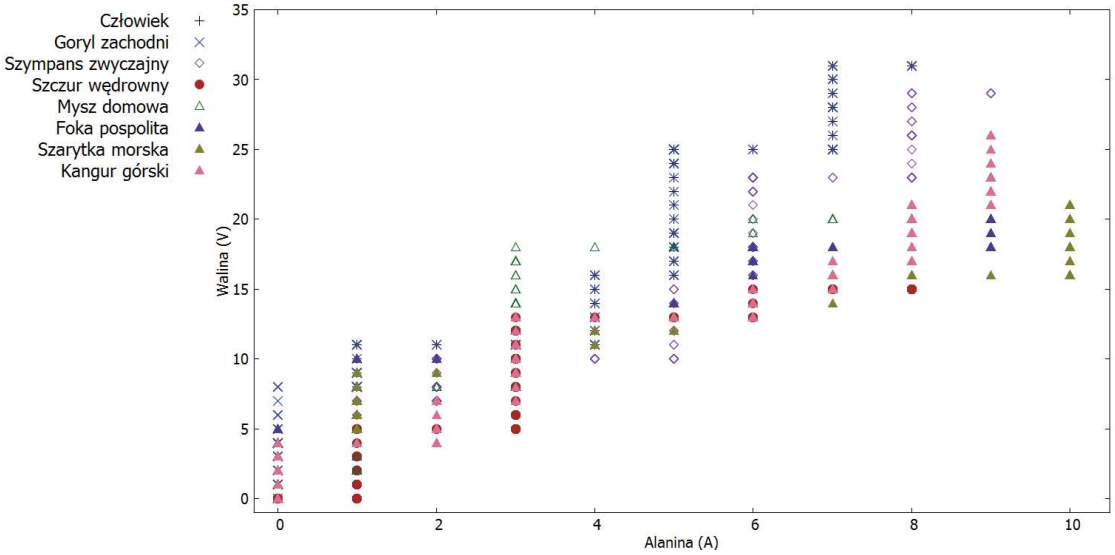
Rysunek 58. Graf 2D - AP dla ND6



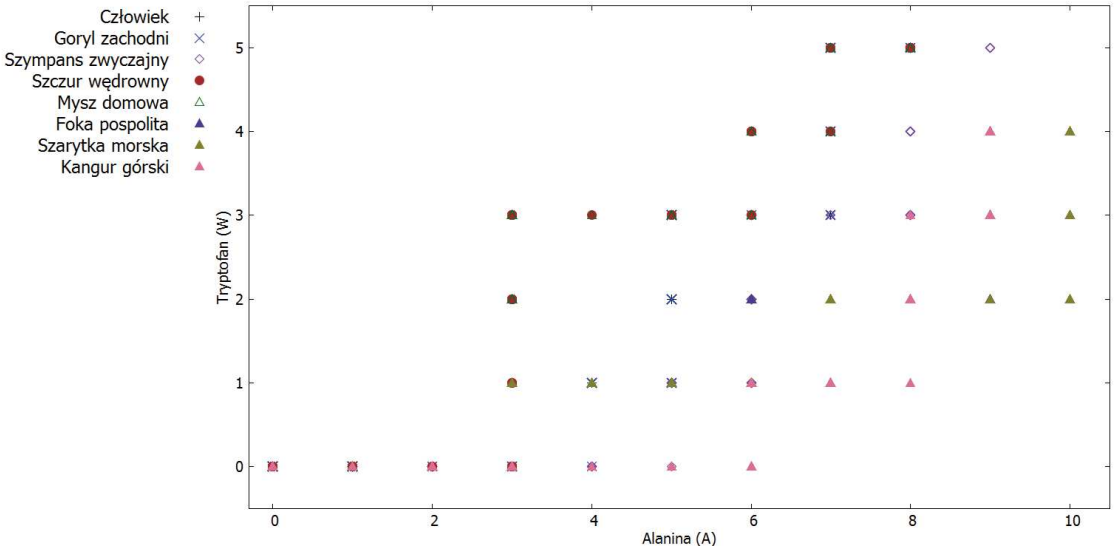
Rysunek 59. Graf 2D - AS dla ND6



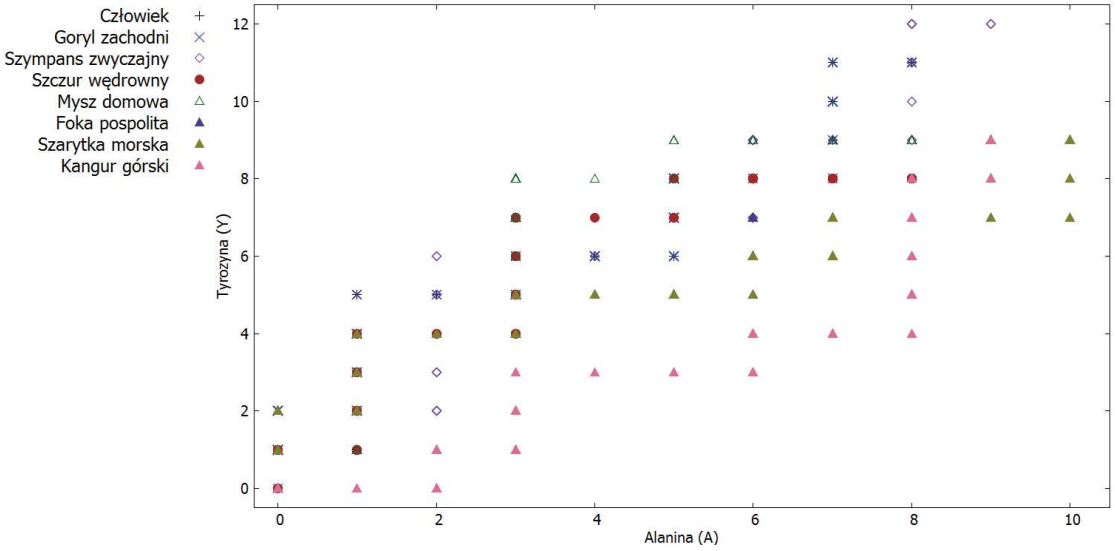
Rysunek 60. Graf 2D - AT dla ND6



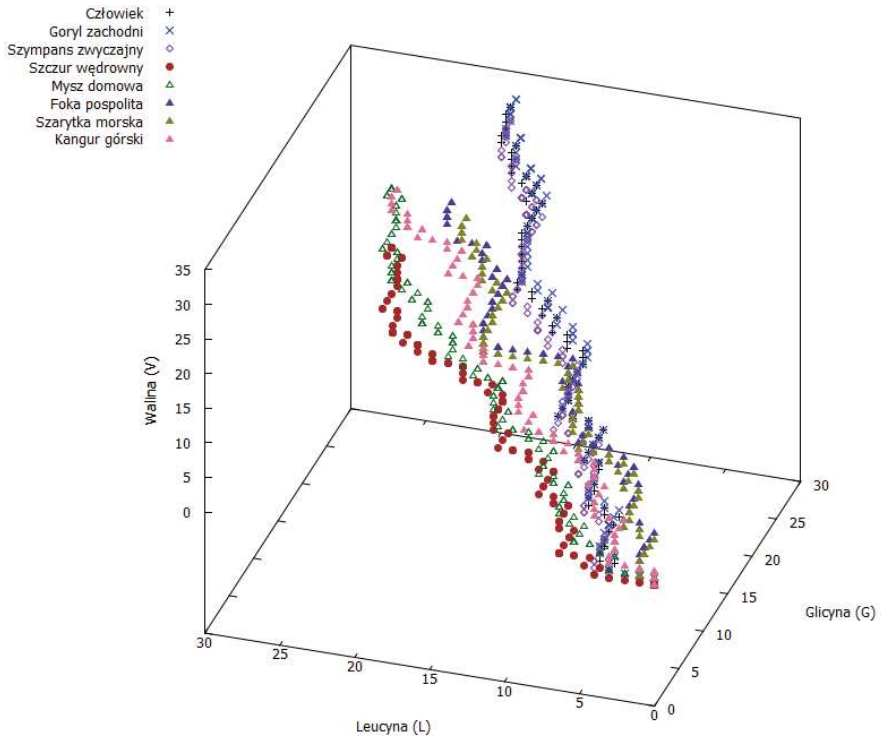
Rysunek 61. Graf 2D - AV dla ND6



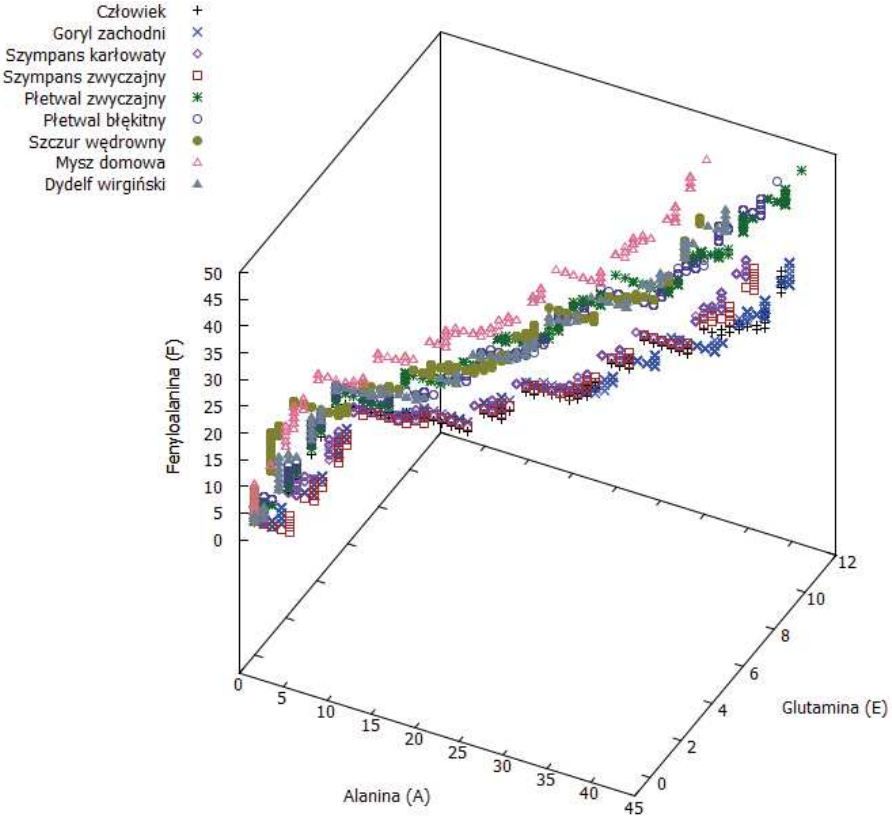
Rysunek 62. Graf 2D - AW dla ND6



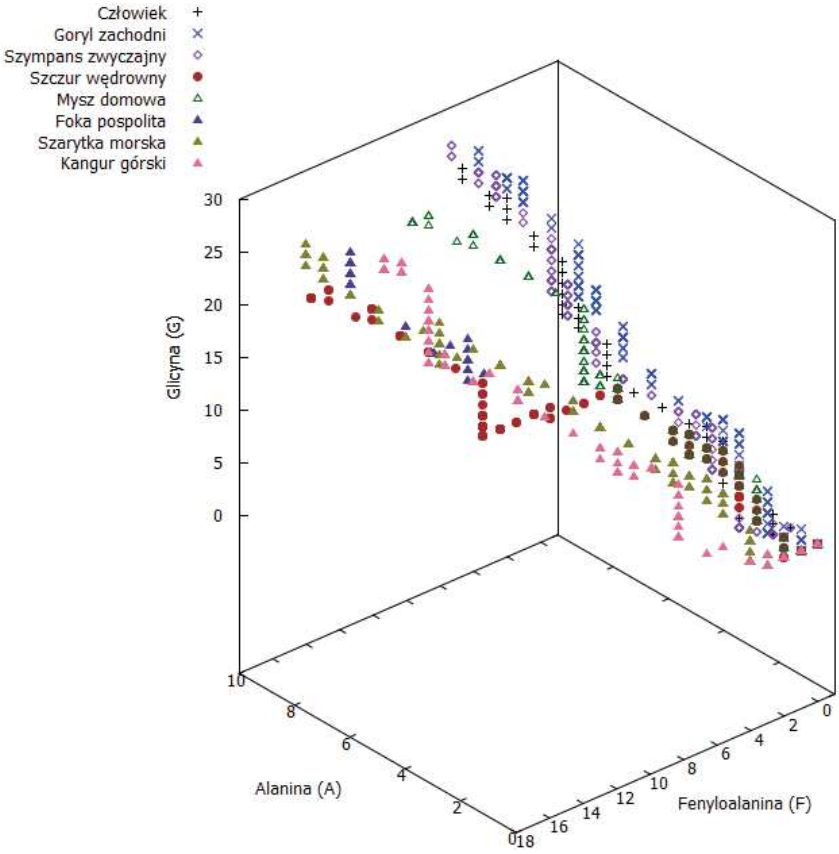
Rysunek 63. Graf 2D - AY dla ND6



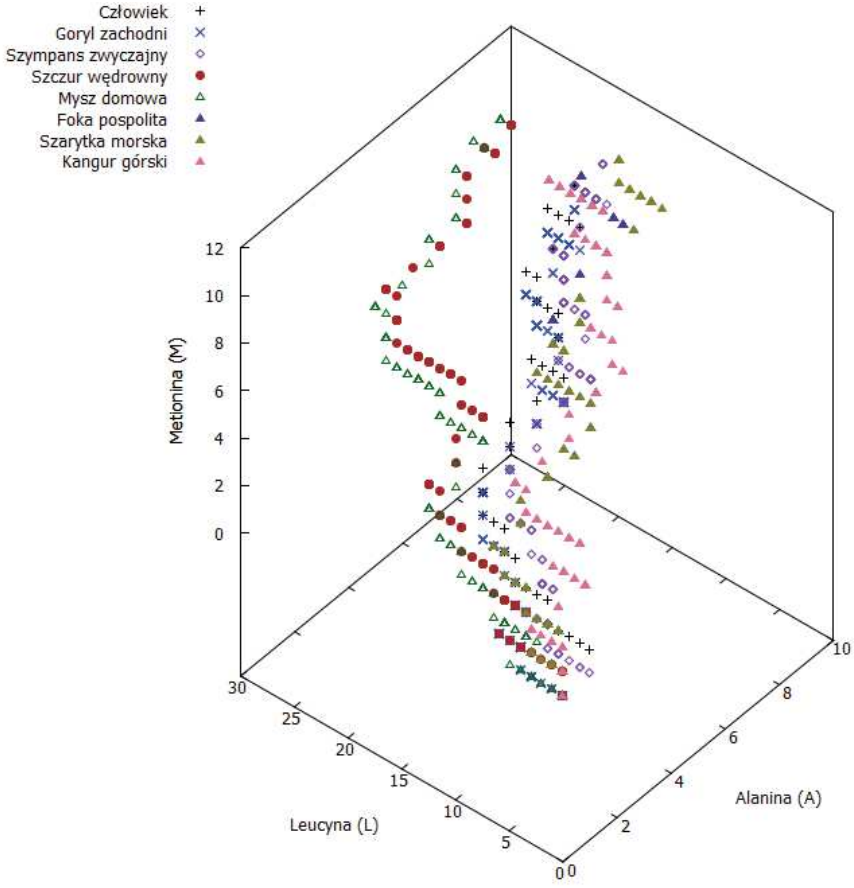
Rysunek 64. Graf 3D - GLV dla ND6



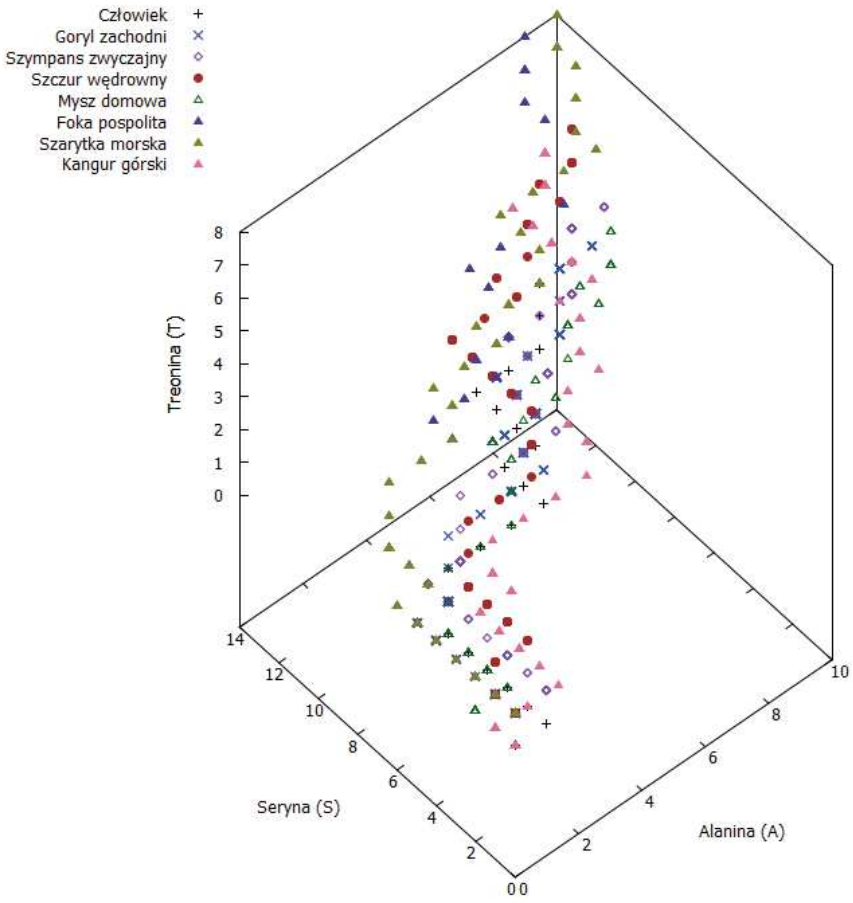
Rysunek 65. Graf 3D - AEF dla ND6



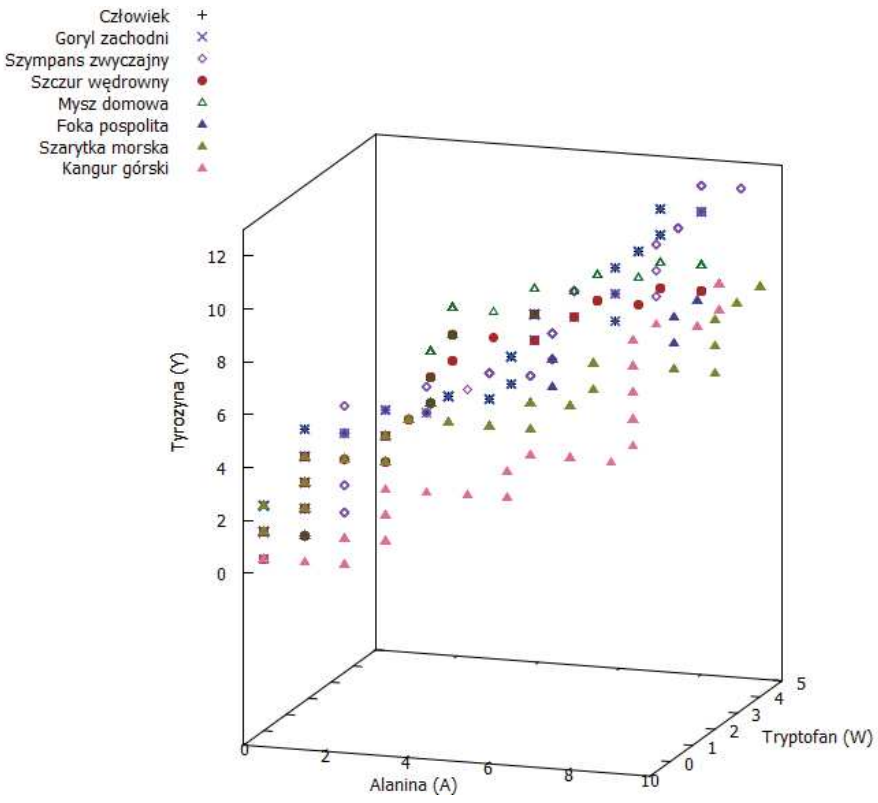
Rysunek 66. Graf 3D - AFG dla ND6



Rysunek 67. Graf 3D - ALM dla ND6



Rysunek 68. Graf 3D - AST dla ND6



Rysunek 69. Graf 3D - AWY dla ND6

3.5. Dane początkowe dla bakulowirusów

Metodę 20 wymiarowej reprezentacji białek wykorzystano również do analizy podobieństwa sekwencji dwunastu gatunków bakulowirusów (67,68). W *tabeli nr 11* podano numer dostępu do sekwencji aminokwasowej dla danego gatunku, która znajduje się w PDB.

Tabela 11. Dane dotyczące użytych sekwencji wirusowych

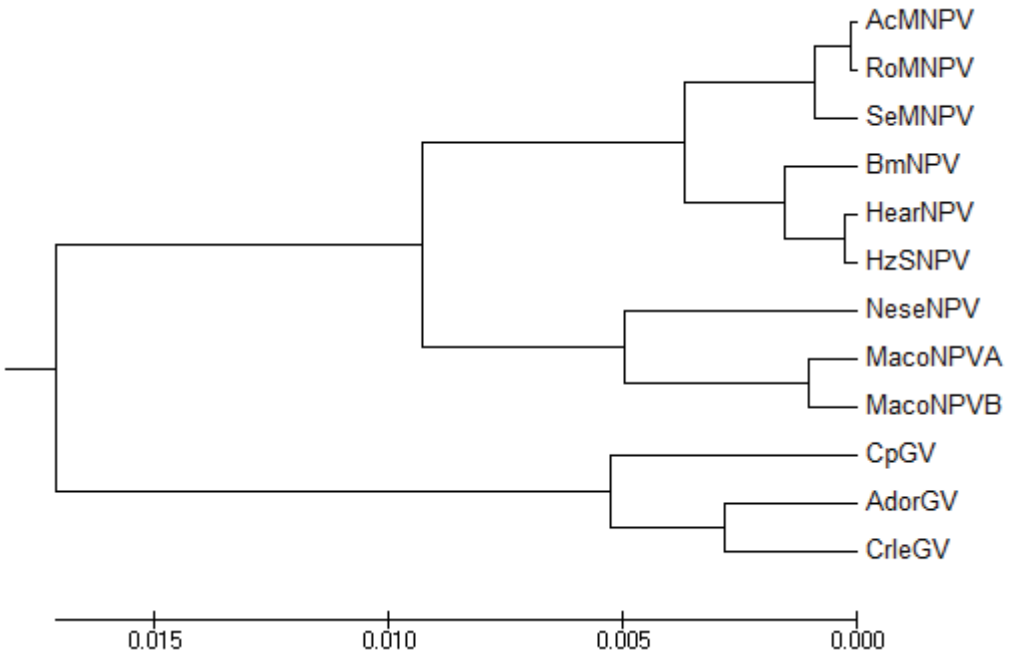
Lp.	Gatunek	Numer dostępu	Dł. Sekwencji
1	AcMNPV (*Autographa californica nucleopolyhedrovirus*)	AAA66725.1	1221
2	BmNPV (*Bombyx mori nucleopolyhedrovirus*)	AAC63764.1	1222
3	RoMNPV (*Rachiplusia ou MNPV*)	AAN28013.1	1221
4	HearNPV (*Helicoverpa armigera nucleopolyhedrovirus)*	AAK57882.1	1253
5	HzSNPV (*Helicoverpa zea single nucleopolyhedrovirus*)	AAL56093.1	1253
6	MacoNPVA (*Mamestra configurata nucleopolyhedrovirus A*)	AAM09201.1	1212
7	MacoNPVB (*Mamestra configurata nucleopolyhedrovirus B*)	AAM95079.1	1209
8	SeMNPV (*Spodoptera exigua multiple nucleopolyhedrovirus*)	AAB96630.1	1222
9	AdorGV (*Adoxophyes orana granulovirus*)	AAP85713.1	1138
10	CpGV (*Cydia pomonella granulovirus*)	AAK70750.1	1131
11	CrleGV (*Cryptophlebia leucotreta granulovirus*)	AAQ21676.1	1128
12	NeseNPV (*Neodiprion sertifer nucleopolyhedrovirus*)	AAQ96438.1	1143

Za pomocą nowej metody wyznaczono macierz podobieństwa i jej reprezentację przy użyciu drzewa filogenetycznego (*rysunek nr 70*).

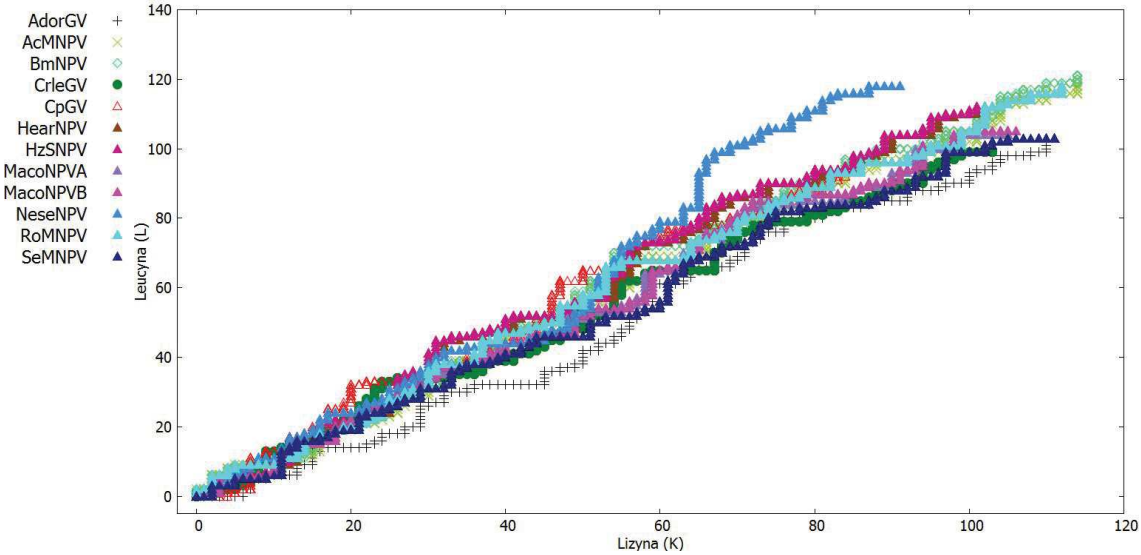
Tabela 12. Macierz podobieństwa dla bakulowirusów (wartości pomożone przez 1000)

Gatunek	AcMNPV	BmNPV	RoMNPV	HearNPV	HzSNPV	MacoNPVA	MacoNPVB	SeMNPV	AdorGV	CpGV	CrleGV	NeseNPV
AcMNPV	0	4.55	0.24	7.38	7.86	11.31	13.40	1.93	31.15	44.52	36.82	22.26
BmNPV		0	4.79	2.83	3.31	15.86	17.94	6.49	35.70	49.06	41.36	26.81
RoMNPV			0	7.62	8.11	11.07	13.15	1.69	30.91	44.27	36.58	22.02
HearNPV				0	0.48	18.69	20.78	9.32	38.53	51.88	44.19	29.64
HzSNPV					0	19.17	21.26	9.80	39.01	52.36	44.67	30.12
MacoNPVA						0	2.09	9.38	19.85	33.22	25.52	10.96
MacoNPVB							0	11.46	17.77	31.14	23.44	8.87
SeMNPV								0	29.22	42.58	34.89	20.33
AdorGV									0	13.38	5.67	8.90
CpGV										0	0.771	22.27
CrleGV											0	14.57
NeseNPV												0

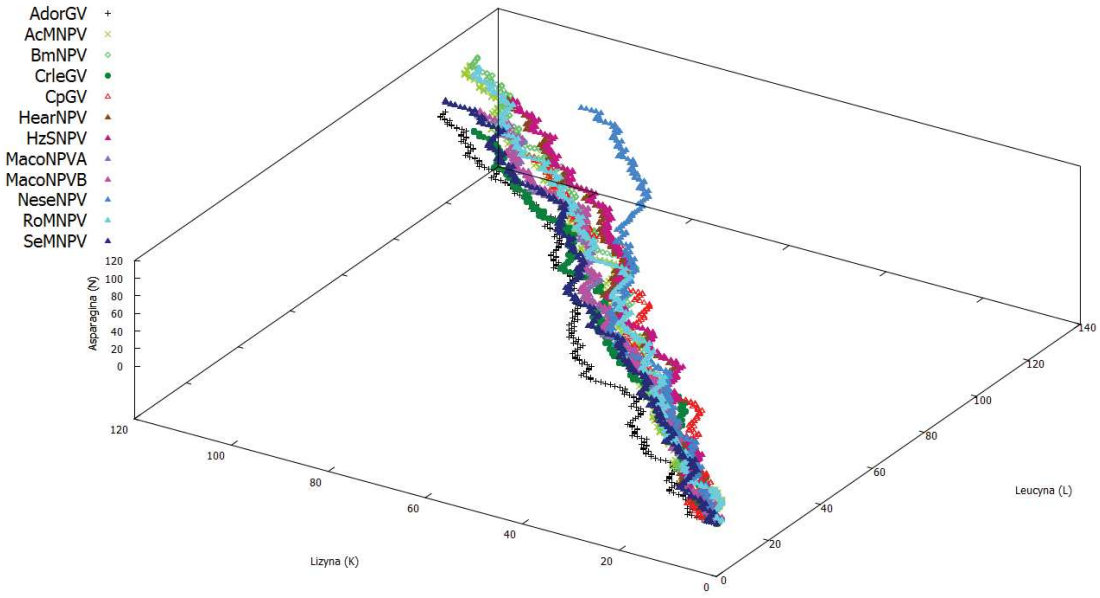
Na podstawie powyższej tabeli (*tabela nr 12*) możemy zauważyć duże podobieństwo pomiędzy pewnymi grupami bakulowirusów, a mianowicie (AcMNPV, RoMNPV, SeMNPV), (BmNPV, HearNPV, HzSNPV), (NeseNPV, MacpNPVA, MacoNPVB) oraz (CpGV, AdorGV, CrleGV). Największe różnice pomiędzy bakulowirusami możemy zauważyć dla sekwencji NeseNPV (duże wartości w tabeli) dla rozważanego białka. Inni autorzy również uzyskali podobne wyniki (32). Podobieństwo pomiędzy gatunkami zostało także przedstawione za pomocą wykresów 2D (*rysunek 71*) i 3D (*rysunek 72*).



Rysunek 70. Drzewo filogenetyczne dla bakulowirusów



Rysunek 71. Graf 2D - KL dla bakulowirusów



Rysunek 72. Graf 3D - KLN dla bakulowirusów

3.6. Dane początkowe dla ND5 – 22 różne gatunki

W celu porównania metody 20 wymiarowej reprezentacji białek z metodą Dynamic Time Warping (69) przeprowadzono analizę podobieństwa sekwencji dehydrogenazy NADH podjednostki 5 (ND5) dla dwudziestu dwóch różnych gatunków. Dane zostały pobrane z PDB.

W *tabeli nr 13* podano numer dostępu do sekwencji aminokwasowej dla danego gatunku, która znajduje się w bazie PDB. Każdy gatunek ma przypisaną liczbę porządkową, która później charakteryzuje dany gatunek w macierzy podobieństwa. Macierz podobieństwa została przedstawiona w innej formie ze względu na duży zestaw danych.

Wielkości opisujące miarę podobieństwa w macierzy zostały zaokrąglone do 3 miejsca po przecinku (*tabela nr 14*).

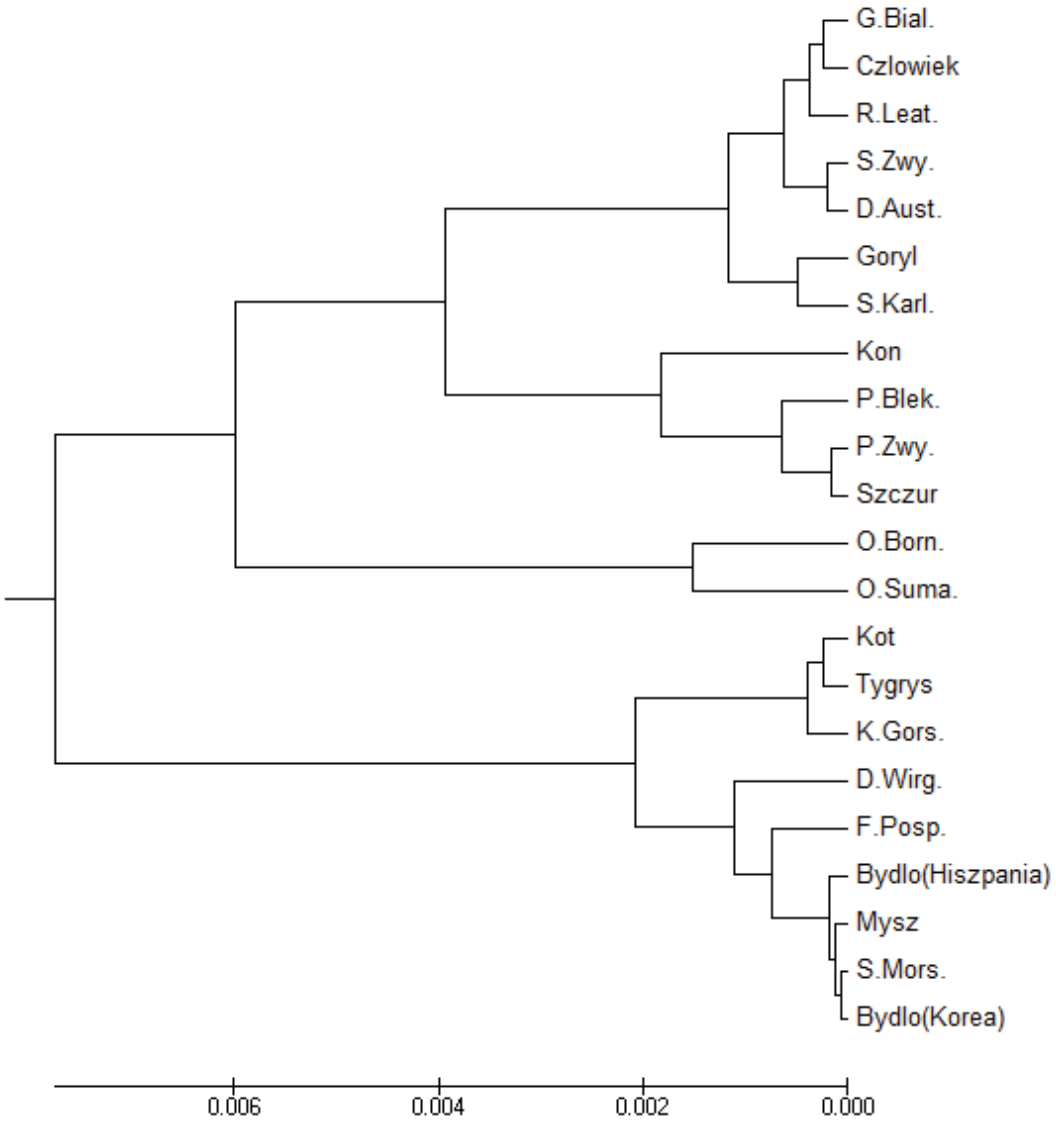
Tabela 13. Dane dotyczące użytych sekwencji ND5

Lp.	Gatunek	Numer dostępu	Dł. sekwencji
1	Płetwal błękitny (*Balaenoptera musculus*)	NP_007066	606
2	Orangutan borneański (*Pongo pygmaeus*)	NP_008235	603
3	Kot domowy (*Felis catus*)	NP_008261	606
4	Szympans zwyczajny (*Pan troglodytes*)	NP_008196	603
5	Płetwal zwyczajny (*Balaenoptera physalus*)	NP_006899	606
6	Gibon białoręki (*Hylobates lar*)	NP_007832	603
7	Goryl zachodni (*Gorilla gorilla*)	NP_008222	603
8	Szarytka morska (*Halichoerus grypus*)	NP_007079	609
9	Foka pospolita (*Phoca vitulina*)	NP_006938	609
10	Człowiek (*Homo sapiens*)	AP_000649	603
11	Koń domowy (*Equus caballus*)	ADQ55101	604
12	Mysz domowa (*Mus musculus*)	NP_904338	607
13	Dydelf wirginijski (*Didelphis virginiana*)	NP_007105	602
14	Szympans Karłowaty (*Pan paniscus*)	NP_008209	603
15	Dziobak australijski (*Ornithorhynchus anatinus*)	NP_008053	604
16	Szczur wędrowny (*Rattus norvegicus*)	AP_004902	610
17	Rhino leatherjacket (*Pseudalutarius nasicornis*)	YP_002520019	611
18	Orangutan sumatrzański (*Pongo abelii*)	NP_007845	601
19	Kangur górski (*Macropus robustus*)	NP_007404	602
20	Tygrys syberyjski (*Panthera tigris altaica*)	ADK73290	606
21	Bydło domowe (*Bos taurus*) (Korea)	YP_209215	606
22	Bydło domowe (*Bos taurus*) (Hiszpania)	AKK32014	606

Tabela 14. Macierz podobieństwa ND5

S(i,j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.000	0.017	0.014	0.008	0.001	0.006	0.004	0.009	0.008	0.007	0.004	0.009	0.011	0.005	0.007	0.001	0.006	0.014	0.013	0.013	0.009	0.009
2		0.000	0.030	0.009	0.018	0.010	0.013	0.026	0.024	0.010	0.021	0.025	0.028	0.012	0.009	0.018	0.011	0.003	0.029	0.030	0.026	0.026
3			0.000	0.021	0.012	0.020	0.018	0.005	0.006	0.020	0.009	0.005	0.003	0.019	0.021	0.013	0.019	0.027	0.001	0.000	0.005	0.004
4				0.000	0.009	0.001	0.004	0.017	0.015	0.001	0.012	0.016	0.019	0.003	0.000	0.009	0.002	0.006	0.020	0.021	0.017	0.017
5					0.000	0.008	0.005	0.007	0.006	0.008	0.003	0.007	0.009	0.006	0.009	0.000	0.007	0.015	0.011	0.012	0.008	0.008
6						0.000	0.002	0.015	0.014	0.000	0.011	0.015	0.017	0.001	0.001	0.007	0.001	0.007	0.019	0.019	0.015	0.016
7							0.000	0.013	0.011	0.003	0.008	0.013	0.015	0.001	0.003	0.005	0.002	0.010	0.017	0.017	0.013	0.013
8								0.000	0.001	0.016	0.004	0.000	0.002	0.014	0.016	0.008	0.015	0.023	0.004	0.004	0.000	0.000
9									0.000	0.014	0.003	0.001	0.003	0.012	0.015	0.006	0.013	0.021	0.005	0.006	0.002	0.002
10										0.000	0.011	0.015	0.018	0.002	0.001	0.008	0.001	0.007	0.019	0.020	0.016	0.016
11											0.000	0.004	0.006	0.009	0.012	0.003	0.010	0.018	0.008	0.009	0.005	0.005
12												0.000	0.002	0.014	0.016	0.008	0.015	0.022	0.004	0.004	0.000	0.000
13													0.000	0.016	0.018	0.010	0.017	0.025	0.002	0.002	0.002	0.002
14														0.000	0.002	0.006	0.001	0.009	0.018	0.018	0.014	0.014
15															0.000	0.008	0.002	0.006	0.020	0.021	0.016	0.017
16																0.000	0.007	0.015	0.012	0.012	0.008	0.008
17																	0.000	0.008	0.018	0.019	0.015	0.015
18																		0.000	0.026	0.027	0.023	0.023
19																			0.000	0.001	0.004	0.003
20																				0.000	0.004	0.004
21																					0.000	0.000
22																						0.000

Za pomocą nowej metody wyznaczono macierz podobieństwa i jej reprezentację przy użyciu drzewa filogenetycznego (*rysunek nr 73*).



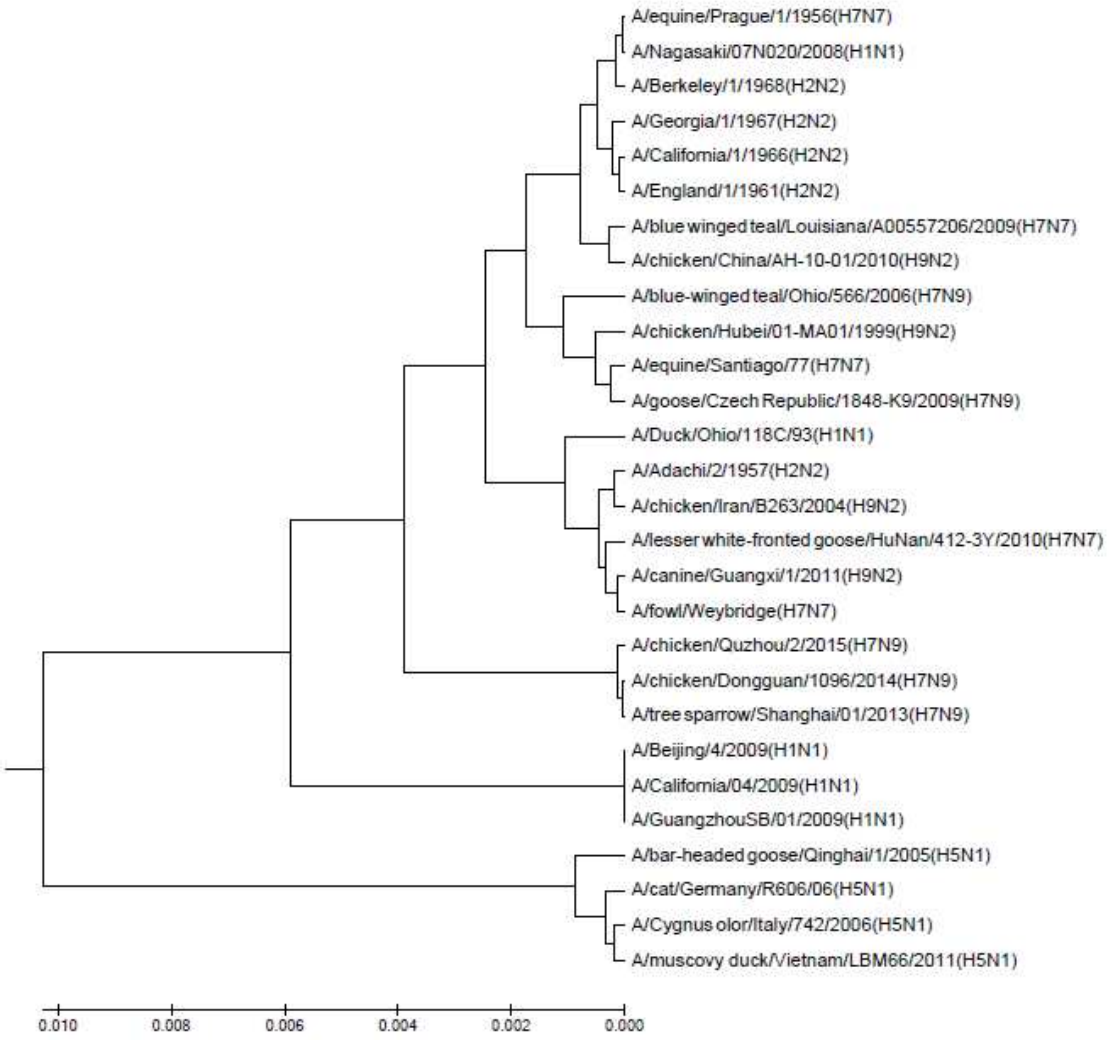
Rysunek 73. Drzewo filogenetyczne dla ND5 – 22 różne gatunki

3.7. Dane początkowe dla wirusa grypy typu A – 28 różnych gatunków

W celu porównania metody 20 wymiarowej reprezentacji białek z metodą Dynamic Time Warping przeprowadzono analizę podobieństwa sekwencji wirusa grypy typu A dla dwudziestu ośmiu różnych gatunków. Dane zostały pobrane z PDB. W *tabeli nr 15* podano numer dostępu do sekwencji aminokwasowej dla danego wirusa, która znajduje się w bazie PDB. W tym przypadku nie została pokazana macierz podobieństwa ze względu na jej wymiary. Podobieństwo między gatunkami zostało przedstawione za pomocą drzewa filogenetycznego (*rysunek nr 74*).

Tabela 15. Dane dotyczące użytych sekwencji wirusa grupy typu A

Lp.	Gatunek	Numer dostępu	Dł. sekwencji
1	A/Adachi/2/1957(H2N2)	BAD16637.1	469
2	A/bar-headed_goose/Qinghai/1/2005(H5N1)	BAM85828.1	449
3	A/Beijing/4/2009(H1N1)	ACR67256.1	469
4	A/Berkeley/1/1968(H2N2)	BAD16641.1	469
5	A/blue-winged_teal/Ohio/566/2006(H7N9)	ABS89412.1	470
6	A/California/1/1966(H2N2)	AAO46235.1	469
7	A/California/04/2009(H1N1)	AEE69012.1	469
8	A/cat/Germany/R606/06(H5N1)	ABF61763.1	449
9	A/chicken/Dongguan/1096/2014(H7N9)	AJJ96855.1	465
10	A/Cygnus_olor/Italy/742/2006(H5N1)	ABF50822.1	449
11	A/chicken/Quzhou/2/2015(H7N9)	AKI82227.1	465
12	A/Duck/Ohio/118C/93(H1N1)	AAF77041.1	469
13	A/blue_winged_teal/Louisiana/A00557206/2009 (H7N7)	ALT67567.1	470
14	A/canine/Guangxi/1/2011(H9N2)	AEK07935.1	469
15	A/chicken/China/AH-10-01/2010(H9N2)	AEE73586.1	466
16	A/chicken/Hubei/01-MA01/1999(H9N2)	AEO92432.1	466
17	A/chicken/Iran/B263/2004(H9N2)	ACD47112.1	469
18	A/England/1/1961(H2N2)	AAO46220.1	469
19	A/equine/Prague/1/1956(H7N7)	AAC57418.1	469
20	A/equine/Santiago/77(H7N7)	AAQ90293.1	469
21	A/fowl/Weybridge(H7N7)	AAA43425.1	471
22	A/Georgia/1/1967(H2N2)	AAO46244.1	469
23	A/goose/Czech_Republic/1848-K9/2009(H7N9)	ACX53685.1	470
24	A/GuangzhouSB/01/2009(H1N1)	ACR49238.1	469
25	A/Nagasaki/07N020/2008(H1N1)	ADC45738.1	470
26	A/lesser_white-fronted_goose/HuNan/412-3Y/2010 (H7N7)	AIW60686.1	471
27	A/muscovy_duck/Vietnam/LBM66/2011(H5N1)	BAM36161.1	449
28	A/tree_sparrow/Shanghai/01/2013(H7N9)	AGW82590.1	465



Rysunek 74. Drzewo filogenetyczne dla wirusa grypy typu A – 28 różnych gatunków

Przedstawiona powyżej metoda 20–wymiarowa dynamiczna reprezentacja sekwencji białkowych jest niewątpliwie wygodnym i niezawodnym narzędziem do rozwiązywania wielu problemów w medycynie i biologii, w których niezbędna jest analiza podobieństwa sekwencji białek. W szczególności metoda ta umożliwia tworzenie drzew filogenetycznych, a jej zaletą jest brak ograniczeń w długości sekwencji.

Zarówno graficzne reprezentacje DNA i białek wciąż motywują naukowców do tworzenia nowych reprezentacji w postaci: 2D, 3D a nawet wielowymiarowych. Naukowcy próbują stworzyć unikatowe metody mające na celu wydobycie nieznanych do teraz informacji ukrytych w kodzie genetycznym. Interesujące prace można znaleźć: (70–72).

4. ZAŁĄCZNIKI

4.1. Zasada działania programów

Poszczególne kroki programów:

- wprowadzenie sekwencji aminokwasowej białka,
- sprawdzanie długości sekwencji białka,
- konwersja znaków na opowiadający im kod ASCII (zmiana z litery na liczbę),
- przeprowadzenie metody: A Walk, której wynikiem jest zestaw współrzędnych dla badanego białka w przestrzeni 20 wymiarowej,
- transponowanie macierzy zestawu współrzędnych,
- obliczanie środka masy każdego aminokwasu,
- wyznaczenie układu współrzędnych w punkcie środka masy,
- obliczanie momentów bezwładności,
- implementacja metody Jacobiego.

4.2. 20-wymiarowa dynamiczna reprezentacja sekwencji białkowych – implementacja w programie Matlab

```
close all
clear all
clc
format
% używane zmienne
a = input('Podaj sekwencję:');
c = double(a);
k = size(c);
d = k(1,2); %długość sekwencji
sequence=zeros(1,20);
s_m = zeros (1,20);
u_n = zeros (d,20);
s_r = zeros (1,20);
I_p = zeros (1,20);
I_b = zeros (1,95);
t = zeros (1,d);
w = zeros (1,20);
wyniki = zeros (d,20);
wyniki_m = zeros (1,d);
index = 1;
index2 = 0;

for i=1:d
    if (c(i) == 97)
        sequence(1,1) = sequence(1,1)+1;
    end
    if (c(i) == 99)
        sequence(1,2) = sequence(1,2)+1;
    end
    if (c(i) == 100)
        sequence(1,3) = sequence(1,3)+1;
    end
    if (c(i) == 101)
```

```
        sequence(1,4) = sequence(1,4)+1;
    end
    if (c(i) == 102)
        sequence(1,5) = sequence(1,5)+1;
    end
    if (c(i) == 103)
        sequence(1,6) = sequence(1,6)+1;
    end
    if (c(i) == 104)
        sequence(1,7) = sequence(1,7)+1;
    end
    if (c(i) == 105)
        sequence(1,8) = sequence(1,8)+1;
    end
    if (c(i) == 107)
        sequence(1,9) = sequence(1,9)+1;
    end
    if (c(i) == 108)
        sequence(1,10) = sequence(1,10)+1;
    end
    if (c(i) == 109)
        sequence(1,11) = sequence(1,11)+1;
    end
    if (c(i) == 110)
        sequence(1,12) = sequence(1,12)+1;
    end
    if (c(i) == 112)
        sequence(1,13) = sequence(1,13)+1;
    end
    if (c(i) == 113)
        sequence(1,14) = sequence(1,14)+1;
    end
    if (c(i) == 114)
        sequence(1,15) = sequence(1,15)+1;
    end
    if (c(i) == 115)
        sequence(1,16) = sequence(1,16)+1;
```

```
    end
    if (c(i) == 116)
        sequence(1,17) = sequence(1,17)+1;
    end
    if (c(i) == 118)
        sequence(1,18) = sequence(1,18)+1;
    end
    if (c(i) == 119)
        sequence(1,19) = sequence(1,19)+1;
    end
    if (c(i) == 121)
        sequence(1,20) = sequence(1,20)+1;
    end
         wyniki (i,1:20)= sequence;  %wszytskie współrzędne

end;
%sprawdzenie
s = sum(wyniki(d,1:20));

%środki masy
dane = (wyniki)';
u_1 = sum(dane(1,:));
u_2 = sum(dane(2,:));
u_3 = sum(dane(3,:));
u_4 = sum(dane(4,:));
u_5 = sum(dane(5,:));
u_6 = sum(dane(6,:));
u_7 = sum(dane(7,:));
u_8 = sum(dane(8,:));
u_9 = sum(dane(9,:));
u_10 = sum(dane(10,:));
u_11 = sum(dane(11,:));
u_12 = sum(dane(12,:));
u_13 = sum(dane(13,:));
u_14 = sum(dane(14,:));
u_15 = sum(dane(15,:));
u_16 = sum(dane(16,:));
```

```
u_17 = sum(dane(17,:));
u_18 = sum(dane(18,:));
u_19 = sum(dane(19,:));
u_20 = sum(dane(20,:));
s_m(1,1) = sum(dane(1,:))/d;
s_m(1,2) = sum(dane(2,:))/d;
s_m(1,3) = sum(dane(3,:))/d;
s_m(1,4) = sum(dane(4,:))/d;
s_m(1,5) = sum(dane(5,:))/d;
s_m(1,6) = sum(dane(6,:))/d;
s_m(1,7) = sum(dane(7,:))/d;
s_m(1,8) = sum(dane(8,:))/d;
s_m(1,9) = sum(dane(9,:))/d;
s_m(1,10)= sum(dane(10,:))/d;
s_m(1,11)= sum(dane(11,:))/d;
s_m(1,12)= sum(dane(12,:))/d;
s_m(1,13)= sum(dane(13,:))/d;
s_m(1,14)= sum(dane(14,:))/d;
s_m(1,15)= sum(dane(15,:))/d;
s_m(1,16) = sum(dane(16,:))/d;
s_m(1,17) = sum(dane(17,:))/d;
s_m(1,18) = sum(dane(18,:))/d;
s_m(1,19) = sum(dane(19,:))/d;
s_m(1,20) = sum(dane(20,:))/d;

% układ współrzędnych w punkcie środka masy
for i= 1:d
    u_n(i,1:20) = wyniki(i,1:20) - s_m(1,1:20);
end

nowe = (u_n)';

% kwadraty współrzędnych
x_1(1,1:d) = nowe(1,1:d).^2;
w(1,1) = sum(x_1(1,1:d));
y_2(1,1:d) = nowe(2,1:d).^2;
w(1,2) = sum(y_2(1,1:d));
```

```
z_3(1,1:d) = nowe(3,1:d).^2;
w(1,3) = sum(z_3(1,1:d));
a_4(1,1:d) = nowe(4,1:d).^2;
w(1,4) = sum(a_4(1,1:d));
b_5(1,1:d) = nowe(5,1:d).^2;
w(1,5) = sum(b_5(1,1:d));
c_6(1,1:d) = nowe(6,1:d).^2;
w(1,6) = sum(c_6(1,1:d));
d_7(1,1:d) = nowe(7,1:d).^2;
w(1,7) = sum(d_7(1,1:d));
e_8(1,1:d) = nowe(8,1:d).^2;
w(1,8) = sum(e_8(1,1:d));
f_9(1,1:d) = nowe(9,1:d).^2;
w(1,9) = sum(f_9(1,1:d));
g_1(1,1:d) = nowe(10,1:d).^2;
w(1,10) = sum(g_1(1,1:d));
h_11(1,1:d) = nowe(11,1:d).^2;
w(1,11) = sum(h_11(1,1:d));
i_12(1,1:d) = nowe(12,1:d).^2;
w(1,12) = sum(i_12(1,1:d));
j_13(1,1:d) = nowe(13,1:d).^2;
w(1,13) = sum(j_13(1,1:d));
k_14(1,1:d) = nowe(14,1:d).^2;
w(1,14) = sum(k_14(1,1:d));
l_15(1,1:d) = nowe(15,1:d).^2;
w(1,15) = sum(l_15(1,1:d));
m_16(1,1:d) = nowe(16,1:d).^2;
w(1,16) = sum(m_16(1,1:d));
n_17(1,1:d) = nowe(17,1:d).^2;
w(1,17) = sum(n_17(1,1:d));
o_18(1,1:d) = nowe(18,1:d).^2;
w(1,18) = sum(o_18(1,1:d));
p_19(1,1:d) = nowe(19,1:d).^2;
w(1,19) = sum(p_19(1,1:d));
r_20(1,1:d) = nowe(20,1:d).^2;
w(1,20) = sum(r_20(1,1:d));
```

```
% momenty bezwładności na przekątnej
I_p(1,1) = sum(w(1,2:20));                    %xx
I_p(1,2) = (w(1,1)+ sum (w(1,3:20)));         %yy
I_p(1,3) = (sum(w(1,1:2))+ sum(w(1,4:20)));   %zz
I_p(1,4) = (sum(w(1,1:3))+ sum(w(1,5:20)));   %aa
I_p(1,5) = (sum(w(1,1:4))+ sum(w(1,6:20)));   %bb
I_p(1,6) = (sum(w(1,1:5))+ sum(w(1,7:20)));   %cc
I_p(1,7) = (sum(w(1,1:6))+ sum(w(1,8:20)));   %dd
I_p(1,8) = (sum(w(1,1:7))+ sum(w(1,9:20)));   %ee
I_p(1,9) = (sum(w(1,1:8))+ sum(w(1,10:20)));  %ff
I_p(1,10) = (sum(w(1,1:9))+ sum(w(1,11:20))); %gg
I_p(1,11) = (sum(w(1,1:10))+ sum(w(1,12:20))); %hh
I_p(1,12) = (sum(w(1,1:11))+ sum(w(1,13:20))); %ii
I_p(1,13) = (sum(w(1,1:12))+ sum(w(1,14:20))); %jj
I_p(1,14) = (sum(w(1,1:13))+ sum(w(1,15:20))); %kk
I_p(1,15) = (sum(w(1,1:14))+ sum(w(1,16:20))); %ll
I_p(1,16) = (sum(w(1,1:15))+ sum(w(1,17:20))); %mm
I_p(1,17) = (sum(w(1,1:16))+ sum(w(1,18:20))); %nn
I_p(1,18) = (sum(w(1,1:17))+ sum(w(1,19:20))); %oo
I_p(1,19) = (sum(w(1,1:18))+ w(1,20));        %pp
I_p(1,20) = sum(w(1,1:19));                   %rr

przekatna = (I_p)';
% mnożenie współrzędnych yx (2,1)

    for j= 2:20
        for k = 1:j-1
            for i = 1:d
                %if k < i
                    index2 = index2+1;
                    t = nowe(j,i)*nowe(k,i);
                    wyniki_m (index,index2)= t;  %wszytskie t
                %end
            end
            index2 = 0;
            index = index +1;
        end
```

```
    end
%całkowite momenty bezwładności
m_p = zeros (1,d);
%yx
for i = 1:190
    m_p(1,i) = -sum(wyniki_m(i,1:d));
end
momenty = (m_p)';
```

4.3. Metoda Jacobiego – implementacja w programie Matlab

```
close all
%clear
clc
%Wprowadz macierz kwadratową - symetryczną
%A=input('Podaj macierz A:');
B=A;
epsilon=0.0001
test = size(A)
rzad = rank(A)
i =0;
j =0;
index=0;
AT = A'
if A == A'
     n = 'Macierz jest symetryczna';
     disp(n)
else A ~= A'
         m = 'Macierz nie jest symetryczna';
         disp(m)
end
przekatna = diag(A)

for i = 1:rzad
    B(i,i)=0;
end
disp(B)
[M,I] = max(abs(B(:)))
[q,p] = ind2sub(size(B),I);
eta = (A(q,q)- A(p,p))/(2*A(p,q))
A(p,p)
A(q,q)
t = sign(eta)/(abs(eta)+sqrt((eta^2)+1))
c=1/sqrt(t^2+1)
s=t*c
% Obliczam A1
Qn = eye(rzad);
```

```
Qn(p,q)=s;
Qn(q,p)=-s;
Qn(q,q)=c;
Qn(p,p)=c;
disp(Qn)
An=Qn'*A*Qn
A1=An;
% Obliczam W1
W=eye(rzad);
Wn=W*Qn
% warunek zakończenia obliczeń
for i = 1:rzad
    A1(i,i)=0;
end
disp(A1)
[sp,I1] = max(abs(A1(:)));
[q1,p1] = ind2sub(size(A1),I1);
disp(sp)
sn=max(abs(diag(An)));
e=sp/sn

while e>epsilon
    index=index+1
    n=('Potrzebna kolejna iteracja');
    disp(n)
    disp(A1)
    disp(An)
         for i = 1:rzad
    A1(i,i)=0;  %wcześniej przeliczone (nic się nie zmienia)
         end
         disp (A1)
[M1,I2] = max(abs(A1(:)))
[q2,p2] = ind2sub(size(A1),I2);
eta2 = (An(q2,q2)- An(p2,p2))/(2*An(p2,q2))
An(p2,p2)
An(q2,q2)
t2 = sign(eta2)/(abs(eta2)+sqrt((eta2^2)+1))
```

```
c2=1/sqrt(t2^2+1)
s2=t2*c2
Qn2 = eye(rzad);
Qn2(p2,q2)=s2;
Qn2(q2,p2)=-s2;
Qn2(q2,q2)=c2;
Qn2(p2,p2)=c2;
disp(Qn2)
An1=Qn2'*An*Qn2
An2=An1;
% Obliczam W1
W2=eye(rzad);
Wn2=Wn*Qn2
disp(Qn2)
% warunek zakończenia obliczeń
for i = 1:rzad
    An1(i,i)=0;
end
disp(An1);
[sp2,I3] = max(abs(An1(:)));
[q3,p3] = ind2sub(size(An1),I1);
disp(sp)
sn2=max(abs(diag(An2)));
e=sp2/sn2
A1=An2
An=An2;
Wn=Wn2;
disp(Wn)
Eig=diag(An2);
%e=0;
end
```

BIBLIOGRAFIA

1. Henikoff, S., Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89,** 10915–10919 (1992).

2. Hamori, E., Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.* **258,** 1318–1327 (1983).

3. Nandy, A. A new graphical representation and analysis of DNA sequence structure: I. methodology and application to globin genes. *Curr. Sci.* **66,** 309–314 (1994).

4. Novič, M., Randić, M. Representation of proteins as walks in 20-D space. *SAR QSAR Env. Res* **19,** 317–337 (2008).

5. Nandy, A., Ghosh, A., Nandy, P. Numerical characterization of protein sequences and application to voltage-gated sodium channel alpha subunit phylogeny. *In Silico Biol.* **9,** 77–87 (2009).

6. Bielińska-Wąż, D. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. *J. Math. Chem.* **49,** 2345–2407 (2011).

7. Randič, M., Nović, M., Plavšić, D. Milestones in graphical bioinformatics. *Int. J. Quantum Chem.* **113,** 2413–2446 (2013).

8. Bielińska-Wąż, D., Clark, T., Wąż, P., Nowak, W., Nandy, A. 2D-dynamic representation of DNA sequences. *Chem. Phys. Lett.* **442,** 140–144 (2007).

9. Bielińska-Wąż, D., Nowak, W., Wąż, P., Nandy, A., Clark, T. Distribution moments of 2D-graphs as descriptors of DNA sequences. *Chem. Phys. Lett.* **443,** 408–413 (2007).

10. Bielińska-Wąż, D., Wąż, T., Clark, T. Similarity studies of DNA sequences using genetic methods. *Chem. Phys. Lett.* **445,** 68–73 (2007).

11. Wąż, P., Bielińska-Wąż, D., Nandy, A. Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences. *J. Math. Chem.* **52,** 132–140 (2014).

12. Aram, V., Iranmanesh, A. 3D-dynamic representation of DNA sequences. *MATCH Commun. Math. Comput. Chem.* **67,** 809–816 (2012).

13. Wąż, P., Bielińska-Wąż, D. 3D-dynamic representation of DNA sequences. *J. Mol. Model.* **20,** 2141 (2014).

14. Wąż, P., Bielińska-Wąż, D. Non-standard similarity/dissimilarity analysis of DNA sequences. *Genomics* **104,** 464–471 (2014).

15. Yao, Y. H., Dai, Q., Li, C., He, P. A., Nan, X. Y., Zhang, Y. Z. Analysis of similarity/dissimilarity of protein sequences. *Proteins. Struct. Funct. Bioinf.* **73,** 864–71 (2008).

16. Yao, Y., Yan, S., Han, J., Dai, Q., He, P. A. A novel descriptor of protein sequences and its application. *J. Theor. Biol.* **347,** 109–17 (2014).

17. Hou, W., Pan, Q., He, M. A new graphical representation of protein sequences ant its applications. *Physica A* **444,** 996–1002 (2016).

18. Wąż, T., Bielińska-Wąż, D. Moments of interia of spectra and distribution moments as molecular descriptors. *MATCH Commun. Math. Comput. Chem.* **70,**

851–865 (2013).

19. Jagiełło, K., Puzyn, T., Wąż, P., Bielińska-Wąż, D. Moments of interia of spectra as descriptors for QSAR/QSPR, in: 1. Gutman (Ed.). *Top. Chem. Graph Theory.* 151–162 (2014).

20. Randić, M. 2-D graphical representation of proteins based on virtual genetic code. *SAR QSAR Env. Res* **15,** 147–157 (2004).

21. Randić, M., Butina, D., Zupan, J. Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **419,** 528–532 (2006).

22. Randić, M. 2-D graphical representation of proteins based on physico-chemical properties of amino acids. *Chem. Phys. Lett.* **444,** 176–180 (2007).

23. Li, C., Xing, L., Wang, X. 2D-graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.* **41,** 217–22 (2008).

24. Wu, Z. C., Xiao, X., Chou, K. C. 2D-MH: a web-server for generating graphs representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J. Theor. Biol.* **267,** 29–34 (2010).

25. He, P. A., Zhang, Y. P., Yao, Y. H., Tang, Y. F., Nan, X. Y. The graphical representation of protein sequences based on the physiochemical properties and its applications. *J. Comput. Chem.* **31,** 2136–42 (2010).

26. Ghosh, A., Nandy, A. Graphical representation and mathematical characterization of protein sequences and applications to viral proteins. *Adv. Protein Chem. Struct. Biol. Protein Struct. Dis.* **83,** 1–42 (2011).

27. Liao, B., Liao, B., Lu, X., Cao, Z. A novel graphical representation of protein sequences and its application. *J. Comput. Chem.* **32,** 2539–2544 (2011).

28. Yu, J. F., Sun, X., Wang, J. H. A novel 2D graphical representation of protein sequence based on individual amino acid. *Int. J. Quantum Chem.* **111,** 2835–2843 (2011).

29. Xie, X., Zheng, L., Yu, Y., Liang, L., Guo, M., Song, J., Yuan, Z. Protein sequence analysis based on hydropathy profile of amino acids. *J. Zhejiang Univ. Sci. B* **13,** 152–8 (2012).

30. He, P. A., Wei, J., Yao, Y., Tie, Z. A novel graphical representation of proteins and its application. *Phys. A Stat. Mech. its Appl.* **391,** 93–99 (2012).

31. Qi, Z. H., Feng, J., Qi, X. Q., Li, L. Application of 2D graphic representation of protein sequence based on Huffman tree method. *Comput. Biol. Med.* **42,** 556–563 (2012).

32. Yao, Y. H., Kong, F., Dai, Q., He, P. A. A sequence-segmented method applied to the similarity analysis of long protein sequence. *Match* **70,** 431–450 (2013).

33. Liu, Y., Li, D., Lu, K., Jiao, Y., He, P. A. P-H Curve, a Graphical Representation of Protein Sequences for Similarities Analysis. *Match-Communications Math. Comput. Chem.* **70,** 451–466 (2013).

34. Yao, Y. H., Yan, S., Xu, H., Han, J., Nan, X., He, P. A., Dai, Q. Similarity /Dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. *Evol. Bioinforma.* **10,** 87–96 (2014).

35. Li, Z., Geng, C., He, P. A., Yao, Y. A novel method of 3D graphical representation and similarity analysis for proteins. *MATCH Commun. Math. Comput. Chem.* **71,**

213–226 (2014).

36. Gupta, M. K., Niyogi, R., Misra, M. A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method. *MATCH Commun. Math. Comput. Chem.* **72,** 519–532 (2014).

37. Ma, T. T., Liu, Y. X., Dai, Q., Yao, Y. H., He, P. A. A graphical representation of protein based on a novel iterated function system. *Phys. A* **403,** 21–28 (2014).

38. Chen, Y., Li, K. S., Chang, S., Yang, L. A new 3D graphical representation for similarity/dissimilarity studies of protein sequences. *Comp. Model. New. Technol.* **18,** 296–303 (2014).

39. Qi, Z. H., Jin, M. Z., Li, S. L., Feng, J. A protein mapping method based on physicochemical properties and dimension reduction. *Comput. Biol. Med.* **57,** 1–7 (2015).

40. Clark, T. QSAR and QSPR based solely on surface properties? in *J. Mol. Graph. Model.* **22,** 519–525 (2004).

41. Güssregen, S., Matter, H., Hessler, G., Müller, M., Schmidt, F., Clark, T. 3D-QSAR based on quantum-chemical molecular fields: Toward an improved description of halogen interactions. *J. Chem. Inf. Model.* **52,** 2441–2453 (2012).

42. El Kerdawy, A., Güssregen, S., Matter, H., Hennemann, M., Clark, T. Quantum mechanics-based properties for 3D-QSAR. *J. Chem. Inf. Model.* **53,** 1486–1502 (2013).

43. Agüero-Chapín, G., Antunes, A., Ubeira, F. M., Chou, K. C., González-Díaz, H. Comparative study of topological indices of macro/supramolecular RNA complex networks. *J. Chem. Inf. Model.* **48,** 2265–2277 (2008).

44. Dea-Ayuela, M. A., Pérez-Castillo, Y., Meneses-Marcel, A., Ubeira, F. M., Bolas-Fernández, F., Chou, K. C., González-Díaz, H. HP-Lattice QSAR for dynein proteins: Experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a Leishmania infantum sequence. *Bioorganic Med. Chem.* **16,** 7770–7776 (2008).

45. Vilar, S., González-Díaz, H., Santana, L., Uriarte, E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *J. Comput. Chem.* **29,** 2613–2622 (2008).

46. Cruz-Monteagudo, M., González-Díaz, H., Borges, F., Dominguez, E. R., Cordeiro, M. N. D. S. 3D-MEDNEs: An alternative 'in silico' technique for chemical research in toxicology. 2. Quantitative proteome-toxicity relationships (QPTR) based on mass spectrum spiral entropy. *Chem. Res. Toxicol.* **21,** 619–632 (2008).

47. Pérez-Montoto, L. G., Santana, L., González-Díaz, H. Scoring function for DNA-drug docking of anticancer and antiparasitic compounds based on spectral moments of 2D lattice graphs for molecular dynamics trajectories. *Eur. J. Med. Chem.* **44,** 4461–4469 (2009).

48. Vilar, S., González-Díaz, H., Santana, L., Uriarte, E. A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer. *J. Theor. Biol.* **261,** 449–458 (2009).

49. González-Díaz, H., Pérez-Montoto, L. G., Duardo-Sanchez, A., Paniagua, E., Vázquez-Prieto, S., Vilas, R., Dea-Ayuela, M. A., Bolas-Fernández, F., Munteanu, C. R., Dorado, J., Costas, J., Ubeira, F. M. Generalized lattice

graphs for 2D-visualization of biological information. *J. Theor. Biol.* **261,** 136–147 (2009).

50. Perez-Bello, A., Munteanu, C. R., Ubeira, F. M., Lopes De Magalhães, A., Uriarte, E., González-Díaz, H. Alignment-free prediction of mycobacterial DNA promoters based on pseudo-folding lattice network or star-graph topological indices. *J. Theor. Biol.* **256,** 458–466 (2009).

51. González-Díaz, H., Dea-Ayuela, M. A., Pérez-Montoto, L. G., Prado-Prado, F. J., Agëro-Chapín, G., Bolas-Fernández, F., Vazquez-Padrón, R. I., Ubeira, F. M. QSAR for RNases and theoretic-experimental study of molecular diversity on peptide mass fingerprints of a new Leishmania infantum protein. *Mol. Divers.* **14,** 349–369 (2010).

52. Schaschke, C. Dictionary of Chemical Engineering. *Oxford Univ. Press.* 53 (2014).

53. Hilvert, D. Design of Protein Catalysts. *Annu. Rev. Biochem.* **82,** 447–470 (2013).

54. Gaucher, E., Cox, V. Encyclopedia of Astrobiology. *Springer Berlin Heidelb.* 1348 (2015).

55. Pauling, L., Corey, R. B., Branson, H. R. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.* **37,** 205–211 (1951).

56. Edwards, A. M., Watson, J. D., Golovin, A., Laskowski, R. A., Henrick, K., Thornton, J. M., Joachimiak, A. Structural bioinformatics: from protein structure to function. *Evol. Methods Macromol. Crystallogr.* 165–179 (2007).

57. Hall, B. G. Łatwe drzewa filogenetyczne. *Wydaw. Uniw. Warsz.* 66–69 (2008).

58. Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

59. Higgs, P. G., Attwood, T. K. Bioinformatyka i ewolucja molekularna. *Bioinformatyka i Ewol. Mol.* 90–93 (2011).

60. Dayhoff, M., Schwartz, R. A Model of Evolutionary Change in Proteins. *Atlas protein Seq. Struct.* 345–352 (1978).

61. Randić, M., Zupan, J., Balaban, A. T., Vikić-Topić, D., Plavšić, D. Graphical representation of proteins. *Chem. Rev.* **111,** 790–862 (2011).

62. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.* **18,** 2163–2170 (1990).

63. Yu, Z. G., Anh, V., Lau, K. S. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theor. Biol.* **226,** 341–348 (2004).

64. Randić, M., Zupan, J., Balaban, A. T. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **397,** 247–252 (2004).

65. Randić, M., Zupan, J., Vikić-Topić, D. On representation of proteins by star-like graphs. *J. Mol. Graph. Model.* **26,** 290–305 (2007).

66. Czerniecka, A., Bielińska-Wąż, D., Wąż, P., Clark, T. 20D-dynamic representation of protein sequences. *Genomics* **107,** 16–23 (2016).

67. Nie, Z. M., Zhang, Z. F., Wang, D., He, P. A., Jiang, C. Y., Song, L., Chen, F., Xu,

J., Yang, L., Yu, L. L., Chen, J., Lv, Z. B., Lu, J. J., Wu, X. F., Zhang, Y. Z. Complete sequence and organization of Antheraea pernyi nucleopolyhedrovirus, a dr-rich baculovirus. *BMC Genomics* **8,** 248 (2007).

68. Herniou, E. A., Olszewski, J. A., O'Reilly, D. R., Cory, J. S. Ancient Coevolution of Baculoviruses and Their Insect Hosts. *J. Virol.* **78,** 3244–3251 (2004).

69. Hou, W., Pan, Q., Pengc, Q., He, M. A new method to analyze protein sequence similarity using Dynamic Time Warping. *Genomics* **109,** 123–130 (2017).

70. Yang, L., Zhang, W. A Multiresolution Graphical Representation for Similarity Relationship and Multiresolution Clustering for Biological Sequences. *J. Comput. Biol.* **24,** 299–310 (2017).

71. Ping, P., Zhu, X., Wang, L. Similarities/Dissimilarities analysis of protein sequences based on PCA-FFT. *J. Biol. Syst.* **25,** 1–17 (2017).

72. Hu, H., Li, Z., Dong, H. Graphical representation and similarity analysis of protein sequences based on fractal interpolation. *TCBB* **14,** 182–192 (2017).