# ALGORYTMY KONSTRUOWANIA DENDROGRAMÓW STOSOWANYCH PRZY ANALIZIE FILOGENETYCZNEJ MIKROORGANIZMÓW

**Filip Zdziennicki, Anna Misiewicz**

Instytut Biotechnologii Przemysłu Rolno-Spożywczego im. prof. Wacława Dąbrowskiego

Zakład Mikrobiologii

ul. Rakowiecka 36, 02-532 Warszawa

filip.zdziennicki@ibprs.pl

**Streszczenie**

Bioinformatyka jest obecnie dynamicznie rozwijającą się dziedziną nauki, łączącą dwie dotychczas niezwiązane ze sobą dziedziny – biologię oraz informatykę. Jej gwałtowny rozwój wynika z konieczności szybkiej i rzetelnej analizy dużej liczby danych otrzymywanych podczas badań. Dane te są niejednokrotnie złożone i wzajemnie powiązane, co wymusza zastosowanie odpowiednich narzędzi informatycznych oraz sprzętu o dużej mocy obliczeniowej. Jednym z takich narzędzi jest filogenetyka molekularna, która daje możliwość stosowania kilku algorytmów konstruowania dendrogramów będących graficzną prezentacją wyników analizy. W ciągu ostatnich lat nastąpił rozwój algorytmów konstruowania dendrogramów, poczynając od UPGMA, przez NJ, po metody znacznie bardziej szczegółowe, takie jak MP czy ML. Na rozwój ten wpływa potrzeba opracowania jak najbardziej optymalnej metody, która uwzględni złożoność procesów biologicznych stojących za danymi. Podstawą matematyczną każdego z opisanych algorytmów jest analiza skupień. Pozwala ona na grupowanie danych ze zbioru w mniejsze podzbiory na podstawie ich podobieństwa.

**Słowa kluczowe**: metoda największej wiarygodności (ML), metoda największej oszczędności (MP), metoda przyłączania sąsiadów (NJ), metoda średnich połączeń (UPGMA)

**DENDROGRAM CONSTRUCTING ALGORITHMS APPLIED IN PHYLOGENETIC ANALYSIS OF MICROORGANISMS**

**Summary**

Nowadays bioinformatics is a dynamically developing field of science linking together two separated fields of biology and informatics. It is an answer to need of having fast and reliable analysis of large amount of data obtained during research. Such data is often complicated and requires adequate informatics tools and equipment with large computing power. One of these tools is molecular phylogenetics, which possibility to use a few dendrogram construction algorithms. Dendrogram is a graphical presentation of obtained results. Over the last years there is observed development of dendrogram constructing algorithms starting from UPGMA, then NJ to more specialized methods like MP or ML. Development of algorithms is forced by need of obtained most optimal method modeling biological process standing behind data. Mathematical base of each described algorithms is cluster analysis. It enables grouping data from set into smaller subsets according to similarity of data.

**Key words**: maximal likelihood, maximal parsimony, neighbor joining, UPGMA

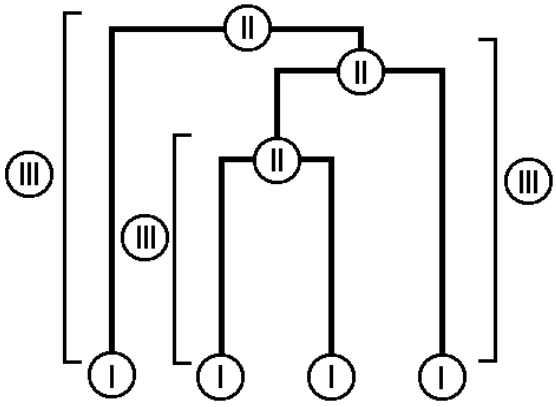
## WSTĘP

W prowadzonych badaniach dotyczących identyfikacji, charakterystyki i różnicowania mikroorganizmów wykorzystuje się między innymi metody molekularne. Do analizy wyników badań stosowane są różne narzędzia bioinformatyczne, w tym filogenetyka molekularna, której celem jest porównywanie kilku genomów, aby znaleźć związki ewolucyjne między nimi. Wywodzi się ona z klasycznych sposobów uszeregowania organizmów zgodnie z ich podobieństwami i różnicami. Badane są zmienne cechy mikroorganizmów. Po wprowadzeniu fenetyki, uznającej konieczność oceny dużej liczby cech wpływających na właściwą klasyfikację, możliwe było przedstawienie ich w postaci numerycznej, a następnie poddanie analizie matematycznej. Obiektywne dane molekularne uzyskane doświadczalnie można było sprowadzić do postaci numerycznej i opracować statystycznie. Zmienność DNA może być więc określana bezpośrednio, poprzez analizy sekwencyjne, lub pośrednio – np. przez różne wyniki elektroforegramów uzyskanych w analizie restrykcyjnej.

Powszechną metodą prezentacji oceny zmienności jest stosowanie dendrogramów. Dendrogram jest to diagram w postaci drzewa, prezentujący zależności pomiędzy obiektami

na podstawie przyjętej zależności. Jedną z jego odmian jest filogram, nazywany inaczej drzewem filogenetycznym, często stosowaną odmianą jest także kladogram. Podstawową różnicą pomiędzy tymi odmianami jest brak uwzględnienia czasu pojawienia się nowych linii ewolucyjnych w przypadku kladogramu. Istnieje kilka metod budowy dendrogramów.

Drzewa filogenetyczne można podzielić na drzewa ukorzenione oraz na nieukorzenione. Drzewa nieukorzenione wskazują wzajemne zależności pomiędzy obiektami, lecz nie wskazują pochodzenia poszczególnych obiektów. W przypadku drzew ukorzenionych można wyróżnić trzy podstawowe elementy. Przykładowy dendrogram został przedstawiony na rysunku 1. Cyfrą I oznaczono na nim liście, które odpowiadają analizowanym obiektom. Węzły łączące poszczególne obiekty zostały oznaczone cyfrą II. Węzły wskazują także miejsce, poniżej którego obiekty nie mają już części wspólnych i nastąpiło wydzielenie się jednego obiektu z drugiego. Odległości pomiędzy kolejnymi węzłami są nazywane gałęziami. Na rysunku zostały one oznaczone cyfrą III. Gałęzie opisują stopień podobieństwa pomiędzy obiektami w kolejnych węzłach. Im większa jest ta odległość, tym większa jest różnica pomiędzy obiektem–przodkiem a obiektem–potomkiem.




**Rysunek 1.** Budowa drzewa filogenetycznego: I – liście, II – węzły, III – gałęzie (materiały własne)
*Construction of the phylogenetic tree: I – leaves, II – nodes, III – branches (own materials)*

Na rysunku 2. pokazano przekształcenie jednego drzewa nieukorzenionego w pięć drzew ukorzenionych. Drzewa dla uproszczenia zostały zaprezentowane jako kladogramy, później przekształcone w filogramy. Warto zauważyć, że dla czterech obiektów na rysunku, oznaczonych jako A, B, C i D, możliwe jest utworzenie trzech różnych drzew nieukorzenionych, a dla każdego z nich pięć drzew ukorzenionych. Miejsca, w których dokonana została transformacja drzewa nieukorzenionego w konkretne drzewo ukorzenione, zostały oznaczone na rysunku cyframi od 1 do 5. Zależność pomiędzy liczbą drzew

nieukorzenionych a liczbą drzew ukorzenionych można przedstawić za pomocą dwóch wzorów:

$$R_n = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$, dla n≥2,

$$U_n = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$, dla n≥3,

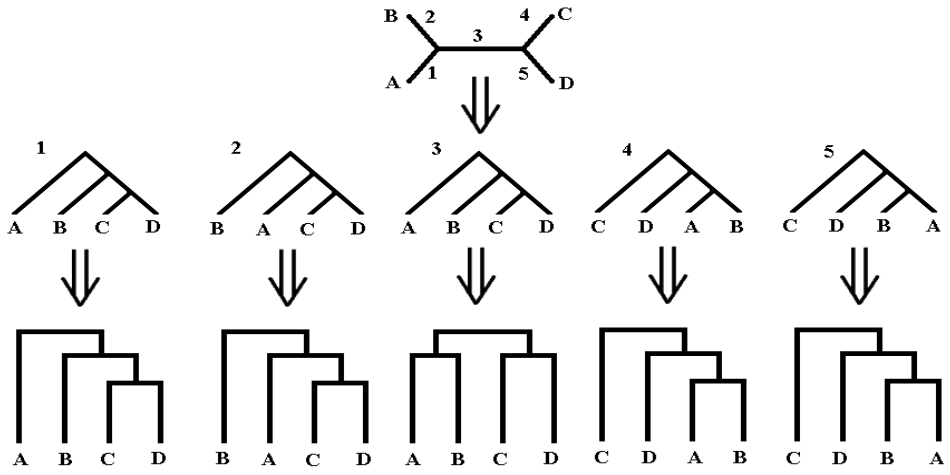
gdzie:

$R_n$ – liczba możliwych drzew ukorzenionych,

$U_n$ – liczba możliwych drzew nieukorzenionych,

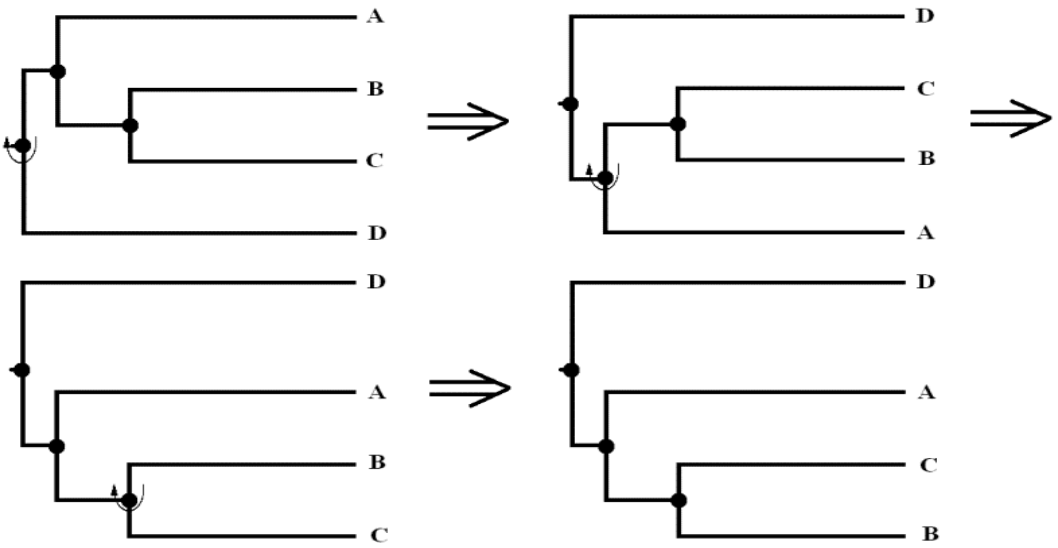
n – liczna obiektów (liści).

Z powyższych wzorów wynika, że liczba drzew nieukorzenionych z n liśćmi jest równa liczbie drzew ukorzenionych z n-1 liśćmi [Felsentein 2004].




**Rysunek 2.** Przekształcenie drzewa nieukorzenionego w drzewa ukorzenione (materiały własne)

*Transforming of unrooted tree into rooted one (own materials)*




**Rysunek 3.** Przekształcenia topologiczne wybranego drzewa filogenetycznego (materiały własne)

*Topological transformation of the phylogenetic tree (own materials)*

Ważną cechą drzewa filogenetycznego jest jego topologia, czyli kształt, forma oraz układ. Na topologię drzew filogenetycznych ma wpływ występowanie i wzajemne położenie węzłów. Warto zauważyć, że gałęzie mogą się swobodnie obracać w węzłach, dzięki czemu otrzymywanych jest wiele, zależnie od liczby węzłów, drzew filogenetycznych. Zależności pomiędzy poszczególnymi węzłami w gałęziach nie ulegają zmianie, co za tym idzie – zależności pomiędzy obiektami w danej gałęzi także nie ulegają zmianie. Należy zauważyć, że zmiany topologii mogą mieć wpływ na zależności pomiędzy obiektami znajdującymi się w różnych gałęziach. Sposób wyboru najlepszego drzewa został opisany w dalszej części tekstu.

Na rysunku 2. przedstawiono dendrogram wraz z jego kolejnymi przekształceniami, polegającymi na obrocie gałęzi wokół osi przeprowadzonej wzdłuż węzła. Widoczne są cztery możliwe topologie, które w różny sposób przedstawiają to samo drzewo. Drzewo filogenetyczne jest porównywane do drzewa genealogicznego z tą różnicą, że drzewa genealogiczne nie są drzewami binarnymi w przeciwieństwie do drzew filogenetycznych. Wynika to z założenia, że w procesie ewolucji nowe gatunki wyodrębniały się pojedynczo z gatunku początkowego. Graficzna postać, jaką jest drzewo filogenetyczne, ułatwia analizę wyników badań, dlatego dendrogramy są powszechnie stosowanym narzędziem służącym do ich prezentacji.

## 1. Analiza skupień

Drzewa filogenetyczne sporządzane są na podstawie przekształcenia macierzy zawierającej dane będące przedmiotem analizy [Posada 2009]. Najczęściej wykorzystywane są hierarchiczne drzewa filogenetyczne. Do ich utworzenia opracowano kilka algorytmów, opierających się na analizie skupień, która jest matematyczną podstawą analiz. Zasadą działania tej metody jest grupowanie elementów we względnie jednorodne klasy. Podstawą grupowania w większości algorytmów jest zbiór podobnych albo wspólnych cech określonych na wstępie analizy. Cechy te określa się jako podobieństwo pomiędzy elementami. Dzięki funkcji podobieństwa wiązanych ze sobą jest coraz więcej obiektów, poczynając od tych wykazujących się największym podobieństwem. Są one zbierane w skupienia elementów, coraz bardziej różniące się od siebie [Lewis 2001]. W końcowym etapie wszystkie obiekty ze zbioru początkowego zostają połączone ze sobą. W miejscach, gdzie zostały uformowane nowe pojedyncze skupienia, powstają na wykresie węzły, dzięki którym możliwe jest odczytanie odległości, w której odpowiednie elementy zostały ze sobą powiązane [Murtagh 1984]. Jeśli dane prezentowane na wykresie mają wyrazistą strukturę, to znaczy widoczne są skupienia podobnych do siebie obiektów, to taka struktura znajdzie odzwierciedlenie na hierarchicznym drzewie w postaci oddzielnych gałęzi. Pomyślna analiza za pomocą algorytmu łączącego obiekty daje możliwość wykrywania gałęzi i ich interpretacji [Jobling i in. 2004]. Do aglomeracji danych w skupienia wykorzystywane są miary rozbieżności lub odległości pomiędzy obiektami. Najbardziej bezpośrednim sposobem obliczenia odległości między obiektami w przestrzeni wielowymiarowej jest obliczenie odległości euklidesowej. Jeśli mamy przestrzeń dwu- lub trójwymiarową, miara ta wyznacza rzeczywistą odległość geometryczną między obiektami w przestrzeni. W przypadku analizy filogenetycznej oraz użytego do niej wybranego algorytmu możliwe jest wykorzystanie miary pochodnej odległości euklidesowej – w zależności od potrzeb badacza. Niemniej jednak najczęściej stosowaną odległością jest odległość euklidesowa. Oblicza się ją następująco:

$$\text{Odległość } (x, y) = [\Sigma_i (x_i - y_i)^2]^{1/2},$$

gdzie:

x, y – obiekty,

$x_i - y_i$ – współrzędne obiektów x, y.

Innym rozwiązaniem możliwym do zastosowania przy obliczaniu odległości jest kwadrat odległości euklidesowej. Dzięki podniesieniu odległości euklidesowej do kwadratu możliwe jest przypisanie obiektom bardziej oddalonym większej wagi.

Odległość $(x, y) = \Sigma_i (x_i - y_i)^2$,

gdzie:

x, y – obiekty,

$x_i - y_i$ – współrzędne obiektów x, y.

Następną możliwością wyznaczenia odległości potrzebnej do analizy jest odległość miejsca, w literaturze anglojęzycznej znana jako Manhattan lub City block. Odległość ta jest sumą różnic mierzonych wzdłuż wymiarów. Uzyskane wyniki są w większości przypadków podobne do wyników otrzymanych z zastosowaniem odległości euklidesowej, aczkolwiek wpływ pojedynczych dużych różnic, czyli przypadków odstających, jest tłumiony.

Odległość $(x, y) = \Sigma_i |x_i - y_i|$,

gdzie:

x, y – obiekty,

$x_i - y_i$ – współrzędne obiektów x, y.

Istnieją także inne odległości możliwe do zastosowania w analizie skupień, aczkolwiek wymienione powyżej są podstawowymi odległościami stosowanymi przy tworzeniu dendrogramów na użytek analizy porównawczej organizmów na podstawie ich budowy molekularnej. Istotne jest podkreślenie potrzeby ustandaryzowania danych liczbowych w celu otrzymania wiarygodnych i łatwych do zinterpretowania danych. Dane powinny być wyrażone w tych samych jednostkach oraz opisywać ten sam lub podobny zbiór cech.

### 1.1 UPGMA

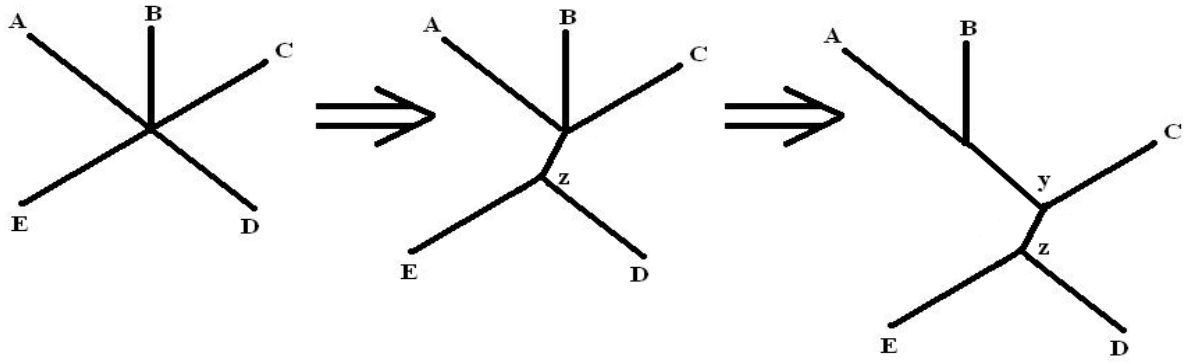
Za najprostszą metodę konstrukcji dendrogramów uważana jest metoda średnich połączeń – UPGMA (Unweighted Pair-Group Method using arithmetic Averages). Wynika to z prostoty stosowanego w niej algorytmu, a co za tym idzie – z krótkiego czasu potrzebnego do opracowania danych oraz otrzymania wyniku w postaci odpowiedniego dendrogramu [Sokal, Michener 1958; Sokal, Sneath 1963; Nei 1975]. Drzewo konstruowane jest na podstawie algorytmu iteracyjnego, składającego się z wielokrotnie powtórzonej czynności grupowania obiektów, polegającej na łączeniu krawędzią każdych dwóch obiektów, które dzieli najmniejszy dystans spośród wszystkich obliczonych. Następnie czynność ta jest powtarzana i następują kolejne przyłączenia obiektów lub ich zgrupowań aż do wyczerpania danych i uzyskania pełnego obrazu drzewa [Fitch 1971; Nei 1975]. Wynikiem tego algorytmu jest dendrogram, który jest ukorzeniony. Istotnym ograniczeniem UPGMA jest błędne założenie, stosowane w większości analiz, o stałym tempie ewolucji każdego gatunku. Każde

odstępstwo od stałego tempa mutacji prowadzi do uzyskania błędnego grafu. Ze względu na swoje ograniczenia opisany powyżej algorytm jest coraz rzadziej stosowany w filogenetyce. Nadal znajduje jednak zastosowanie w bioinformatyce, przy analizie danych mikromacierzowych.

### 1.2 Neighbor-Joining

Obecnie dużym uznaniem cieszy się metoda najbliższego sąsiada, czyli NJ (Neighbor-Joining). Jest ona oparta na innej, poprzedzającej ją metodzie, tj. ME (Minimal Evolution). Metoda ME także wykorzystuje algorytm iteracyjny i polega na wyznaczeniu długości gałęzi łączących badane populacje, a następnie ich zsumowaniu dla wszystkich topologii uznanych za wiarygodne. Biorąc pod uwagę założenie, że ewolucja zachodzi zawsze możliwie najkrótszą z dostępnych dróg, najbardziej prawdopodobnym drzewem jest to, którego suma długości wszystkich gałęzi jest najmniejsza [Cavalli-Sforza, Edwards 1967]. Warunkiem koniecznym do uzyskania wiarygodnego dendrogramu jest poddanie analizie dostatecznie dużej liczby danych, co wiąże się z dużym nakładem czasu [Nei 1996]. Zaproponowano więc, aby analizę rozpoczynać od konstruowania drzewa za pomocą metody NJ. Efektem tego było wyznaczenie innych drzew o zbliżonej budowie wraz z sumami długości gałęzi. Po porównaniu wszystkich uzyskanych w ten sposób drzew wybierane jest to o najmniejszej wartości. Wynikiem tej metody najczęściej jest drzewo tożsame z drzewem wyjściowym uzyskanym metodą ME lub różniące się od niego nieznacznie [Rzhetsky, Nei 1993].

Argumentem przemawiającym za szerokim zastosowaniem metody NJ jest czas potrzebny do otrzymania wyniku. Metoda ta jest *de facto* uproszczoną wersją metody ME, a więc czas wykonania algorytmu jest istotnie krótszy. Istotą metody jest znalezienie dwóch obiektów o najmniejszej dzielącej je odległości. Są to dwa obiekty połączone jednym węzłem w drzewie nieukorzenionym, nazywane sąsiadami. Na rysunku 4. pokazano etapy powstawania drzewa ukorzenionego zawierającego 5 obiektów (A, B, C, D i E). Początkowo wszystkie obiekty są połączone wspólnym węzłem, za pomocą macierzy odległości porównywane są odległości pomiędzy poszczególnymi obiektami. Dwa najbardziej podobne obiekty, czyli z najmniejszą łączącą je odległością (sąsiedzi), są łączone osobnym węzłem i w następnym kroku traktowane jako jeden obiekt. Na rysunku były to obiekty E i D. Następnie obiekty E i D traktowane jako jeden (z) poddawane są ponownie analizie odległości, z której wynika, że sąsiadem z jest obiekt C i zostaje on połączony osobnym węzłem (y) z węzłem łączącym E i D (z).



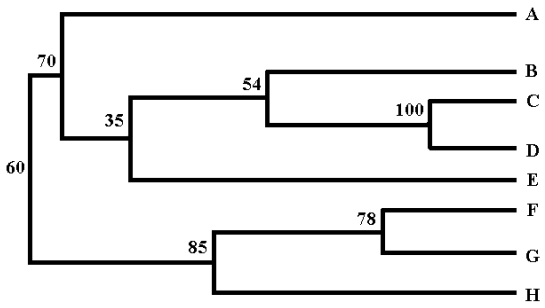

**Rysunek 4.** Schemat algorytmu NJ (materiały własne)

*Schematic NJ algorithm (own materials)*

Pełny dendrogram jest tworzony poprzez powtarzanie przyłączeń kolejnych sąsiadów do poprzedniej pary obiektów, traktowanych w kolejnym powtórzeniu jako jeden obiekt. Zauważyć można, że w przypadku drzewa obrazującego zależność dla czterech lub więcej obiektów, oznaczoną jako N, minimalna liczba par wynosi dwa, a maksymalna N/2 dla liczby parzystej N oraz (N-1)/2 w przypadku nieparzystej [Saitou, Nei 1987].

## 2. Określenie dokładności skonstruowanego drzewa. Metoda bootstrap

Metoda bootstrap jest rozwinięciem wcześniej stosowanych metod i wynika z potrzeby szacowania rozgałęzień w dendrogramach. Generowane drzewa przedstawiają jedno z wielu możliwych rozwiązań, które pomimo swojej matematycznej poprawności może być różne od drzewa rzeczywistego. Spowodowane jest to faktem, że odległości mierzone pomiędzy obiektami podlegają przypadkowej zmienności. Dlatego właśnie obecnie węzły w dendrogramach opatrzone są wartościami bootstrap, jak to zostało przedstawione na rysunku 5. Wartości te są wyznaczone poprzez znajdywanie dopasowań obiektów, którymi w przypadku analizy genetycznej są odpowiednie sekwencje, a później konstruowanie odpowiadających im dendrogramów [Soltis, Solis 2003]. Następnie taki zbiór drzew, składający się z od 100 do 1000 elementów, jest podawany analizie pod względem uzyskanej topologii. Wartość bootstrap jest wartością procentową liczby drzew posiadających taką samą topologię jak drzewo oryginalne, poddawane analizie bootstrap do wszystkich wygenerowanych drzew. Wynikiem opisanej powyżej metody jest drzewo konsensusowe, zawierające rozgałęzienia uszeregowane według malejących wartości bootstrap.




**Rysunek 5**. Drzewo filogenetyczne z podanymi wartościami bootstrap w węzłach (materiały własne)

*A phylogenetic tree with bootstrap values indicated at the nodes (own materials)*

### 2.1 Maximum parsimony. Metoda największej oszczędności

Innym podejściem do zwiększenia wiarygodności otrzymywanych drzew filogenetycznych jest zastosowanie metody największej oszczędności. Głównym celem metody jest uzyskanie takiego drzewa, które będzie obrazowało zależności między wszystkimi elementami zbioru danych wraz z uwzględnieniem warunku, jakim jest najmniejsza konieczna liczba zmian ewolucyjnych pozwalająca wytłumaczyć otrzymany wynik. Dzięki temu możliwe jest włączenie do analizy informacji dotyczących zmiennego tempa powstawania różnych rodzajów mutacji. Drzewo powstałe w wyniku użycia metody MP jest zawsze nieukorzenione i wskazuje względne zależności między elementami zbioru danych, jednak nie daje żadnych wskazówek odnośnie do czasu dywergencji [Steel, Penny 2000]. Z drugiej strony, jeśli zastosowane kryterium parsymonii prowadzi do uzyskania kilku równorzędnych drzew, nie istnieje metoda pozwalająca na wybór najbardziej optymalnego spośród nich, a więc w takich przypadkach konieczne jest zastosowanie innych narzędzi wspomagających analizę. Pomimo tych utrudnień metoda największej oszczędności jest stosowana przez naukowców z racji bardzo prostego algorytmu, który prowadzi do zadowalającego wyniku [Higgs, Attwood 2005].

### 2.2 Maximum Likelihood. Metoda największej wiarygodności

Jest to metoda podobna pod względem działania do metody bootstrap. Jej celem jest weryfikacja uzyskanego dendrogramu na podstawie przyjętego kryterium dla uzyskania rzetelnych wyników. Metodę ML (Maximum Likelihood) opracowano na podstawie założenia, że wszystkie drzewa mogą być rozpatrywane jako alternatywne rozwiązania [Felsenstein 1981]. Każde drzewo filogenetyczne można opisać za pomocą parametrów takich

jak: topologia, długość gałęzi oraz matematyczny model powstawania mutacji. Zadaniem metody ML jest znalezienie takiego układu powyższych trzech parametrów, aby maksymalizować prawdopodobieństwo uzyskania odpowiedniego wykresu dla danego zbioru danych [Strimmer, von Haeseler 1996; Pevsner 2009; Kishino, Hasegawa 1989].

## PODSUMOWANIE

Bioinformatyka w dziedzinie filogenetyki molekularnej daje możliwość stosowania wielu narzędzi umożliwiających skuteczną analizę otrzymanych wyników, a także ich graficzną interpretację. W tabeli 1. zestawiono opisane wcześniej algorytmy konstruowania dendrogramów, z uwzględnieniem ich mocnych oraz słabych stron. Mnogość dostępnych metod sprawia, że analityk, dokonując wyboru konkretnej metody, powinien przede wszystkim wziąć pod uwagę wynik, jaki chce osiągnąć, a ponadto czynniki takie jak: czas, sprzęt oraz przedmiot analizy.

**Tabela 1**. Zestawienie algorytmów konstruowania dendrogramów
*Summary of algorithms for constructing dendrograms*

| Metoda | Zalety | Wady |
|---|---|---|
| UPGMA | • bardzo prosta<br>• szybka | • wrażliwa na różne tempo ewolucji |
| Neighbor-Joining | • bardzo szybka<br>• akceptuje linie wykazujące różne tempo ewolucji | • daje tylko jedno możliwe drzewo<br>• mocno zależna od rodzaju zastosowanego modelu ewolucji |
| Maximal Parsimony | • jedyna w pełni kladystyczna metoda<br>• można identyfikować obszary problematyczne<br>• nie redukuje informacji z sekwencji<br>• sprawdza różnorodne drzewa | • bardzo powolna nawet dla niedużych sekwencji<br>• nie zakłada modelu ewolucji<br>• nie dostarcza informacji o długości gałęzi |
| Maximal Likelihood | • niższa wariancja (mniejszy wpływ błędu próby)<br>• dobre podstawy statystyczne<br>• sprawdza różne topologie | • bardzo wolna i wymaga dużej mocy komputera<br>• rezultaty zależą od zastosowanego modelu ewolucji |

## PIŚMIENNICTWO


1 Cavalli-Sforza L. L., Edwards A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet., 19, 233-257

2 Felsenstein J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol, 17, 368-376

3 Felsentein J. (2004). Inferring Phylogenies. Sunderland, MA: Sinauer Associates

4 Fitch W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst. Zool., 20, 406-416

5 Higgs P. G., Attwood T. K. (2005). Bioinformatics and Molecular Evolution. Oxford: Blackwell Science Ltd.

6 Jobling M. A., Hurles M., Tyler-Smith C. (2004). Human evolutionary genetics. Origins, Peoples & Disease. New York: Garland Science

7 Kishino H., Hasegawa M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol., 29, 170-179

8 Lewis P. O. (2001). Phylogenetic systematics turns over a new leaf. Trends. Ecol., 16 (1), 30-37

9 Murtagh F. (1984). Counting dendrograms: A survey. Discrete Appl. Math., 7, 191-199

10 Nei M. (1975). Molecular Population Genetics and Evolution. North Holland. Amsterdam & New York

11 Nei M. (1996). Phylogenetic analysis in molecular evolutionary genetics. Ann. Rev. Genet., 30, 371-403

12 Pevsner J. (2009). Bioinformatics and Functional Genomics. 2nd Edition. Hoboken: Willey-Blackwell

13 Posada D. (2009). Bioinformatics for DNA sequence analysis. Humana Press, New York

14 Rzhetsky A., Nei M. (1993). Theoretical foundation of the minimum – evolution method of phylogenetic inference. Mol. Biol. Evol., 10, 1073-1095

15 Saitou N., Nei M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol., 4, 406-425

16 Sokal R. R., Michener C. D. (1958). A statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin, 38, 1409-1438

17 Sokal R. R., Sneath P. H. A. (1963). Principles of Numerical Taxonomy. San Francisco: Freeman

18 Soltis P. S., Soltis D. E. (2003). Applying the Bootstrap in Phylogeny Reconstruction. Stat Scie, 18, 256-267

19 Steel M., Penny D. (2000). Parsimony, Likelihood, and the Role of Models in Molecular Phylogenetics. Mol. Biol. Evol., 7 (6), 839-850

20 Strimmer K., von Haeseler A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol., 13, 964-969