POLITECHNIKA ŚLĄSKA
Wydział Automatyki, Elektroniki i Informatyki
Instytut Automatyki

Metody integracji w analizie danych wielodziedzinowych badań biologii molekularnej dla poszukiwania biomarkerów chorób cywilizacyjnych

Autoreferat

Anna Papież

Promotor

prof. dr hab. inż. Joanna Polańska

2019
Gliwice, POLSKA

© 2019
Anna Papież
Wszelkie prawa zastrzeżone

Finansowanie

Praca doktorska była realizowana przy wsparciu następujących projektów:

- "DoktoRIS - Program stypendialny na rzecz innowacyjnego Śląska", współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego
- POIG.02.03.01-24-099/13 finansowanie oraz infrastruktura: "GeCONiI - Górnośląskie Centrum Obliczeń Naukowych i Inżynieryjnych"
- grant Harmonia Narodowego Centrum Nauki numer DEC-2013/08/M/ST6/00924: "BioRadInt"
- grant OPUS Narodowego Centrum Nauki numer UMO-2015/19/B/ST6/01736: "BiTIMS"
- grant PBS Narodowego Centrum Badań i Rozwoju numer PBS3/A7/29/2015/ID-247184 "MOLTEST BIS"

Spis treści

Finansowanie **1**

Spis treści **2**

1 Wprowadzenie **3**
1.1 Motywacja 3
1.2 Cel pracy 4

2 Metody **5**
2.1 Identyfikacja efektu paczki metodą programowania dynamicznego . . . 5
2.2 Integracja wieloplatformowych danych w ramach omiki 6
2.3 Integracja wielodziedzinowa 9
2.4 Analiza integracyjna międzytkankowa 10

3 Wyniki **11**
3.1 Identyfikacja efektu paczki 11
3.2 Integracja danych transkryptomicznych 14
3.3 Integracja wielodziedzinowa 17
3.4 Integracja międzytkankowa 19

4 Wnioski **21**

Bibliografia **23**

Rozdział 1

Wprowadzenie

1.1 Motywacja

Łączenie informacji pochodzących z eksperymentów pozyskiwanych technikami wysokoprzepustowymi w biologii molekularnej jest zadaniem, z którym mierzą się coraz liczniejsze grupy naukowców. Nieustający wzrost ilości danych dostępnych za pośrednictwem wielu repozytoriów wprowadza potrzebę podnoszenia efektywności algorytmów przetwarzania, gdyż duże ilości istotnych informacji giną w natłoku gromadzonych wyników badań. Potrzeba przetwarzania dużych ilości danych w naukach biologicznych i medycznych pociąga za sobą rozwój algorytmów statystycznych i eksploracji danych w celu fuzji i walidacji eksperymentów biomedycznych.

W dzisiejszych czasach wciąż pozostaje ogrom wiedzy do zgłębienia w temacie mechanizmów molekularnych stojących za chorobami. Wiedza ta jest niezwykle istotna, zwłaszcza w kontekście stale rozwijającej się dziedziny medycyny spersonalizowanej. Indywidualne planowanie terapii jest naglącym problemem, biorąc pod uwagę rosnącą częstość występowania chorób cywilizacyjnych.

Choroby serca i układu krwionośnego, nowotwory i cukrzyca rozpowszechniły się w szybkim tempie wraz ze wzrostem zamożności w społeczeństwach wysoko rozwiniętych. Obecnie choroby te stają się również wiodącymi przyczynami śmierci w krajach rozwijających się. Dlatego też badania w dziedzinach diagnostyki, prognostyki i leczenia są kluczowe dla podnoszenia poziomu życia, jak również jego wydłużania w skali globalnej.

Badania w zakresie różnych dziedzin biologii molekularnej: omik, mają za cel ustalenie przyczyn wielu chorób śmiertelnych. Neologizm *omika* pochodzi od przyrostka występującego w nazwach poszczególnych dziedzin: genomiki, transkryptomiki, proteomiki, metabolomiki, itp. Obserwuje się również tendencję we współczesnej medycynie do poszukiwania mechanizmów poprzez łączenie informacji z różnych omik oraz ich interakcji, zamiast postrzegana wyników eksperymentów z pojedynczych dziedzin jako głównego źródła wiedzy.

Obecnie statystyczne metody projektowania eksperymentów pozwalają na planowanie złożonych studiów badawczych przy zachowaniu kontroli nad źródłami zakłóceń oraz zmienności. Na równi jednak istotnym jest również stosowanie rozwiązań,

które są dostosowane do typu eksperymentu i poprawiają jakość otrzymanych wyników. Ponadto, wykorzystanie zaawansowanych narzędzi statystycznych w połączeniu z przeglądami literaturowymi oraz bioinformatycznych baz danych pozwala na zwiększenie efektywności wnioskowania i odkrywania nowej wiedzy.

Pomimo istnienia bogatego zbioru prac naukowych poświęconych powyższej tematyce, ciągle rosnące w ilość zbiory danych w repozytoriach bionformatycznych powodują zapotrzebowanie na stały rozwój technik optymalizacji procesu analizy danych. W związku z tym, niniejsza praca została poświęcona implementacji oraz badaniom nad opracowaniem procedur integracyjnej analizy danych pochodzących z różnych platform oraz dziedzin biologii molekularnej, a które zostały pozyskane w wyniku zastosowania technik wysokoprzepustowych. Opracowane metody mają za zadanie umożliwić kompleksową analizę w celu poszukiwania biomarkerów współczesnych chorób oraz ich interakcji. Omówione tu procedury dotyczą analizy na wielu etapach, od wstępnego przetwarzania i filtracji poprzez końcowe etapy łączonych analiz pomiędzy eksperymentami z różnych platform oraz omik.

1.2 Cel pracy

Celem niniejszej pracy było opracowanie metod integracyjnej analizy danych pochodzących z wysokoprzepustowych technik biologii molekularnej dla celów poszukiwania biomarkerów chorób cywilizacyjnych. Zastosowana metodologia składa się z omówienia istniejących technik dla łączonej oraz komparatywnej analizy danych oraz propozycji nowych metod integracyjnej analizy eksperymentów z różnych omik. Oczekiwanym wynikiem prac jest opracowanie narzędzi dostosowanych do łączonej analizy danych oraz wyników z wielu platform i dziedzin biologii molekularnej.

Uwzględniając wyżej wymienione cele pracy, sformułowano następujące tezy rozprawy:

1. Właściwe wstępne przetwarzanie danych pozyskanych technikami wysokoprzepustowymi biologii molekularnej oraz korekta efektu paczki zapobiega utracie wartościowych informacji uzyskanych na podstawie analizy wyników eksperymentu.

2. Wprowadzenie kompleksowych rozwiązań dla analizy pokrewnych eksperymentów w obrębie jednej dziedziny pozyskanych za pośrednictwem różnych platform zapewnia poprawę jakości wnioskowania statystycznego oraz zadań klasyfikacji.

3. Statystyczna integracja danych z różnych dziedzin biologii molekularnej oraz różnych tkanek stanowi narzędzie do walidacji wyników pojedynczego eksperymentu oraz prowadzi do kompleksowego scharakteryzowania nowych mechanizmów stojących za procesami w biologii molekularnej.

Rozdział 2

Metody

2.1 Identyfikacja efektu paczki metodą programowania dynamicznego

Narzędzia do korekty efektu paczki umożliwiają filtrację czynników zakłócających sygnał w zbiorach danych i w ten sposób są źródłem poprawy wyników analiz poprzez zwiększenie nacisku na zmienność biologiczną (Labaj et al., 2017). W wielu eksperymentach informacja o czynnikach wywołujących efekt paczki jest nieznana, jednak najpopularniejsze narzędzia do korekty wymagają tej informacji do działania. Dlatego w ramach niniejszej rozprawy zaproponowano i przetestowano nowy algorytm identyfikacji efektu paczki (Papiez et al., 2018) na kilku zestawach danych eksperymentalnych. Obejmują one badania na mikromacierzach DNA, spektrometrię masową i pomiary RNA-seq.

Zaproponowana metoda identyfikacji z użyciem programowania dynamicznego wymaga wyznaczenia wskaźnika jakości, który reprezentuje każdą próbkę. Przykładowo, w przypadku mikromacierzy może to być średnia intensywność, dla spektrometrii masowej całkowity ładunek jonów, w danych z sekwencjonowania - mediana zliczeń.

Identyfikacja efektu paczki w tym rozumieniu może być zdefiniowana jako podział szeregu posortowanych danych na grupy w taki sposób, aby suma bezwzględnych odchyłek wskaźników jakości wewnątrz grupy (paczki) była jak najmniejsza. Zadanie to jest rozwiązywanie za pomocą algorytmu programowania dynamicznego (Bellman, 1961; Jackson et al., 2005). Indeksy próbek oznacza się jako $i = 1, 2, \ldots, N$. Podział na podgrupy wymaga zdefiniowania K paczek, $B_1, B_2, \ldots B_K$, gdzie $k - ta$ paczka zawiera indeksy $B_k = B(i, i+1, \ldots, j) = i, i+1, \ldots, j$. Wskaźnik jakości jest oznaczony jako QI_i. Bezwzględna odchyłka wskaźnika jakości wewnątrz paczki to:

$$AbsDev(B_k) = \sum_{l \in B_k} |QI_l - \overline{QI}_{B_k}| \tag{2.1}$$

Minimalizowany wskaźnik dla algorytmu programowania dynamicznego to suma odchyłek bezwzględnych

$$I(K) = \sum_{k=1}^{K} AbsDev(B_k) \tag{2.2}$$

Optymalny podział $B_1^{opt}, B_2^{opt}, \ldots B_K^{opt}$ prowadzi do wartości minimalnej sumy bezwzględnych odchyłek wskaźników odpowiadających wszystkim paczkom:

$$I_{1\ldots N}^{opt}(K) = min_{partitions}^{1\ldots N}[\sum_{k=1}^{K} AbsDev(B_k)] \tag{2.3}$$

W celu sformułowania rekurencji w programowaniu dynamicznym, obliczany jest optymalny cząstkowy wskaźnik dla zakresu próbek $1, 2, \ldots, j$:

$$OCI_{1\ldots j}(k) = min_{partitions}^{1\ldots j}[\sum_{\chi=1}^{K} AbsDev(B_\chi)] \tag{2.4}$$

Równanie Bellmana dla procedury rekurencyjnej można zapisać następująco:

$$OCI_{1\ldots j}(k+1) = min_{i=1\ldots j-1}[OCI_{1\ldots i-1}(k) + AbsDev(B(i, i+1, \ldots, j))] \tag{2.5}$$

Iterowanie powyższego równania Bellmana prowadzi do uzyskania optymalnego podziału $B_1^{opt}, B_2^{opt}, \ldots B_K^{opt}$ oraz optymalnej wartości sumy bezwzględnych odchyłek $I_{1\ldots N}^{opt}(K)$. Algorytm wymaga, by paczka składała się z co najmniej trzech próbek dla umożliwienia wyliczeń miar rozrzutu. Liczba paczek dobierana jest na podstawie dystrybucji statystyki δ w metodzie gPCA (Reese et al., 2013). Jeżeli otrzymana δ nie jest istotna statystycznie, uznaje się, że efekt paczki w danym zbiorze danych jest pomijalny.

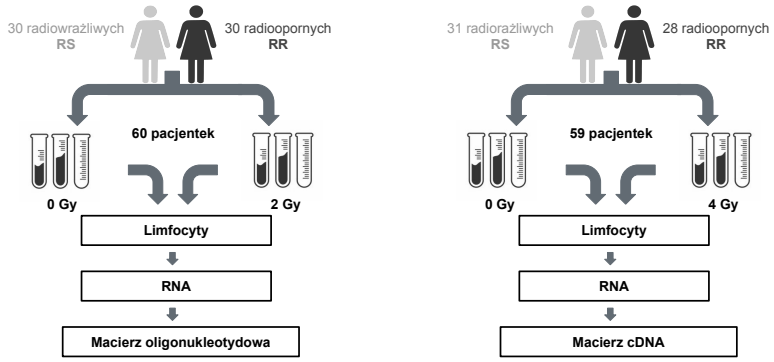
Algorytm identyfikacji efektu paczki przy użyciu programowania dynamicznego został zaimplementowany oraz udostępniony w postaci pakietu języka R BatchI (Papiez et al., 2018).

2.2 Integracja wieloplatformowych danych w ramach omiki

W niniejszej rozprawie przeanalizowano dwa zestawy danych o poziomach ekspresji genów pozyskane z wykorzystaniem rożnych platform mikromacierzowych. Poniżej przedstawiono różne aspekty łączenia danych w celu pozyskania lepszej jakościowo

oraz ilościowo informacji w stosunku do analiz prowadzonych na pojedynczych, mniej licznych zbiorach danych.

Dane o poziomach ekspresji otrzymano na drodze niezależnych eksperymentów mikromacierzowych przeprowadzonych w celu identyfikacji mechanizmów radiowrażliwości w pacjentkach cierpiących na raka piersi, które zostały poddane radioterapii (Yarnold et al., 2005). W obu przypadkach, z próbek krwi pacjentek wyodrębniono limfocyty, które następnie podzielono na grupę kontrolną oraz drugą, którą napromieniono. Jeden eksperyment przeprowadzono na mikromacierzach oligonukleotydowych, drugi na macierzach typu cDNA.

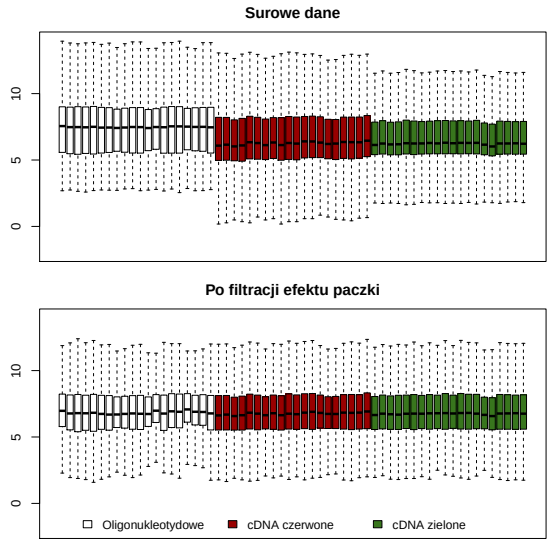


Rysunek 2.1: Schemat przedstawiający porównanie eksperymentów. W obu przypadkach wykorzystano etykietowanie pacjentek RR (radiooporne) i RS (radiowrażliwe). Różnice w schemacie można zaobserwować w przypadku dawki promieniowania oraz platformy mikromacierzowej.

W celu sprowadzenia danych z dwóch eksperymentów, przeprowadzono korektę efektu paczki, aby sprowadzić dane do wspólnej przestrzeni. Następne dla mikromacierzy cDNA, która jest platformą dwukanałową, dokonano uśrednienia informacji na obu kanałach, aby umożliwić porównanie i łączenie wyników w aspekcie biologicznym (Papiez et al., 2014).

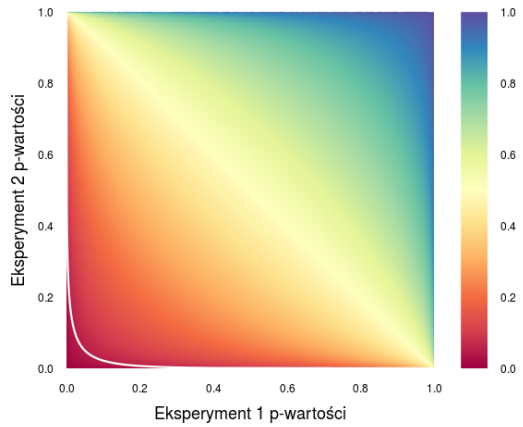
Dane poddano łączonej analizie różnicowania pomiędzy próbkami radioopornymi i radiowrażliwymi na trzy sposoby (Papiez et al., 2015):

- Restrykcyjny
  Dane wstępnie przetworzono oraz analizowano pod kątem testów na różnicowanie niezależnie w dwóch zbiorach. W wyniku tego uzyskano zestawy genów różnicujących na poziomie 0.05, a następnie brano pod uwagę część wspólną zbiorów, jako ostateczną listę genów.
- Arraymining
  Dane analizowano niezależnie w dwóch zbiorach, a jako kryterium różnicowania przyjęto ranking na podstawie metod eksploracji danych zaimplementowanych w serwisie Arraymining (Glaab et al., 2009). Miara powstała poprzez połączenie rang z czterech algorytmów: metody empirycznej Bayesa, cząstkowych najmniejszych kwadratów, lasów losowych oraz analizy istotności dla mikromacierzy, stanowiła o rankingu ostatecznej listy genów.



Rysunek 2.2: Przykładowe dystrybucje próbek przed i po korekcji efektu paczki.

- Integracyjny
  Integracja oparta jest o algorytm łączenia p-wartości metodą ważonych Z-transformacji (Zaykin, 2011). Dane z dwóch zbiorów poddane są testom na różnicowanie, a następnie p-wartości dla każdej cechy zostają połączone na drodze Z-transformacji i łączona p-wartość staje się miarą różnicowania. Ostateczna lista genów wyłoniona jest na podstawie łączonych p-wartości.



Rysunek 2.3: Ilustracja łączenia p-wartości metodą Z-transformacji.

Efektywność list genów uzyskanych na trzy sposoby została zbadana w zadaniu separowalności próbek radiowrażliwych i radioopornych. Zadanie klasyfikacji sformułowano dla modelu regresji logistycznej oraz maszyny wektorów podpierających.

Następnie na drodze wielokrotnej kroswalidacji stratyfikowanej przeprowadzono klasyfikację próbek ze względu na odpowiedź na promieniowanie. Do tego celu wyłoniono profile odpowiedzi (Papiez et al., 2019b):

- reakcja na napromieniowanie regulowane w górę
- reakcja na napromieniowanie regulowane w dół
- potencjalnie dozymetryczne regulowane w górę
- potencjalnie dozymetryczne regulowane w dół
- aktywowane wysoką dawką regulowane w górę
- aktywowane wysoką dawką regulowane w dół

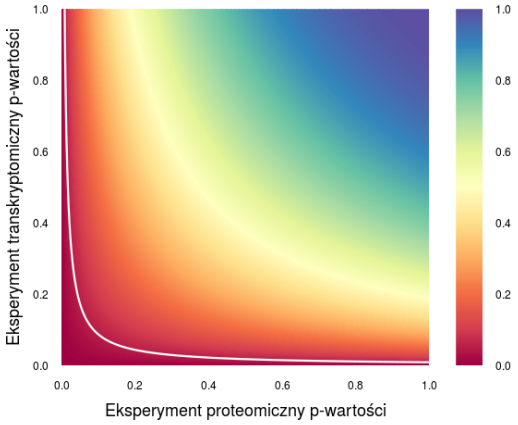
Do analizy wybrano geny, które nie różnicowały w kontrolach w obu eksperymentach. Następnie przeprowadzono interpolację pomiędzy dawką 2 i 4 Gy w celu ujednolicenia danych z dwóch eksperymentów. W interpolacji uwzględniono profile odpowiedzi, tak że w grupie potencjalnie dozymetrycznych dokonano interpolacji liniowej do 2 Gy, natomiast w profilu oznaczonym jako reakcja na napromieniowanie zachowano wartość ekspresji z poziomu 4 Gy. Cechy wybierane do modelu najczęściej w kroswalidacji zostały dodatkowo zwalidowane wykorzystując metody wyboru cech Monte Carlo (Krol, 2015).

2.3 Integracja wielodziedzinowa

Statystyczna integracja p-wartości w analizie danych wieloplatformowych okazała się być adekwatnym podejściem, zatem zastosowano je również do łączonej analizy z różnych omik (Papiez et al., 2019a). W tej części łączenie wykorzystano do eksperymentów z dwóch dziedzin: transkryptomiki i proteomiki. Dane dotyczyły pracowników zakładu produkcji jądrowej, który zmarli na skutek choroby niedokrwiennej serca. Celem badania było zgłębienie wiedzy na temat mechanizmów wywołania choroby poprzez ekspozycję na promieniowanie. Pierwszym eksperymentem było scharakteryzowanie białek w sercu na drodze spektrometrii masowej (Azimzadeh et al., 2017). Drugi eksperyment przeprowadzono na podzbiorze próbek z pierwszego i wykonano sekwencjonowanie RNA-seq.

W zbiorze danych z proteomiki problemem na wstępie była korelacja zaabsorbowanej dawki promieniowania oraz wieku. Z tego względu przeprowadzono analizę regresji krokowej pod kątem czynników wieku i dawki, w celu oznaczenia cech jako zależnych od wieku lub dawki promieniowania. Następnie wśród białek oznaczonych jako zależne od dawki wyłoniono różnicujące pomiędzy wysokimi dawkami promieniowania oraz kontrolami.

W danych RNA-seq również przeprowadzono wnioskowanie statystyczne w celu identyfikacji genów różnicujących na podstawie rozkładu ujemnego dwumianowego. Następnie przeprowadzono integrację p-wartości dla odpowiadających sobie par gen-białko z dwóch eksperymentów. Do tego zadania wykorzystano metodę łączenia p-wartości Fishera, ze względu na brak symetrii w dystrybucjach danych.



Rysunek 2.4: Ilustracja metody łączenia p-wartości Fishera.

2.4 Analiza integracyjna międzytkankowa

Zebrano dane dotyczące egzosomów z czterech rodzajów komórek:

- ludzkich fibroblastów
- ludzkich komórek śródbłonka tętnicy wieńcowej
- ludzkich komórek nabłonka gruczołu sutkowego (MCF10A)
- ludzkich leukocytów

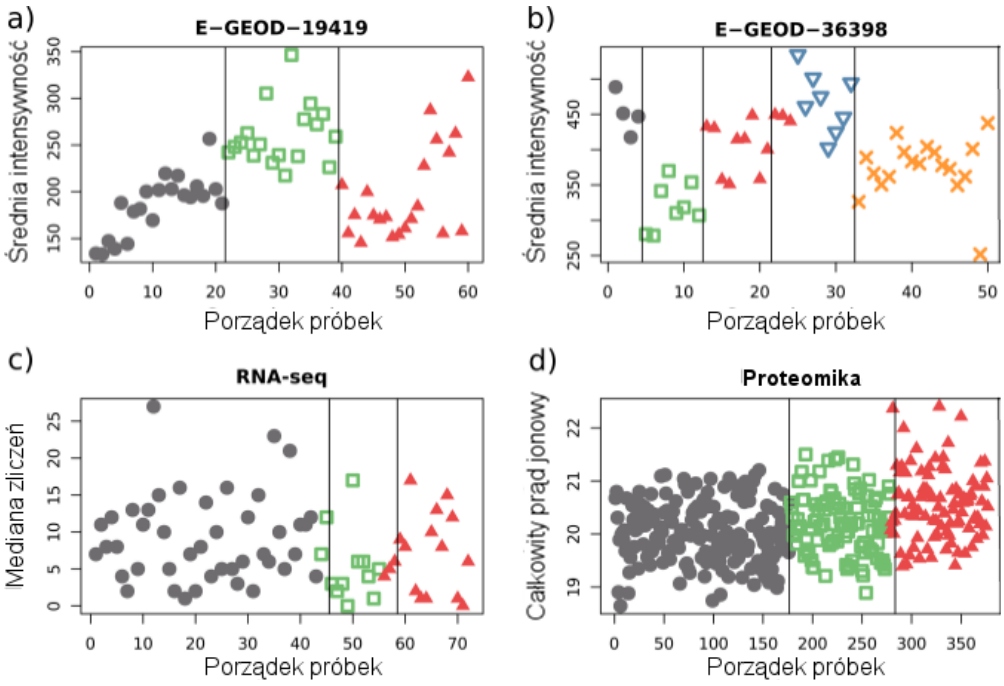
Egzosomy podzielono na cztery grupy dawek: 0 Gy controls, 1 Gy, 2 Gy, 6 Gy, a następnie przeprowadzono eksperyment spektrometrii masowej w celu identyfikacji białek egzosomalnych. Dla zestawu białek wspólnych we wszystkich rodzajach komórek przeprowadzono analizę białek różnicujących, klasteryzację hierarchiczną oraz analizę podobieństwa (Frank et al., 2007).

Rozdział 3

Wyniki

3.1 Identyfikacja efektu paczki

W pierwszej kolejności testowano zaproponowany algorytm na czterech zbiorach danych ze znaną strukturą paczek: dwa zbiory danych mikromacierzowych E-GEOD-19419 i E-GEOD-36398, dane RNA-seq oraz dane ze spektrometrii mas. Badano stopień odtworzenia istniejącego podziału ze względu na paczki, jak również wpływ na poprawę jakości danych po identyfikacji oraz korekcji efektu paczki przy użyciu algorytmu ComBat (Johnson et al., 2007). Oryginalny podział na grupy porównano za pomocą indeksu Dice'a oraz zilustrowano na Rysunku 3.1.

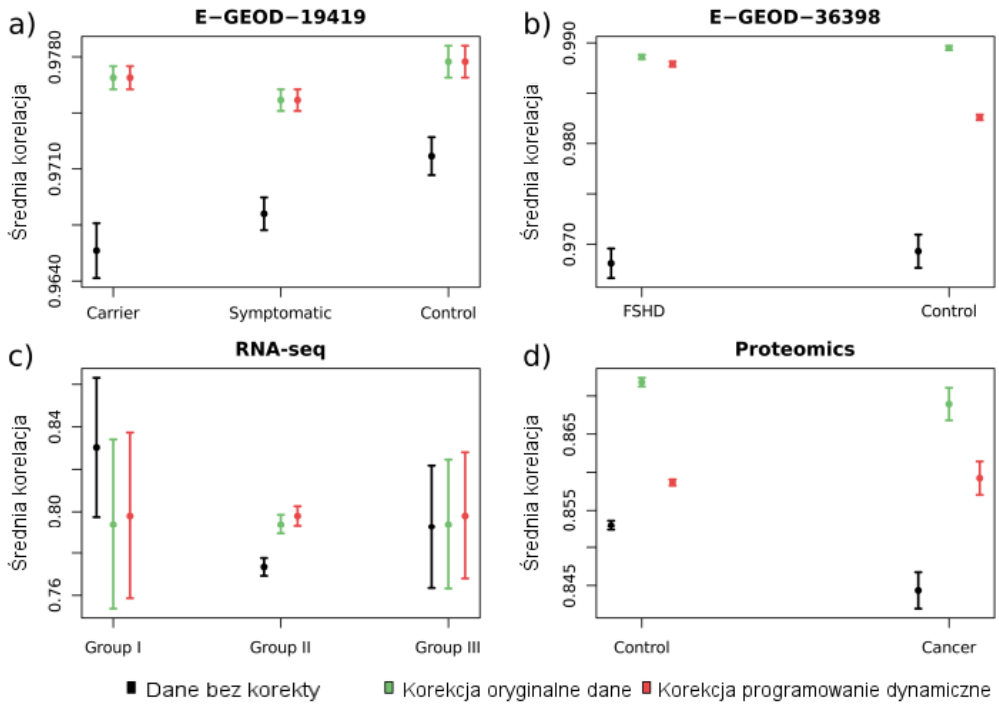


Rysunek 3.1: Podział zbiorów danych na paczki w zbiorach o znanej wcześniej strukturze. Oryginalny podział odzwierciedlają kolory i kształty, natomiast pionowe linie są wynikiem działania algorytmu BatchI.

- Dane mikromacierzowe
  W zbiorze danych E-GEOD-19419 odtworzono oryginalny podział w 100%. W zbiorze średni wązony indeks Dice'a wynosi 94.05%.

- RNA-seq
  W danych z sekwencjonowania średni ważony indeks Dice'a wynosi 93.02%.
- Spektrometria masowa
  W danych ze spektrometrii masowej indeks średni ważony indeks Dice'a wynosi 99.78%.

Po korekcji efektu paczki badano korelację wewnątrz grup, przy założeniu, że po prawidłowym usunięciu efektu paczki, powinna ona wzrosnąć. W większości badanych grup zaobserwowano wzrost korelacji wewnątrz grup względem danych przed korektą (Rysunek 3.2). Jedynie w danych z sekwencjonowania widoczna jest tendencja spadkowa w części grup ze względu na niezbilansowanie grup pod kątem liczności. Efektywność korekcji jest zauważalna również poprzez redukcję zmienności wywołanej efektem paczki do ogółu zmienności, wyrażonej przez współczynnik δ (Tabela 3.1.



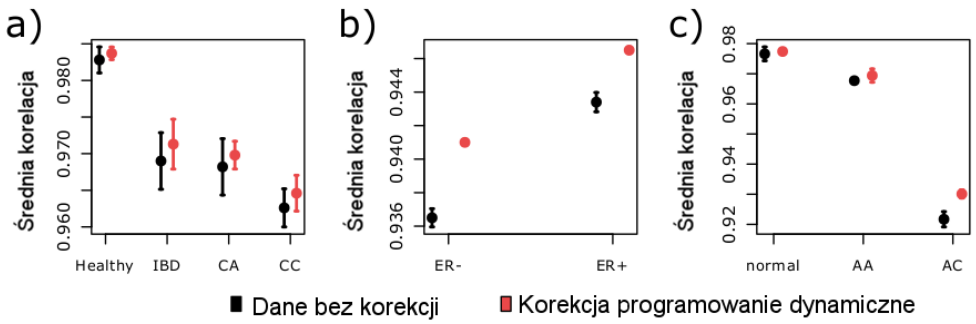
Rysunek 3.2: 95% przedziały ufności dla średniej korelacji wewnątrz paczek.

Następnie przeanalizowano trzy zestawy danych z eksperymentów mikromacierzowych E-GEOD-2034, E-GEOD-4183 oraz E-GEOD-10927, w których struktura paczek nie była znana *a priori*. Pierwszy eksperyment dotyczył raka piersi, drugi raka jelita grubego, natomiast ostatni raka kory nadnerczy. Widoczna była poprawa korelacji w stosunku do danych bez korekcji we wszystkich trzech zbiorach (Rysunek 3.3).

Ponadto, wyniki przeanalizowano pod kątem informacji biologicznej pozyskanej na podstawie genów różnicujących wyłonionych z danych po korekcie efektu paczki. Analizę funkcjonalną przeprowadzono na podstawie wskaźnika zawartości informacji (Information Content: IC) ontologii genowych, które są nadreprezentowane przez geny różnicujące z trzech zbiorów danych (Rysunek 3.4. Im wyższy wskaźnik IC, tym

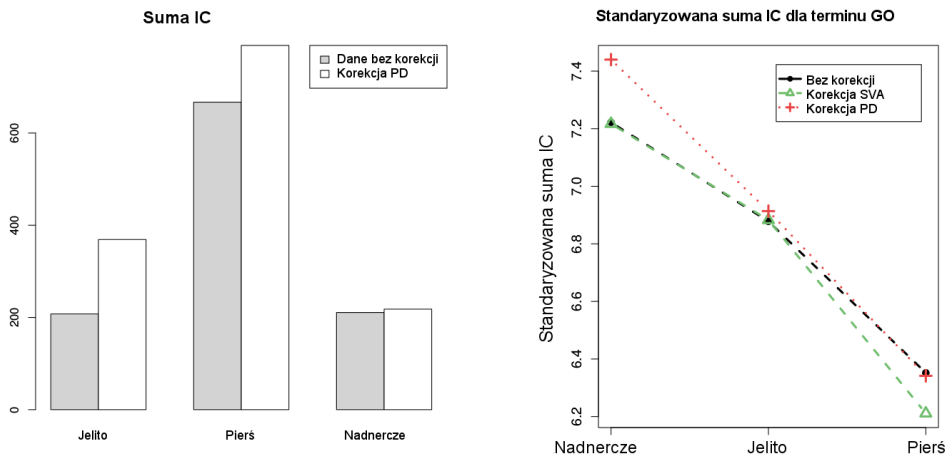
Tablica 3.1: Procent zmienności wywołanej efektem paczki w stosunku do całej zmienności obserwowanej w analizowanych zbiorach danych.

E-GEOD-19419	**Korekcja oryginalnych paczek**	**Korekcja BatchI**
Zmienność [%]	69.23	69.23
δ	0.9271	0.9271
p-wartość	4.69E-08	4.78E-08
E-GEOD-36398	**Korekcja oryginalnych paczek**	**Korekcja BatchI**
Zmienność [%]	48.15	50.14
δ	0.9991	0.9989
p-value	2.24E-07	2.90E-07
RNA-seq	**Korekcja oryginalnych paczek**	**Korekcja BatchI**
Zmienność [%]	65.12	67.23
δ	0.2765	0.6175
p-value	4.87E-01	9.38E-02
Proteomika	**Korekcja oryginalnych paczek**	**Korekcja BatchI**
Zmienność [%]	23.82	24.56
δ	0.6645	0.6671
p-value	7.32E-01	7.15E-01



Rysunek 3.3: 95% przedziały ufności dla średniej korelacji wewnątrz grup w zbiorach danych o nieznanej strukturze paczek.

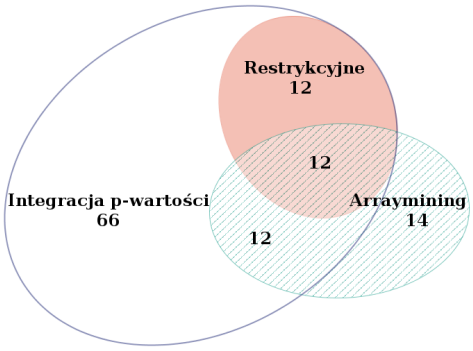
bardziej szczegółowa informacja otrzymana z danego terminu ontologicznego. Wskaźnik IC jest również standaryzowany ze względu na liczność genów powiązanych z daną ontologią. Z niniejszej analizy wynika, że korekta efektu paczki nie powoduje wzrostu jakości informacji biologicznej otrzymanej na temat dokładniej przebadanych chorób (rak piersi, jelita), natomiast zysk informacji jest większy w przy rzadziej występujących chorobach (rak kory nadnerczy).



Rysunek 3.4: Porównanie wskaźnika Information Content dla trzech eksperymentów mikromacierzowych.

3.2 Integracja danych transkryptomicznych

Dane z dwóch eksperymentów mikromacierzowych zostały przeanalizowane z wykorzystaniem trzech metod łączenia: restrykcyjnej, Arraymining oraz integracyjnej. Na diagramie Venna przedstawiono pokrycie genów zidentyfikowanych jako różnicujące z użyciem tych trzech podejść (Rysunek 3.5).

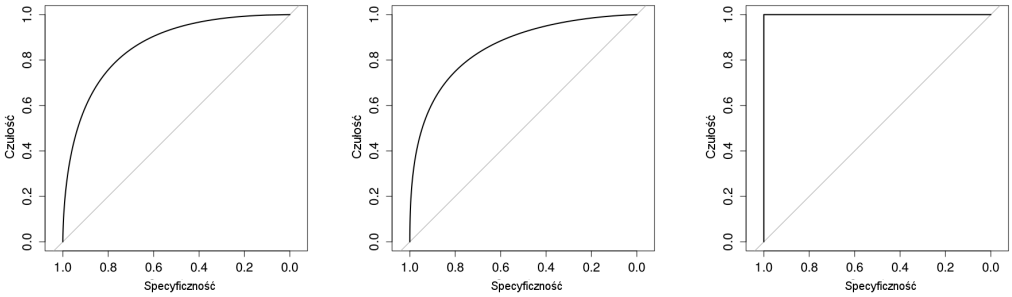


Rysunek 3.5: Diagram Venna przedstawiający liczności list genów otrzymanych trzema technikami łączenia danych.

Wysoka liczba genów otrzymanych metodą integracji statystycznej p-wartości pokazuje, że jest to metoda korzystniejsza od łączenia wyników list genów otrzymanych na podstawie ustalonych progów odcięcia.

Przeprowadzono analizę separowalności zbiorów pacjentek radiowrażliwych i radioopornych z wykorzystaniem list genów otrzymanych przy użyciu trzech podejść. Badanie wykonano przy użyciu modeli regresji logistycznej oraz maszyny wektorów podpierających. Wyniki zaprezentowano na Rysunkach 3.6 oraz 3.7. Zarówno

krzywe ROC, jak i statystyki dodatniej (PPV) i ujemnej (NPV) wartości predykcyjnej (Tabela 3.2) pokazują, że całkowita separowalność dwóch grup pacjentek była możliwa jedynie z wykorzystaniem cech uzyskanych metodą statystycznej integracji p-wartości.

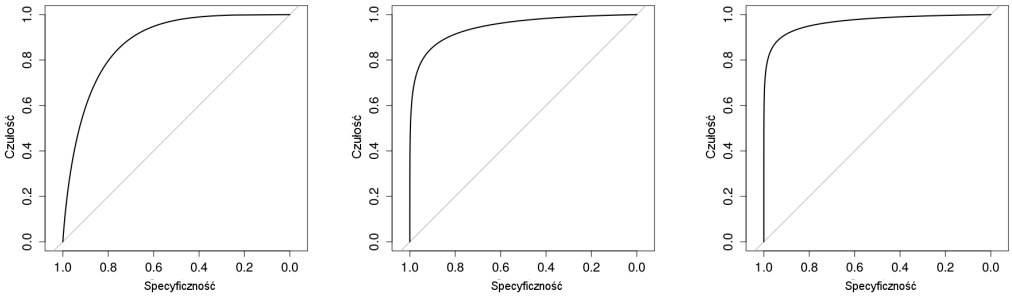


(a) Restrykcyjne AUC = 86.2%

(b) Arraymining AUC = 85.6%

(c) Integracyjne AUC = 100.0%

Rysunek 3.6: Krzywe ROC dla separowalności w modelu regresji logistycznej.



(a) Restrykcyjne AUC = 87.9 %

(b) Arraymining AUC = 94.6 %

(c) Integracyjne AUC = 96.7 %

Rysunek 3.7: Krzywe ROC dla separowalności w maszynie wektorów podpierających.

	Regresja logistyczna		**SVM**	
	PPV [%]	NPV [%]	PPV [%]	NPV [%]
Restrykcyjna	86.67	74.32	88.33	91.52
Arraymining	70.13	90.47	92.98	91.94
Integracyjna	100.00	100.00	98.18	93.75

Tablica 3.2: Dodatnie i ujemne wartości predykcyjne dla modelu regresji logistycznej i maszyny wektorów podpierających (SVM).

Otrzymane sygnatury były również przebadane ze względu na ich funkcje biologiczne. Geny, które wyłoniono jedynie przy użyciu podejścia integracyjnego biorą udział w procesach odpowiedzi na promieniowanie oraz powiązanych z nowotworem, między innymi w ścieżkach JAK-STAT, receptora interakcji cytokin, receptora komórkowego T.

W następnej kolejności badano możliwość klasyfikacji na podstawie biomarkerów odpowiedzi na promieniowanie. W tym celu połączono zbiory danych poprzez interpolację dawek ze względu na profil odpowiedzi. Liczby genów zaliczających się do sześciu wyszczególnionych profilów zawarto w Tabeli 3.3.

Tablica 3.3: Liczba genów według profilu odpowiedzi na dawkę promieniowania.

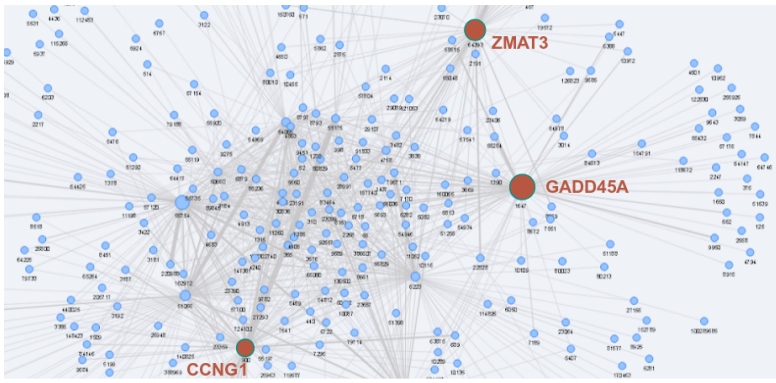
Liczba genów w profilach					
Reakcja na promnieniowanie		Potencjalnie dozymetryczne		Aktywowane wys. dawką	
W górę-Bez zmian	610	W górę-W górę	117	Bez zmian-W górę	48
W dół-Bez zmian	1067	W dół-W dół	969	Bez zmian-W dół	319

Następnie na drodze wielokrotnej kroswalidacji stratyfikowanej z modelem regresji logistycznej porównano wyniki klasyfikacji z zastosowaniem zaproponowanej metody transformacji danych w oparciu o profile odpowiedzi na promieniowanie z danymi oryginalnymi. Wyniki uśrednionych dodatniej oraz ujemnej wartości predykcji, a także dokładności klasyfikacji wskazują na większą efektywność metody z uwzględnieniem profilów. Zastosowanie metody dopasowanej do typu analizowanych danych spowodowało znaczącą poprawę wyników.

Tablica 3.4: Wyniki wielokrotnej kroswalidacji stratyfikowanej. *Dolny* oraz *Górny* odnoszą się do dolnej oraz górnej granicy przedziałów ufności dla średniej.

Oryginalne dane			
	Średnia [%]	Dolny [%]	Górny [%]
PPV	86.71	86.13	87.29
NPV	89.32	88.76	89.89
Dokładność	87.73	87.44	88.02
Dane interpolowane			
PPV	93.11	92.78	93.45
NPV	94.38	94.08	94.67
Dokładność	93.56	93.39	93.72

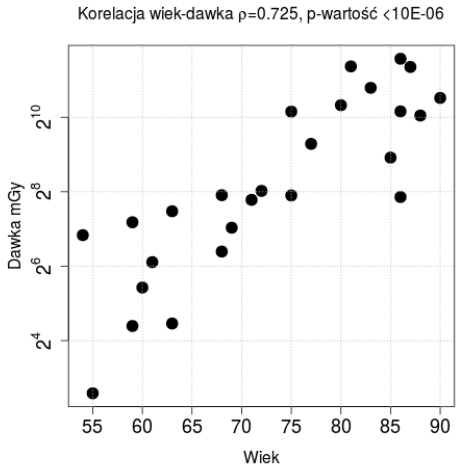
W kolejnych iteracjach klasyfikacji, najczęściej występującymi genami w modelu były GADD45A, ZMAT3 i NAMPT. Użyteczność tych potencjalnych biomarkerów dla zadań klasyfikacji potwierdzono niezależną metodą opartą na selekcji cech Monte Carlo (MCFS), gdzie największe sieci oddziaływań zaobserwowano w przypadku genów GADD45A, ZMAT3 and CCNG1 (Rysunek 3.8). Niezależna identyfikacja tych cech obiema metodami nie tylko potwierdza efektywność zaproponowanej metody, ale także zwraca uwagę w stronę analizy sieci powiązań genów, w przeciwieństwie do poszukiwania pojedynczych biomarkerów.



Rysunek 3.8: Fragment sieci interakcji genowych wyznaczonej na podstawie selekcji cech Monte Carlo.

3.3 Integracja wielodziedzinowa

Dane transkryptomiczne i proteomiczne z próbek górników zostały połączone poprzez statystyczną integrację p-wartości. Uprzednio jednak, niezbędne było wstępne przetworzenie danych proteomicznych, aby zidentyfikować białka powiązane z dawką ze względu na wysoką korelację czynników wieku i dawki (Rysunek 3.9).



Rysunek 3.9: Wykres przedstawiający powiązanie czynników wieku oraz dawki promieniowania w próbkach Mayak.

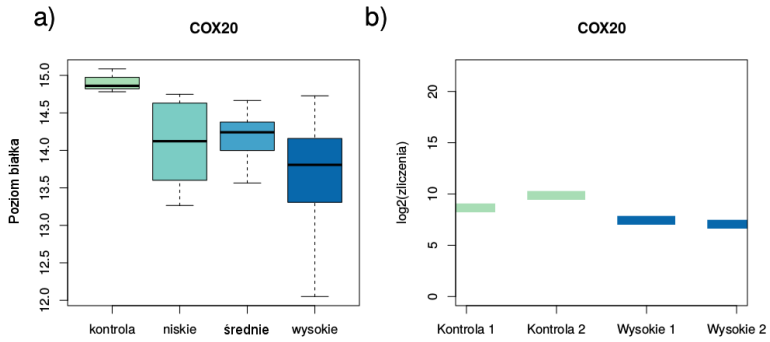
Przeprowadzono regresję krokową i w ten sposób otrzymano 582 białka (ze 1,281), gdzie dominującym źródłem zmienności była dawka promieniowania. 225 zidentyfikowano natomiast jako zależne od wieku oraz 212, gdzie zmienność była opisana zależnością wiek:dawka. Przeanalizowano ścieżki sygnałowe, w których uczestniczą białka z poszczególnych grup (Tabela 3.5). W grupie białek zależnych od dawki pojawiają się ścieżki potwierdzone wcześniej w literaturze (Azimzadeh et al., 2017) jako powiązane z reakcją na promieniowanie: PPAR signaling, Glycolysis, Fatty acid metabolism oraz TCA cycle.

Tablica 3.5: Ścieżki sygnałowe KEGG nadreprezentowane przez białka w grupach zależnych od wieku, od dawki oraz zależności wiek:dawka.

Zależne od wieku	**Zależne od dawki**	
Fatty acid elongation	PI3K-Akt signaling pathway	Ribosome
Tryptophan metabolism	Pathogenic Escherichia coli infection	Carbon metabolism
	Protein processing in endoplasmic reticulum	Glyoxylate and dicarboxylate metabolism
	Biosynthesis of amino acids	Arrhythmogenic right ventricular cardiomyopathy
	Proteasome	Pyruvate metabolism
Zależne od wieku i dawki	Tight junction	Butanoate metabolism
Metabolic pathways	Glycolysis/Gluconeogenesis	Adrenergic signaling in cardiomyocytes
Cardiac muscle contraction	Peroxisome	AMPK signaling pathway
Propanoate metabolism	Leukocyte transendothelial migration	Vasopressin-regulated water reabsorption
Valine, leucine and isoleucine degradation	Fatty acid metabolism	Beta-Alanine metabolism
Hypertrophic cardiomyopathy	ECM-receptor interaction	Antigen processing and presentation
Dilated cardiomyopathy	PPAR signaling pathway	Phagosome
Oxidative phosphorylation	Fatty acid degradation	TCA cycle
	Porphyrin and chlorophyll metabolism	2-Oxocarboxylic acid metabolism
	Focal adhesion	

Do integracji z danymi transkryptomicznymi wykorzystano białka, które zidentyfikowano jako zależne od dawki oraz różnicowały istotnie grupę kontrolną od grupy wysokich dawek w teście Dunnetta (307 białek). W analizie danych RNA-seq otrzymano 979 transkryptów różnicujących w górę oraz 895 różnicujących w dół. Następnie, z uwzględnieniem kierunku różnicowania (Rysunek 3.10) dokonano integracji p-wartości dla transkryptów oraz białek metodą Fishera.

Następnie porównano wyniki integracji Fishera z podejściem restrykcyjnym, gdzie uwzględniono wyniki różnicowania dwóch eksperymentów analizowanych osobno. Nadreprezentowane ścieżki w podejściu restrykcyjnym dotyczyły ogólnych procesów zachodzących w sercu (Tabela 3.6), natomiast w podejściu integracyjnym potwierdzono mechanizmy specyficzne dla występowania choroby wieńcowej w odpowiedzi



Rysunek 3.10: Przykładowa para białko transkrypt, w której kierunek różnicowania między wysoką dawką a kontrolą jest zgodny.

na promieniowanie.

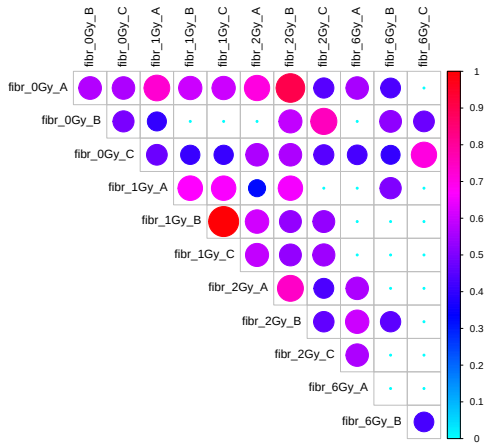
Tablica 3.6: Ścieżki sygnałowe KEGG nadreprezentowane przez pary gen-białko w dwóch podejściach do analizy: restrykcyjnym oraz integracyjnym.

Podejście restrykcyjne	**Podejście integracyjne**	
Proteasome	Glycolysis / Gluconeogenesis	Beta-Alanine metabolism
Ribosome	Oxidative phosphorylation	Metabolic pathways
Proteoglycans in cancer	Citrate cycle (TCA cycle)	Tryptophan metabolism
Pathogenic Escherichia coli infection	Bacterial invasion of epithelial cells	Arginine and proline metabolism
Propanoate metabolism		Lysine degradation
	Phagosome	PPAR signaling pathway
	Vasopressin-regulated water reabsorption	Proximal tubule bicarbonate reclamation
	Ascorbate and aldarate metabolism	Terpenoid backbone biosynthesis
	Valine, leucine and isoleucine degradation	Glyoxylate and dicarboxylate metabolism
	Histidine metabolism	Fatty acid degradation
	Pyruvate metabolism	Carbon metabolism

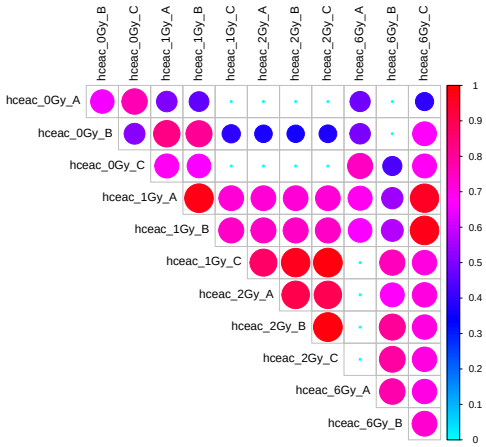
3.4 Integracja międzytkankowa

W danych ze spektrometrii masowej egzosomów czterech typów komórek zidentyfikowano 161 białek wspólnych dla wszystkich grup. Następnie w oparciu o ten zbiór przeanalizowano podobieństwo między tkankami oraz dawkami. Wyniki potwierdzają, że wiodącym czynnikiem w różnicowaniu białek jest typ komórek,a w dalszej kolejności dawka. Zastosowana metryka podobieństwa wskazuje, że dwie próbki odstające w

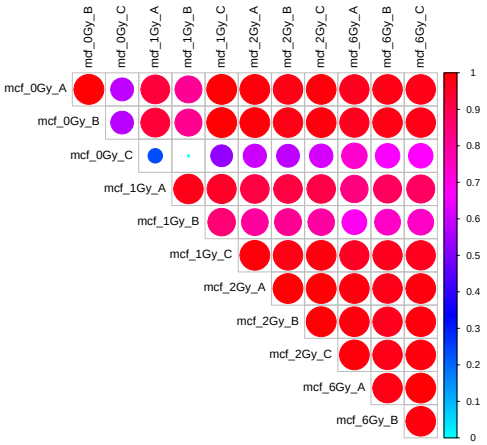
grupie fibroblastów wykazują wysokie podobieństwo względem gruczołu sutkowego. Natomiast w tętnicy wieńcowej miara podobieństwa jest wysoka w ramach grup dawkowych. Najniższe wartości podobieństwa mogą być obserwowane w próbkach fibroblastów.



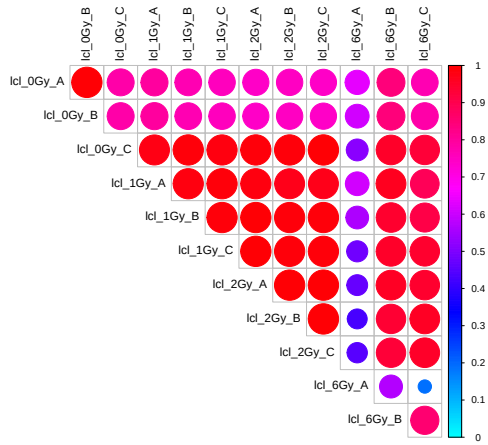
(a) Fibroblasty



(b) Tętnica wieńcowa



(c) Gruczoły sutkowe



(d) Leukocyty

Rysunek 3.11: Diagramy podobieństwa dla poszczególnych typów komórek.

Rozdział 4

Wnioski

Celem prac przedstawionych w rozprawie było zaproponowanie integracyjnych metod statystycznych oraz eksploracji danych jako posiadających potencjał w poszukiwaniu biomarkerów chorób cywilizacyjnych. Cel ten został osiągnięty na wielu płaszczyznach.

Wpierw zademonstrowano konieczność stosowania algorytmów identyfikacji oraz korekcji efektu paczki w danych pozyskanych technikami wysokoprzepustowymi. Ponadto zaproponowano skuteczną metodę identyfikacji efektu paczki dla danych, które można posortować według czasu przetwarzania próbek lub innego czynnika. Identyfikacja oparta jest na algorytmie programowania dynamicznego, a liczba paczek jest ustalana z użyciem statystyki gPCA. Efektywność algorytmu została potwierdzona zarówno na zbiorach danych, gdzie oryginalny podział był podany *a priori*, jak i na zbiorach z nieznaną strukturą paczek. Analiza funkcjonalna i literaturowa wykazała dodatkowe informacje na temat badanych procesów uzyskane na podstawie danych ze zidentyfikowanym oraz skorygowanym efektem paczki.

Łączenie danych w ramach jednej dziedziny pozwala na zwiększenie mocy testowania statystycznego oraz uzyskanie lepszych jakościowo wyników wnioskowania. Wyzwaniem w tym wypadku pozostaje sprowadzenie danych pozyskanych z różnych platform do wspólnej przestrzeni obliczeniowej oraz biologicznej. W pracy dokonano analizy danych transkryptomicznych pochodzących od pacjentek cierpiących na raka piersi poddanych radioterapii. Eksperymenty zostały przeprowadzone na dwóch różnych rodzajach miromacierzy i wymagały ujednolicenia ze względu na różną liczbę kanałów. Dane przekształcono do wspólnej przestrzeni za pomocą korekcji efektu paczki. Następnie pokazano, że statystyczna integracja p-wartości pozwala na uzyskanie sygnatury zapewniającej pełną separowalność pacjentek radiowrażliwych oraz radioopornych. Na koniec zaproponowano metodę selekcji cech na podstawie profilów odpowiedzi na promieniowanie, co pozwoliło uzyskać lepsze wyniki klasyfikacji metodą wielokrotnej kroswalidacji stratyfikowanej. Cechy dominujące w modelach zostały potwierdzone jako cechy o największej liczbie zależności w odpowiedzi na promieniowanie zarówno w przeglądzie literaturowym, jak i niezależną metodą selekcji Monte Carlo. Podsumowując, wykorzystanie połączenia technik eksploracji oraz statystycznej integracji danych z analizą funkcjonalną jest skuteczną procedurą dla analizy

danych wieloplatformowych w celu zbadania mechanizmów odpowiedzi na dawkę promieniowania w nowotworze piersi.

Analiza integracyjna danych transkryptomicznych i proteomicznych umożliwiła potwierdzenie oraz pogłębienie wiedzy o mechanizmach choroby wieńcowej indukowanej promieniowaniem jonizującym. Rozróżnienie białek, których regulacja zależy od dawki od białek zależnych od wieku, pozwoliło na dokładniejszą analizę mechanizmów regulacji oraz lepszą efektywność integracji z danymi RNA-seq. Statystyczna integracja p-wartości potwierdziła wcześniej zidentyfikowane procesy, m.in. ścieżkę sygnałową PPAR. Łączona analiza danych podkreśla istotność stosowania niestandardowych metod analizy w łączeniu danych z biologii molekularnej dla zwiększenia szansy wyciągnięcia właściwych wniosków na temat badanych chorób.

Przypadek analizy proteomiki egzosomów wielotkankowych wskazuje na znaczenie rozwijania metod przetwarzania złożonych danych. Analiza podobieństwa między typami komórek pozwoliła na ocenę głównych czynników różnicowania napromieniowanych próbek. Wyniki te mogą być punktem odniesienia dla planowania dalszych eksperymentów dotyczących wzorców proteomicznych w egzosomach.

Algorytmy oraz metody opracowane w ramach tej rozprawy doktorskiej stanowią nowatorskie podejście do analiz statystycznych oraz eksploracji danych w wysokoprzepustowych eksperymentach biologii molekularnej. Algorytm identyfikacji efektu paczki metodą programowania dynamicznego jest oryginalnym narzędziem, udostępniony społeczności naukowej w postaci implementacji w pakiecie R BatchI. Równocześnie procedury zaproponowane do analizy integracyjnej wieloplatformowej, wielodziedzinowej oraz międzytkankowej nie były do tej pory opracowane ani wykorzystane w zaproponowanej tu formie. Pomyślne wyniki uzyskane przy ich zastosowaniu są wartościowym wkładem w nieustający rozwój metod analizy danych biomedycznych.

Bibliografia

Azimzadeh, O., Azizova, T., Merl-Pham, J., Subramanian, V., Bakshi, M. V., Moseeva, M., Zubkova, O., Hauck, S. M., Anastasov, N., Atkinson, M. J., et al. (2017). A dose-dependent perturbation in cardiac energy metabolism is linked to radiation-induced ischemic heart disease in Mayak nuclear workers. *Oncotarget*, 8(6):9067.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284.

Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., and Pevzner, P. A. (2007). Clustering millions of tandem mass spectra. *Journal of proteome research*, 7(01):113–122.

Glaab, E., Garibaldi, J. M., and Krasnogor, N. (2009). Arraymining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC bioinformatics*, 10(1):358.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

Krol, L. (2015). Distributed Monte Carlo feature selection: extracting informative features out of multidimensional problems with linear speedup. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, pages 463–474. Springer.

Labaj, W., Papiez, A., Polanski, A., and Polanska, J. (2017). Comprehensive analysis of mile gene expression data set advances discovery of leukaemia type and subtype biomarkers. *Interdisciplinary Sciences: Computational Life Sciences*, 9(1):24–35.

Papiez, A., Azimzadeh, O., Tapio, S., and Polanska, J. (2019a). Integrative multiomics study for validation of mechanisms in radiation-induced ischemic heart disease. *PloS ONE*.

Papiez, A., Badie, C., and Polanska, J. (2019b). Machine learning techniques combined with dose profiles indicate radiation response biomarkers. *International Journal of Applied Mathematics and Computer Science*, 29(1).

Papiez, A., Finnon, P., Badie, C., Bouffler, S., and Polanska, J. (2014). Integrating expression data from different microarray platforms in search of biomarkers of radiosensitivity. In *IWBBIO*, pages 484–493.

Papiez, A., Kabacik, S., Badie, C., Bouffler, S., and Polanska, J. (2015). Statistical integration of p-values for enhancing discovery of radiotoxicity gene signatures. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 503–513. Springer.

Papiez, A., Marczyk, M., Polanska, J., and Polanski, A. (2018). Batchi: Batch effect identification in high-throughput screening data using a dynamic programming algorithm. *Bioinformatics*.

Reese, S. E., Archer, K. J., Therneau, T. M., Atkinson, E. J., Vachon, C. M., de Andrade, M., Kocher, J.-P. A., and Eckel-Passow, J. E. (2013). A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. *Bioinformatics*, page btt480.

Yarnold, J., Ashton, A., Bliss, J., Homewood, J., Harper, C., Hanson, J., Haviland, J., Bentzen, S., and Owen, R. (2005). Fractionation sensitivity and dose response of late adverse effects in the breast after radiotherapy for early breast cancer: long-term results of a randomised trial. *Radiotherapy and oncology*, 75(1):9–17.

Zaykin, D. V. (2011). Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, 24(8):1836–1841.