



Wykład 7



Porównanie wielu różnych sekwencji

TO JEST TASEKWENCJA

TAMTA JEST TEZ SEKWENCJA

ITOTEZJESTSEKWENCJA

TO JESTTA SEKWENCJA

| | | | | | | | | | | | | | |

TAMTA JEST TEZ SEKWENCJA

| | | | | | | | | | | | | | |

ITO JEST SEKWENCJA



Test bardzo wielu sekwencji

Q5E940_BOVIN	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_HUMAN	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_MOUSE	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_RAT	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_CHICK	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_RANSY	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--SALE	76
Q7ZUG3_BRARE	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0 ICTPU	-----HPREDRATWESNYFLKIIQLLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_DROME	-----HWRENKAAMEAQYFIKYVLEFDEFPKCFIVGADNVGSKOMQOIRMSLRGK-AVYLMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_DICDI	-----HSGAG-SKREKLFIEKATELFTTIDKMIYAEADYVGSQLOKIRKSIRGI-GAYLMGKKTMMRKAIRGHELENNH--PALE	75
Q54LP0_DICDI	-----HSGAG-SKRENVFIEKATELFTTIDKMIYAEADYVGSQLOKIRKSIRGI-GAYLMGKKTMMRKAIRGHELENNH--PALE	75
RLA0_PLAFL	-----HAKLSKQKQKQMYIEKLSLIDQYKILIVHVDNVGSHOMASYRKSIRGK-ATILMGKKTMMRKAIRGHELENNH--PALE	76
RLA0_SULAC	-----HIGLAYITTKKIAKWEYDVAELTSEKLTHTITIIANIEGFPADKLHEIEKKLRGK-ADIKVTKMHLFNIALKHAG-----VDYK	79
RLA0_SULTO	-----HRIMAYITQERKIAKWEIEVKELEKLRHTITIIANIEGFPADKLHEIEKKMIGM-AEIKVTKMHLFNIALKHAG-----LDVS	80
RLA0_SULSO	-----HKRLALALQKRYASWLEIEVKELEKLRHTITIIANIEGFPADKLHEIEKKLRGK-ATIEVTKMHLFNIALKHAG-----IDIE	80
RLA0_AERPE	MSVYSLVGMHYKREKPIPEWETLMLRELEELFSEKRYVFLADLTGTFVYVRYEKKLWKKYFMMYAKKRIILRAMEAAGLE---LDDN	86
RLA0_PYRAE	-----MMLAIGKREYVTRQYFARYKIVSEATELLQKRYVYFLFDLHGLSRIILHEIYRILRYGVIKIKVTLFKIAFTKYTG---IPAE	85
RLA0_METAC	-----MAEERHTEHIPQWKKDEIENIKELIQSKYVGHVGIIEGILATKMKIRREDLDV-AVLEKVSRTLLERALNQLG-----ETIP	78
RLA0_METHA	-----MAEERHTEHIPQWKKDEIENIKELIQSKYVGHVGIIEGILATKMKIRREDLDV-AVLEKVSRTLLERALNQLG-----ESIP	78
RLA0_ARCFU	-----MAAVRS---PPEYVRAVEEKEMISSKVVVAIVSFRNVFASOMKIRREFRGK-AEIKVVKETLLEBALDAG-----GDYL	75
RLA0_METKA	MAVKAKGQPPSGYEKVAEWKRRREVKELKLMDEYENYGLVDLEGIPAPQLQEIIRAKLRERDIIRMSRNTLHRIALEEKLDER---PELE	88
RLA0_METTH	-----MAHVAEWKKEEVQELNDLIGYEVVGIANLADIPARLOKMRQTLRDS-ALIRMSKETLISLALAKAGREL---ENVD	74
RLA0_METTL	-----MITAESHKIAFWKIEEVNKEKLLKNGQIYALVDHMEVPAVQLQEIIRAKLRERDIIRMSRNTLHRIALEEKLDER---PELE	82
RLA0_METVA	-----MIDAKSEHKIAFWKIEEVNKEKLLKNGQIYALVDHMEVPAVQLQEIIRAKLRERDIIRMSRNTLHRIALEEKLDER---PELE	82
RLA0_METJA	-----METKYKAVYAPWIEEVNKEKLLKNGQIYALVDHMEVPAVQLQEIIRAKLRERDIIRMSRNTLHRIALEEKLDER---PELE	81
RLA0_PYRAB	-----MAHVAEWKKEEVQELNDLIGYEVVGIANLADIPARLOKMRQTLRDS-ALIRMSKETLISLALAKAGREL---ENVD	77
RLA0_PYRBO	-----MAHVAEWKKEEVQELNDLIGYEVVGIANLADIPARLOKMRQTLRDS-ALIRMSKETLISLALAKAGREL---ENVD	77
RLA0_PYRFU	-----MAHVAEWKKEEVQELNDLIGYEVVGIANLADIPARLOKMRQTLRDS-ALIRMSKETLISLALAKAGREL---ENVD	77
RLA0_PYRKO	-----MAHVAEWKKEEVQELNDLIGYEVVGIANLADIPARLOKMRQTLRDS-ALIRMSKETLISLALAKAGREL---ENVD	76
RLA0_HALMA	MSAESERKTETIPEWQEQEVDIVMIESYESVGVVNIAGIPKQLODMRRDLHGT-AELRVSRTLLERALDDVD-----DGLE	79
RLA0_HALVO	MSSEVVRQTEVIPQWKRREVDDELVDYIESYESVGVVGVAGIPKQLODMRRDLHGT-AELRVSRTLLERALDDVD-----DGLE	79
RLA0_HALSA	MSAEEQRTEEVPEWQEQEVDIVMIESYESVGVVNIAGIPKQLODMRRDLHGT-AELRVSRTLLERALDDVD-----DGLE	79
RLA0_THEAC	-----MKEYSQQKKELVNEITRIKASRSVAIVDTAGIRTRQIDDEGKIRGK-INLEKIKETLLFKALENLGD-----EKLS	72
RLA0_THEVO	-----HRKINPKKKEIVSELAADITKSKAVAIVDIKGVKIRMODIRAKNRDKYKIKVVKETLLFKALENLGD-----EKLT	72
RLA0_FICTO	-----MTEPAQWEIDFVKNELEINSRKVAIVSIIKGLRHHFKQIEMSIRDK-ARIEVSRARLLRLAIENTGK-----NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

autor: Miguel Andrade, http://en.wikipedia.org/wiki/Multiple_sequence_alignment



Dopasowanie wielu sekwencji (Multiple Sequence Alignment)

- Bardziej wiarygodne
 - rozstrzyga sytuacje niejednoznaczne dla dwóch sekwencji
 - wskazuje regiony o dużym podobieństwie
- Zastosowania
 - poszukiwanie wzorców w danej rodzinie białek
 - tworzenie drzew filogenetycznych
 - wykrywanie homologii nowej sekwencji
 - przewidywanie struktury przestrzennej nowej sekwencji



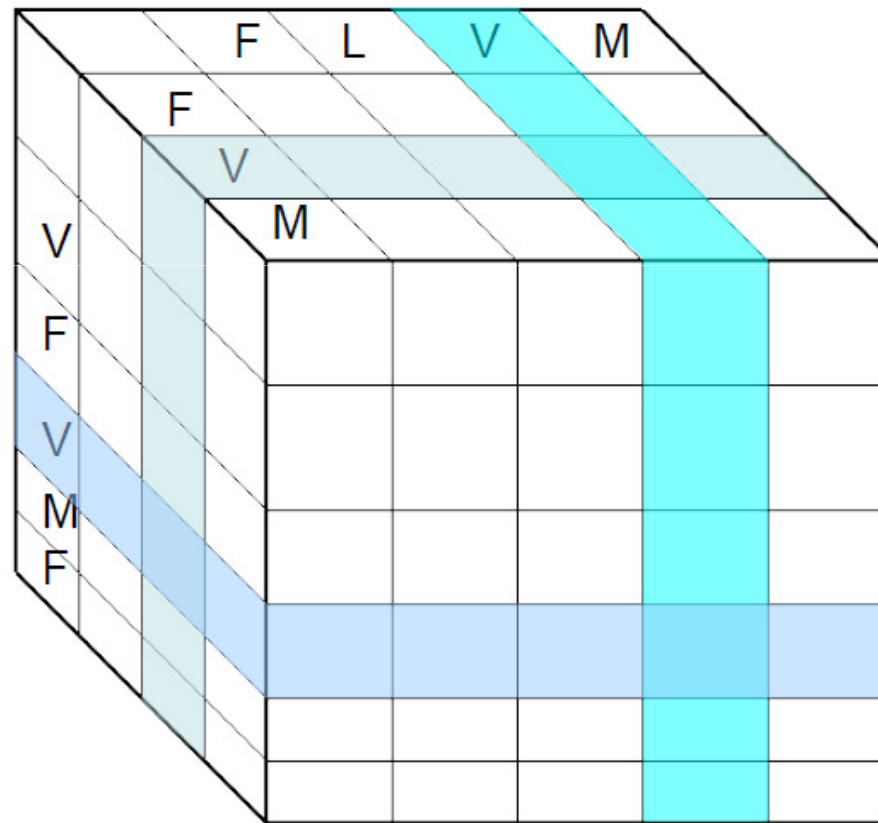
MSA – jak to zrobić?

- Programowanie dynamiczne w n wymiarach?
np. 3 sekwencje (macierz 3-wym.)

FLVM
FVM
VFVMF

- Możliwe ruchy:

X
Y
X+Y
Z
Y+Z
X+Z
X+Y+Z



Niepraktyczne, spotka nas przekleństwo wymiarowości!



Metody

- Programowanie dynamiczne
- Dopasowanie hierarchiczne (klastrowanie)
- Ukryte modele Markowa (Hidden Markov Models – HMM)
- Metody uczenia maszynowego
- Algorytmy genetyczne
- Metody wykorzystujące wiedzę filogenetyczną



Klastering czyli uczenie bez nadzoru

Odległość pomiędzy elementami

Pomiędzy elementami zbioru danych x (wektor p -cech) wyznaczane są wartości funkcji podobieństw lub niepodobieństwa d (dissimilarity; częściej)

$$D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_{ij}, x_{i'j})$$

Najczęściej jest to kwadrat odległości:

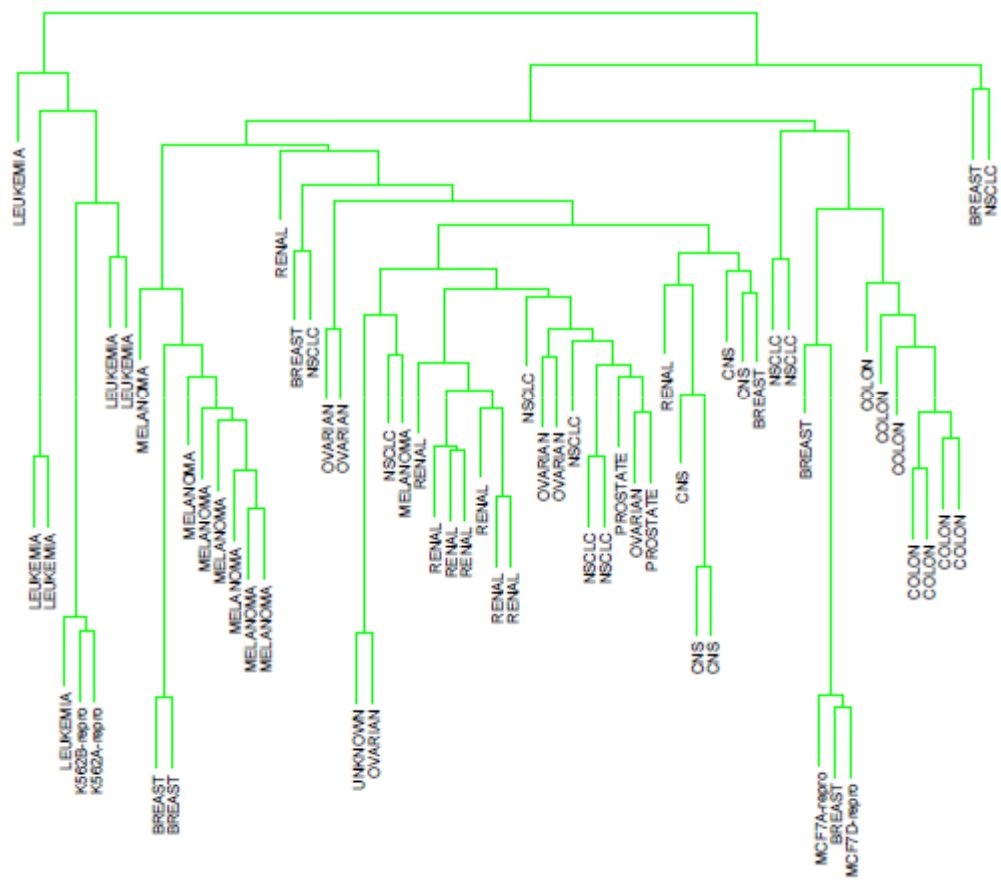
$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

Jeżeli wpływ cech jest niezrównoważony to można zastosować sumę ważoną z cech (ale $w_j = 1$ nie oznacza jednakowego wpływu bo zależy od rozkładu pomiędzy cechami):

$$D(x_i, x_{i'}) = \sum_{j=1}^p w_j \cdot d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^p w_j = 1$$



Z góry do dołu (klastering)





Klasteryzacja k-średnich

Inicjalizacja:

$m = 0$ % nr iteracji,

$D_m = \infty$ % średni błąd kwantyzacji w m-tej iteracji)

Wybierz N początkowych wektorów kodujących r_i

Określ maksymalny błąd kwantyzacji ε

Iteracja:

* Dopóki $D_m > \varepsilon$:

Podziel wektory danych na N grup. Wektor x_j jest przypisywany do i -tej grupy, gdy $d(x_j, r_i) \leq d(x_j, r_k)$ dla wszystkich pozostałych centroidów r_k

Wyznacz nowe centroidy r_i

Wyznacz średni błąd kwantyzacji D_m i sprawdź czy $D_m < \varepsilon$

Jeśli nie jest: $m=i+1$ i wróć do *



Przykład

1. Zbiór danych o dwóch cechach:
np. $D = \{(1,12), (17,1), (2,14), (5,17), (12,3), (20,2)\}$
2. Decydujemy się na podział na K (np. 2) klastry
(wybór K istotny problem, będzie omówiony w przyszłości)
3. Definiujemy co to jest odległość pomiędzy obiektami
4. Losujemy (lub inaczej wybieramy) K centroidów:
np. $c1 = \{12,3\}$, $c2 = \{20,2\}$
5. Obliczamy w kolejnych iteracjach nowe centroidy:
Znajdujemy punkty, które są bliższe $c1$ ($c2$) i obliczamy średnią ze *wszystkich* tych punktów (albo punkt z D najbliższy tej średniej). Ten punkt ustanawiamy nowym centroidem $c1$ ($c2$)
6. Powtarzamy (4) aż zaobserwujemy zbieganie obliczeń w kolejnych krokach iteracji



Ewolucja wyniku w algorytmie K -średnich

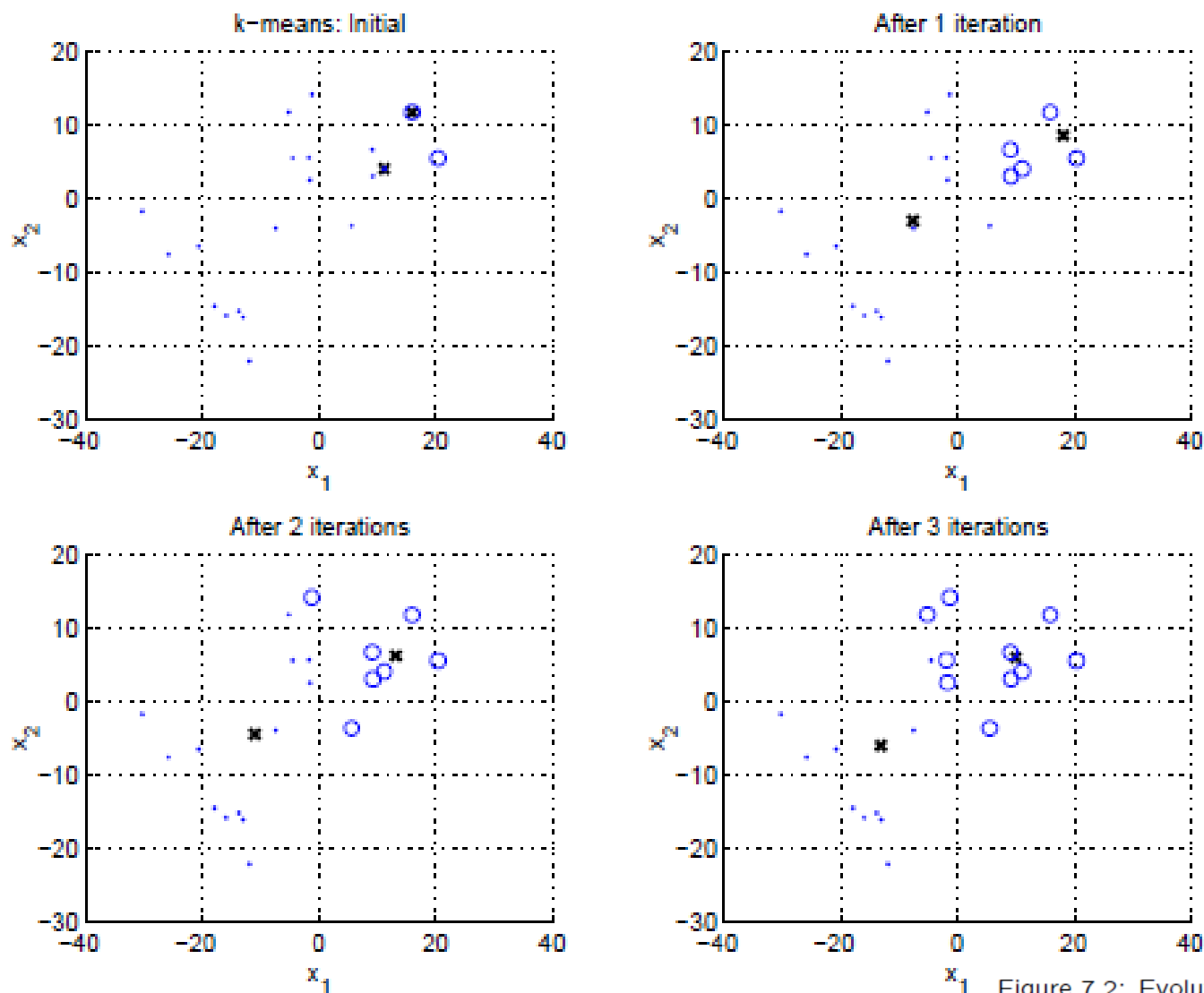


Figure 7.2: Evolution of k -means. Crosses indicate center positions. Data points are marked depending on the closest center. From: E. Alpaydm. 2004.



Wpływ pre-processingu w metodzie k-średnich

Standaryzacja wg. cechy

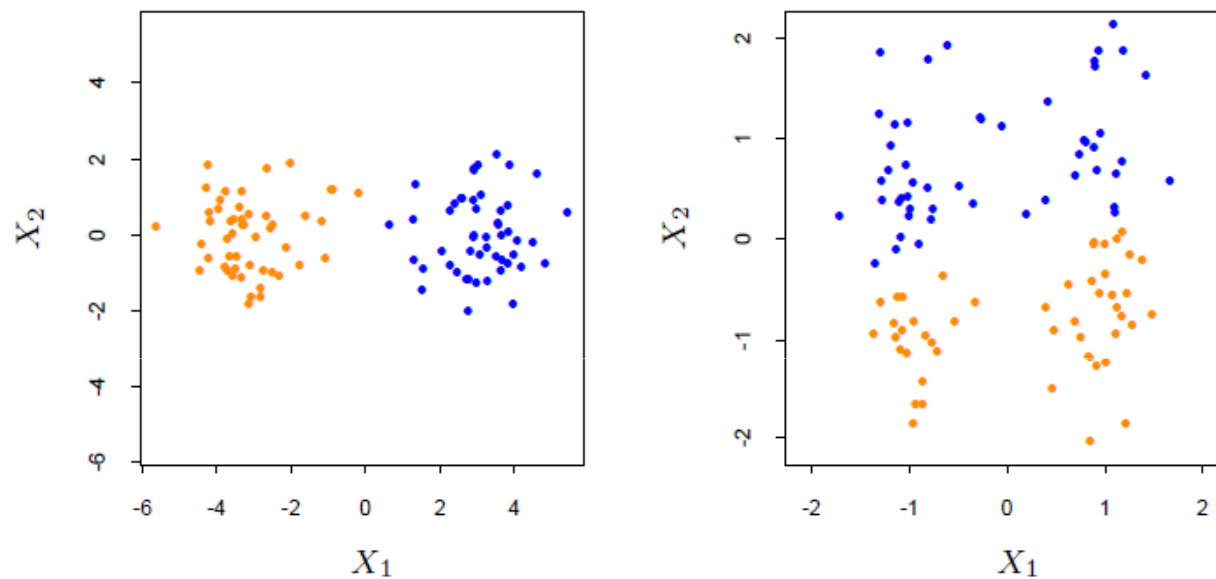


FIGURE 14.5. Simulated data: on the left, K -means clustering (with $K=2$) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights $1/[2 \cdot \text{var}(X_j)]$. The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.



<https://www.youtube.com/watch?v=BVFG7fd1H30>



Z dołu-do-góry („bottom-up”)

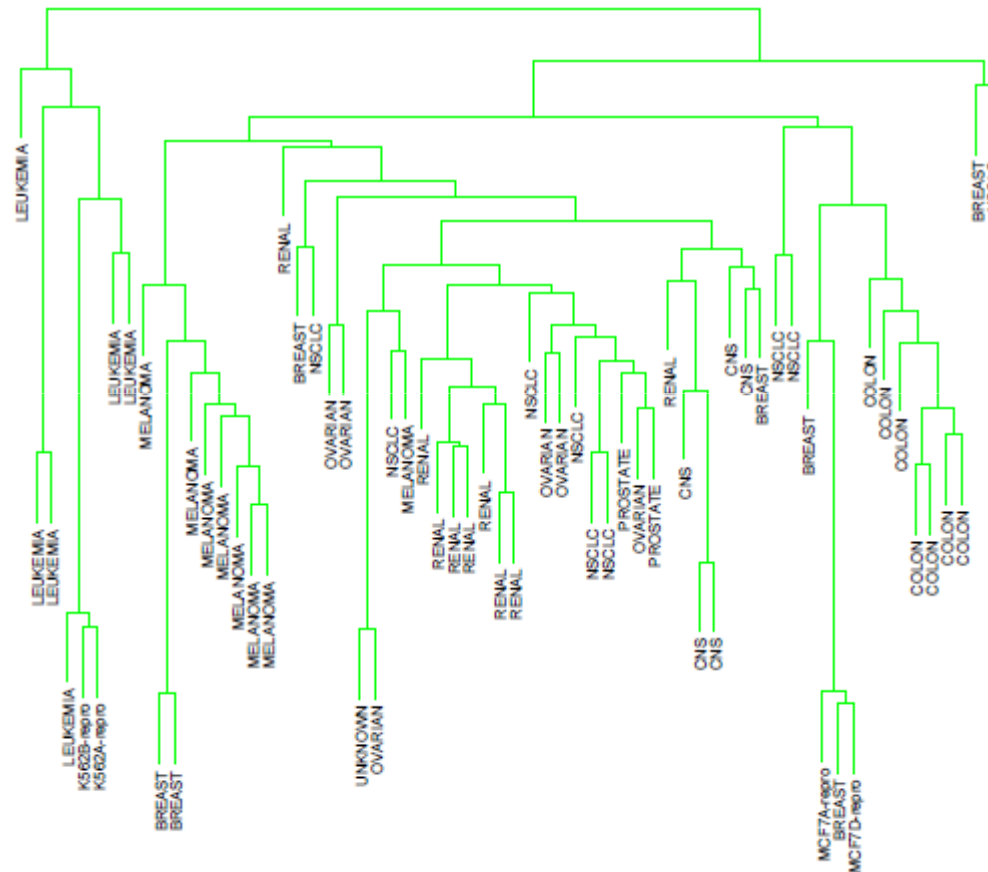


FIGURE 14.12. Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.



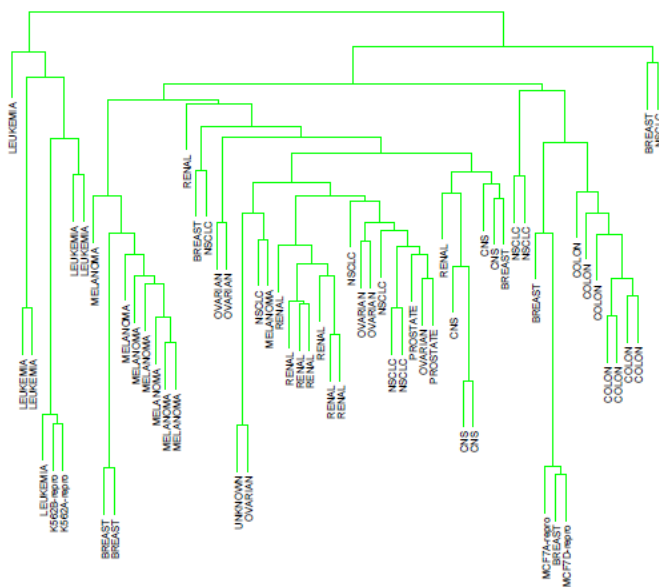
M_d	x^1	x^2	x^3	x^4	x^5
x^1	0	11	8	9	8
x^2	11	0	13	14	13
x^3	8	13	0	9	8
x^4	9	14	9	0	9
x^5	8	13	8	9	0



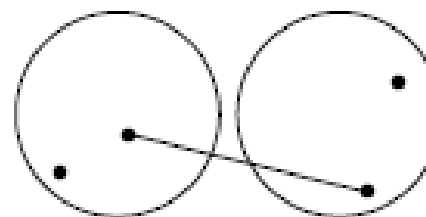
Metody scalania w dendrogramie

Sposób wybierania reprezentacji podgrupy

- **Pojedynczego łączenia (single linkage)** – maksimum podobieństwa



$$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

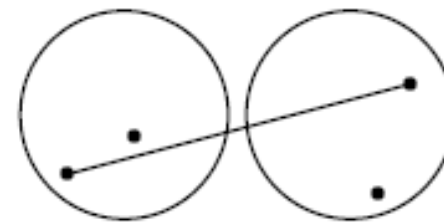
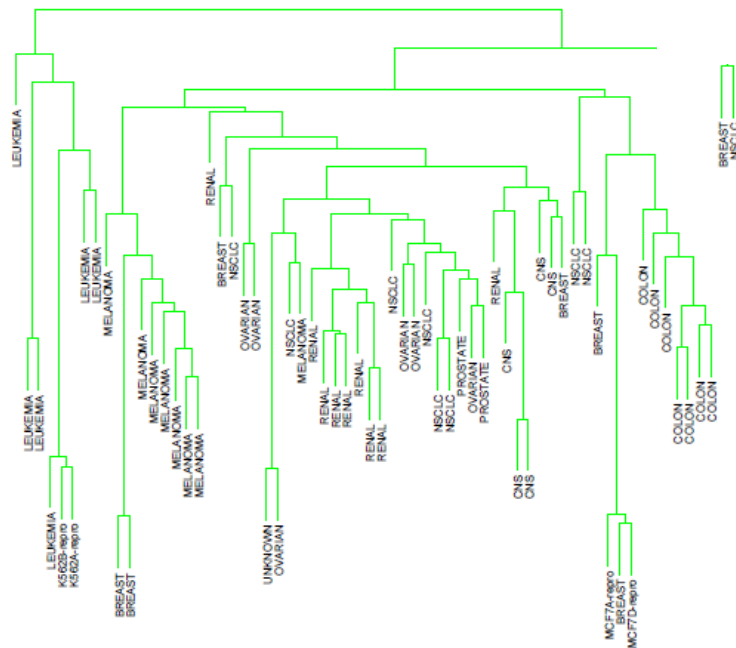




Metody scalania w dendrogramie

- **Pełnego łączenia (complete linkage)** – minimum podobieństwa

$$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}$$

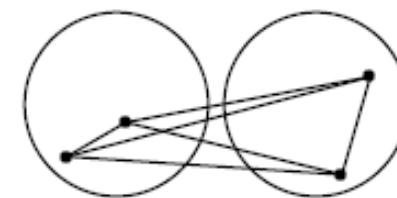
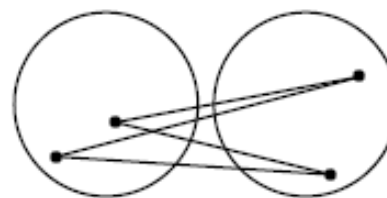
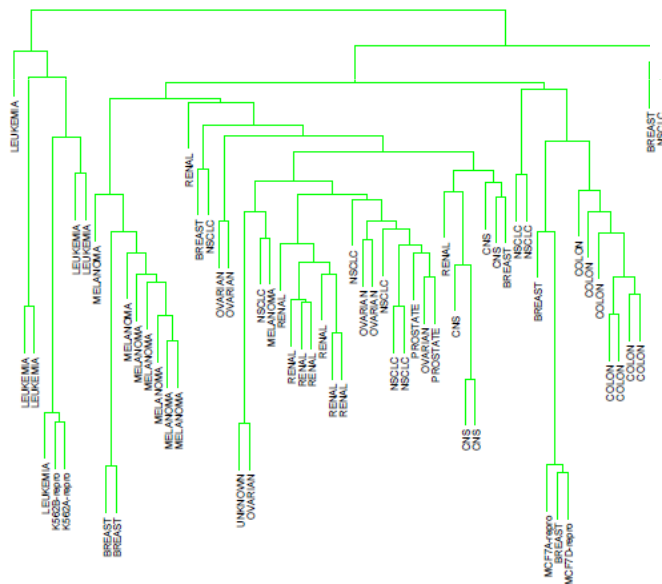




Metody scalania w dendrogramie

- Średniego podobieństwa w grupie (average linkage)

$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

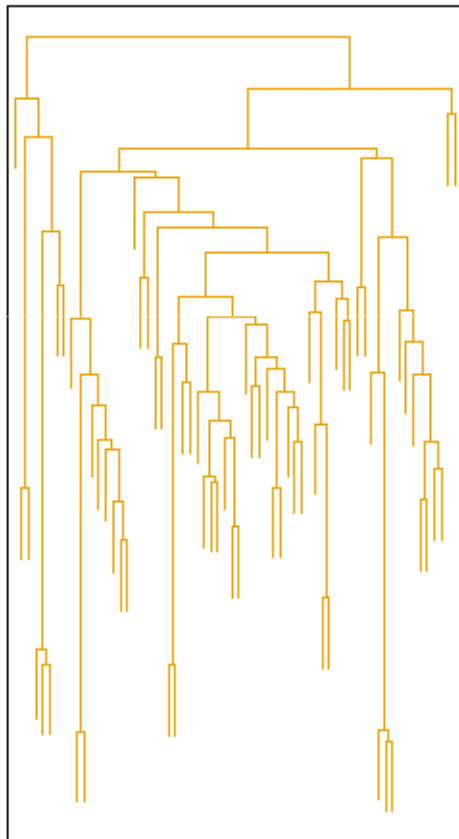


centroid: average inter-similarity (d) group-average: average of all similarities

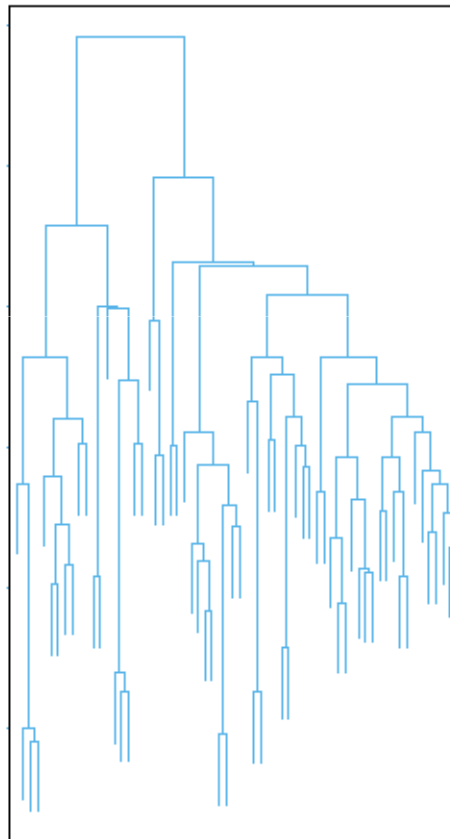


Porównanie dendrogramów

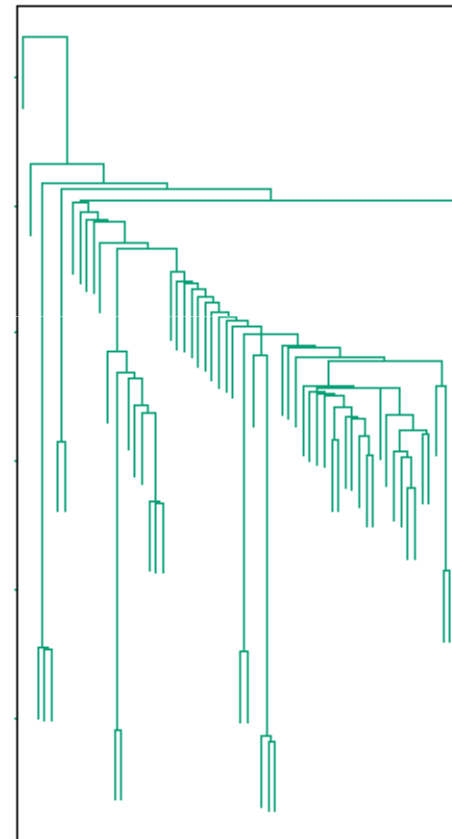
Average Linkage



Complete Linkage



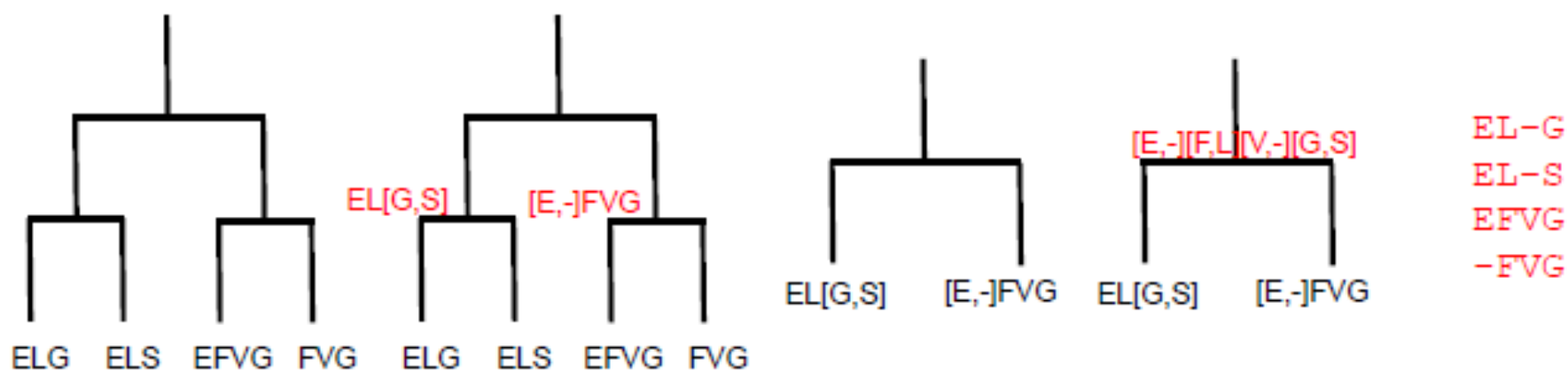
Single Linkage





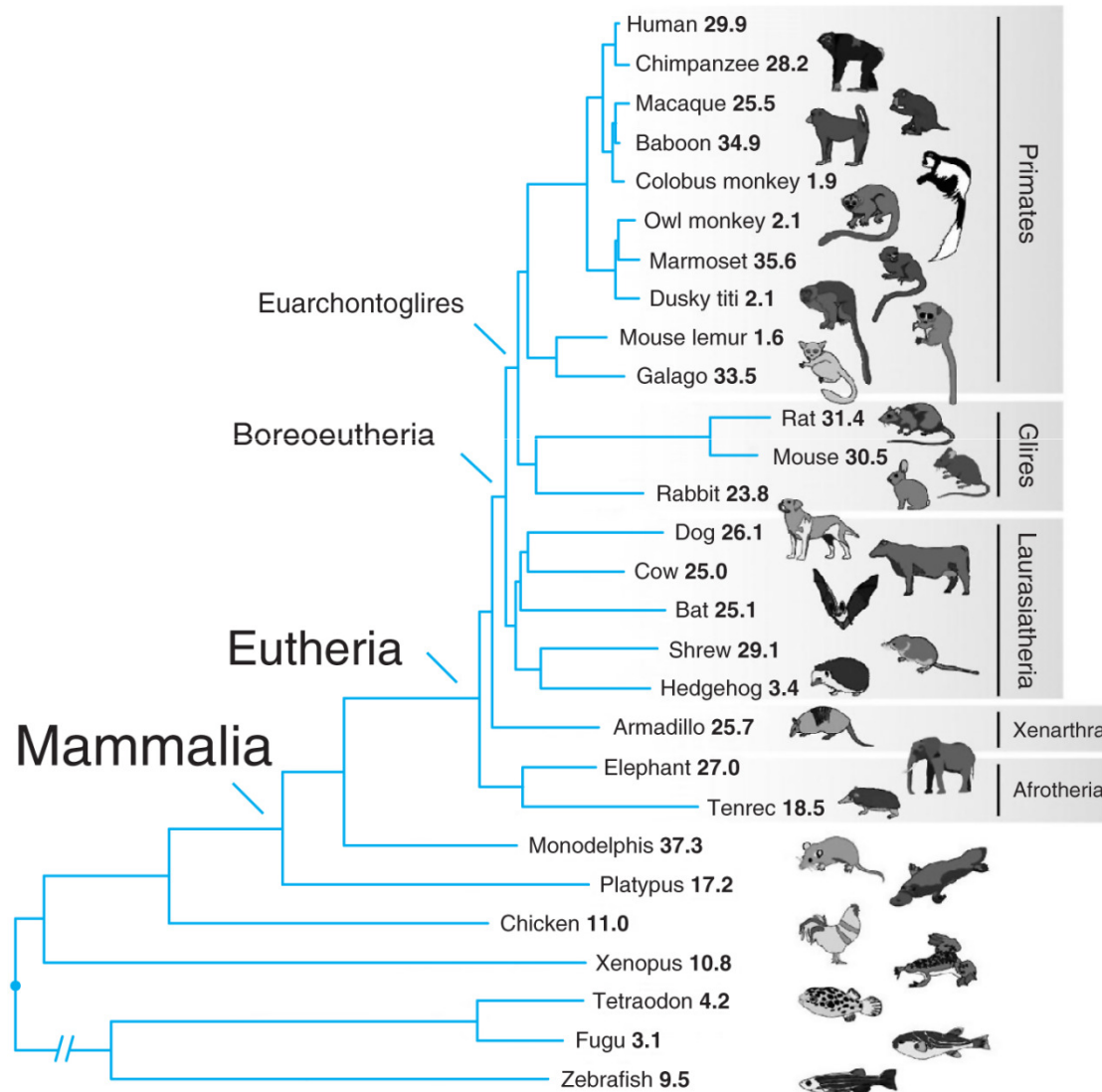
Łączenie dopasowań par sekwencji *ClustalW*

- Łącz dopasowania dwóch sekwencji wg drzewa podobieństwa
 - zaczynając od najbardziej podobnych (od liści do korzenia drzewa, *bottom-up*)
 - programowanie dynamiczne
 - raz utworzona przerwa zostaje
 - sumy częściowe w każdym stanie uśrednione (z Wągami zależnymi od podobieństwa)





Drzewo filogenetyczne

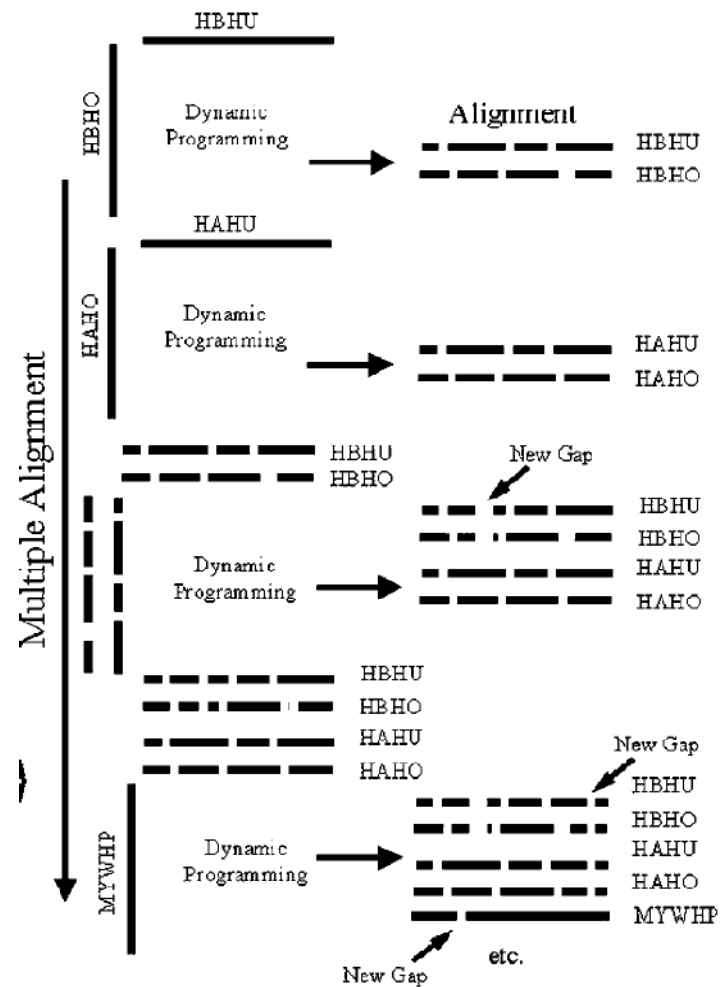
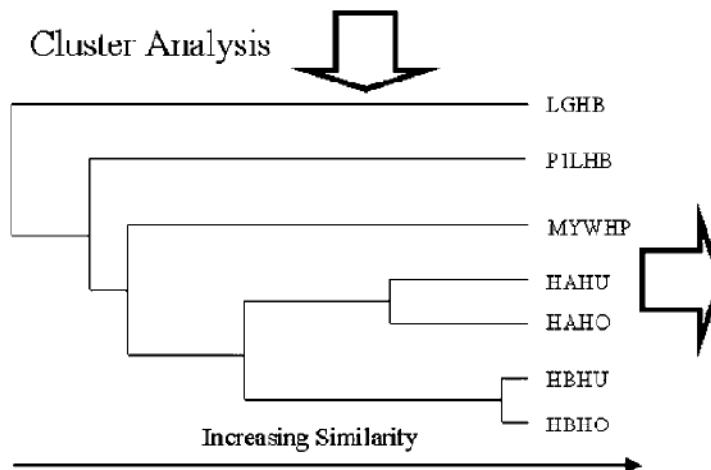




Dopasowanie progresywne - podsumowanie

	HAHU	HBHU	HAHO	HBHO	MYWHP	PILHB	LGHB
HAHU							
HBHU	21.1						
HAHO	32.9	19.7					
HBHO	20.7	39.0	20.4				
MYWHP	11.0	9.8	10.3	9.7			
PILHB	9.3	8.6	9.6	8.4	7.0		
LGHB	7.1	7.3	7.5	7.4	7.3	4.3	

Cluster Analysis





EFEKTY

Dopasowanie wielokrotne poprawia dokładność dopasowania sekwencji o niskim podobieństwie.

Metody hierarchiczne nie dają gwarancji znalezienia jednego optymalnego dopasowania dla całego zestawu sekwencji



Narzędzia on-line

Clustal Omega

<http://www.ebi.ac.uk/Tools/msa/clustalo/>

Wizualizacja MSA

https://www.ncbi.nlm.nih.gov/projects/msaviewer/?appname=ncbi_multialign&openuploaddialog