Integracja dużych zbiorów danych za pomocą sieci korelacyjnych

Aneta Sawikowska

XENSTATS

1) Xenstats Sp. z o.o.

2) Wydział Metod Matematycznych i Statystycznych,
Uniwersytet Przyrodniczy w Poznaniu,

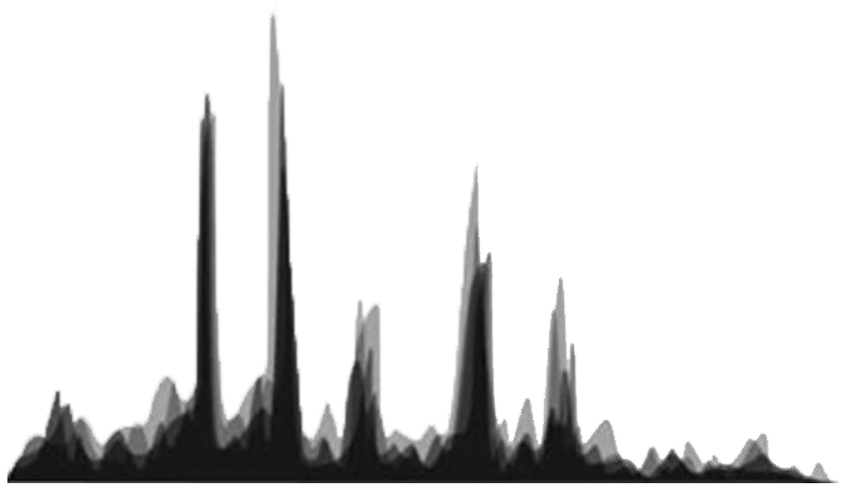
3) Instytut Genetyki Roślin,
Polska Akademia Nauk

STWUR 26.11.2019

Plan prezentacji

1. Dane, dla których tworzy się sieci.
2. Wieloetapowe przetwarzanie danych.
3. Analiza statystyczna doświadczenia wieloczynnikowego.
4. Analiza sieci korelacyjnych.
5. Wizualizacja sieci.
6. Omówienie pakietu WGCNA na przykładowym skrypcie.
7. Podsumowanie

Dane:

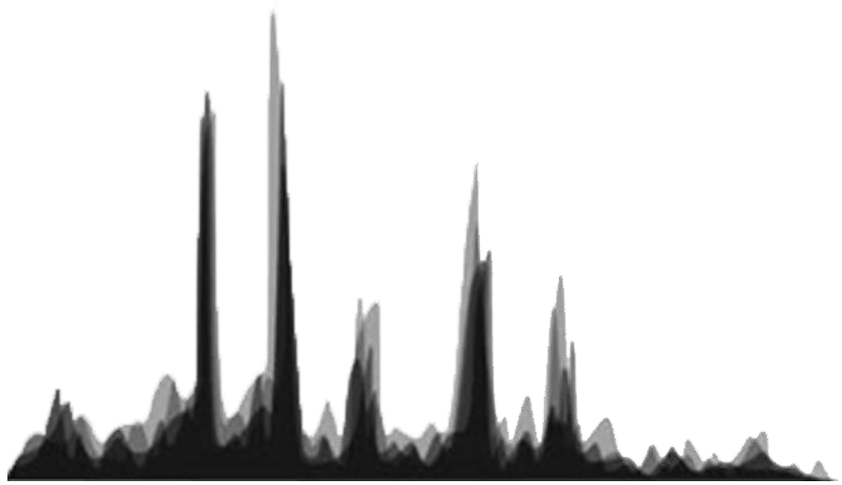


- metabolity pierwotne
- metabolity wtórne
- białka

w liściach jęczmienia pod wpływem suszy.

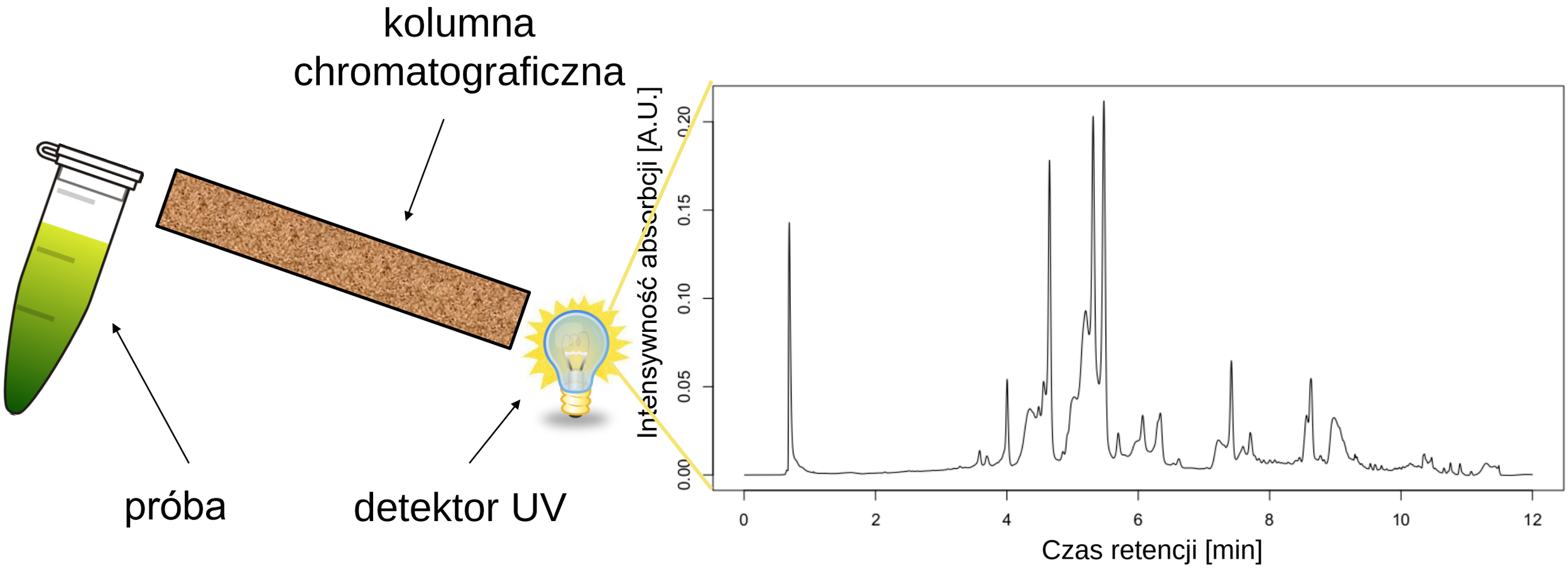
Pomiary metabolitów pierwotnych wykonano za pomocą chromatografu gazowego połączonego ze spektrometrem mas, metabolitów wtórnych za pomocą ultrasprawnego chromatografu cieczowego z detektorem UV oraz białek za pomocą dwukierunkowej elektroforezy.

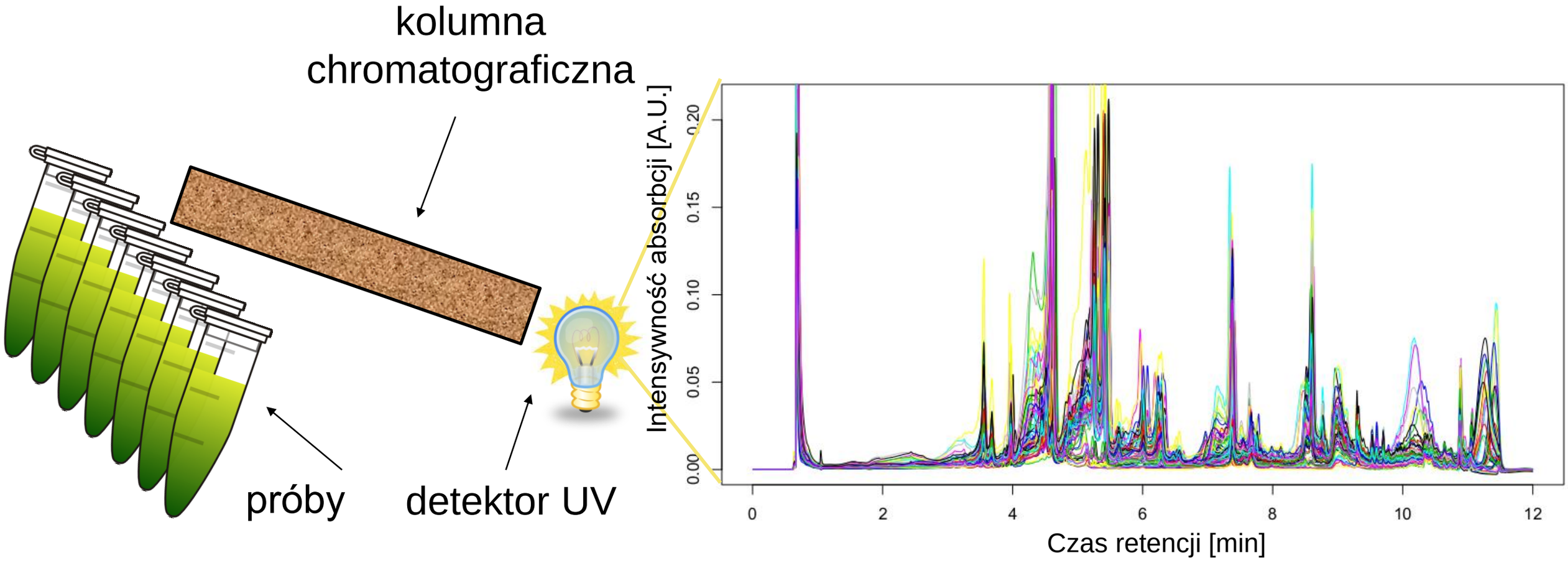
Dane:



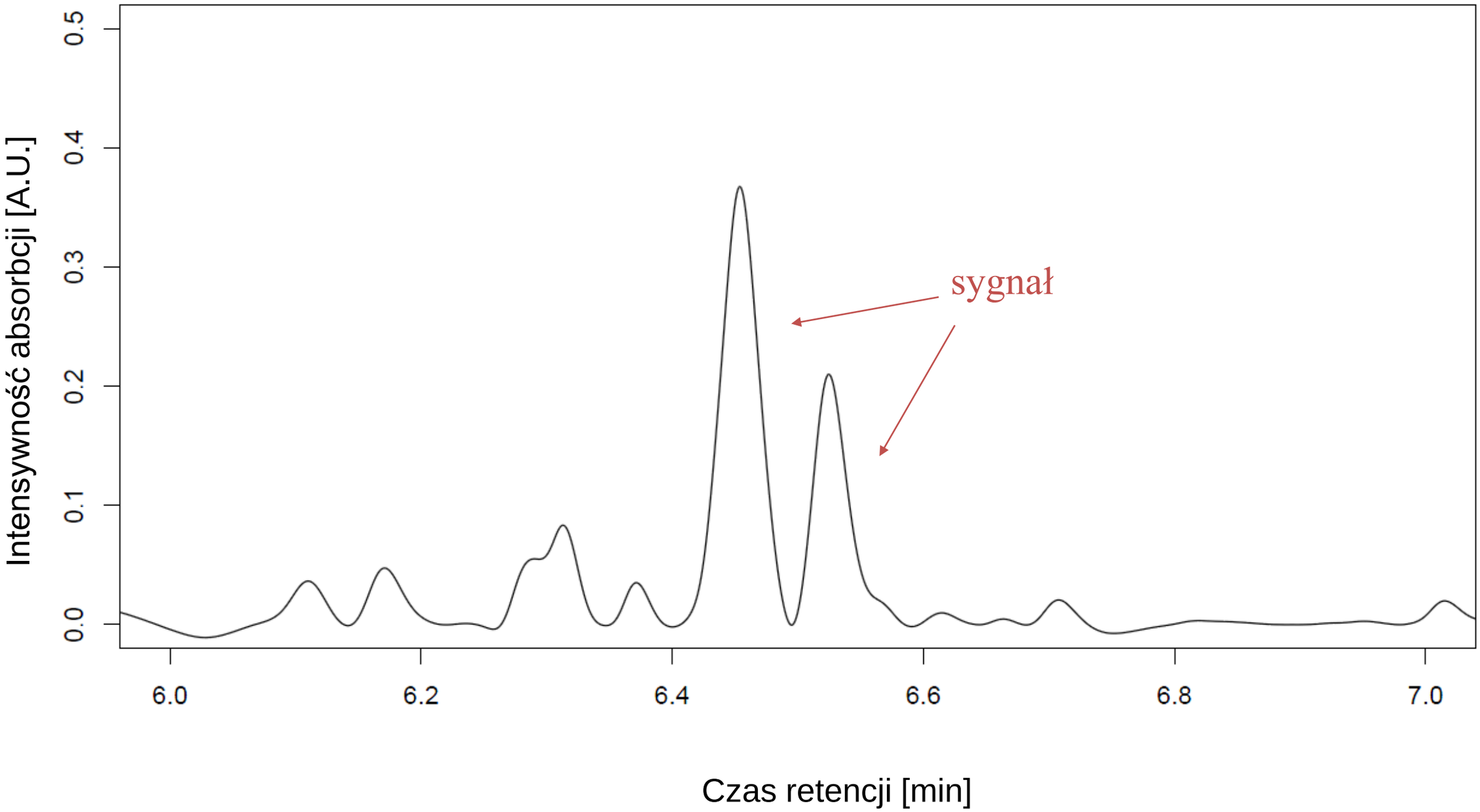
- 100 linii jęczmienia
- kontrola i susza
- 2 stadia rozwojowe rośliny (2 punkty czasowe)
- 3 powtórzenia biologiczne

15 mln obserwacji – metabolity wtórne


kolumna
chromatograficzna
próba
detektor UV
Intensywność absorbcji [A.U.]
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
Czas retencji [min]


kolumna
chromatograficzna
próby
detektor UV
Intensywność absorbcji [A.U.]
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
Czas retencji [min]

Przykładowy chromatogram uzyskany metodą UPLC/UV



Maresi λ=330 nm

NORMALIZACJA

a)

Intensywność absorbcji [A.U.]

0.0 0.2 0.4 0.6 0.8

USUNIĘCIE LINII BAZOWEJ

b)

Zróżnicowana intensywność absorbcji [A.U.]

−0.02 0.00 0.02

WYRÓWNANIE CZASU RETENCJI

c)

Zróżnicowana intensywność absorbcji [A.U.]

−0.02 0.00 0.02

DETEKCJA SYGNAŁÓW

d)

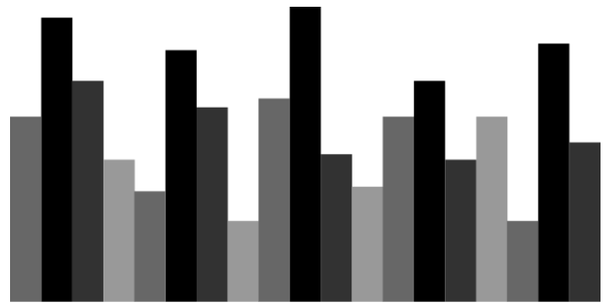
Wartość zintegrowanego sygnału

0.0 0.5 1.0

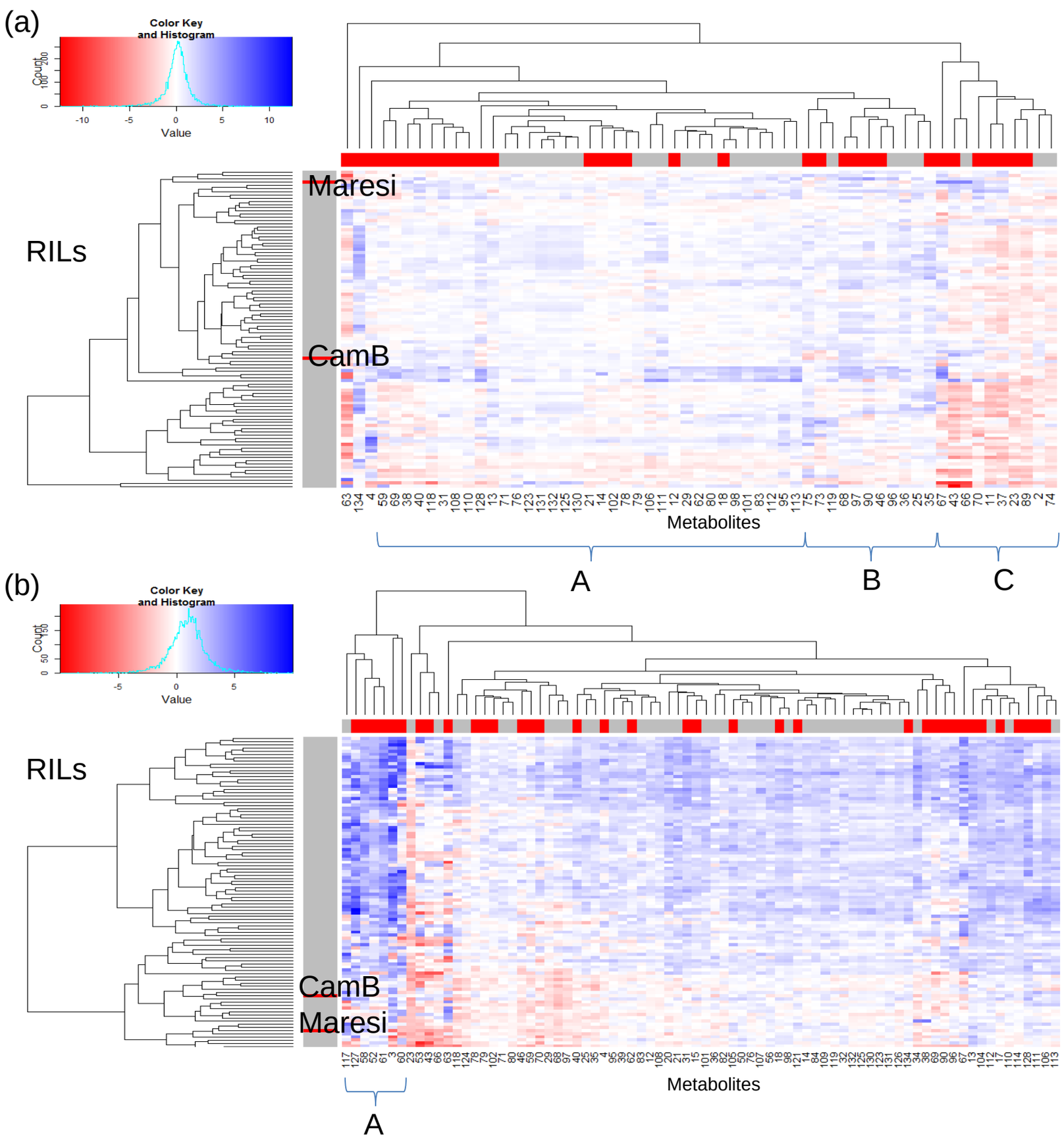
2 4 6 8 10

Czas retencji [min]

Analiza statystyczna:

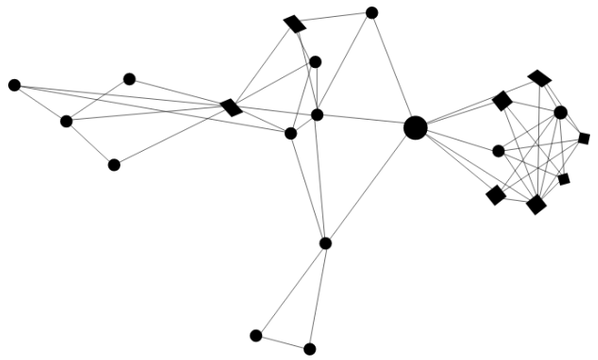


1. Logarytmowanie danych.
2. 3-czynnikowa analiza wariancji;
   pakiety: agricolae, lme4.
3. Selekcja sygnałów wykazujących istotne efekty za pomocą testu F z poprawką Bonferroniego;
   funkcja: var.test.
4. Grupowanie hierarchiczne, mapy ciepła, analiza zmienności i analiza korelacji na podstawie istotnych sygnałów;
   pakiety: pvclust, gplots, funkcje: hclust, pvclust, heatmap.2.



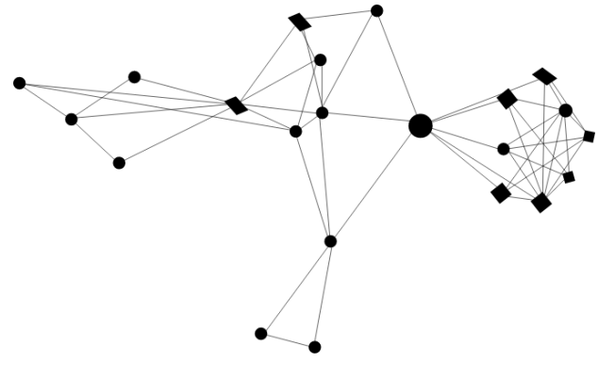
Mapy ciepła przedstawiające efekt suszy dla wszystkich linii i wszystkich metabolitów wykazujących istotny wpływ interakcji linia × traktowanie suszą, zaobserwowane w punkcie czasowym (a) T1 (b) T2.

Analiza sieci korelacyjnych:



1. Konstrukcja sieci korelacyjnej poprzez pakiet WGCNA:
   - wyznaczenie macierzy korelacji,
   - podniesienie do potęgi bezwzględnej wartości macierzy korelacji,
   - wyznaczenie TOM (*topological overlap matrix*),
   - detekcja modułów poprzez grupowanie hierarchiczne oraz algorytm *dynamic tree cut*,
   - znalezienie *hubów*, czyli wierzchołków z największą liczbą połączeń, najbardziej skorelowanych cech z innymi cechami.

Analiza sieci korelacyjnych:



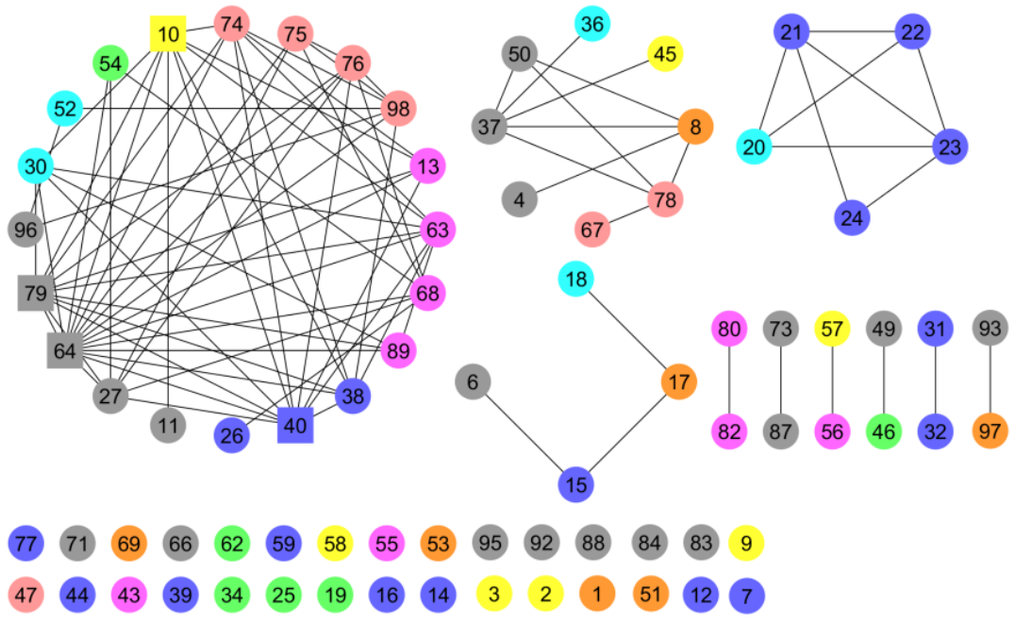
2. Konstrukcja sieci różnicowej na podstawie testu opartego na transformacji Fishera Z z poprawką Bonferroniego, $p<0.01/z$, gdzie z oznacza liczbę wszystkich par metabolitów.
3. Wizualizacja sieci za pomocą programu Cytoscape;

   R:

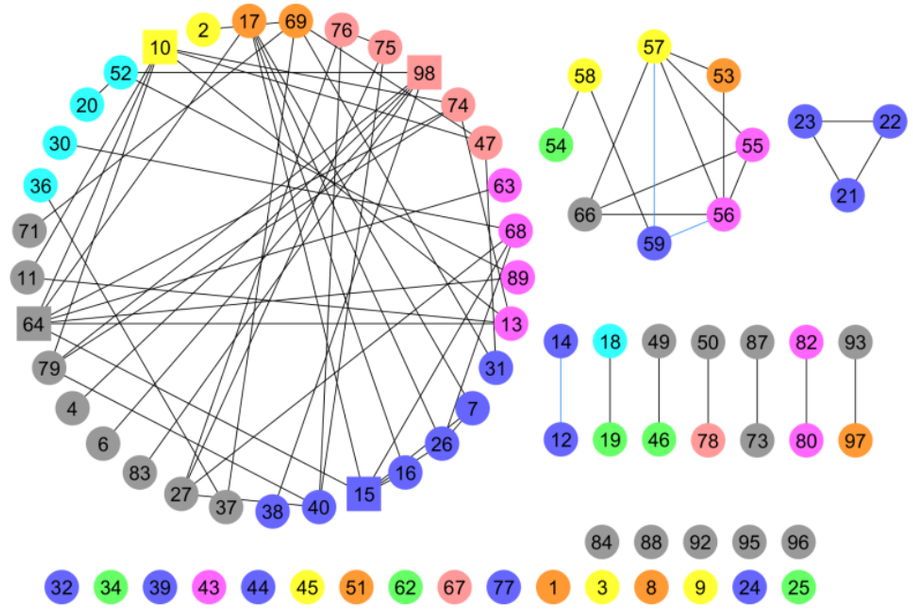
   pakiet: corrr, funkcja: network_plot;

   pakiet: qgraph, funkcja: qgraph.

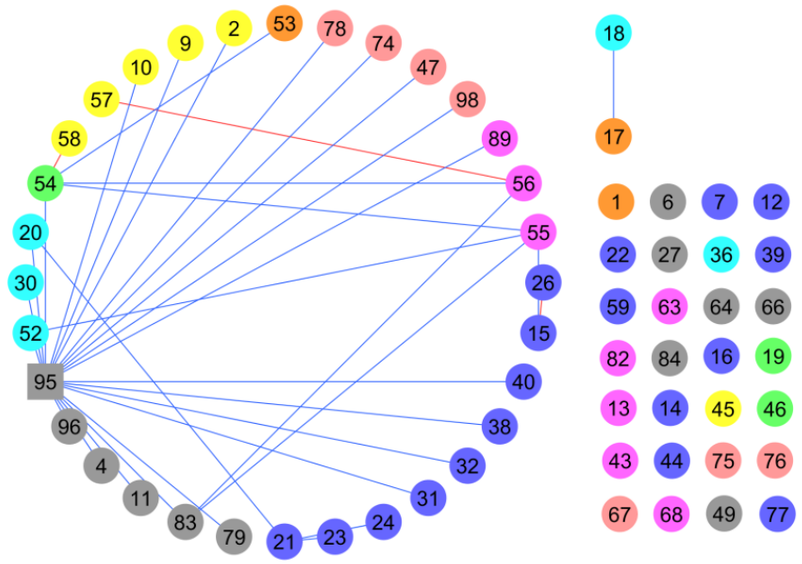
(a) **Metabolity pierwotne, kontrola**



(b) **Metabolity pierwotne, susza**



(c) **Metabolity pierwotne, kontrola vs susza**

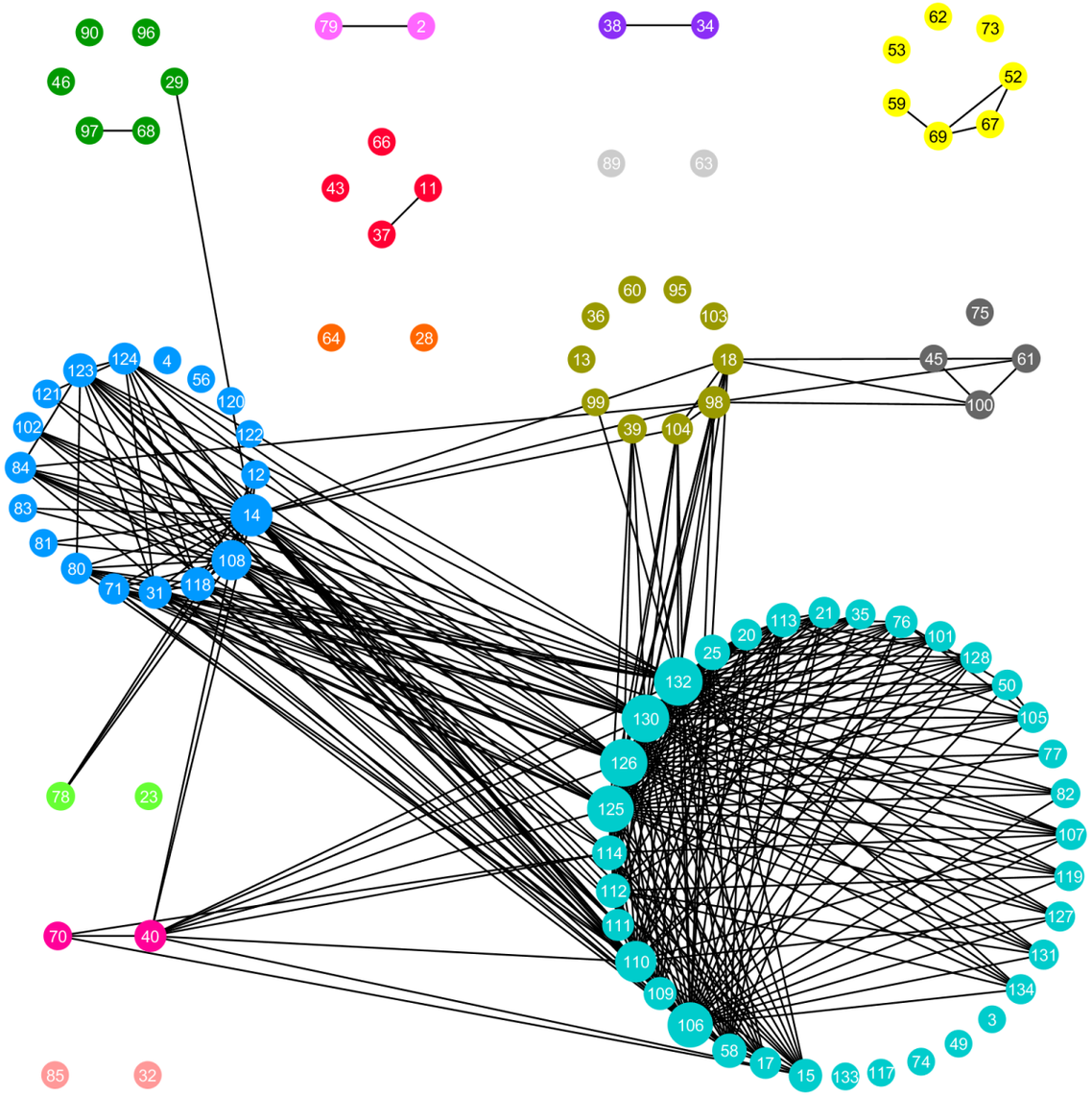


Grupy metabolitów pierwotnych

- amino acids
- carbohydrates
- lipids
- other carboxylic acids
- other nitrogen compounds
- sugar acids
- TCA
- unknown-other

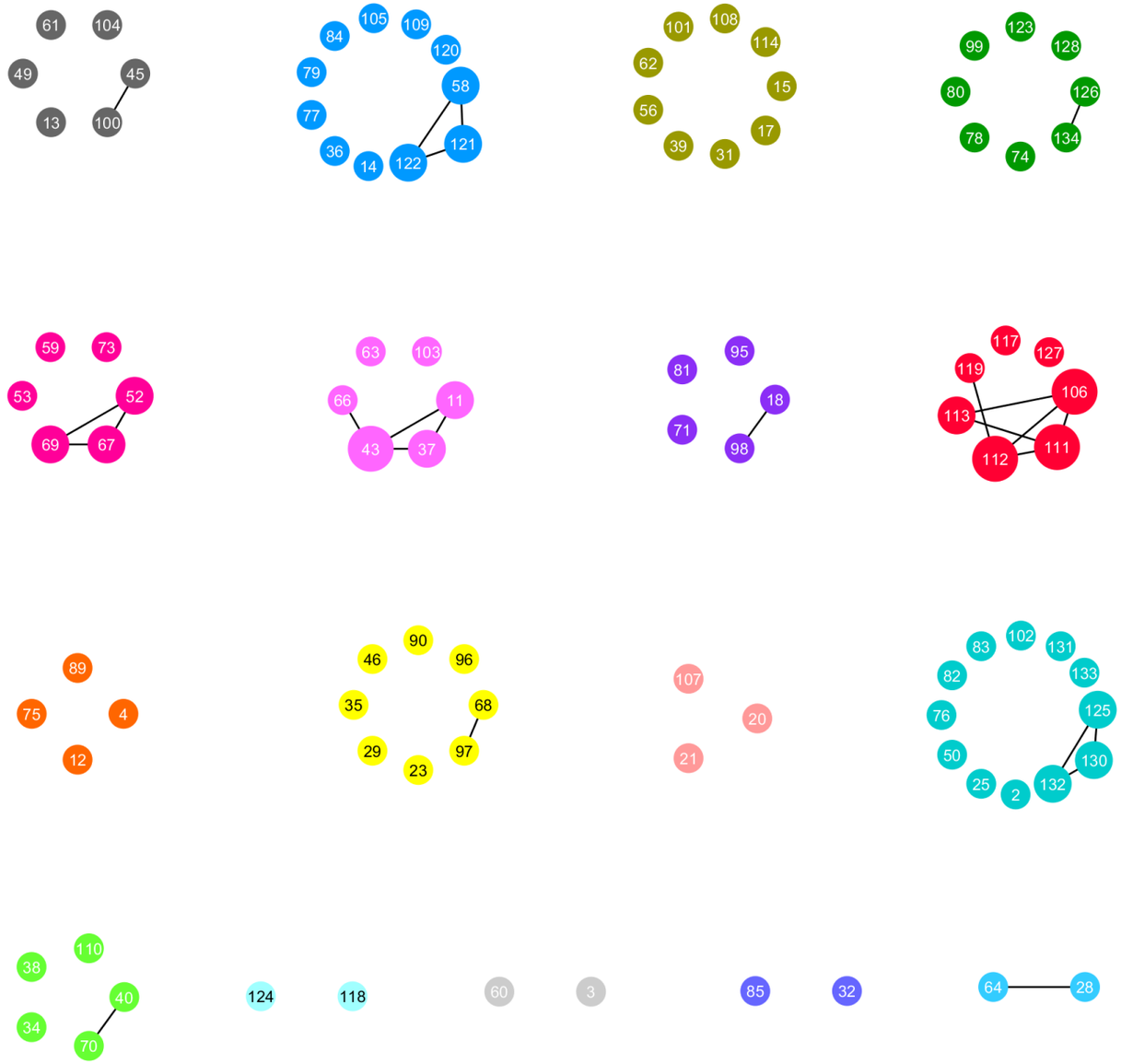
(a)

Metabolity wtórne, kontrola, punkt czasowy T2



(b)

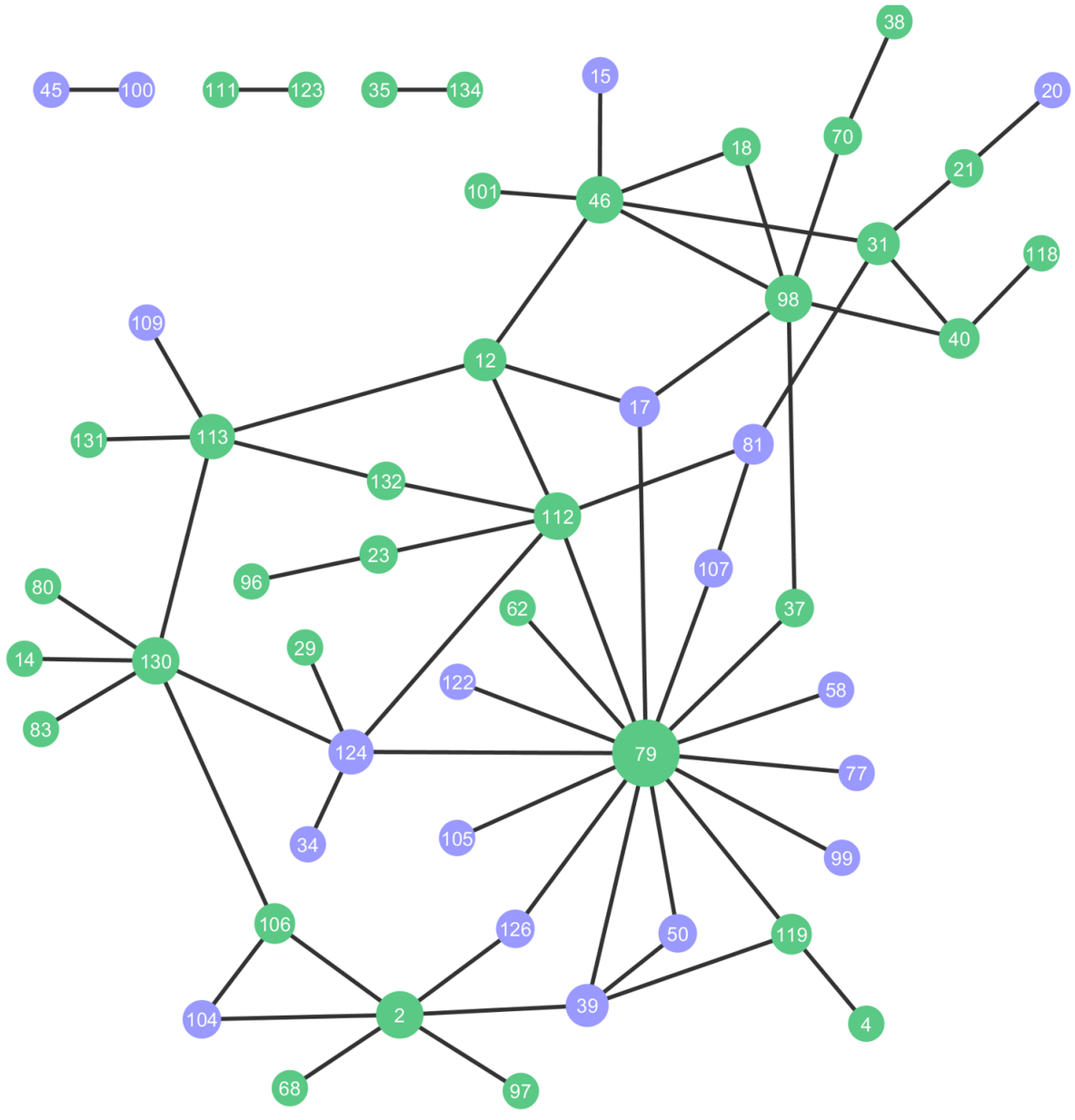
Metabolity wtórne, susza, punkt czasowy T2



Pokazane zostały krawędzie, które odpowiadają elementom z macierzy TOM o wartości większej niż 0.2. Wszystkie krawędzie odpowiadają korelacjom dodatnim.

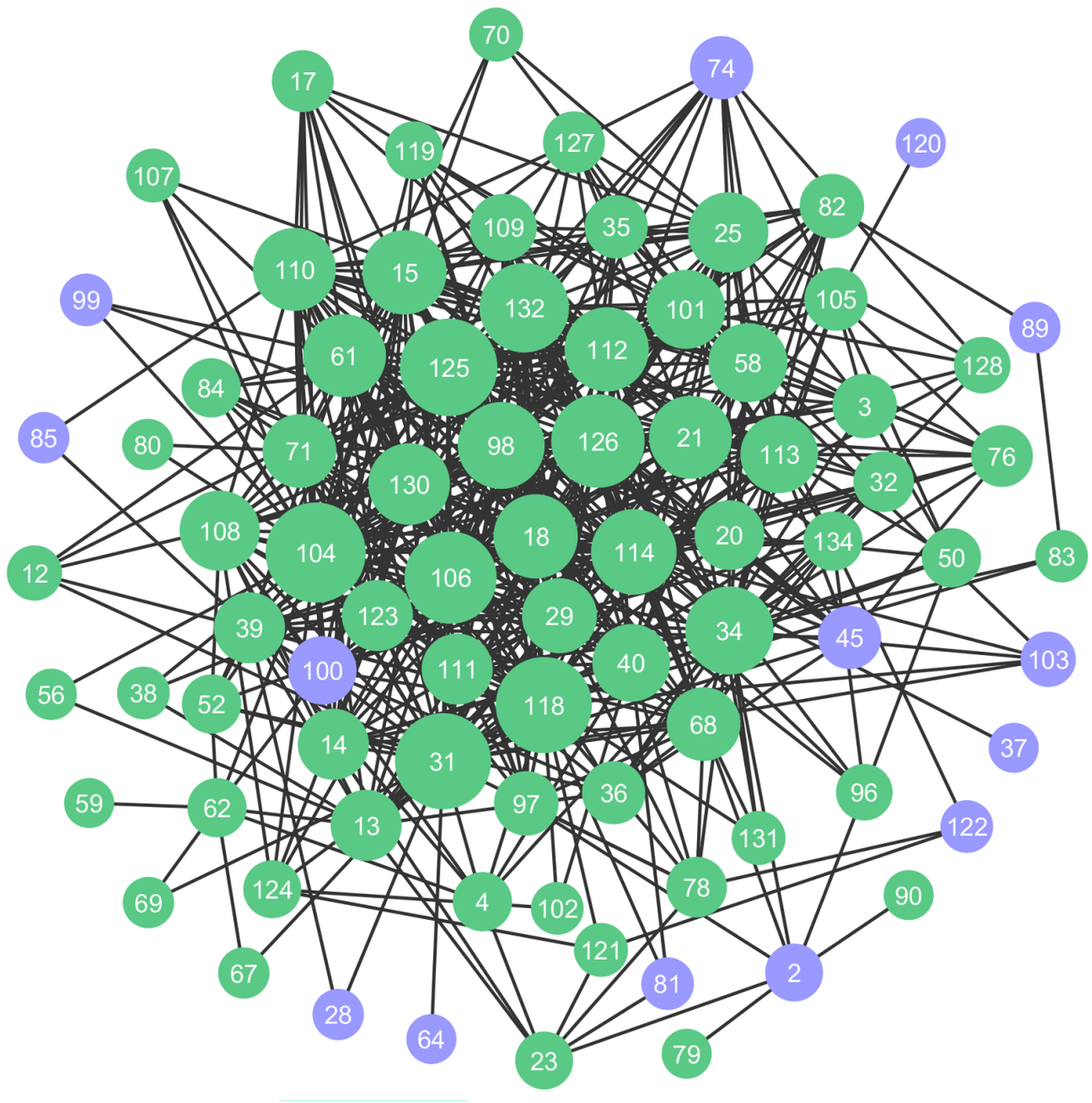
(a)

Porównanie kontroli do suszy, punkt czasowy T1



(b)

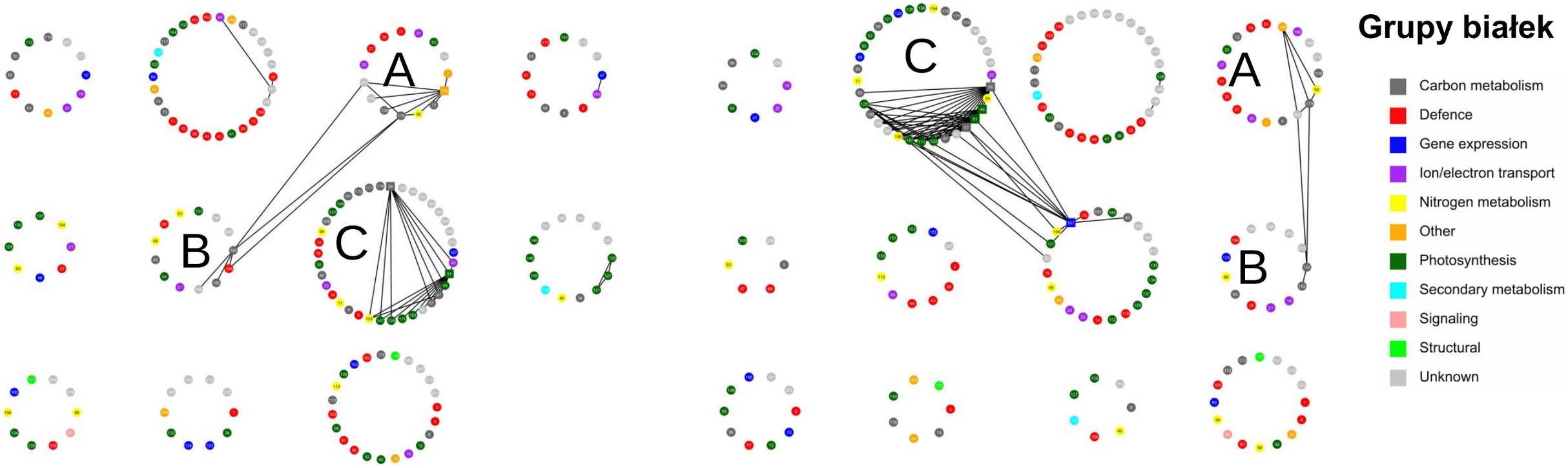
Porównanie kontroli do suszy, punkt czasowy T2



Korelacyjne sieci różnicowe. Metabolity, dla których interakcja linia × traktowanie suszą była istotna zaznaczone są na zielono.

(a) **Białka, kontrola**

(b) **Białka, susza**



Sieci korelacyjne białek w kontroli i suszy. Pokazane zostały krawędzie, które odpowiadają elementom z macierzy TOM o wartości większej niż 0.15. Moduły o największej liczbie białek (>60%) występującej zarówno w kontroli jak i w suszy są oznaczone literami A,B,C.

Tutorials for the WGCNA package

Peter Langfelder and Steve Horvath

Dept. of Human Genetics, UC Los Ageles (PL, SH), Dept. of Biostatistics, UC Los Ageles (SH)

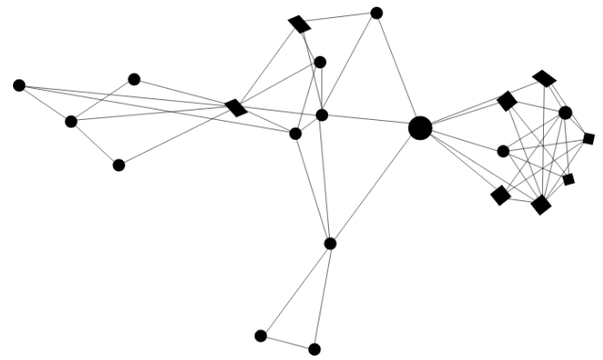
Peter (dot) Langfelder (at) gmail (dot) com, SHorvath (at) mednet (dot) ucla (dot) edu

This page provides a set of tutorials for the WGCNA package. We illustrate various aspects of data input, network construction, module detection, relating modules and genes to external information etc. Before going through the tutorials, please make sure you have installed (the newest version of) the WGCNA package and all packages it depends on. Please refer to the main WGCNA page and the installation instructions for details.

We provide three introductory tutorials (I - III), each split into smaller sections for easier reading, and we link to more advanced tutorials that describe research analyses in which we used WGCNA.

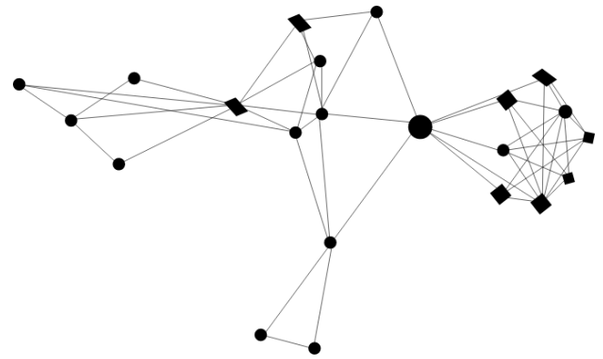
- The first tutorial guides the reader through an analysis of a single empirical gene expression data set. Results obtained in this tutorial are described in the example analysis section of the main paper. We highly recommend that the reader work through this tutorial before moving on to the second tutorial.
- The second tutorial introduces consensus module analysis that closely parallels the single data set analysis presented in Tutorial I, but contains the extra ingredient of analyzing the consensus of two related but different sets.

Podsumowanie



1. Sieć korelacyjna: analiza statystyczna (redukcja cech do cech istotnych statystycznie ze względu na badany czynnik), konstrukcja sieci, wizualizacja sieci dostosowana do potrzeb.
2. Wszystkie przedstawione metody, algorytmy, pakiety, funkcje można dostosować do dowolnie dużych danych np. biologicznych, chemicznych.

Bibliografia



Piasecka A., Sawikowska A., Kuczyńska A., Ogrodowicz P., Mikołajczak K., Krystowiak K., Gudyś K., Guzy-Wróbelska J., Krajewski P., Kachlicki P. (**2017**). Drought related secondary metabolites of barley (*Hordeum vulgare L.*) leaves and their association with mQTLs. Plant J, 89: 898–913.

Swarcewicz B., Sawikowska A., Marczak Ł., Łuczak M., Ciesiołka D., Krystkowiak K., Kuczyńska A., Piślewska-Bednarek M., Krajewski P., Stobiecki M. (**2017**). Effect of drought stress on metabolite contents in barley recombinant inbred line population revealed by untargeted GC–MS profiling. Acta Physiol Plant 39:158.

Rodziewicz P., Chmielewska K., Sawikowska A., Marczak Ł., Łuczak M., Bednarek P., Mikołajczak K., Ogrodowicz P., Kuczyńska A., Krajewski P., Stobiecki M. (**2019**). Identification of drought responsive proteins and related proteomic QTLs in barley. Journal of Experimental Botany 70: 2823-2837.

Surma M., Kuczyńska A., Mikołajczak K., Ogrodowicz P., Adamski T., Ćwiek-Kupczyńska H., Sawikowska A., Pecio A., Wach D., Józefaciuk G., Łukowska M., Zych J., Krajewski P. (**2019**). Barley varieties in semi-controlled and natural conditions - response to water shortage and changing environment. J Agro and Crop Sci 205: 295-308.

Kim jesteśmy

Xenstats sp. z o.o. jest specjalistyczną firmą bioinformatyczną.

- Zapewniamy kompleksowe rozwiązania w zakresie planowania, prowadzenia badań, analizy statystycznej i interpretacji wyników.
- Oferujemy najwyższej jakości specjalistyczne analizy statystyczne przygotowane w R.
- Wykonujemy analizę danych RNA-seq, danych metabolomicznych i proteomicznych, prowadzimy konsultacje oraz warsztaty.
- Nasi eksperci to naukowcy z dziedziny matematyki, bioinformatyki, statystyki, biologii i chemii.

XENSTATS

Warsztaty z programowania w R

https://www.xenstats.com/szkolenia

Wprowadzenie do R

dla początkujących

Termin: 21-22.02.2020 10:00-15:00

Analiza danych w R

dla początkujących

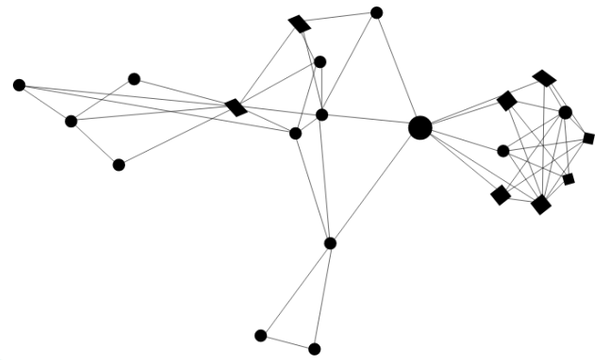
Termin: 6-7.03.2020 10:00-15:00

Analiza danych RNA-seq w R

dla zaawansowanych

Termin: 20-21.03.2020 10:00-15:00

Dziękuję za uwagę



Kontakt

Dr Aneta Sawikowska
aneta.sawikowska@xenstats.com

https://www.xenstats.com
LinkedIn: https://www.linkedin.com/in/aneta-sawikowska-06743417b/