

Politechnika Warszawska

WYDZIAŁ ELEKTRONIKI
I TECHNIK INFORMACYJNYCH



Instytut Informatyki

Praca dyplomowa magisterska

na kierunku Informatyka
w specjalności Inżynieria systemów informatycznych

Badanie widm metabolitów z grupy lipidów i ich pochodnych na
podstawie wyników opublikowanych w bazie danych HMDB

Jakub Mariusz Kierejsza

Numer albumu 261423

promotor
prof. dr hab. inż. Jan Mulawka

WARSZAWA 2020

Badanie widm metabolitów z grupy lipidów i ich pochodnych na podstawie wyników opublikowanych w bazie danych HMDB

Streszczenie. Metabolomika jest jedną z dziedzin metod bioinformatyki zajmującą się analizą ilościową i jakościową metabolitów (produktów przemiany materii w organizmach żywych). Powstała ona już w latach 50-tych ubiegłego wieku. Jednak dopiero rozwój technologiczny w informatyce i chemii analitycznej pozwolił na dokładniejsze badanie związków metabolicznych. Ilość danych, które taka analiza generuje, okazuje się zbyt duża, aby człowiek mógł sprawnie się nimi posługiwać. Powstała przez to potrzeba zaprojektowania i zrealizowania programu, który usprawni proces prawidłowej analizy jakościowej i porównawczej z wykorzystaniem spektrometru mas, która używa danych z ogólnodostępnej internetowej bazy HMDB. Głównym celem niniejszej pracy jest usprawnienie procesu postępowania podczas wyznaczania parametrów rejestracji metabolitów przez badaczy, chemików i analityków oraz zbadanie zależności pomiędzy utraconymi cząsteczkami obojętnymi w widmach lipidów i ich pochodnych. Problem wyznaczania parametrów rejestracji polegał na trudności w dostępie do wielkiej ilości danych z bazy HMDB i szybkim manipulowaniu nimi, jak również wykonywaniu odpowiednich transformacji. Oba problemy rozwiązano za pomocą odpowiedniego narzędzia informatycznego zaimplementowanego w języku python.

Słowa kluczowe: HMDB, utracone cząsteczki obojętne, lipidy, python, django, mongodb, MRM, metabolit, metabolomika

Investigation of metabolite spectra of lipids and lipid-like molecules based on data from Humane Metabolome Database

Abstract. Metabolomics is one of bioinformatics branches. Its primary objective is to study quantitative and qualitative analysis of metabolites (products of metabolism in live organisms). The concept of metabolomics appears as early as in the 1950s. Only later technological advancement in informatics and analytical chemistry allowed for more complex research about chemical metabolites. Unfortunately data quantity generated during such analysis turns out to be too much for human to use efficiently. This created the need to design and develop program, which will improve and streamline process of quantitative and comparative mass spectrometer analysis which use data from HMDB. The main purpose of this work is simplification of method for determining metabolite registration parameters used by researches, chemists and analysts as well as analysis of relationship between neutral loss compounds in lipids and lipid-like molecules. The problem in previous method of determining metabolite registration parameters was the difficulty in accessing to large amounts of HMDB data as well as performing relevant transformations on them. Both problems were solved with use of appropriate tool, which was implemented with use of python programming language.

Keywords: HMDB, neutral loss, lipids, python, django, mongodb, MRM, metabolite, metabolomics



.....
miejscowość i data

.....
imię i nazwisko studenta

.....
numer albumu

.....
kierunek studiów

OŚWIADCZENIE

Świadomy/-a odpowiedzialności karnej za składanie fałszywych zeznań oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie, pod opieką kierującego pracą dyplomową.

Jednocześnie oświadczam, że:

- niniejsza praca dyplomowa nie narusza praw autorskich w rozumieniu ustawy z dnia 4 lutego 1994 roku o prawie autorskim i prawach pokrewnych (Dz.U. z 2006 r. Nr 90, poz. 631 z późn. zm.) oraz dóbr osobistych chronionych prawem cywilnym,
- niniejsza praca dyplomowa nie zawiera danych i informacji, które uzyskałem/-am w sposób niedozwolony,
- niniejsza praca dyplomowa nie była wcześniej podstawą żadnej innej urzędowej procedury związanej z nadawaniem dyplomów lub tytułów zawodowych,
- wszystkie informacje umieszczone w niniejszej pracy, uzyskane ze źródeł pisanych i elektronicznych, zostały udokumentowane w wykazie literatury odpowiednimi odnośnikami,
- znam regulacje prawne Politechniki Warszawskiej w sprawie zarządzania prawami autorskimi i prawami pokrewnymi, prawami własności przemysłowej oraz zasadami komercjalizacji.

Oświadczam, że treść pracy dyplomowej w wersji drukowanej, treść pracy dyplomowej zawartej na nośniku elektronicznym (płycie kompaktowej) oraz treść pracy dyplomowej w module APD systemu USOS są identyczne.

.....
czytelny podpis studenta

Spis treści

1. Wstęp	11
1.1. Rozwój metabolomiki	11
1.2. Baza danych HMDB	11
1.3. Widma masowe	11
1.4. Zakres i cel pracy	12
2. Postawienie i propozycja rozwiązania problemu	14
2.1. Aplikacja wyznaczająca parametry rejestracji	14
2.1.1. Aplikacja webowa	14
2.1.2. Architektura aplikacji	15
2.2. Badanie zależności pomiędzy utraconymi cząsteczkami obojętnymi	15
2.2.1. Utracone cząsteczki obojętne	16
3. Wybór technologii	17
3.1. Baza danych	17
3.1.1. Baza relacyjna czy nierelacyjna	17
3.1.2. MongoDB	17
3.2. Język programowania	18
3.3. Framework	18
3.3.1. Warstwa reprezentacyjna	18
3.4. Serwer	18
3.4.1. Maszyna wirtualna	18
3.4.2. Gunicorn	18
3.4.3. Nginx	19
4. Opis tworzonej aplikacji	20
4.1. Dane w bazie HMDB	20
4.1.1. Metabolity	20
4.1.2. Widma	20
4.2. Transformacja i procesowanie danych	20
4.3. Struktura bazy aplikacyjnej	21
4.4. Backend	25
4.4.1. Struktura django	25
4.4.2. Widoki	26
4.5. Frontend	27
4.5.1. Szablony HTML	27
4.5.2. Javascript	27
4.5.3. Wygląd i sposób działania	28
5. Przeprowadzone testy	34
5.1. Testy automatyczne	34
5.1.1. Struktura testów automatycznych	34
5.1.2. Opis testów jednostkowych	35

0. Spis treści

5.2.	Testy manualne	35
5.2.1.	Testy bazy danych	36
5.2.2.	Testy frontendu	36
5.2.3.	Testy użytkownika	36
5.3.	Wyniki testów	37
5.3.1.	Wyniki testów manualnych	37
5.3.2.	Wyniki pierwszych testów użytkownika	37
5.3.3.	Wyniki drugich testów użytkownika	38
6.	Badania	39
6.1.	Badane dane	39
6.1.1.	Lipidy i ich pochodne	39
6.1.2.	Transformacja i procesowanie danych	39
6.2.	Metoda asocjacyjna	40
6.2.1.	Algorytm apriori	41
7.	Wyniki badań	43
7.1.	Liczliwość utraconych cząsteczek obojętnych na pełnym zbiorze związków	43
7.2.	Liczliwość utraconych cząsteczek obojętnych w grupie lipidów i ich pochodnych	48
7.3.	Podsumowanie podliczeń	52
7.4.	Reguły asocjacyjne	53
7.5.	Wnioski	60
8.	Eksperymentowanie z utworzoną aplikacją	62
8.1.	Przykład 1 - wyznaczanie parametrów rejestracji kwasu dikowoilochinowego	62
8.1.1.	Manualne wyznaczenie parametrów	62
8.1.2.	Wyznaczenie parametrów z wykorzystaniem utworzonego oprogramowania	65
8.1.3.	Porównanie	66
8.2.	Przykład 2 - wyznaczanie parametrów rejestracji 1-Metylohistaminy	67
8.2.1.	Manualne wyznaczenie parametrów	67
8.2.2.	Wyznaczenie parametrów z wykorzystaniem utworzonego oprogramowania	70
8.2.3.	Porównanie	71
8.3.	Przykład 3 - identyfikacja Karwonu	72
8.3.1.	Identyfikacja związku przy pomocy strony HMDB	72
8.3.2.	Identyfikacja związku z wykorzystaniem utworzonego oprogramowania	74
8.3.3.	Porównanie	75
9.	Zakończenie	76
9.1.	Zrealizowanie celu pracy	76
9.2.	Aplikacja webowa	76
9.3.	Analiza utraconych cząsteczek neutralnych	77
9.4.	Dalsze kierunki badań	77
Bibliografia	79	
Wykaz symboli i skrótów	80	
Spis rysункów	80	

Spis tabel	82
-------------------	----

1. Wstęp

W dzisiejszych czasach medycyna i biologia jest nieodłącznie związana z nowymi technologiami, zwłaszcza w dziedzinie komputerów i chemii analitycznej. To właśnie ona pozwala na nowe odkrycia, dokładniejsze badania, ale też przyspiesza i ułatwia dotychczasową pracę w wielu ośrodkach badawczych i firmach. Coraz lepsze aparatury pomiarowe oraz publicznie dostępne bazy danych spowodowały gwałtowny rozwój, ale też przyczyniły się do natłoku informacji, które należy usystematyzować za pomocą odpowiednich programów. Jedną z takich dziedzin jest metabolomika, w której wiele osób pokłada wielkie nadzieje jako przyszłość diagnostyki medycznej.

1.1. Rozwój metabolomiki

Podstawowe przeświadczenie, że produkty ludzkiego metabolizmu zawierają информацию o stanie organizmu, było znane od bardzo dawna. Już w starożytnych Chinach używano mrówek, by sprawdzić stężenie glukozy w moczu, której obecność mogła być symptomem cukrzycy. Pierwsza wzmianka o profilu metabolicznym pojawiła się w późnych latach 40-tych, lecz dopiero w 1971 roku rozwój technologii pozwolił na ilościowy pomiar metabolitów [1], co pozwoliło na dokładniejsze badania. W 2005 roku powstała pierwsza baza danych zawierająca informacje o ludzkich metabolitach: METLIN. Zawiera ona dane spektrometryczne ponad 500 tysięcy związków, a liczba ta ciągle się zwiększa [2]. W 2015 roku zaprezentowano profilowanie metabolomu w czasie rzeczywistym [3].

1.2. Baza danych HMDB

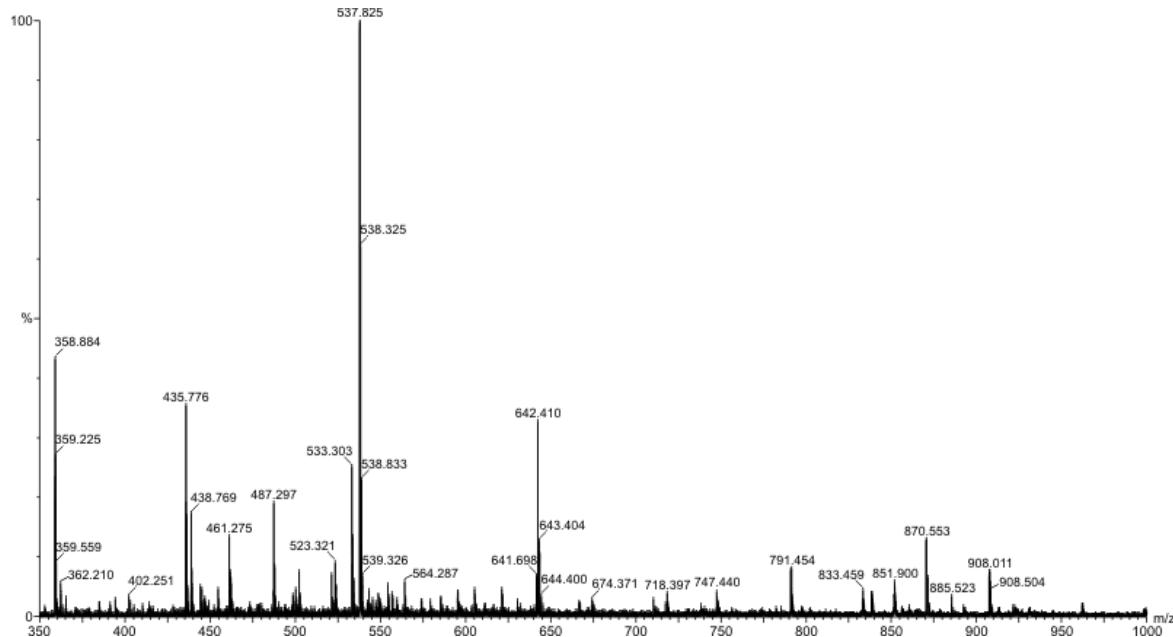
HMDB (Human Metabolome Database) [4] [5] [6] [7] jest nieodpłatną bazą zawierającą szczegółowe informacje na temat metabolitów znajdujących się w ludzkim ciele. Została ona utworzona w 2007 roku podczas projektu Human Metabolome Project sponsorowanego przez Genome Canada. Została ona zaprojektowana w celu utrzymywania trzech rodzajów informacji: chemicznych, klinicznych i biochemicznych. Obecnie (2020 rok) baza zawiera informacje na temat 114 177 metabolitów. Stanowi ona nieocenione źródło danych przydatnych w medycznych badaniach podstawowych oraz ogólnie w diagnostyce medycznej.

1.3. Widma masowe

Znaczna większość niniejszej pracy opiera się na widmach masowych zapisanych w bazie danych HMDB. Widmo masowe jest to wykres intensywności dla stosunku masy do ładunku (m/z). Otrzymuje się je przy użyciu spektrometru mas. Analizowana substancja jest uderzana jonami pozytywnymi lub negatywnymi, które przyczepiają się do cząsteczek powodując ich rozpad na cząsteczkę obojętną oraz odpowiednio naładowaną cząsteczkę, która może podlec dalszemu rozpadowi. Cząsteczka z dołączonym dodatkowym jonom

1. Wstęp

jest rejestrowana przez czujniki i pokazana na wykresie jako charakterystyczna dla niej wartość stosunku jej masy do jej ładunku. Intensywność jest wartością znormalizowaną do procentów, gdzie najsilniejszy sygnał stanowi wartość 100%. Reprezentuje ona częstotliwość, z jaką dana cząsteczka została wykryta przez czujnik.



Rysunek 1. Przykład widma masowego.

1.4. Zakres i cel pracy

Niejednorodność bazy danych HMDB (zawiera ona także dane na temat metabolitów: roślin, leków, związków toksycznych, narkotyków i innych niesyntezowanych przez człowieka związków) ogranicza w znaczącym stopniu jej stosowalność.

Pierwszym celem tej pracy jest przyspieszenie i usprawnienie procesu postępowania podczas wyznaczania parametrów rejestracji metabolitów wykonywanej przez badaczy, chemików i analityków. Problem polegał na trudności w dostępie do wielkiej ilości danych z bazy HMDB i szybkim manipulowaniu nimi, jak również wykonywaniu odpowiednich transformacji. Problem ma być rozwiązany za pomocą odpowiedniego narzędzia informatycznego zaimplementowanego w języku python. Stworzone narzędzie ma pozwalać na wyszukiwanie metabolitów oraz określanie dla nich charakterystycznych parametrów takich jak stosunku masy do ładunku, dla których mają być dokonane pomiary za pomocą spektrometru mas. Narzędzie to będzie pozwalało na wykonywanie analiz jakościowych i porównawczych bez stosowania kosztownych substancji chemicznych jako wzorcowych, a także pozwoli badać nowo odkryte metabolity, co znacznie przyśpieszy rozwój diagnostyki medycznej.

Drugim celem jest zbadanie zależności pomiędzy utraconymi cząsteczkami obojętnymi w widmach metabolitów z grupy lipidów i ich pochodnych. Utracone cząstki obo-

jetne (ang. neutral loss) są rzadko badaną grupą. Jest to spowodowane faktem, że widma masowe nie zawierają jednoznacznej informacji o występujących po rozpadzie cząstkach obojętnych. Ich analiza może pozwolić odkryć dodatkowe zależności i informacje dotyczące metabolitów z grupy lipidów i ich pochodnych.

2. Postawienie i propozycja rozwiązania problemu

2.1. Aplikacja wyznaczająca parametry rejestracji

Pierwszym celem tej pracy było ułatwienie postępowania przy wyszukiwaniu i spersonowaniu odpowiednich parametrów stosunku masy do ładunku. Rozmowa z użytkownikiem i głębsza analiza problemu wyłoniła dodatkowe wymagania, które rozwiązanie powinno spełniać.

Do wykonania prawidłowej analizy jakościowej i porównawczej przy użyciu spektrometru mas potrzebne są odpowiednie parametry stosunku masy do ładunku oraz parametry spektrometru, które były użyte podczas ich uzyskania. Po wprowadzeniu ustawień do maszyny wykonuje się analizę, a następnie weryfikuje wyniki z zewnętrznymi parametrami. Niestety, mimo że baza danych HMDB zawiera wszystkie potrzebne informacje, to nie ma ona łatwego sposobu na ich uzyskanie. Jest to robione ręcznie poprzez wyszukanie badanego związku, a następnie wejściu w każde kolejne widmo spektralne i przekopiowanie szukanych parametrów jeden po drugim. Są one umieszczane w formie wykresu, który ułatwia szybką analizę wizualną, ale nie pozwala na łatwe w obsłudze kopiowanie wartości. Jest to monotonna i czasochłonna praca. Przez to zaistniała potrzeba stworzenia narzędzia, które wspomoże i ułatwi wyszukanie i spersonowanie danych do późniejszego porównania.

Szczegółowe wymagania powodują, że proponowane rozwiązanie powinno spełniać następujące postulaty:

- mieć możliwość wyszukiwania metabolitów po nazwie, klasie macierzystej, klasie, podklasie, miejscu występowania i przedziale mas,
- wyświetlać parametry rejestracji widm MS-MS w intuicyjnej i łatwej do skopiowania formie,
- mieć możliwość filtrowania parametrów w zależności od ich intensywności,
- sortować parametry od największej intensywności do najmniejszej,
- oddzielać wyniki w zależności od trybu jonizacji, a w przypadku jak taki tryb nie jest określony to o tym informować,
- działać szybko i intuicyjnie,
- grupować parametry po widmach, z których zostały odczytane,
- być łatwo dostępne z dowolnego komputera.

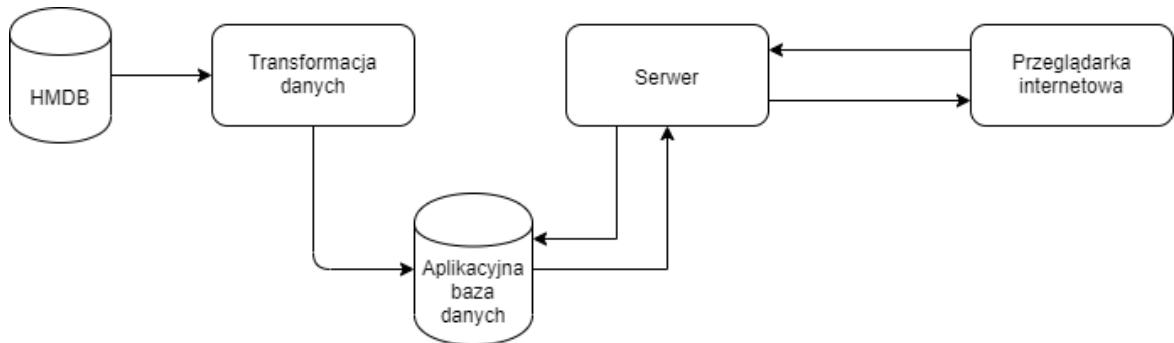
Biorąc pod uwagę powyższe wymagania, najlepszym rozwiązaniem jest aplikacja webowa.

2.1.1. Aplikacja webowa

Aplikacja webowa (zwana również aplikacją internetową) jest to program komputerowy pracujący na serwerze, który udostępnia użytkownikowi poprzez sieć komputerową interfejs w oknie przeglądarki internetowej. Aby skorzystać z takiego rozwiązania, po-

trzebny jest dostęp do globalnej sieci internetowej, poprzez którą przeglądarka komunikuje się z serwerem aplikacji. Główną zaletą takiego rozwiązania jest jego dostępność. Użytkownik nie potrzebuje mieć dostępnej całej bazy danych lokalnie, tylko potrzebuje konkretnych związków i ich parametrów. Takie rozwiązanie też pozwoli na dalszy rozwój aplikacji poprzez dodanie nowych modułów, których dystrybucja będzie wymagała jedynie aktualizacji programu na maszynie serwera.

2.1.2. Architektura aplikacji



Rysunek 2. Diagram architektury aplikacji.

Na rysunku 2 jest zaprezentowana architektura proponowanej aplikacji. Baza danych HMDB nie ma możliwości wykonywania zapytań bezpośrednio do niej, więc trzeba dane ściągnąć i umieścić w aplikacyjnej bazie danych. W trakcie tego procesu dane zostaną też przefiltrowane i przetworzone tak, aby najważniejsze informacje były, jak najłatwiej dostępne. Następnie aplikacja w przeglądarce komunikuje się z serwerem i odpytuje go w celu otrzymania już szczegółowych fragmentów danych, które potem przesyła użytkownikowi i wyświetla w przeglądarce internetowej. Aplikacja służy tylko do odczytu i wyszukiwania, więc użytkownik nie ma możliwości ingerencji w stan bazy danych. Dzięki temu nie wymaga ona żadnej formy autoryzacji i może być dostępna dla wszystkich.

2.2. Badanie zależności pomiędzy utraconymi cząsteczkami obojętnymi

Drugim celem pracy jest zbadanie zależności pomiędzy utraconymi cząsteczkami obojętnymi w widmach metabolitów. Baza HMDB zawiera informacje o widmach ponad 114 tysięcy metabolitów, które mogą zostać pogrupowane na wiele sposobów. Badanie zależności w widmach związków, które nie są ze sobą powiązane, może produkować błędne wnioski, zatem zaistniała potrzeba wyspecyfikowania konkretnej grupy związków, które będą przebadane. Taka grupa musi być wystarczająco liczna, aby metody eksploracji danych mogły wykryć istniejące zależności. Ta praca skupiać się będzie głównie na związkach metabolicznych z grupą lipidów i ich pochodnych. Do zbadania zależności pomiędzy utraconymi cząsteczkami obojętnymi została wybrana analiza koszykowa. Jest to metoda eksploracji danych, której zadaniem jest znalezienie reguł asocjacyjnych

2. Postawienie i propozycja rozwiązania problemu

pomiędzy występującymi elementami. Ta metoda bardzo dobrze pasuje do podanego problemu, ponieważ każde widmo może być potraktowane jako transakcja zawierająca w sobie wszystkie utracone cząsteczki obojętne. Dzięki niej będzie widoczne, jak często poszczególne cząsteczki występują razem i jaki mogą mieć na siebie wpływ.

2.2.1. Utracone cząsteczki obojętne

Utracone cząsteczki obojętne (ang. neutral loss) są to związki, które odłączyły się od cząsteczek badanej substancji podczas analizowania jej przy użyciu spektrometru mas. Podczas takiej analizy rozproszona substancja jest bombardowana elektronami o zadanej energii mierzonej w elektronowoltach oraz odpowiedniej polaryzacji. Po zderzeniu elektron przyłącza się do związku, a przekazana energia powoduje jego rozpad na cząsteczkę z elektronem (naładowaną pozytywnie lub negatywnie) oraz cząsteczkę obojętną. Naładowany związek jest następnie wychwytywany przez czujniki spektrometru, który jest potem reprezentowany jako wierzchołek na widmie. Informacje o utraconej cząsteczce obojętnej nie są zapisywane w widmie. Taki sposób zapisu informacji powoduje, że badanie i analiza utraconych cząsteczek obojętnych jest bardzo trudna. Jest jednak możliwe uzyskanie informacji na ich temat poprzez obliczenie odległości pomiędzy poszczególnymi wierzchołkami widma. Niestety taki sposób niesie ze sobą też dużo błędnych informacji, ponieważ nie da się jednoznacznie określić, która odległość jest prawidłowa.

3. Wybór technologii

Z gotową propozycją rozwiązania trzeba było wybrać odpowiednie technologie, w których zostanie ona zrealizowana. W całym projekcie są cztery główne punkty, które należy wziąć pod uwagę: aplikacyjną bazę danych, framework aplikacji, warstwę reprezentacyjną oraz ustawienia serwera. Wybrane narzędzia też muszą odpowiednio wspomagać metody eksploracji danych oraz ułatwiać ich przetwarzanie i analizę.

3.1. Baza danych

3.1.1. Baza relacyjna czy nierelacyjna

Pierwszą ważną decyzją jest wybór bazy danych. Obecnie istnieją dwa rodzaje baz danych: relacyjne i nierelacyjne. Głównym warunkiem, który musi zostać spełniony, jest szybki odczyt i wyszukiwanie na dużym zbiorze danych. Bazy relacyjne wymagają, by dane były w postaci normalnej. Powoduje to, że w danych nie występuje redundancja, ale otrzymanie potrzebnego nam złączenia byłoby kosztowne. W przypadku nierelacyjnych baz danych można odpowiednio spreparować dokumenty tak, by jedno zapytanie zwracało wszystkie interesujące użytkownika informacje, bez potrzeby wykonywania zaawansowanych operacji. Według rankingu baz danych pod względem popularności wykonanego przez DB-Engines [8] MongoDB [9] jest najbardziej popularną nierelacyjną bazą danych i właśnie ona została wybrana w proponowanym rozwiązaniu.

3.1.2. MongoDB

MongoDB jest nieodpłatną dokumentową bazą danych, która cechuje się skalowalnością i elastycznością. Dokumenty są zapisywane w formacie JSON, co pozwala na wygodne przetrzymywanie całych struktur. Baza danych MongoDB ukazała się w 2009 roku. Posiada ona sterowniki do ponad dziesięciu różnych języków programowania, więc nie ogranicza wyboru technologii, w jakiej napisany byłby serwer. Dodatkowo MongoDB dostarcza wygodne narzędzia do zarządzania bazą danych, które pozwalają dokładnie ją kontrolować.

Ta baza danych jest dobrym rozwiązaniem dla przedstawionego w tym projekcie problemu. Aplikacja webowa potrzebuje tylko ściśle określonego zbioru danych, który będzie zaprezentowany jako jedna kolekcja. Pozwoli to na wykonanie pojedynczego zapytania i otrzymanie wszystkich interesujących użytkownika informacji bez potrzeby wykonywania żadnych złączeń. Taka struktura też jest bardzo korzystna do przeprowadzania badań. Pozwala ona na spreparowanie zbioru danych pod konkretną metodą, co przyspieszy jej wykonanie poprzez wyeliminowanie wykonywania niepotrzebnych złączeń. Dodatkowo baza danych HMDB oferuje dane w postaci XML, dlatego można je łatwo przefiltrować i przekształcić do formatu JSON używając skryptu.

3. Wybór technologii

3.2. Język programowania

Kolejnym krokiem jest wybór języka programowania, w którym napisana będzie aplikacja internetowa. Aplikacja nie jest krytyczna czasowo, więc jedynymi wymaganiami są wygoda rozwoju aplikacji oraz dostęp do narzędzi ułatwiających jej pracę. Właśnie python jest wysokopoziomowym językiem interpretowanym, który spełnia te wymagania. Python oferuje też szereg bibliotek i narzędzi, które ułatwiają eksplorację danych.

3.3. Framework

Django jest to wolny i otwarty framework służący do tworzenia aplikacji webowych. Projekt ten został opublikowany w 2005 roku i od tego czasu został wykorzystany w wielu popularnych stronach takich jak pinterest czy instagram. Dostarcza on automatycznie wygenerowany panel administracyjny oraz własny serwer do testowania aplikacji. Definiuje on strukturę aplikacji oraz mechanizm działania, a także umożliwia łatwy dostęp do najpopularniejszych baz danych przy pomocy sprzążonych bibliotek.

Do połączenia frameworku z bazą danych MongoDB została wykorzystana biblioteka django [10]. Udostępnia ona mechanizmy, które pozwalają na korzystanie z tej bazy danych bez zmiany w mapowaniu obiektowo-relacyjnym, które dostarcza django.

3.3.1. Warstwa reprezentacyjna

Django dostarcza możliwość podłączenia dodatkowych bibliotek do tworzenia interfejsów graficznych aplikacji internetowych takich jak react. Jednak w przypadku przedstawionego problemu wbudowane mechanizmy są zupełnie wystarczające. Pozwalają one na tworzenie i łączenie szablonów HTML oraz upraszczają komunikację między interfejsem a serwerem.

3.4. Serwer

Mimo wielu wbudowanych narzędzi django jest tylko frameworkiem aplikacji i nie udostępnia serwera produkcyjnego, na którym ta aplikacja może działać.

3.4.1. Maszyna wirtualna

Serwer zostanie oparty na maszynie wirtualnej na serwerach wydziału Chemii Politechniki Warszawskiej. Taka konfiguracja ułatwia zarządzanie oraz utrzymanie wielu współdziałających programów. Dodatkowo odseparowuje środowiska, co powoduje, że nie występują takie problemy, jak na przykład konflikty dotyczące wersji bibliotek.

3.4.2. Gunicorn

Gunicorn [11] jest to serwer WSGI napisany w języku programowania python. Dostarcza też zintegrowaną obsługę dla aplikacji napisanych w frameworku django. Został on użyty jako serwer aplikacji ustawiony za odwrotnym serwerem proxy.

3.4.3. Nginx

Jako odwrotny serwer proxy zostało wykorzystane oprogramowanie nginx [12]. Jest to popularne narzędzie wykorzystywane na całym świecie. Nginx dodatkowo dostarcza funkcjonalności równoważenia obciążenia.

Użycie odwrotnego serwera proxy pozwala stworzyć pojedynczy punkt dostępu dla wielu aplikacji działających równocześnie.

4. Opis tworzonej aplikacji

4.1. Dane w bazie HMDB

Baza danych HMDB udostępnia dane w różnych formatach w zależności od ich rodzaju. Znajdują się tam dane o metabolitach i proteinach, dane o ich strukturach oraz dane dotyczące ich widm. Problem przedstawiony w tej pracy dotyczy jedynie widm i danych o metabolitach i proteinach. Oba rodzaje są zapisane w formacie XML z tą różnicą, że dane metaboliczne znajdują się wszystkie w jednym pliku, a dane widm są podzielone po jednym pliku na widmo.

4.1.1. Metabolity

Dane dotyczące metabolitów zawierają 60 pól, jednak aplikacja korzysta tylko z niektórych z nich. W tabeli 1 zostały opisane pola, które wykorzystuje aplikacja.

Tabela 1. Wykorzystane pola metabolitów.

Nazwa	Opis
accession	id w bazie HMDB
name	nazwa związku
monoisotopic-molecular-weight	monoizotopowa masa związku
taxonomy	obiekt informacje na temat klasyfikacji związku
spectra	obiekt zawierający listę widm spektralnych powiązanych z danym związkiem
biological-properties	obiekt zawierający listy z lokalizacją związku w produktach metabolizmu, komórce oraz tkankach

4.1.2. Widma

W bazie HMDB są dostępne widma spektroskopii magnetycznego rezonansu jądro-wego NMR (ang. Nuclear Magnetic Resonance), spektroskopii masowej MS (ang. mass spectroscopy) oraz z tandemowego urządzenia łączącego techniki chromatografii gazowej i spektroskopii masowej. Zgodnie z wymaganiami aplikacja korzysta tylko z danych ze spektroskopii masowej z tandemowego spektrometru mas MS-MS. Zawierają one 27 pól, ale tak samo, jak w przypadku metabolitów aplikacja wykorzystuje tylko niektóre z nich. Zostały one zaprezentowane w tabeli 2.

4.2. Transformacja i procesowanie danych

Pościągnięciu odpowiednich plików z danymi należało je następnie przeprocesować. Głównym zadaniem tego kroku jest zamiana danych z formatu XML na format JSON, który będzie łatwy do przeczytania przez bazę danych MongoDB. Do stworzenia tego skryptu został wykorzystany język programowania python z biblioteką *xmldict*, która czyta dane

Tabela 2. Wykorzystane pola widm.

Nazwa	Opis
collision-energy-voltage	ustawione napięcie podczas wykonywania pomiaru
ionization-mode	tryb jonizacji pozytywny lub negatywny
ms-ms-peaks	lista zmierzonych wierzchołków zawierająca masy i intensywność pomiarów

i tworzy z nich słownik. Podczas procesowania danych postępowanie to ujednolica pola zawierające listy. Jest to konieczne, ponieważ w pliku XML obiekty nie posiadają znacznika, czy pole zawiera jeden element, czy wiele. Powoduje to niejednolitość w typie pola. Konieczne też były drobne zmiany w nazewnictwie. Jest to spowodowane konfliktem ze słowami kluczowymi języka programowania python (np. *class*) lub znakami specjalnymi, które nie mogą być w nazwie zmiennych (np. “-”. Niestety konieczność zmiany nazw spowodowała, że konieczne było skopiowanie struktury, co podwaja ilość zajętej pamięci podręcznej.

Biblioteka *xmltodict* podczas parsowania danych nie rozpoznaje automatycznie typu czytanych danych. Powoduje to, że wszystkie liczby są zapisywane jako ciąg znaków. Dlatego następnym krokiem jest ponowne przejście przez strukturę i przekształcenie ciągów znaków na liczby. Dodatkowo proces zawęża dane do tylko takich, które są używane w aplikacji. Potem przy pomocy biblioteki *json* struktura jest zapisywana w docelowym formacie.

Kroki w procesie dla obu rodzajów danych (metabolitów i widm) są takie same. Jedyna różnica jest taka, że metabolity znajdują się wszystkie w jednym pliku, a widma są podzielone na oddzielne pliki.

4.3. Struktura bazy aplikacyjnej

Aplikacja łączy się z bazą danych z wykorzystaniem konta aplikacyjnego. Wykorzystany framework django tworzy też standardową strukturę bazy, która dzieli się na trzy części.

Pierwsza część jest związana z autoryzacją, użytkownikami i uprawnieniami. Są to wszystkie kolekcje zaczynające się od przedrostka *auth_*, które zostały wygenerowane automatycznie przez django. W tabeli 3 opisane jest zastosowanie każdej z nich. Zaprezentowane w tej pracy rozwiązanie nie używa tych automatycznie wygenerowanych tabel. Jest to spowodowane tym, że aplikacja nie posiada żadnych ról użytkowników, tylko ma być otwarta do wykorzystania dla wszystkich.

Druga część bazy danych związana jest z informacjami pomocniczymi dla django.

4. Opis tworzonej aplikacji

Tabela 3. Kolekcje administracyjne.

Nazwa	Opis
auth_group	nazwy grup, do których przypisuje się użytkowników
auth_group_permissions	lista uprawnień przypisanych do grupy
auth_permission	lista wszystkich uprawnień
auth_user	lista użytkowników
auth_user_groups	lista przypisująca użytkownika do grupy
auth_user_user_permissions	lista uprawnień przypisanych do użytkownika

Wszystkie te kolekcje zaczynają się od przedrostka *django_*. Zawierają one informacje na temat migracji, sesji użytkowników czy dziennika administracyjnego.

Ostatnia część z kolekcjami zaczynającymi się od przedrostka *hmdb_*, zawiera informację na temat przetworzonych już danych z bazy HMDB. Znajdują się tutaj cztery kolekcje, z których dwie są pomocnicze, a dwie już z właściwymi danymi.

Hmdb_biolocation jest kolekcją pomocniczą, która zawiera wyłuskaną listę substancji, gdzie dany związek występuje. Te informacje są wykorzystywane jako pomoc podczas wyszukiwania związków przez użytkownika.

Hmdb_met_names jest kolekcją pomocniczą zawierającą tylko dane metabolitów, po których użytkownik ma możliwość wyszukiwania. Tabela 4 zamiera dokładny opis struktury trzymanych dokumentów. Ta kolekcja posiada założone indeksy na każdym polu, po którym użytkownik może wyszukiwać związki.

Tabela 4. Struktura kolekcji *hmdb_met_names*.

nazwa	nazwa związku
met_id	indeks metabolitu w kolekcji <i>hmdb_metabolite</i>
super_class	nadklasa w taksonomii związku
main_class	główna klasa w taksonomii związku
sub_class	podkласa w taksonomii związku
biospecimen_locations	lista substancji gdzie występuje dany związek
monisotopic_molecular_weight	monoizotopowa masa związku

Kolekcja *hmdb_metabolites* posiada dane szczegółowe dotyczące metabolitów. Znajduje się w niej 114 tysięcy metabolitów, z których każdy posiada 45 pól. Są to:

- version - wersja bazy HMDB,
- creation_date - data utworzenia dokumentu metabolitu,
- update_date - data aktualizacji dokumentu,
- accession - lista id na stronie HMDB; jest to lista, ponieważ w historii HMDB były przypadki redundancji rekordów pod inną nazwą; po ostatnich zmianach zawiera tylko jedną wartość,

- status - status metabolitu,
- secondary_acccessions - lista zamiennych id na stronie HMDB, które prowadzą do tego samego związku,
- name - nazwa związku,
- cs_description - opis związku,
- description - dodatkowy opis charakteryzacji substancji,
- synonyms - synonimy nazwy związku,
- chemical_formula - wzór chemiczny,
- average_molecular_weight - średnia waga molekularna,
- monoisotopic_molecular_weight - monoizotopowa waga związku,
- iupac_name - nazwa związku w Międzynarodowej Unii Chemii Czystej i Stosowanej,
- traditional_iupac - tradycyjna nazwa związku,
- cas_registry_number - oznaczenie numeryczne przypisane związkowi przez organizację Chemical Abstracts Service,
- smiles - simplified molecular-input line-entry system (SMILES), jednoznaczny zapis struktury chemicznej w formacie ASCII,
- inchi - międzynarodowy identyfikator związku,
- inchikey - hasz s tworzony z identyfikatora inchi,
- taxonomy - taksonomia związku; zawiera:
 - direct_parent - bezpośredni rodzic związku,
 - kingdom - królestwo,
 - super_class - klasa nadrzędna,
 - main_class - klasa główna,
 - sub_class - podklasa,
 - molecular_framework - szkielet molekularny,
 - alternative_parents - lista innych rodziców związku,
 - substituents - lista zamienników,
 - external_descriptors - lista zewnętrznych deskryptorów,
- state - stan skupienia,
- experimental_properties - lista właściwości otrzymanych w sposób eksperymentalny; każdy element zawiera nazwę oraz wartość,
- predicted_properties - lista oczekiwanych właściwości; każdy element zawiera nazwę oraz wartość,
- spectra - lista id widm powiązanych ze związkiem,
- biological_properties - właściwości biologiczne; zawiera:
 - cellular_locations - lista miejsc występowania związku w komórce,
 - biospecimen_locations - lista miejsc występowania związku w produktach metabolizmu,
 - tissue_locations - lista miejsc występowania związku w tkankach,

4. Opis tworzonej aplikacji

- pathways - lista szlaków metabolicznych,
- normal_concentrations - lista obiektów ze stężeniami związku dla zdrowych osobników,
- abnormal_concentrations - lista obiektów ze stężeniami związku dla osobników z podanym stanem medycznym,
- diseases - lista chorób, lub powodów dla którego mogą występować zaburzenia w stężeniu danego związku,
- kegg_id, chebi_id, chemspider_id, pubchem_compound_id, foodb_id, drugbank_id, phenol_explorer_compound_id, meta_cyc_id, wikipedia_id, knapsack_id, bigg_id, metlin_id, pdb_id - id związku dla odpowiadających zewnętrznych systemów,
- synthesis_reference - odniesienie do publikacji dotyczącej syntezy,
- general_references - lista publikacji dotycząca danego związku,
- protein_associations - lista powiązanych białek.

Hmdb_spectra jest ostatnią z kolekcji zawierających dane dotyczące bazy HMDB. Posiada ona prawie 460 tysięcy widm. Struktura wszystkich widm masowych jest jednolita. Mimo to wiele widm ma niekompletne dane lub dane odbiegające od standardów. Ta praca nie zajmuje się ich korekcją, tylko zostawia to ocenie użytkownika. Każdy dokument składa się z:

- notes - notatki z dodatkowymi informacjami,
- sample_concentration - stężenie próbki,
- solvent - rozpuszczalnik,
- sample_mass - masa próbki,
- sample_assessment - oszacowanie jakości próbki,
- spectra_assessment - oszacowanie jakości widma,
- sample_source - źródło pochodzenia próbki,
- collection_date - data pobrania próbki,
- instrument_type - rodzaj wykorzystanej aparatury,
- peak_counter - liczba wierzchołków w widmie,
- created_at - data utworzenia dokumentu,
- updated_at - data ostatniej aktualizacji dokumentu,
- mono_mass - masa monoizotopowa związku,
- collision_energy_level - poziom energii kolizji,
- collision_energy_voltage - napięcie kolizji ustawione na urządzeniu,
- ionization_mode - tryb jonizacji. Może być albo pozytywny, albo negatywny,
- sample_concentration_units - jednostki stężenia próbki,
- sample_mass_units - jednostki masy próbki,
- predicted - wartość boolowska, określająca czy widmo było eksperymentalne czy oszacowane,
- structure_id - numer porządkowy struktury,

- `splash_key` - klucz SPLASH, który jednoznacznie określa widmo,
- `database_id` - id metabolitu, dla którego jest dane widmo,
- `references` - lista publikacji zawierająca dane widmo spektralne,
- `ms_ms_peaks` - liczba wierzchołków widma, składająca się z ich intensywności i masy.

Ostatnią kolekcją jest *sequence*. Jest to pomocnicza kolekcja, która zawiera dokumenty, które służą jako odpowiedniki sekwencji liczbowych w SQL. Jest ona potrzebna, ponieważ MongoDB nie udostępnia takiego mechanizmu w standardzie.

4.4. Backend

4.4.1. Struktura django

Django posiada zdefiniowaną strukturę, która ułatwia zarządzanie projektem. Taki projekt składa się z aplikacji, z których każda jest oddzielnym tworem. Jednym obowiązkowym elementem, który musi posiadać każda aplikacja, jest plik *urls.py*. Jego zadaniem jest nawigowanie i komunikacja pomiędzy aplikacjami przy użyciu adresu URL oraz przekierowywanie zapytań do funkcji, które je obsługują.

Aplikacja główna, która nazywa się tak samo jak utworzony projekt, może być własną odrębną aplikacją, ale głównie używana jest jako jeden punkt dostępu dla całego projektu. Poza plikiem *urls.py* musi ona jeszcze zawierać dwa dodatkowe pliki: *settings.py* i *wsgi.py*. Plik *settings.py* jest to plik konfiguracyjny. Zdefiniowane są w nim połączenia do bazy danych, użyte aplikacje (zarówno stworzone na potrzeby projektu, jak i dostarczone przez deweloperów django), oprogramowanie pośrednie, ustawienia lokalizacji i wiele innych. Plik *wsgi.py* służy do zdefiniowania WSGI, interfejsu aplikacyjnego, do którego serwer przekierowuje zapytania.

Standardowa aplikacja w django składa się z następujących plików:

- *apps.py* - klasa konfiguracyjna definiująca istnienie aplikacji,
- *admin.py* - plik, w którym definiuje się jakie modele powinny być możliwe do edycji w panelu administracyjnym,
- *models.py* - plik, w którym definiuje się wszystkie modele w bazie danych używane przez aplikację; przy większych modelach, są one rozbijane na oddzielne pliki, a następnie importowane z powrotem do *models.py*,
- *tests.py* - w tym pliku zdefiniowane są testy jednostkowe,
- *views.py* - plik definiujący logikę widoków.

Dodatkowo aplikacje posiadają folder z plikami statycznymi i szablonami stron, które zostały opisane w podrozdziale 4.5.

4.4.2. Widoki

Widokiem w django nazywane są wszystkie funkcje, które obsługują przesłane za pytanie HTTP. Opisywana w tym dokumencie aplikacja posiada tylko dwa takie widoki: *reg_param* oraz *reg_param_get_async*.

Reg_param jest najprostszym widokiem, którego jedynym zadaniem jest zwrócić szablon HTML. Jest to wejściowy punkt aplikacji.

Reg_param_get_async jest widokiem, który nie zwraca widoku, tylko dane w formacie JSON. Jest on użyty w celu asynchronicznej obsługi zapytań, co pozwala zmieniać dane na stronie bez potrzeby przeładowywania jej całości. W zależności od zawartości pytania widok przekazuje informacje do poniższych funkcji:

1. *reg_parm_get_biospecimen* - zwraca listę wszystkich możliwych miejsc występowania związku,
2. *reg_parm_get_names* - zwraca nazwy metabolitów na podstawie podstawowego wyszukiwania,
3. *reg_parm_get_names_advanced* - zwraca nazwy metabolitów na podstawie zaawansowanego wyszukiwania,
4. *reg_parm_get_metabolites* - zwraca informacje o konkretnym metabolicie.

Reg_parm_get_biospecimen odwołuje się do kolekcji *hmdb_biolocation* i zwraca całą jej zawartość w postaci listy. Jest to następnie wykorzystane w warstwie reprezentacyjnej jako jeden z wyborów w wyszukiwaniu zaawansowanym.

Reg_parm_get_names przyjmuje jeden argument, na podstawie którego zwraca wszystkie wyniki z kolekcji *hmdb_met_names*, których nazwa zawiera część wpisywanej frazy.

Reg_parm_get_names_advanced działa tak samo, jak *reg_parm_get_names*, tylko dodatkowo przyjmuje jeszcze opcjonalne argumenty takie jak klasa nadziedna, klasa główna, podklasa, lokacja związku oraz zakres monoizotopowej wagi molekularnej. Wystarczy tylko jeden z argumentów, by funkcja zwróciła dopasowane nazwy związków.

Reg_parm_get_metabolites jest funkcją, która przyjmuje listę ID metabolitów i na jej podstawie zwraca informację z listą parametrów rejestracji metabolitów. Struktura każdego elementu zawarta jest w klasie *MetaboliteRegistration* i wygląda następująco:

- name - nazwa metabolitu,
- m_1 - monoizotopowa masa molekularna,
- accession - id związku w bazie danych HMDB,
- spectra_params - słownik z trzema listami z rodzajami widm, w zależności od ich trybu jonizacji: pozytywnego, negatywnego, lub nieznanego; są posortowane po ustawnionym napięciu, a każdy z nich ma formę:
 - ionization_mode - tryb jonizacji danego widma,
 - e - napięcie ustalone na spektrometrze mas podczas wykonywania pomiarów,
 - reg_param - lista parametrów rejestracji metabolitów, posortowana po intensywności, składająca się z:

- q2_3 - wartości stosunku masy do ładunku pojedynczego wierzchołka widma,
- intensity - intensywność pojedynczego wierzchołka.

4.5. Frontend

Warstwa reprezentacyjna aplikacji składa się trzech elementów: szablonów HTML, plików CSS i kodu javascript. Szablony HTML odpowiadają za rozkład elementów na obecnie odwiedzanej stronie. Pliki CSS opisują wygląd poszczególnych elementów HTML, a pliki z kodem javascript dostarczają funkcje obsługujące logikę po stronie przeglądarki.

4.5.1. Szablony HTML

Django w swoim frameworku dostarcza możliwość tworzenia szablonów HTML, które przyspieszają i ułatwiają tworzenie widoków aplikacji. Jednym z takich mechanizmów jest struktura bloków tekstu, które można nadpisywać. Pozwala to na stworzenie szablonu, który będzie rozszerzany o dodatkowe elementy. Poniżej jest zaprezentowany przykład użycia takiego mechanizmu.

```
{% block nazwa_bloku %}  
  <!-- Kod html. Nadpisuje blok o tej samej nazwie -->  
{% endblock %}
```

Taka funkcjonalność pozwala ponownie wykorzystywać napisany kod, co ułatwia jego czytelność i zmniejsza nakład pracy. Z jej wykorzystaniem strona aplikacji została podzielona na trzy części.

Base_generic.html jest podstawowym szablonem, który zawiera wyłącznie podstawowe znaczniki HTML oraz definiuje bloki dla domyślnego układu strony, gdzie panel nawigacyjny jest po lewej stronie ekranu. Dodatkowo ładuje pliki statyczne z globalnymi stylami.

Navbar.html rozszerza bazowy szablon o menu aplikacji. Zawiera on wylistowanie linków do poszczególnych widoków.

Reg_param.html jest głównym szablonem aplikacji. Rozszerza on *navbar.html* o główną zawartość strony oraz dodaje pliki ze skryptami oraz stylami dla tego widoku. Głównym jej zadaniem jest zdefiniowanie początkowej struktury strony oraz przygotowanie jej do zmian, które nastąpią w trakcie działania aplikacji. Wygląd strony został opisany i zaprezentowany w podrozdziale 4.5.3.

4.5.2. Javascript

Plik statyczny z kodem javascript zawiera funkcje, które kontrolują logikę warstwy reprezentacyjnej. Obsługują one wykryte zdarzenia takie jak załadowanie się strony, przyciśnięcie elementu lub najechanie na element kursorem. Ich głównym zadaniem jest zmiana w ułożeniu i wyglądzie strony bez potrzeby jej całkowitego przeładowania.

4. Opis tworzonej aplikacji

Dodatkowo wykonują asynchroniczne zapytania do aplikacji serwera, co pozwala na komfortowe używanie strony, bez potrzeby czekania na odpowiedź serwera.

Po załadowaniu strony dodawane są do elementów pierwsze detektory zdarzeń. Przyciski zmiany trybu wyszukiwania ukrywają obecnie wyświetlane elementy i zastępują je drugą opcją. Przycisk wyszukiwania zbiera wpisane lub wybrane przez użytkownika dane i wysyła je z wykorzystaniem biblioteki *jQuery* do widoku w backendzie. Dane są przesyłane przy pomocy zapytania GET z danymi w formacie json, które w zserializowanej postaci są zawarte w adresie url. W oczekiwaniu na odpowiedź serwera funkcje usuwają wszystkie elementy z poprzedniej tabelki wyników, jeśli już wcześniej była wypełniona oraz wyświetlają napis z informacją, że dane są ładowane. Po poprawnym załadowaniu danych przygotowana tabelka jest wypełniana nazwami metabolitów, które następnie są kolorowane dla łatwiejszego wizualnego rozróżnienia wierszy. Dodatkowo do tabelki podpięta jest obsługa zdarzeń, która podświetla wiersz pod kursorem, a po wcisnięciu wywołuje kolejne asynchroniczne zapytanie do serwera. Serwer otrzymuje informacje w postaci listy ID wybranych związków.

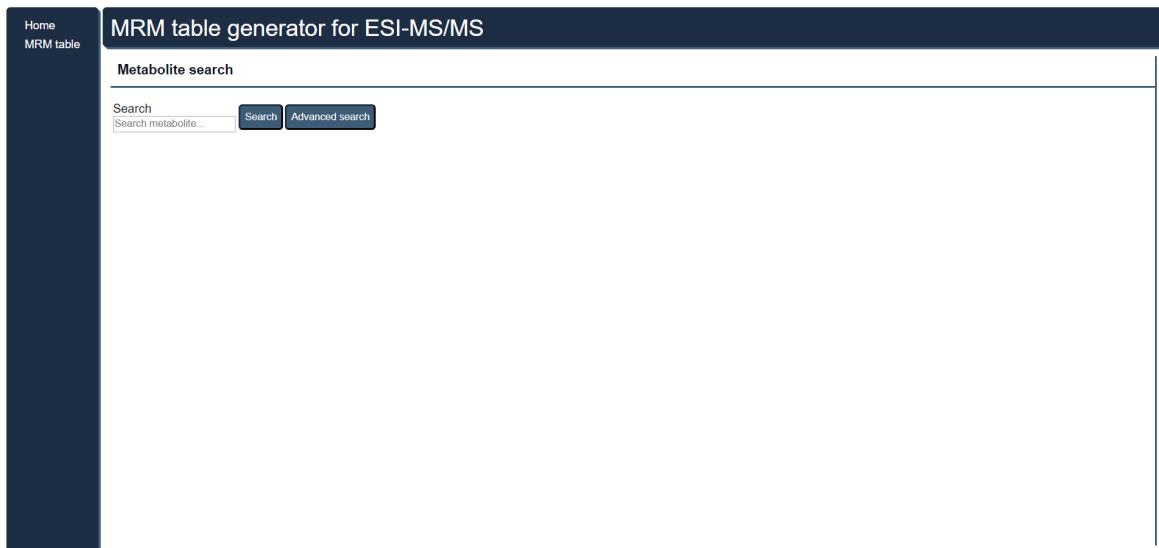
Aplikacja po otrzymaniu informacji zwrotnej z danymi na temat wybranego związku odsłania i wypełnia dodatkową sekcję strony. Iterując przez zwrócone dane, tworzy i uzupełnia tabelki, oddzielając widma pustymi wierszami. Na samej górze tworzy część z ogólnymi informacjami na temat danego związku oraz udostępnia link do związku na oficjalnej bazie danych HMDB. Wyświetlone tabelki są podzielone w zależności od trybu jonizacji, a wyświetlane w nich widma są posortowane po napięciu ustawnionym na spektrometrze podczas badania w kierunku rosnącym, a następnie po intensywności wierzchołków w kierunku malejącym. Tabelki następnie są kolorowane z pominięciem wierszy rozdzielających. Tak wypełnione tabelki mogą być filtrowane po intensywności, by nie wyświetlać informacji, które są jedynie zakłóceniami w pomiarze.

Plik z funkcjami javascript zawiera też serię pomocniczych funkcji, których zadaniem jest zwiększenie czytelności kodu i pozwala na jego ponowne wykorzystanie.

4.5.3. Wygląd i sposób działania

Interfejs jest bardzo ważnym elementem aplikacji. Jest to miejsce styku użytkownika z programem. Zostało ono zaprojektowane, aby było intuicyjne i proste w użyciu. Osoba korzystająca z aplikacji powinna bez problemów móc odszukać informacje, które potrzebuje. Na rysunku 3 jest pokazany początkowy układ strony, na który natrafia użytkownik od razu po otwarciu aplikacji w przeglądarce internetowej. Pionowy granatowy prostokąt po lewej stronie stanowi panel nawigacyjny, pozwalający na poruszanie się pomiędzy modułami programu poprzez wybór przycisku z nazwą. Obecnie istnieje tylko jeden moduł o skróconej nazwie "MRM table", który jest zaprezentowanym w tej pracy rozwiązaniem. Na górze ekranu znajduje się granatowy pasek z pełną nazwą wybranego modułu.

Większą część ekranu, na białym tle, stanowi panel do wyszukiwania metabolitów. Jest on pokazany na rysunku 4. Składa się on z pola do wpisania nazwy szukanych metabolitów



Rysunek 3. Początkowy układ strony.



Rysunek 4. Podstawowy tryb wyszukiwania.

oraz przycisku wyszukiwania (przycisk “Search”) i przycisku zmiany na zaawansowany tryb wyszukiwania (przycisk “Advanced search”). Po przejściu na zaawansowany tryb wyszukiwania, w panelu pojawiają się dodatkowe opcjonalne pola pozwalające bardziej zawięźci zakres wyszukiwania:

- Super class - klasa macierzysta związku,
- Main class - główna klasa związku,
- Sub class - podklasa związku,
- Biospecimen location - miejsce występowania związku,
- Mass (“min” “max”) - przedział monoizotopowej masy molekularnej.

Panel z zaawansowanymi polami wyszukiwania pokazany jest na rysunku 5.

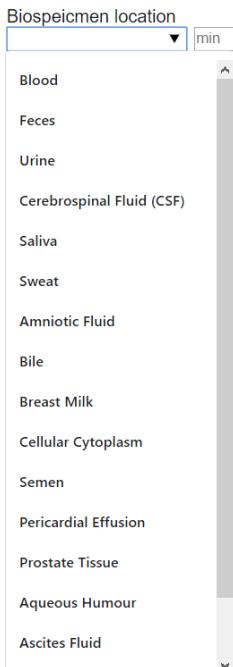
This screenshot shows the advanced search interface. It features six input fields labeled "Name", "Super class", "Main class", "Sub class", "Biospecimen location", and "Mass". Below these fields are two buttons: "Search" and "Simple search".

Rysunek 5. Zaawansowany tryb wyszukiwania.

Jak zaznaczono wcześniej, program służy do wyszukiwania, z bazy danych umieszczonej na serwerze, parametrów rejestracji metabolitów dokonanych i zarejestrowanych wcześniej przy użyciu spektrometru mas. Aby wyszukać parametry rejestracji metabolitów, należy na początku znaleźć w bazie danych dokładną nazwę związku, która może

4. Opis tworzonej aplikacji

być bardzo skomplikowana. Dany związek można znaleźć przy pomocy wyszukiwania podstawowego lub zaawansowanego.



Rysunek 6. Lista miejsc występowania związków.

W panelu wyszukiwania podstawowego należy wpisać jego nazwę lub jej część w pole tekstowe, a następnie wcisnąć klawisze enter, lub przycisk z napisem “Search”. Program łączy się z serwerem i wyświetla panel z wynikami nazw metabolitów pasujących do podanego w wyszukiwaniu wzorca.

(R)C(R)S-S-Propylcysteine sulfoxide
(S)C(S)S-S-Methylcysteine sulfoxide
(gamma-Glutamyl-gamma-glutamyl)-S-methylcysteine
3-Mercaptolactate-cysteine disulfide
4'-Phosphopantethenoylcysteine
Acetylcysteine
Ajocysteine
Alanyl-Cysteine
Arginyl-Cysteine
Asparaginyl-Cysteine
Aspartyl-Cysteine
Captopril-cysteine disulfide
Cysteine-S-sulfate
Cysteineglutathione disulfide
Cysteinyl-Cysteine
D-Cysteine
D-Pantethenoyl-L-cysteine
Farnesylcysteine
Gamma-Glutamyl-Se-methylselenocysteine
Geranylgeranylcycteine
Glutaminylcysteine
Glutamylcysteine

Rysunek 7. Przykładowa tabela ze znalezionymi metabolitami.

Wyniki wyszukiwania w trybie podstawowym mogą być bardzo liczne. Do zawężenia wyszukiwania właściwego związku służy panel wyszukiwania zaawansowanego. Wpisując w odpowiednie pola znane dane związku, zmniejszamy liczbę wyników nazw metabolitów, ułatwiając tym samym wybór odpowiedniego. Pole tekstowe “Biospecimen location” posiada rozwijaną listę wszystkich dostępnych wyborów (rysunek 6).

(gamma-Glutamyl-gamma-glutamyl)-S-methylcysteine
3-Mercaptolactate-cysteine disulfide
4'-Phosphopantethenoylcysteine
Acetylcysteine

Rysunek 8. Wygląd wiersza po najechaniu na niego.

Cysteinyl-Cysteine
D-Cysteine
D-Pantethenoyl-L-cysteine
Farnesylcysteine

Rysunek 9. Wygląd wcisniętego wiersza.

Home
MRM table

MRM table generator for ESI-MS/MS

Metabolite search	Registration parameters table																								
Search cysteine	Minimal intensity 20																								
(R)C(R)S-S-Propylcysteine sulfoxide	D-Cysteine HMDB® Monoisotopic molecular weight: 121.019749643																								
(S)C(S)S-S-Methylcysteine sulfoxide	-----																								
(gamma-Glutamyl-gamma-glutamyl)-S-methylcysteine	Ionization mode: positive																								
3-Mercaptolactate-cysteine disulfide	<table border="1"> <thead> <tr> <th>Voltage</th> <th>Q1</th> <th>Q2/3</th> <th>Intensity</th> </tr> </thead> <tbody> <tr><td>10</td><td>122.019749643</td><td>76.02209519</td><td>32.5984967</td></tr> <tr><td>10</td><td>122.019749643</td><td>122.0275745</td><td>25.32321241</td></tr> <tr><td>20</td><td>122.019749643</td><td>76.02209519</td><td>41.24619422</td></tr> <tr><td>40</td><td>122.019749643</td><td>42.03437413</td><td>27.11474875</td></tr> <tr><td>40</td><td>122.019749643</td><td>58.9955461</td><td>21.91043386</td></tr> </tbody> </table>	Voltage	Q1	Q2/3	Intensity	10	122.019749643	76.02209519	32.5984967	10	122.019749643	122.0275745	25.32321241	20	122.019749643	76.02209519	41.24619422	40	122.019749643	42.03437413	27.11474875	40	122.019749643	58.9955461	21.91043386
Voltage	Q1	Q2/3	Intensity																						
10	122.019749643	76.02209519	32.5984967																						
10	122.019749643	122.0275745	25.32321241																						
20	122.019749643	76.02209519	41.24619422																						
40	122.019749643	42.03437413	27.11474875																						
40	122.019749643	58.9955461	21.91043386																						
4'-Phosphopantethenoylcysteine	-----																								
Acetylcysteine	Ionization mode: negative																								
Ajocysteine	<table border="1"> <thead> <tr> <th>Voltage</th> <th>Q1</th> <th>Q2/3</th> <th>Intensity</th> </tr> </thead> <tbody> <tr><td>10</td><td>120.019749643</td><td>120.0119244</td><td>53.94033807</td></tr> <tr><td>10</td><td>120.019749643</td><td>86.02420337</td><td>23.14764475</td></tr> <tr><td>20</td><td>120.019749643</td><td>86.02420337</td><td>29.23136202</td></tr> <tr><td>20</td><td>120.019749643</td><td>120.0119244</td><td>28.98198599</td></tr> <tr><td>40</td><td>120.019749643</td><td>32.97989603</td><td>44.40073374</td></tr> </tbody> </table>	Voltage	Q1	Q2/3	Intensity	10	120.019749643	120.0119244	53.94033807	10	120.019749643	86.02420337	23.14764475	20	120.019749643	86.02420337	29.23136202	20	120.019749643	120.0119244	28.98198599	40	120.019749643	32.97989603	44.40073374
Voltage	Q1	Q2/3	Intensity																						
10	120.019749643	120.0119244	53.94033807																						
10	120.019749643	86.02420337	23.14764475																						
20	120.019749643	86.02420337	29.23136202																						
20	120.019749643	120.0119244	28.98198599																						
40	120.019749643	32.97989603	44.40073374																						
Alanyl-Cysteine	-----																								
Arginyl-Cysteine	Ionization mode: na																								
Asparaginyl-Cysteine	<table border="1"> <thead> <tr> <th>Voltage</th> <th>Q1</th> <th>Q2/3</th> <th>Intensity</th> </tr> </thead> <tbody> <tr><td>10</td><td>120.019749643</td><td>76</td><td>100</td></tr> <tr><td>10</td><td>120.019749643</td><td>87</td><td>36.447</td></tr> <tr><td>10</td><td>120.019749643</td><td>105</td><td>35.659</td></tr> <tr><td>10</td><td>120.019749643</td><td>122</td><td>28.86</td></tr> </tbody> </table>	Voltage	Q1	Q2/3	Intensity	10	120.019749643	76	100	10	120.019749643	87	36.447	10	120.019749643	105	35.659	10	120.019749643	122	28.86				
Voltage	Q1	Q2/3	Intensity																						
10	120.019749643	76	100																						
10	120.019749643	87	36.447																						
10	120.019749643	105	35.659																						
10	120.019749643	122	28.86																						
Aspartyl-Cysteine	-----																								
Captopril-cysteine disulfide	-----																								
Cysteine-S-sulfate	-----																								
Cysteineglutathione disulfide	-----																								
Cysteinyl-Cysteine	-----																								
D-Cysteine	-----																								
D-Pantethenoyl-L-cysteine	-----																								
Farnesylcysteine	-----																								
Gamma-Glutamyl-Se-methylselenocysteine	-----																								
Geranylgeranylcysteine	-----																								
Glutaminylcysteine	-----																								
Glutamylcysteine	-----																								
Glycy-Cysteine	-----																								
Histidinyl-Cysteine	-----																								
Homocysteine	-----																								

Rysunek 10. Końcowy układ strony.

Po znalezieniu na liście wyników (rysunek 7) wyszukiwania odpowiedniego metabolitu należy na niego najechać i kliknąć (rysunek 8). Zostanie on wyróżniony na liście przy pomocy ciemnego tła (rysunek 9), a po prawej stronie wyświetli się okno z podstawowymi informacjami o metabolicie (jego nazwa i monoizotopowa masa molekularna) oraz listami jego wszystkich parametrów rejestracji (rysunki 10 i 11). Nazwa metabolitu dodatkowo zawiera link do strony związku na oficjalnej stronie bazdy HMDB (rysunek 12). Listy zawierają następujące parametry rejestracji:

- Voltage - napięcie ustawione na spektrometrze mas,
- Q1 - główny stosunek masy do ładunku związku; jest on zazwyczaj równy monoizotopowej masie molekularnej; w przypadku pozytywnego trybu jonizacji jest ona zwiększena o 1, a w przypadku jonizacji negatywnej jest zmniejszona o 1; jeśli tryb jonizacji nie jest dostępny pozostaje ona równa monoizotopowej masie molekularnej związku; w przypadkach szczególnym ta reguła może się różnić, lecz zostało to pozostawione ocenie według wiedzy użytkownika,
- Q2/3 - stosunek masy do ładunku sygnału,

4. Opis tworzonej aplikacji

- Intensity - intensywność sygnału.

Registration parameters table			
Minimal intensity			
D-Cysteine <small>HMDB</small>			
Monoisotopic molecular weight: 121.019749643			

Ionization mode: positive			
Voltage	Q1	Q2/3	Intensity
10	122.019749643	76.02209519	32.5984967
10	122.019749643	122.0275745	25.32321241
10	122.019749643	105.0010254	10.75613739
10	122.019749643	104.0170098	5.677219893
10	122.019749643	88.03985343	5.538970855
10	122.019749643	86.99046072	4.915702254
10	122.019749643	42.03437413	3.319173898
10	122.019749643	78.03774526	2.408491952
10	122.019749643	58.9955461	2.163853424
20	122.019749643	76.02209519	41.24619422
20	122.019749643	122.0275745	9.815673137
20	122.019749643	42.03437413	9.098381427
20	122.019749643	58.9955461	7.402378074
20	122.019749643	104.0170098	5.503557964
20	122.019749643	105.0010254	4.511600358
20	122.019749643	88.03985343	4.435873859
20	122.019749643	46.9955461	4.409936326
20	122.019749643	71.01330434	3.184085985
20	122.019749643	86.99046072	2.171229866
20	122.019749643	49.01119616	2.010204617
40	122.019749643	42.03437413	27.11474875
40	122.019749643	58.9955461	21.91043386
10	122.019749643	46.9955461	44.50002407

Rysunek 11. Sekcja dotycząca parametrów rejestracji metabolitu.

D-Cysteine HMDB
Monoisotopic molecular weight: 121.019749643

Rysunek 12. Link do zewnętrznej bazy danych HMDB.

Omówione tu listy są podzielone w zależności od trybu jonizacji, a następnie posortowane po napięciu ustawionym na spektrometrze według wartości rosnącej. Na koniec wiersze tabel są posortowane po intensywności wyników w kolejności malejącej. Na górze sekcji (rysunek 11) znajduje się pole do filtrowania wyników po ich intensywności. Domyślnie jest on ustawiony na wartość 20, co oznacza, że pokazują się tylko wiersze, gdzie jest ona równa lub większa niż 20. Można tę wartość zmieniać w zakresie od 0 do 100.

Ionization mode: positive

Voltage	Q1	Q2/3	Intensity
10	122.019749643	76.02209519	32.5984967
10	122.019749643	122.0275745	25.32321241
20	122.019749643	76.02209519	41.24619422
40	122.019749643	42.03437413	27.11474875
40	122.019749643	58.9955461	21.91043386

Rysunek 13. Tabela parametrów rejestracji metabolitów dla jednego trybu jonizacji.

Pojedyncza lista (przykład na rysunku 13) umożliwia proste zaznaczenie i skopiowanie wartości do zewnętrznych programów w celu dalszej analizy.

5. Przeprowadzone testy

Po zaimplementowaniu rozważanej aplikacji, poddana ona zostanie różnym weryfikacjom eksperymentalnym. Aby program został odebrany i zaakceptowany jako poprawne rozwiązanie, muszą zostać na nim przeprowadzone testy, które zweryfikują jego działanie. W trakcie prac nad programem wykonywane były automatyczne testy jednostkowe, a po każdym przeładowaniu danych wykonywane były testy ilościowe i jakościowe rekordów w bazie danych. Dodatkową zaletą testów jednostkowych wykonywanych w trakcie rozwijania aplikacji jest wcześnie wykrycie błędów, co pozwala na natychmiastową poprawę. W tym rozdziale opisana zostanie tylko struktura i forma testów. Wyniki i wnioski zostały zamieszczone w podrozdziale 5.3.

5.1. Testy automatyczne

Framework django udostępnia mechanizm tworzenia i obsługi testów automatycznych. Dostarcza narzędzi, które pozwalają na stworzenie testowego środowiska oraz testowej bazy danych, w którym uruchamiane są testowane elementy.

Przypadki testowe są zawarte w pliku *tests.py*, lecz w obecnej konfiguracji są one tam tylko importowane, a same testy są rozproszone w plikach z nazwami, które odpowiadają testowanemu elementowi. Taki sposób utrzymania pozwala na rozdzielenie i szybką nawigację pomiędzy testami. Aby uruchomić testy, wystarczy użyć administracyjnego pliku django, podając mu jako parametr opcję *test* i nazwę modułu, który chcemy przetestować. W przypadku obecnej konfiguracji jest to *python manage.py test hmdb*. Podczas wykonywania testów, framework django tworzy nową bazę danych, w której umieszcza imitację danych stworzoną na potrzeby poszczególnych testów. Po wykonaniu wszystkich testów baza jest usuwana. Takie działanie pozwala na bezpieczne testowanie danych i komunikację z bazą danych bez obawy zanieczyszczenia bazy produkcyjnej.

5.1.1. Struktura testów automatycznych

Plik o nazwie *test_reg_param.py* zawiera wszystkie przypadki testowe testujące logikę widoków po stronie serwera. Każda klasa testowa musi dziedziczyć po klasie *TestCase* z biblioteki *django.test*. Poniżej pokazany jest kod realizujący podstawowe ustawienie klasy testowej.

```
class RegParamTest(TestCase):
    def setUp(self):
        create_test_data()
    def tearDown(self):
        Metabolite.objects.all().delete()
        Spectra.objects.all().delete()
```

```
MetaboliteNames.objects.all().delete()
```

Klasa testowa poza przypadkami testowymi powinna też posiadać funkcje ustawiającą środowisko przed testami (*def setUp(self)*) oraz funkcję, która wyczyści środowisko testowe z wykorzystanych danych testowych (*def tearDown(self)*). Dla tego widoku dane testowe składają się z 3 rekordów z tabelki odpowiedzialnej za wyszukiwanie, jednego metabolitu i dwóch powiązanych z nim widm: jednego pozytywnego dla napięcia 10 woltów i jednego negatywnego dla napięcia 25 woltów.

5.1.2. Opis testów jednostkowych

Poniżej zostały opisane wszystkie zaimplementowane testy jednostkowe.

Test_mock_data_creation jest pierwszym wykonywanym testem. Jego zadaniem nie jest przetestowanie widoku, lecz sprawdzenie, czy dane testowe się zgadzają. Jest to potencjalne zabezpieczenie przed sytuacją, gdzie konto aplikacyjne nie będzie posiadało wystarczających uprawnień na bazie danych, aby stworzyć dane testowe.

Test_reg_param_view_get_names_async sprawdza, czy asynchroniczne zapytanie do widoku mające zwrócić znalezioną listę nazw wszystkich metabolitów, których nazwa zawiera podaną frazę. Test ten odpowiada wyszukiwaniu metabolitów przy pomocy podstawowego panelu wyszukiwania. Po odebraniu odpowiedzi jest sprawdzany kod statusu, który powinien się równać 200, co oznacza poprawne wykonanie. Następnie sprawdzana jest liczba zwróconych wierszy. Sprawdzane jest, czy zgadza się ona z oczekiwana liczbą rekordów z danych testowych.

Test *test_reg_param_view_get_metabolite_async* ma za zadanie sprawdzenie, czy zapytanie do serwera z prośbą o zwrócenie danych ze szczegółami o parametrach rejestracji metabolitów zwraca poprawne dane. W aplikacji ten test odpowiada wybraniu konkretnego wiersza z listy z pełnymi nazwami metabolitów. Po odebraniu odpowiedzi jest sprawdzany kod statusu, który powinien się równać 200, co oznacza poprawne wykonanie. Kolejnym krokiem jest sprawdzenie, czy został zwrócony pojedynczy metabolit oraz, czy liczba parametrów rejestracji zgadza się dla poszczególnych trybów jonizacji.

Test *test_reg_param_view_get_metabolite_async_404* wykonuje dokładnie tę samą czynność co poprzedni test, lecz próbuje uzyskać informacje na temat metabolitu, który nie istnieje. W takiej sytuacji widok powinien zwrócić odpowiedź z kodem statusu równym 404, co oznacza, że taki metabolit nie został odnaleziony.

Test_reg_param_view_get jest z kolei prostym testem sprawdzającym, czy podstawowa strona aplikacji zwraca odpowiedź z poprawnym kodem statusu.

5.2. Testy manualne

Testy wykonane w sposób manualny polegają natomiast na ręcznym sprawdzeniu działania aplikacji. Są one potrzebne, aby sprawdzić elementy programu, które ciężko zautomatyzować, lub nie potrzebują aż tak częstego sprawdzania.

5.2.1. Testy bazy danych

Głównym elementem testowania bazy danych było sprawdzenie jakości i ilości dokumentów w niej umieszczonych. Proces sprawdzenia polegał na zebraniu statystyk z liczby metabolitów i widm przed i po ich przeprocesowaniu. Liczba dla metabolitów powinna być identyczna, a liczba widm powinna się zgadzać z liczbą widm typu MS-MS.

Testy jakościowe bazy danych zostały przeprowadzone na narzędziu MongoDB Compass [13]. Jest to narzędzie wspomagające przeglądanie, zarządzanie i zbieranie statystyk z bazy danych MongoDB. Dane widm zostały sprawdzone pod względem ich typu oraz ilości wierzchołków. Zarówno w danych metabolitów, jak i w danych widm zostały sprawdzone pola numeryczne i czy ich typ się zawsze zgadzał.

5.2.2. Testy frontendu

Manualne testy frontendu polegają na ręcznym przejściu przez aplikację po stronie warstwy reprezentacyjnej. Proces testowania układu strony polegał na sprawdzeniu wyglądu aplikacji przy różnych rozmiarach okna przeglądarki. Układ strony powinien pozostać taki sam, a wszystkie informacje powinny być czytelne. W przypadku małego ekranu lub zmniejszonego okna przeglądarki strona nie powinna się skalować, tylko powinna mieć możliwość jej przesuwania, żeby zachować czytelność.

Kolejnym etapem testowania manualnego były testy pól tekstowych, które może wypełnić użytkownik. Były one sprawdzane pod względem rodzaju wpisywanych danych. Pola liczbowe nie powinny pozwalać na wprowadzenie innych znaków niż cyfry, znak minusa oraz kropki dziesiętnej. Pole wyboru limitu intensywności po wpisaniu liczby powinno ją ograniczać do zakresu od 0 do 100. Pola tekstowe nie posiadają odgórnich ograniczeń i mogą zawierać każdy rodzaj znaku.

Aplikacja po wysłaniu zapytania, które trwa zbyt długo lub po zerwaniu połączenia internetowego podczas jego wykonywania, powinna zwracać błąd informujący o zdarzeniu. Taki błąd jest zasygnalizowany poprzez pojawiające się okno z informacją.

Ostatnim testem było sprawdzenie płynności działania aplikacji. Aplikacja powinna być zawsze responsywna, aby użytkownik nie tracił nad nią kontroli.

5.2.3. Testy użytkownika

Po skończeniu pierwszej wersji programu został on oddany głównemu użytkownikowi do przetestowania. Celem tego testu było sprawdzenie intuicyjności oraz łatwości użycia interfejsu, jak również zebranie odpowiednich wskazówek co można polepszyć. Użytkownik testujący, jako osoba często spotykająca się z problemem przedstawionym w tej pracy, miał za zadanie przejść przez program, odnajdując interesujące go informacje, bez pomocy dewelopera. W trakcie testu wykonywane były obserwacje zachowania osoby testującej oraz zanotowane zostały miejsca, które sprawiały trudność. Dodatkowo tester na koniec dostarczył swoją opinię i sugestie.

W drugiej iteracji testów użytkownika program został już umieszczony na serwerze, do którego był publiczny dostęp. Pozwoliło to na przetestowanie rozwiązania na większą skalę, a wszystkie ich uwagi i sugestie były zebrane i przeanalizowane.

5.3. Wyniki testów

5.3.1. Wyniki testów manualnych

Podczas manualnych testów bazy danych zostały wykryte nieścisłości w zapisie numerycznym, takie jak zapis liczby w postaci ciągu znaków. Kolejnym sprawdzonym elementem bazy danych były pola z listami. Poprzez niejednoznaczny zapis XML źródłowych danych, istniały metabolity, które w polu zawierały jeden element, który potem był przekształcany na obiekt, podczas gdy inne metabolity zawierały w tym samym polu kilka elementów i były zapisywane jako lista. Ten test pozwolił prześledzić każde pole i wytypować takie, które wymagają dodatkowych przekształceń.

Ręczne testy frontendu wykazały, że przy mniejszych ekranach panel wyszukiwania zaawansowanego jest zupełnie niedostępny i nieczytelny. Problem ten został rozwiązany poprzez dodanie przewijanego panelu wyszukiwania zaawansowanego w przypadku gdy obecny nie mieści się w całości. Kolejnym wykrytym błędem był brak informacji o fakcie wysłania zapytania do serwera. Po kliknięciu przycisku "Search" aplikacja działała poprawnie, ale użytkownik nie miał żadnej informacji zwrotnej o zaistniałej nieprawidłowości.

5.3.2. Wyniki pierwszych testów użytkownika

Pierwszy test użytkownika przebiegł poprawnie. Osoba testująca bez problemów uzyskała interesujące ją informację. Mimo to głównymi uwagami były:

- brak informacji o intensywności,
- brak podziału na tryb jonizacji,
- brak filtrowania po intensywności,
- grupowanie po napięciu powinno być zrobione w jednej tabelce a nie oddzielnych,
- nie ma zewnętrznego łącza do oficjalnej strony HMDB z danym związkiem.

Dzięki tym sugestiom i uwagom powstała druga wersja programu. Do tabelek z informacjami o parametrach rejestracji zostały dołączone dwie kolumny: z napięciem oraz z intensywnością. Tabelki zostały podzielone po trybie jonizacji, a następnie każda z nich posortowana rosnąco według napięcia, po czym malejąco według intensywności. Opcja filtrowania po intensywności była już zaimplementowana, ale użytkownik jej nie zauważył. Została ona przez to przesunięta w bardziej widoczne miejsce. Link do zewnętrznej strony HMDB odpowiadającej wyszukanemu związkowi został ukryty pod wyświetlana nazwą związku i został oznaczony małym napisem oraz symbolem łącza zaprezentowanym na obrazku 12.

5. Przeprowadzone testy

5.3.3. Wyniki drugich testów użytkownika

Drugie testy użytkownika, mimo że przeprowadzone na większą skalę, przyniosły o wiele mniej uwag. Sugestie w większości dotyczyły kolorów i kontrastów dla łatwiejszej nawigacji i orientacji w tabelkach. Propozycje te zostały skorygowane, co poprawiło jakość tworzonej aplikacji.

6. Badania

Widma spektrometryczne metabolitów zawierają informacje dotyczące sposobu rozpadu cząsteczek na mniejsze łańcuchy. Są one zaprezentowane w postaci wykresu intensywności do stosunku masy ładunku. Te wykresy są prawie unikatowe dla związków i pokazują potencjalne ścieżki rozpadu. Jednak widma otrzymywane przy użyciu spektrometru mas nie zawierają jednoznacznych informacji na temat utraconych cząsteczek obojętnych. Obecność różnych cząsteczek obojętnych podczas rozpadu niesie ze sobą informację na temat charakterystyki związku.

6.1. Badane dane

Niniejsza praca korzysta z danych z bazy danych ludzkich metabolitów HMDB. Do przeprowadzenia badań należało wybrać konkretną grupę związków chemicznych. Taka grupa musi zawierać wystarczająco dużą liczbę związków, aby metody eksploracji danych dały prawidłowe wyniki. Dodatkowo informacje o spektrach związków muszą być jak najlepszej jakości. Baza danych HMDB jest otwartą bazą danych, co niestety powoduje, że zapisywane tam widma często są niekompletne lub nieznormalizowane. Analiza bazy danych wykazała, że grupa lipidów i ich pochodnych jest bardzo dobrym reprezentantem. Zawiera ona ponad 90 tysięcy związków, a skumulowana liczba wszystkich ich widm przekroczyła nawet 500 tysięcy. Jednak nadal widma te nie były do końca spójne, więc zostały zawężone tylko do widm, które występowały w każdym związku. Są to widma typu Ms-Ms, wykonane w trybie jonizacji pozytywnej lub negatywnej o energii zderzenia 10, 20 lub 40 eV.

6.1.1. Lipidy i ich pochodne

Lipidy są to często występujące w naturze związki chemiczne. Ich głównym zadaniem w przyrodzie jest magazynowanie energii, udział w przesyłaniu sygnałów oraz tworzenie błon biologicznych. Do tej grupy zaliczane są tłuszcze, woski, sterole, witaminy rozpuszczalne w tłuszczach i wiele innych substancji. Związki te można zdefiniować jako hydrofobowe lub amfifilowe cząsteczki. Pozwala to im na tworzenie pęcherzyków i błon w środowisku wodnym. Lipidy są cząsteczkami niezbędnymi do prawidłowego funkcjonowania organizmu. Baza HMDB zawiera informacje wyłącznie o ludzkich metabolitach, co dodatkowo zawęża badaną grupę do lipidów znajdujących się w produktach ludzkiego metabolizmu.

6.1.2. Transformacja i procesowanie danych

Pierwszym krokiem potrzebnym do przeanalizowania danych jest ich transformacja i procesowanie do postaci wygodnej dla algorytmu. Najpierw zostało wykonane wyliczenie prawdopodobnych utraconych cząsteczek obojętnych. Ta operacja składała się z obliczenia odległości pomiędzy kombinacjami wszystkich wierzchołków w widmie dla

6. Badania

każdego widma rodzaju Ms-Ms, wykonanego w trybie jonizacji pozytywnej lub negatywnej o energii zderzenia 10, 20 lub 40 eV. W taki sposób zostaną obliczone wszystkie potencjalne zmiany w masie pomiędzy zaobserwowanymi naładowanymi cząsteczkami. Masa cząsteczkowa nigdy nie będzie liczbą całkowitą, jednak jest to spowodowane istnieniem różnych izotopów poszczególnych atomów. Mimo to odchylenie od wartości całkowej jest na tyle małe, że często dla uproszczenia są one zaokrąglane. Uzyskane wartości zostały zatem zaokrąglone do wartości całkowitych, co pozwoli na lepsze ich zliczenie oraz znalezienie zależności w ich występowaniu. Niestety podczas tej metody produkowane są też błędne artefakty. Najmniejszą cząsteczką, która może wystąpić podczas takiego rozkładu, jest amoniak (wzór chemiczny NH_3) o wadze cząsteczkowej bliskiej $17u$. Dzięki tej informacji możemy odrzucić wszystkie wyniki, których masa jest poniżej 17. Po tym kroku każde widmo jest w postaci przedstawionej w tabeli 5.

Tabela 5. Struktura dokumentów zawierających dane o utraconych cząsteczkach obojętnych w widmie.

Nazwa	Opis
id	unikalny identyfikator widma
met_id	unikalny identyfikator badanej cząsteczki
collision_energy_voltage	energia zderzenia w eV
ionization_mode	tryb jonizacji
distances	lista wszystkich wyliczonych odległości

Kolejnym krokiem do zawężenia danych było zrobienie podliczenia liczby wystąpień każdej wyliczonej odległości. Były one podzielone w ten sam sposób jak widma (na tryb jonizacji i dla każdej energii kolizji) i zrobione dla wszystkich wyliczonych widm, jak i dla widm z grupy lipidów i ich pochodnych. Wyniki zostały zaprezentowane w rozdziale 7. Najczęściej występujące odległości zostały wykorzystane w dalszych krokach analizy.

6.2. Metoda asocjacyjna

Metoda asocjacyjna jest jedną z wielu metod eksploracji danych. Znajduje ona zastosowanie w dziedzinach takich jak bioinformatyka, uczenie maszynowe oraz produkcja ciągła. Jej głównym zadaniem jest znalezienie zależności pomiędzy powtarzającymi się elementami. Jej wynikiem są reguły asocjacyjne, które zawierają parametry szczegółowo je opisujące. Reguły asocjacyjne nie uwzględniają sekwencji, tylko same wystąpienia elementów. Niech $I = \{i_1, i_2, i_3, \dots, i_n\}$ będzie zbiorem wszystkich obliczonych mas utraconych cząsteczek obojętnych. Bazą transakcyjną nad zbiorem I jest nazywany zbiór transakcji $D = \{T_1, T_2, \dots, T_k\}$ (będący w tym przypadku zbiorem wszystkich widm), w którym każda transakcja ma postać $T = \{id, X\}$, gdzie id w naszym przypadku składa się z id widma, id badanej cząsteczki, energii kolizji i trybu jonizacji, a $X \subseteq I$. Regułą asocjacyjną jest dowolna implikacja $A \Rightarrow B$, gdzie $A \subseteq I$, $B \subseteq I$ i $A \cap B = \emptyset$. Reguły opisane są trzema

wskaźnikami: wsparciem (ang. support), zaufaniem (ang. confidence) i przyrostem (ang. lift). Reguła $A \Rightarrow B$ na zbiorze D ma wsparcie s , jeśli s procent transakcji pokrywa zbiór $A \cup B$. Reguła $A \Rightarrow B$ ma zaufanie c , jeśli wśród widm zawierających A jest c procent zawierających B . Wymienione wskaźniki obliczone są wzorami:

$$\begin{aligned} supp(A) &= \frac{|t \in D; A \subseteq t|}{|D|} \\ supp(A \Rightarrow B) &= \frac{supp(A \cup B)}{|D|} \\ conf(A \Rightarrow B) &= \frac{supp(A \cup B)}{supp(A)} \end{aligned}$$

Trzecim rodzajem wskaźnika jest przyrost. Określa on znaczenie reguły, czyli wpływ poprzednika A na sekwencję B . Wartość przyrostu większa od 1 oznacza, że wystąpienia poprzednika i sekwencji są od siebie mocno zależne, jeśli wartość wynosi 1, są niepowiązane, a jak wartość jest mniejsza niż 1, to są przeciwnie związane. Przyrost jest obliczany wzorem:

$$lift(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A) * supp(B)}$$

6.2.1. Algorytm apriori

Algorytm apriori [14] jest jednym z rozwiązań metody asocjacyjnej. W niniejszej pracy została wykorzystana biblioteka pythonowa Efficient-Apriori [15], która implementuje ten algorytm. Jest on podzielony na dwie części: generację częstych zbiorów oraz wyznaczanie z nich reguł asocjacyjnych.

Generacja częstych zbiorów polega na przeszukaniu całej bazy wszerz i wyznaczeniu wszystkich zbiorów jednoelementowych, których częstotliwość występowania jest większa niż zadany na wejściu algorytmu parametr. Następnie wykorzystuje wszystkie znalezione zbiory jednoelementowe, by z nich utworzyć zbiory dwuelementowe, które następnie są wykorzystywane do stworzenia zbiorów trójelementowych itd. Podczas tworzenia zbiorów dla każdego z nich sprawdzana jest częstotliwość jego wystąpienia i jeśli nie spełnia on zadanych warunków, zostaje on odrzucony i nie jest wykorzystywany do tworzenia kolejnych nadzbiorów. Jest to możliwe, ponieważ jeśli zbiór nie jest zbiorem częstym, to jego nadzbiór też nie będzie zbiorem częstym. Taka czynność przyspiesza działanie algorytmu, ale mimo to na dużych zbiorach danych utworzenie tylu kombinacji jest czasochłonne.

Drugim krokiem algorytmu jest generowanie reguł asocjacyjnych. W tym celu dla każdego utworzonego zbioru częstego tworzone są wszystkie kombinacje jego elementów, dzieląc je na grupę poprzedników i grupę sekwencji. Następnie dla każdej reguły obliczana jest wartość zaufania i jeśli nie spełnia ona minimalnej zadanej wartości, to jest odrzucana. Jeśli reguła spełnia wszystkie zadane wymagania, to jest dla niej obliczana jeszcze wartość

6. Badania

przyrostu. W rezultacie powstaną też reguły przeciwnostawne, czyli $A \Rightarrow B$, jak i $B \Rightarrow A$. W takim przypadku wybierana jest reguła z większym wsparciem.

7. Wyniki badań

W tym rozdziale zostały zaprezentowane wyniki poszczególnych kroków wykonanych badań przeprowadzonych z wykorzystaniem metody opisanej w rozdziale 6. W ich trakcie powstały wykresy liczebności wyliczonych utraconych cząsteczek obojętnych dla całego zbioru dostępnych widm ludzkich metabolitów, jak i dla jego podzbioru lipidów i ich pochodnych. Do interpretacji otrzymanych wyników zostały wykorzystane dane opublikowane w artykule Agrawala [14]. Znajduje się tam tabela z najczęściej otrzymywanymi cząsteczkami obojętnymi, która pomoże zidentyfikować większość wyników.

7.1. Liczebność utraconych cząsteczek obojętnych na pełnym zbiorze związków

Pierwszym krokiem po obliczeniu występujących odległości w widmach było ich pogrupowanie i podliczenie. Każda kombinacja trybu jonizacji i energii kolizji została podliczona oddzielnie. Tabela 6 zawiera listę 20 najczęściej występujących utraconych cząsteczek obojętnych posortowanych w kolejności malejącej dla wszystkich widm metabolitów. Te informacje w postaci wykresów słupkowych zostały też przedstawione na rysunkach 14 - 19.

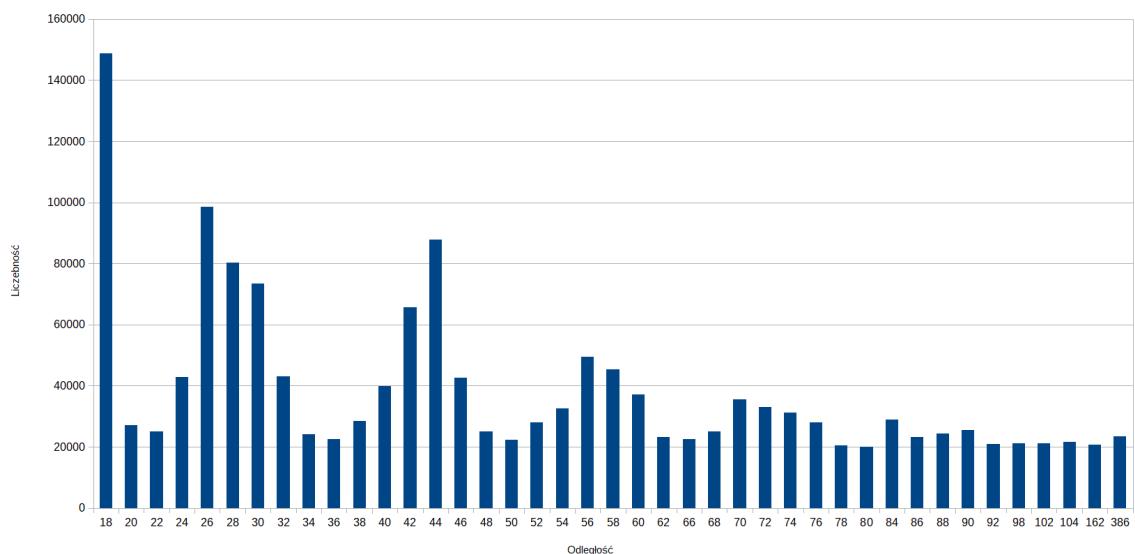
Tabela 6. 20 najczęściej występujących odległości dla każdej kategorii.

Jonizacja negatywna			Jonizacja pozytywna		
10 eV	20 eV	40 eV	10 eV	20 eV	40 eV
18	18	28	18	18	28
26	42	42	28	28	42
44	28	26	40	42	26
28	26	44	26	26	30
30	44	30	42	56	56
42	30	18	56	30	18
56	56	56	98	40	40
58	58	40	30	58	44
32	40	58	154	24	24
24	136	46	58	44	70
46	74	24	24	98	58
40	84	54	44	70	54
60	46	70	252	54	46
70	24	32	20	46	38
72	70	136	54	154	32
54	32	60	46	84	84
74	60	38	70	96	154
84	100	52	100	80	68
38	54	68	60	32	52
52	72	84	138	38	98

7. Wyniki badań

Od razu widać, że cząsteczka 18 pojawia się najczęściej. Odpowiada ona utracie cząsteczek wody w przypadku obecności w związkach grupy hydroksylowej ($-OH$) i karboksylowej ($-COOH$). Zaraz za nią są cząsteczki 28 (CO), 44 (CO_2) i 30 (NO lub CH_2O). Są to już mniej oczywiste związki, ale nadal często występujące w związkach organicznych. Zaskakujący jest jednak brak cząsteczek o nieparzystej masie, szczególnie 17 (NH_3) i 27 (HCN), które zawierają grupę aminową. Wartości te pojawiają się średnio 10 razy rzadziej niż wartości 18 i 28. Świadczy to o tym, że znaczna większość metabolitów ludzkich ma charakter kwaśny.

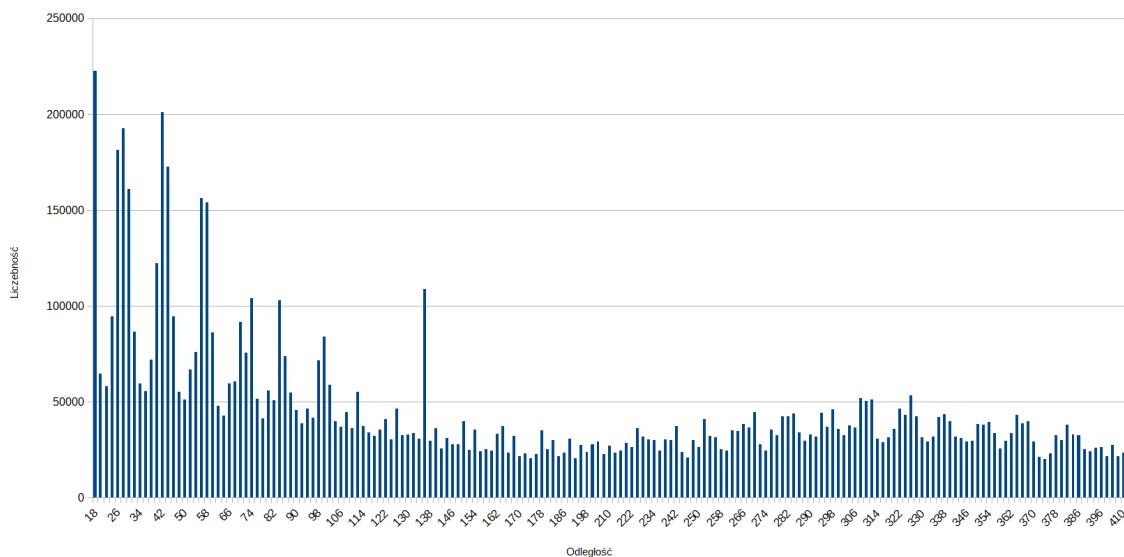
Na rysunku 14 można zaobserwować rozkład liczebności utraconych cząsteczek obojętnych dla widm o jonizacji negatywnej dla energii kolizji 10 eV. Dla czytelności zostały wycięte wszystkie wyliczone związki, które występowały rzadko. Jak widać, wartość 18 występuje znacznie częściej niż inne wartości. Dalej widać dwa wyraźne wzniesienia w przypadku utraty cząsteczek o masie 26 i masie 44.



Rysunek 14. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 10 eV.

Na rysunku 15 jest zaprezentowana liczebność poszczególnych utraconych cząsteczek obojętnych dla widm o jonizacji ujemnej i energii kolizji równej 20 eV. Tak jak wcześniej cząsteczka o masie 18 występuje najczęściej. Dodatkowo wzniesienia otaczające kolejne najwyższe punkty są o bardziej strome. Wynika to z faktu, że zwiększąc energię kolizji, zwiększa się szumy, a także jest większa szansa na wytrącenie się nietypowej cząsteczki obojętnej. Dodatkowo znów widać bardzo częste występowanie par cząsteczek 26 i 28, 42 i 44 oraz 56 i 58, które różnią się tylko dwoma jednostkami masy (zatem 2 atomami wodoru). Interesującym zjawiskiem jest za to bardzo duża liczba cząsteczek o masie 136, której otoczenie jest znacznie rzadsze. Wartość ta odpowiada cząsteczce o wzorze sumarycznym $C_3H_5O_4P$, w której znajduje się grupa fosforylowa. Cząsteczki tej należy

się spodziewać w przypadku rozpadu fosfolipidów, podstawowego budulca membrany komórkowej. Liczebność jej jest zatem uzasadniona. Dodatkowo można zauważyc trzy dodatkowe wznieśienia dookoła punktów 74, 84 i 100.



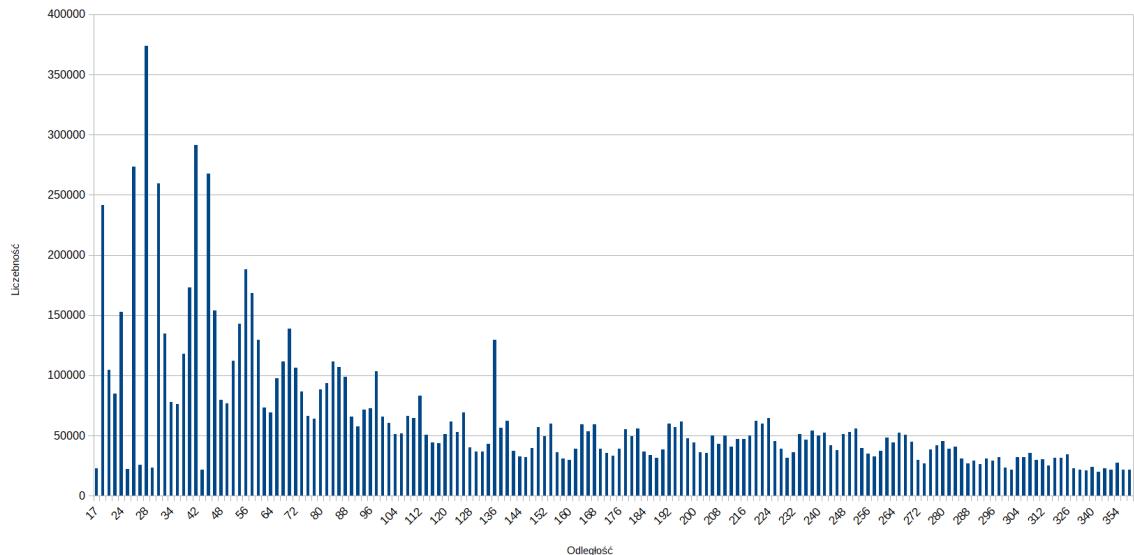
Rysunek 15. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 20 eV.

Rysunek 16 przedstawia podliczenie utraconych cząsteczek obojętnych dla widm o jonizacji negatywnej i energii kolizji 40 eV. Cząsteczka o masie 18 nadal pojawia się często, lecz już znacznie mniej niż w poprzednich grupach. Najczęściej pojawiającą się cząsteczką jest teraz cząsteczka o masie 28. Cząsteczki o masach 26, 42, 44 i 56 też są nadal mocno widoczne. Odległość 136 jest nadal bardzo widoczna na tle swojego otoczenia. Dodatkowo można zaobserwować o wiele więcej małych wznieśień. Większość z nich jest prawdopodobnie spowodowana szumem na wyjściu spektrometru mas, który nie zawsze da się łatwo wyczyścić.

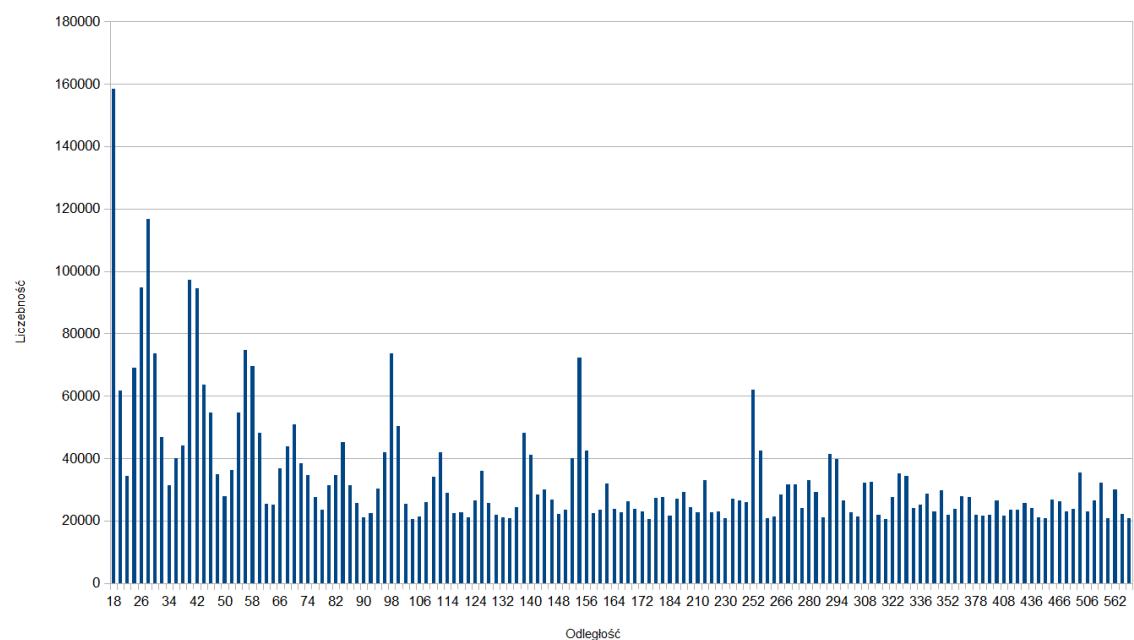
Na rysunku 17 znajduje się wykres liczebności wyliczonych odległości dla pozytywnych widm o niskiej energii kolizji równej 10 eV. Można tu zaobserwować, że w porównaniu do wykresu dla widm o negatywnej jonizacji o niskiej energii kolizji, wierzchołków jest o wiele więcej. Cząsteczka o masie 18 znów jest najczęściej występującym związkiem, czego można się było spodziewać. Kolejne są znów powtarzające się pary 28 i 26, 56 i 58. Widoczną różnicą jest pojawienie się wartości 40 zamiast 44. Dodatkowo wartości 98 (H_3PO_4), 154 i 252 (cukry) też są bardzo widoczne.

Na rysunku 18 reprezentującym utracone cząsteczki obojętne na widmach jonów dodatnich uzyskanych dla energii kolizji 20 eV można zaobserwować znaczny spadek w liczbie wystąpień cząsteczki 18. Jednak mimo to wartości najczęściej występujących cząsteczek zostały takie same. Cząsteczki o dużej masie (154 i 252) też występują wiele rzadziej. Powodem tego jest fakt, że po zderzeniu o większej energii badana w spektrometrze cząsteczka od razu rozpada się na drobniejsze związki.

7. Wyniki badań

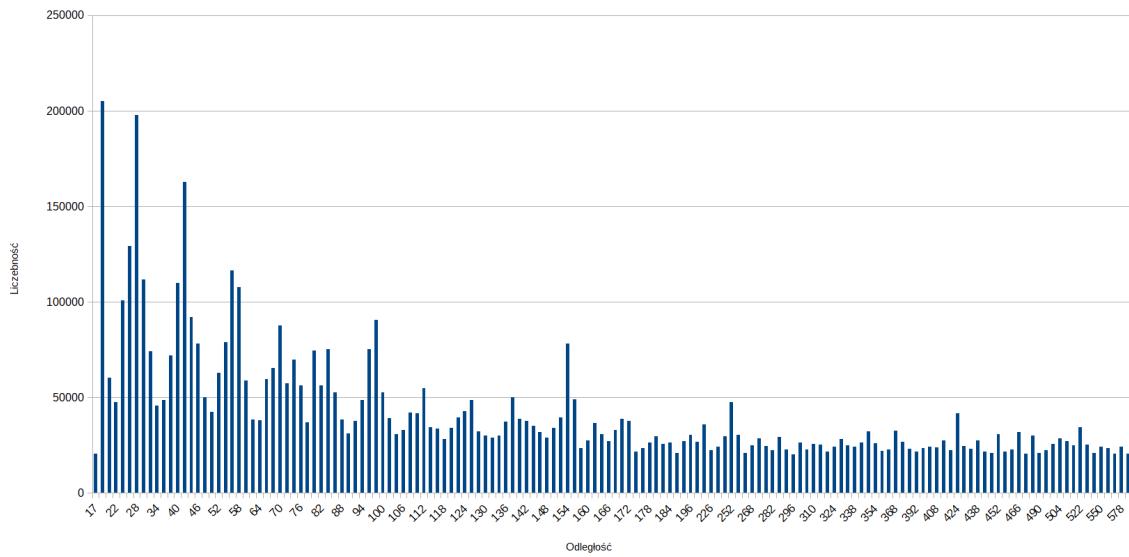


Rysunek 16. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 40 eV.

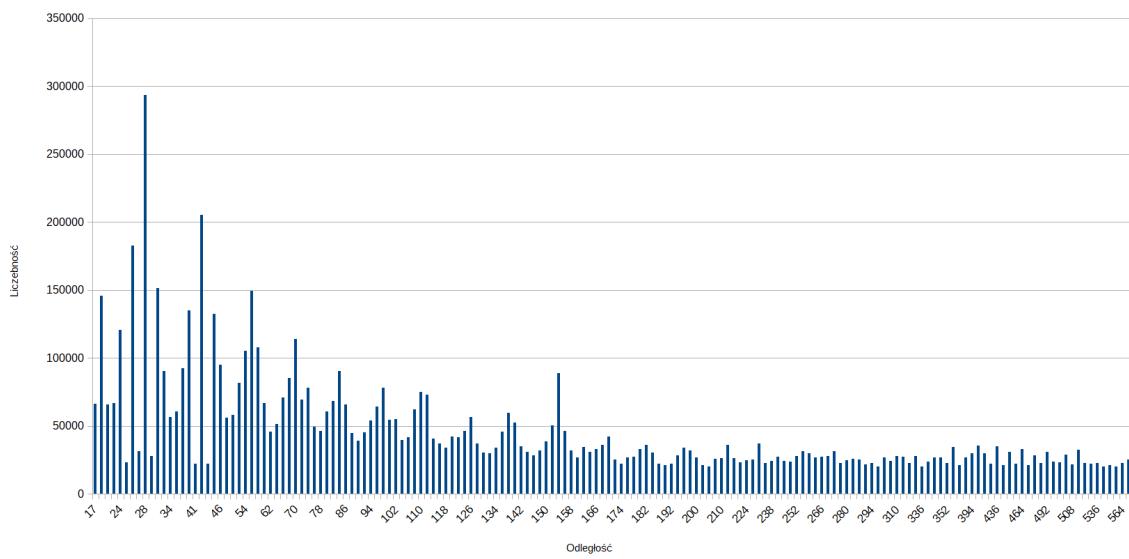


Rysunek 17. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 10 eV.

Na rysunku 19 można zaobserwować podobną tendencję jak w przypadku widm o takiej samej energii kolizji i dodatniej jonizacji. Cząsteczki o masie 18 występują znacznie rzadziej niż cząsteczki o masie 28. Następnie można zaobserwować kolejne bardzo wyraźne punkty takie jak 42, 56, 70, 84, 98 i 154.



Rysunek 18. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 20 eV.



Rysunek 19. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 40 eV.

7.2. Liczebność utraconych cząsteczek obojętnych w grupie lipidów i ich pochodnych

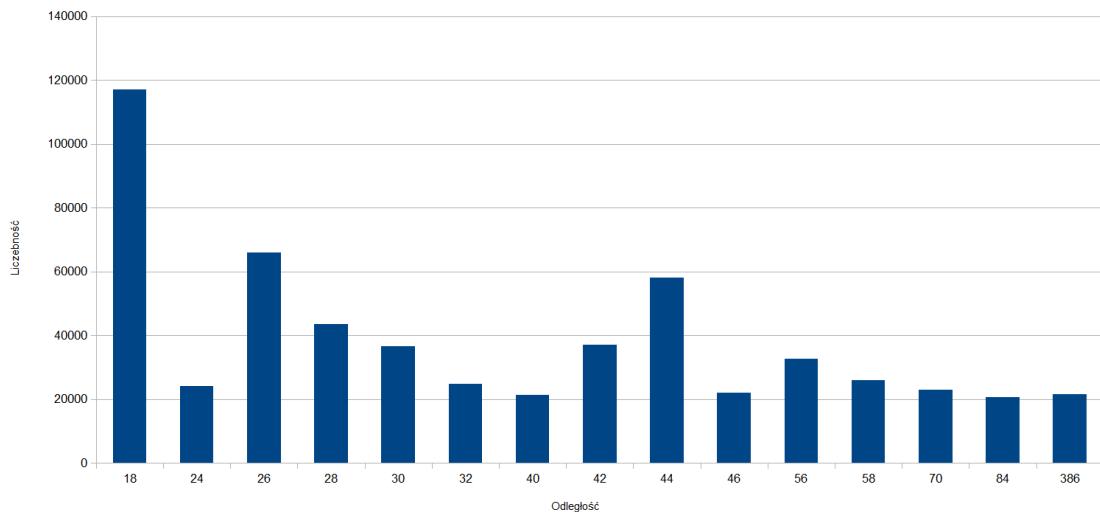
Z całego zbioru obliczonych utraconych cząsteczek obojętnych wybrano tylko takie, które zostały wygenerowane z widm związków chemicznych z grupy lipidów i ich pochodnych. Następnie zostały wykonane podliczenia liczby wystąpień poszczególnych odległości. Tabela 7 zawiera listę 20 najczęściej występujących odległość dla każdej kombinacji trybu jonizacji i energii kolizji. Można zaobserwować, że pojawiające się w niej wartości są bardzo zbliżone do wartości z tabeli 6. Wartości 18, 28, 26, 42 i 44 znów występują najczęściej. Jest to oczekiwany wynik, ponieważ te cząsteczki są częstym końcem łańcucha rozpadu. W trybie jonów ujemnych pojawia się także wartość 136 i jej odpowiednik 138 w trybie jonów dodatnich odpowiadający cząsteczce o wzorze sumarycznym $C_3H_7O_4P$.

Tabela 7. 20 najczęściej występujących odległości dla grupy lipidów i ich pochodnych.

Jonizacja negatywna			Jonizacja pozytywna		
10 eV	20 eV	40 eV	10 eV	20 eV	40 eV
18	18	28	18	18	28
26	42	42	28	28	42
44	28	44	40	42	56
28	56	18	42	56	26
42	26	26	154	58	154
30	44	30	98	98	70
56	58	56	26	40	18
58	30	58	252	154	44
32	136	136	56	26	30
24	40	46	58	70	40
70	84	40	24	24	84
46	74	70	30	96	58
386	100	54	20	30	98
40	70	24	100	84	110
84	46	60	44	80	112
54	24	32	138	74	54
60	60	98	254	44	74
368	98	84	54	54	46
52	32	86	292	46	24
382	86	38	70	38	68

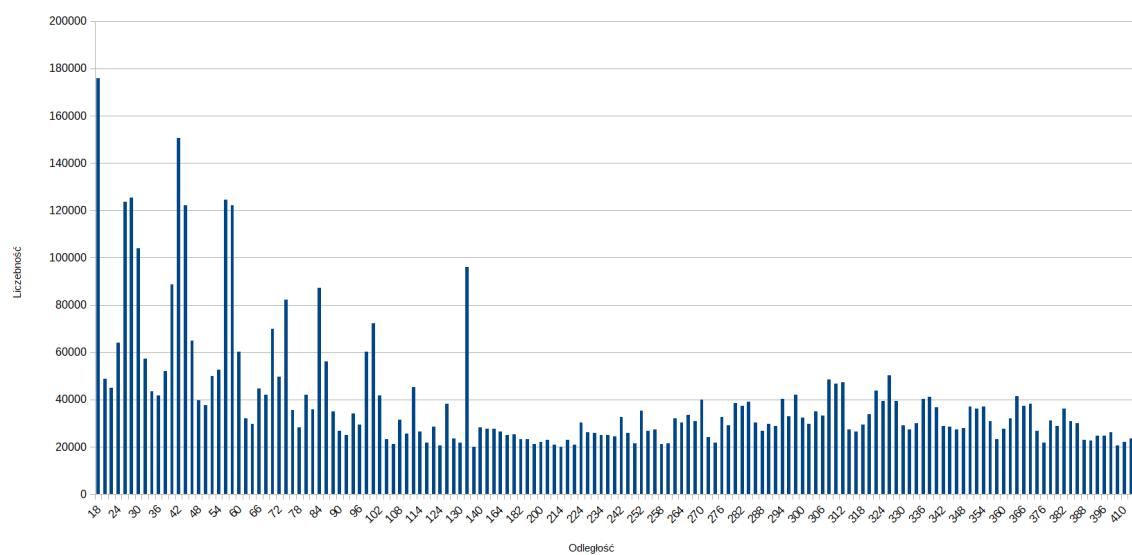
Rysunki 20 - 25 przedstawiają wykresy liczebności występujących odległości na poszczególnych kombinacjach trybu jonizacji i energii kolizji. Na rysunku 20 przedstawione są dane dla jonizacji negatywnej i energii kolizji 10 eV. Wartość 18 jest znów najczęściej występującą cząsteczką. Zaraz za nią występują wartości 26 i 44. Można też zauważyć, że liczba wyników (liczba słupków) jest o wiele mniejsza niż w przypadku widm o tej samej energii kolizji i pozytywnym trybie jonizacji. Powoduje to, że widma uzyskane w trybie

jonów dodatnich są bardziej reprezentatywne, szczególnie dla niskich wartości energii kolizji.



Rysunek 20. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 10 eV.

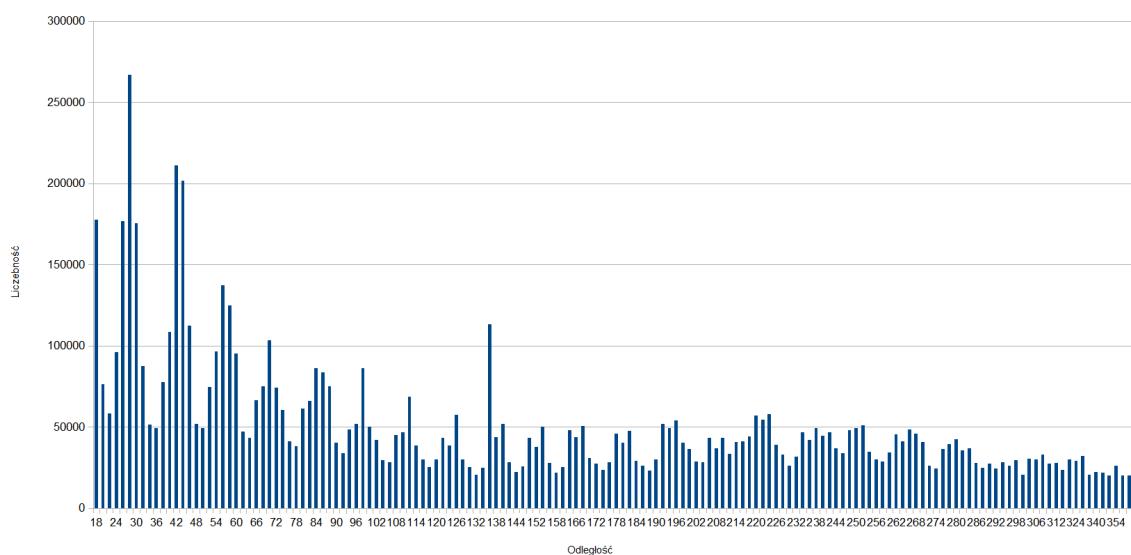
Dla średniej energii kolizji i negatywnego trybu jonizacji obliczone odległości znów tworzą charakterystyczne wzniesienia, co zostało przedstawione na rysunku 21. Najbardziej zaskakujący jest fakt, że wartość 42 pojawiła się częściej niż wartość 26 lub 28. Dodatkowo odległość równa 136 znów wystąpiła wielokrotnie. Zaobserwować też można pojawiające się fosforany (wartości 98 i 100).



Rysunek 21. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 20 eV.

7. Wyniki badań

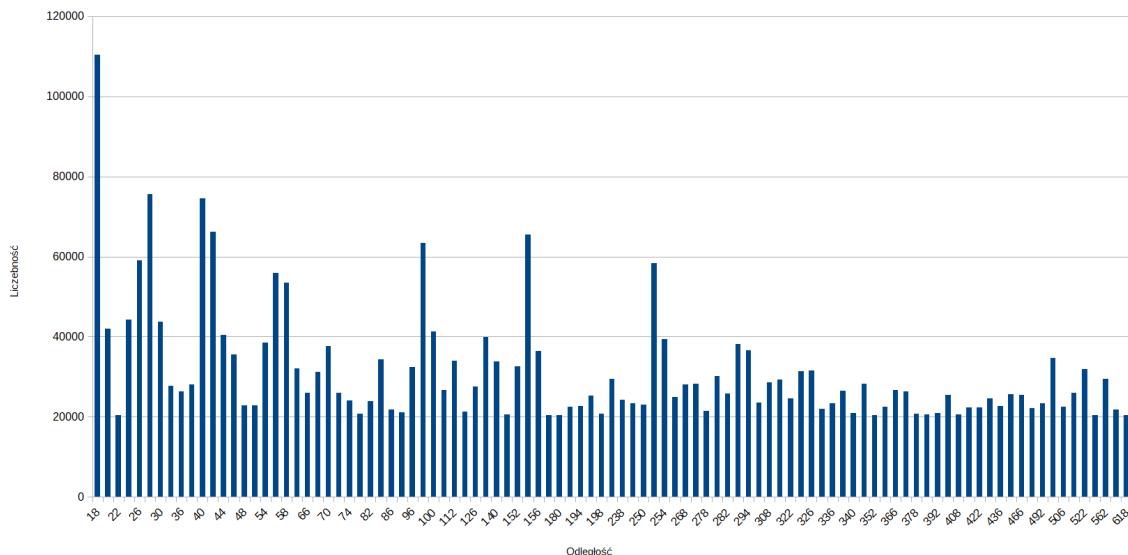
Dla grupy lipidów i ich pochodnych i ich widm o wysokiej energii kolizji można zauważać taką samą zależność, jaką się pojawiła dla grupy wszystkich związków. Na rysunku 22 widać, że liczliwość cząsteczki o wadze 28 jest znów największa. Jest to spowodowane tym, że przy dużej energii kolizji cząsteczki częściej rozpadają się na większe związki, które już nie mogą się rozpaść. Przy niższej energii kolizji ten proces występuje w stopniowych fazach. Wartość 136 znów występuje bardzo często, szczególnie w kontraste z jej otoczeniem. Pozostałe wznieśienia są o wiele wyraźniejsze i można zaobserwować dwie dodatkowe wartości 112 i 126, które nie były tak widoczne dla widm wszystkich związków.



Rysunek 22. Liczliwość utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 40 eV.

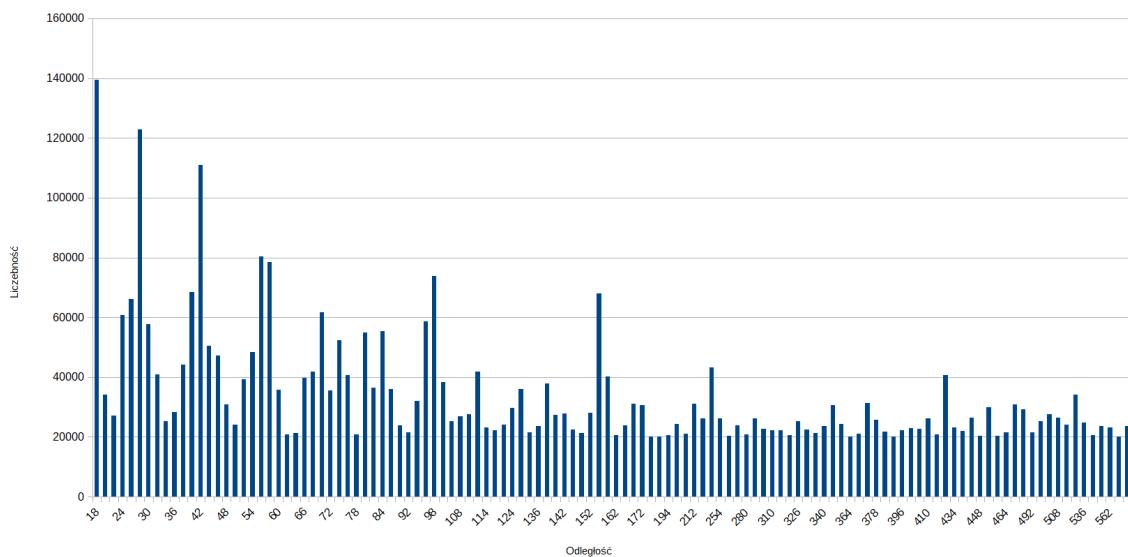
Rysunek 23 przedstawia liczliwość utraconych cząsteczek obojętnych w widmach o jonizacji pozytywnej i energii kolizji 10 eV dla grupy lipidów i ich pochodnych. W przeciwnieństwie do wykresu dla jonizacji negatywnej widać, że przy niskiej energii kolizji nadal można zaobserwować poszczególne cząsteczki obojętne. Cząsteczka o masie 18 znów występuje najczęściej. Dalej widać znów często powtarzające się cząsteczki: 26, 28, 40, 42, 56, 58. Dodatkowo występują jeszcze trzy odległości: 98, 154 i 252. Wartość 98 odpowiada utracie grupy fosforylowej w postaci kwasu fosforowego (H_3PO_4), natomiast 154 może odpowiadać grupie fosforylowej z przyłączonym cztero-węglowym łańcuchem ($C_4H_11O_4P$).

Na rysunku 24 widać, że różnica pomiędzy wystąpieniami cząsteczek o wadze 18 nie jest już tak dominująca. Cząsteczki 28 i 42 są też bardzo wyraźnie zaznaczone na tle ich otoczenia. Można też zaobserwować, że cząsteczki 98 i 154 nadal są bardzo widoczne. Cząsteczka o masie atomowej 252 też jest widoczna, ale widać znaczne obniżenie jej liczliwości. Jest to spowodowane tym, że przy większej energii kolizji duże cząsteczki rozpadają się szybciej. Zaskakująca jest jednak obecność jeszcze jednej cząsteczki: 424.



Rysunek 23. Liczebność ultraconnych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 10 eV.

Jest to bardzo duża cząsteczka, która nie była widoczna na widmach o niższej energii kolizji.

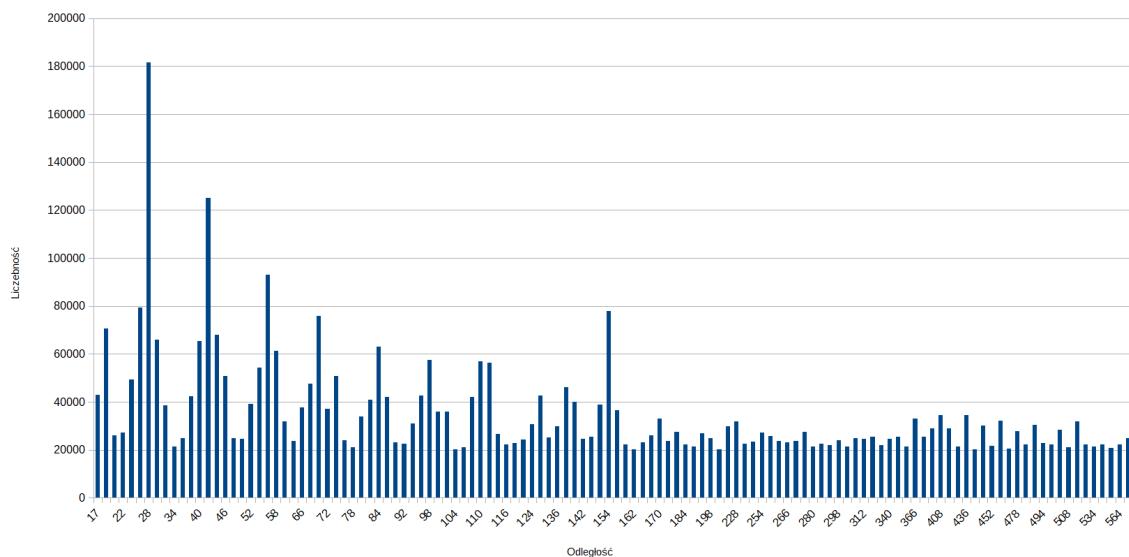


Rysunek 24. Liczebność ultraconnych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 20 eV.

Na rysunku 25 widnieje ostatnie podliczenie ultraconnych cząsteczek obojętnych. Tak jak w każdym widmie o energii kolizji równym 40 eV widać ewidentną przewagę cząsteczki o masie 28 oraz zanik występowania cząsteczki 18. Pozostałe wznowienia są bardziej uwydawnione na poszczególnych wartościach. Są to cząsteczki o masach 42, 56, 70, 84 i 98. Cząsteczka 154 jest jeszcze bardziej uwydawniona niż na poprzednich wykresach. Powyżej

7. Wyniki badań

154 żadna cząsteczka nie jest już tak charakterystyczna, co jest oczekiwany wynikiem. Przy takiej energii kolizji występowanie bardzo złożonych cząsteczek jest małe.



Rysunek 25. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 40 eV.

7.3. Podsumowanie podliczeń

Wszystkie zaprezentowane tutaj podliczenia cechowały się bardzo podobnym kształtem. Cząsteczki o małej wadze występuły wielokrotnie częściej niż większe cząsteczki, a poszczególne związki formowały w swoim otoczeniu charakterystyczne wznieśienie.

Najczęściej występującą cząsteczką jest cząsteczka o masie 18 (H_2O). Jest to spowodowane tym, że jest to jedna z podstawowych cząsteczek, więc pojawia się ona nie tylko w wyniku rozpadu badanego związku, ale też kolejnych związków na jego ścieżce rozpadu. Dodatkowo można zaobserwować grupę często powtarzających się wartości. Odległość 26 (najprawdopodobniej C_2H_2), 28 (CO lub C_2H_4), 42, 44 (CO_2), 56 i 58 (najprawdopodobniej C_3H_6O w przypadku grupy lipidów i ich pochodnych). Słupki w otoczeniu tych punktów różnią się atomami wodoru (np. odległość 28 może być związkiem C_2H_4 , który ma dwa atomy wodoru więcej niż cząsteczka o masie 26 i wzorze chemicznym C_2H_2). Ciekawą obserwacją jest też fakt, jak rzadko występują cząsteczki 17 (NH_3) i 27 (HCN). Oznacza to, że znaczna większość metabolitów ludzkich występujących w naturze ma charakter kwaśny.

Najciekawszymi jednak są odległości, których otoczenie nie opada łagodnie. Dla widm negatywnych jest to cząsteczka 136, a dla widm pozytywnych są to cząsteczki 154 i 252. Są to duże cząstki, więc zaskakujący jest fakt, jak bardzo one są stabilne, że nie rozpadają się pod wpływem kolizji. Cząsteczki te są tracone w przypadku występowania grupy fosforylowej charakterystycznej dla budowy fosfolipidów.

7.4. Reguły asocjacyjne

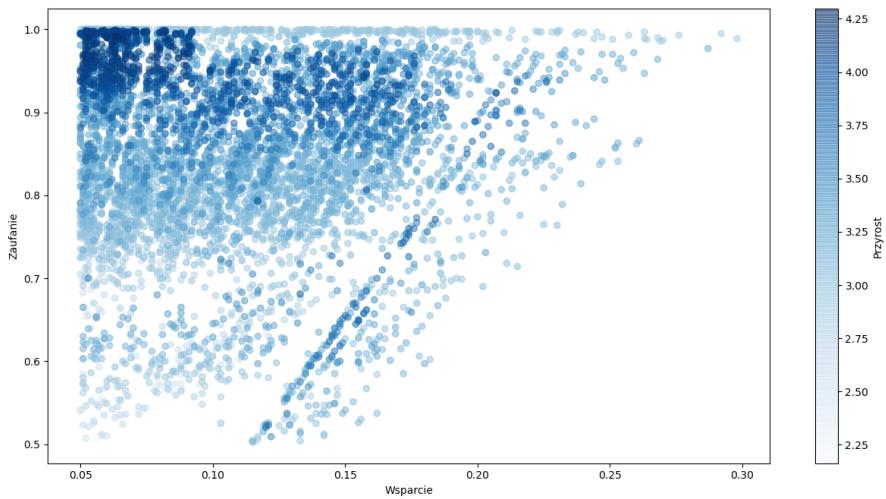
W pracy [14] jest zaprezentowany fakt, że w widmach pozytywnych charakterystyczne utracone cząsteczki obojętne występują o wiele częściej. Z tego powodu reguły asocjacyjne zostały wykonane tylko dla widm o pozytywnej jonizacji. Wyniki reguł asocjacyjnych zostały przedstawione w postaci wykresów punktowych. Na osi poziomej znajduje się wsparcie reguły, czyli częstotliwość, z jaką poprzednik i sekwencja występują razem. Jest to wartość w zakresie od 0 do 1. Na osi pionowej jest zaznaczona wartość zaufania reguły, czyli jaka jest szansa, że jeśli występuje poprzednik, to występuje też sekwencja. Wartość ta zawiera się w zakresie od 0 do 1. Każdy punkt symbolizujący regułę jest też oznaczony kolorem. Kolor oznacza wartość przyrostu i im ciemniejszy kolor, tym większa jego wartość. Przyrost jest współczynnikiem, który pozwala określić siłę (wpływ) danej reguły. Wartość 1 oznacza, że poprzednik i sekwencja nie są skorelowane, a wartości powyżej jedynki określają, jak bardzo zależne są wartości w regule.

Rysunki 26 - 31 przedstawiają wykresy punktowe reguł asocjacyjnych, gdzie poprzednikiem jest większa cząsteczka, a sekwencją mniejsza. Takie reguły mogą pomóc w określeniu, na jakie pomniejsze związki dana cząsteczka potencjalnie najczęściej się rozpada. Niestety liczebność wystąpień cząsteczek 18 i 28 powoduje, że bardzo dużo reguł zawiera je jako sekwencje z bardzo dużym zaufaniem bliskim 1. Wartość występująca w prawie każdym widmie powoduje, że algorytm apriori znajdzie regułę dla każdej kombinacji tego związku z każdym innym. Z tego powodu dla każdej wartości energii kolizji został zrobiony też drugi wykres, który nie zawiera reguł zawierających cząsteczek 18 i 28. Niestety te wykresy zawierają dużą ilość szumu, ponieważ reguły zostały wygenerowane dla wielu rzadko występujących cząsteczek, które mają według algorytmu apriori duże powiązanie z bardzo powszechnymi związkami.

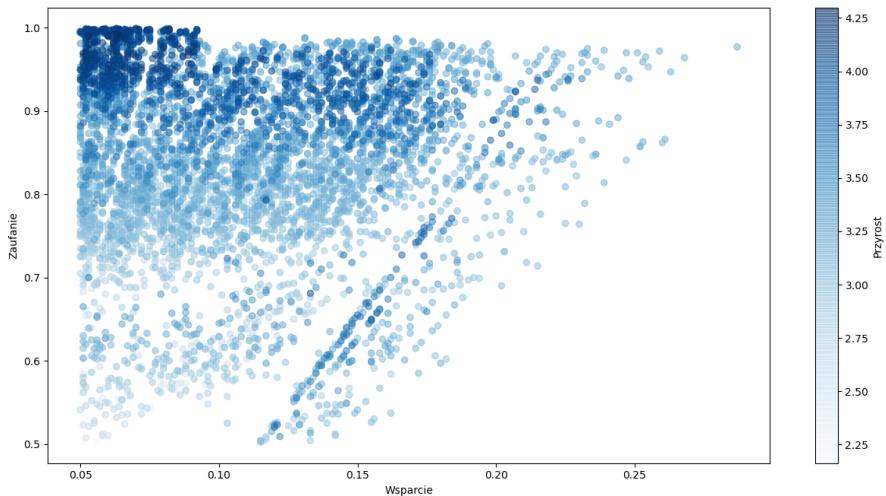
Na rysunkach 32 - 34 znajdują się wykresy punktowe reguł asocjacyjnych, gdzie poprzednikiem jest mniejsza cząsteczka, a sekwencją większa. Te reguły pomagają określić i znaleźć związki, które występują często w parze z mniejszymi cząsteczkami. W przeciwnieństwie do reguł z większym poprzednikiem, a mniejszą sekwencją są one bardziej czytelne. Spowodowane jest to tym, że wartości szumu występują zbyt rzadko by być brane pod uwagę.

Na wszystkich poniższych wykresach można zaobserwować charakterystyczne linie. Są one efektem dopasowania kolejnych sekwencji do tego samego poprzednika. Dzieje się tak dlatego, że wsparcie samego poprzednika jest stałe, a wraz ze zwiększającymi się wspólnymi wystąpieniami poprzednika i sekwencji (wsparcia reguły) zwiększa się też pewność, z jaką razem występują (zaufanie reguły). W takich przypadkach najczęstszymi regułami są te na prawym górnym krańcu takiej linii. Te reguły mają największe wsparcie i największe zaufanie.

Rysunki 26 i 27 przedstawiają wykresy dla reguł w kierunku z większej cząsteczki do mniejszej wykonanych dla widm o energii kolizji 10 eV. Różnice pomiędzy wykresami



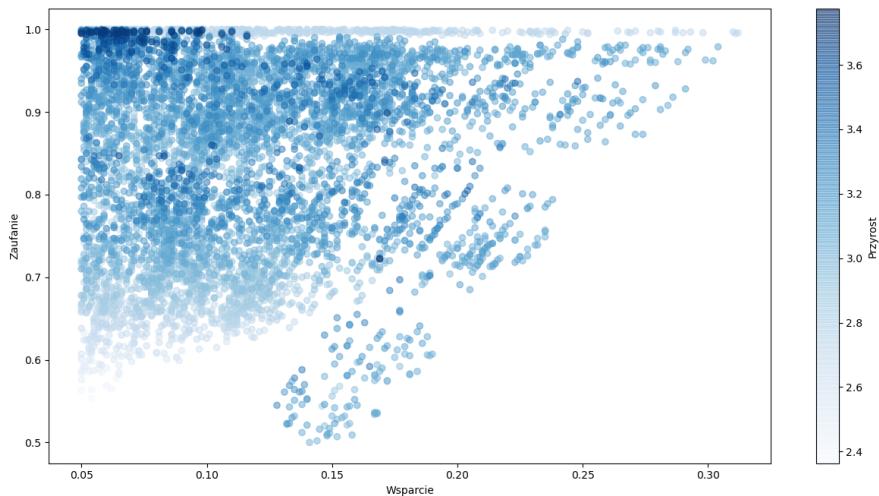
Rysunek 26. Wykres punktowy reguł asocjacyjnych dla ultracony czasteczek obojetnych w kierunku z większej czasteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozitwnej i energii kolizji 10 eV.



Rysunek 27. Wykres punktowy reguł asocjacyjnych dla ultracony czasteczek obojetnych w kierunku z większej czasteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozitwnej i energii kolizji 10 eV z wykluczeniem cząsteczek 18 i 28.

pokazują, jak duży wpływ mają cząsteczki 18 i 28. Wiele reguł gdzie te cząsteczki były sekwencjami miały wartość zaufania bliską 1. Na obu wykresach można też zaobserwować 2 charakterystyczne miejsca o dużej wartości przyrostu: zgrupowanie punktów w lewym górnym rogu oraz linia. Zgrupowanie punktów w lewym górnym rogu jest wynikiem utworzenia reguł z szumem jako poprzednikiem i często występującą wartością jako sekwencją. Zawierają one reguły, z bardzo dużą ilością poprzedników o dużych odległościach, które

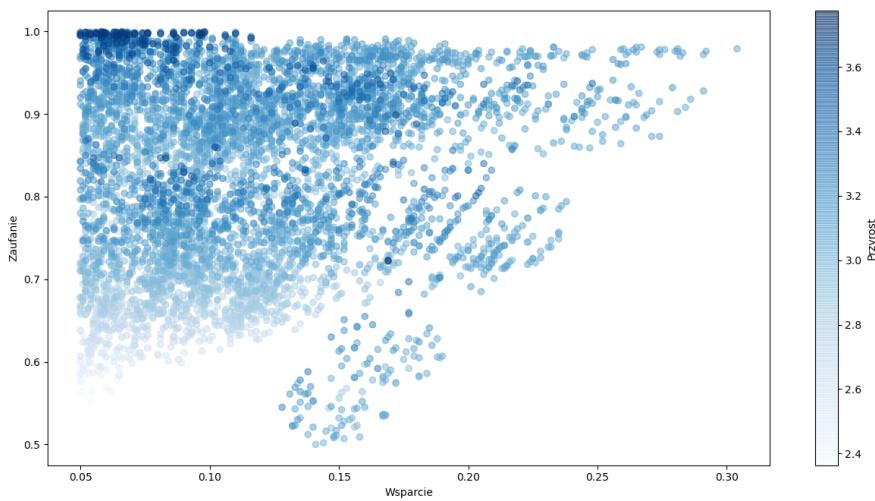
nie występują często w widmach i z sekwencją, która jest jedną z często występujących wartości, lub przypadkową inną wartością szumu, która w tych rzadkich przypadkach występuje zawsze razem. Istotniejszym fragmentem są ciemnoniebieskie linie. Punkty na tych liniach składają się z reguł, których poprzednikiem są najpowszechniejsze duże cząsteczki, które pojawiają się w 10 eV widmach pozytywnych, czyli wartości 154 i 252. Wszystkie te punkty mają bardzo wysoką wartość przyrostu. Większość z nich jest jednak regułami wiążącymi te wartości z wartościami z ich otoczenia (które najczęściej są błędem pomiarowym). Jednak po odfiltrowaniu tych wartości okazuje się, że wartości 252 i 154 są bardzo mocno powiązane z wartością 40, 58 (C_3H_6O), 98 (H_3PO_4) i między sobą. Można z tego wywnioskować, że są to wartości, na jakie rozpada się ten związek i potwierdzają zaproponowane wcześniej tożsamości. Wykonując proste obliczenia matematyczne, można zauważać, że wartość 252 najprawdopodobniej rozpada się na cząsteczki 154 i 98.



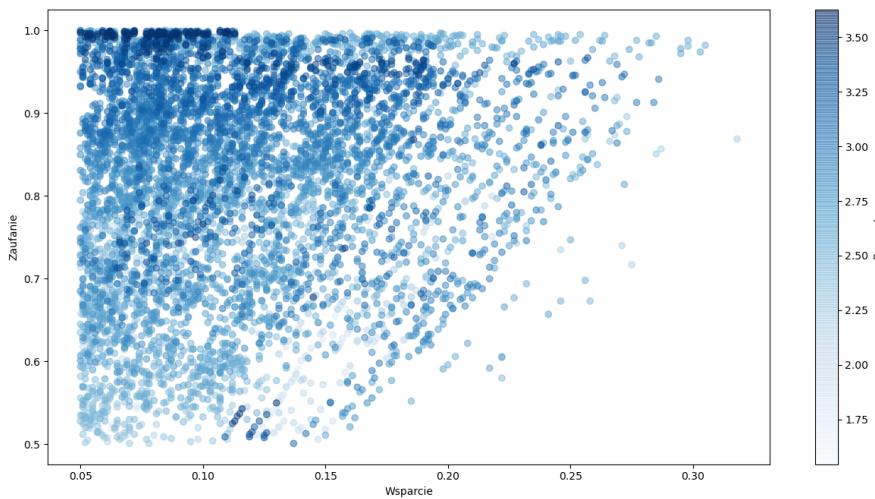
Rysunek 28. Wykres punktowy reguł asocjacyjnych utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV.

Na rysunkach 28 i 29 znajdują się wykresy reguł asocjacyjnych w kierunku z większej cząsteczki do mniejszej zrobionych dla widm o energii kolizji 20 eV. Znowu można zaobserwować duży wpływ cząsteczek 18 i 28. W lewym górnym rogu widać skupisko reguł, które składają się w znacznej większości z szumu. W przeciwnieństwie do poprzednich wykresów te nie zawierają wyróżniające się linii. Jednak po odfiltrowaniu rzadko występujących poprzedników udało się odnaleźć reguły podobne do tych znalezionych w widmach o energii kolizji 10 eV. Jednak bez tej wiedzy sam wykres nie pozwala na łatwe znalezienie tych zależności.

Wykresy reguł asocjacyjnych w kierunku z większej cząsteczki do mniejszej, które zostały wykonane dla widm o energii kolizji 40 eV, znajdują się na rysunkach 30 i 31. Przy

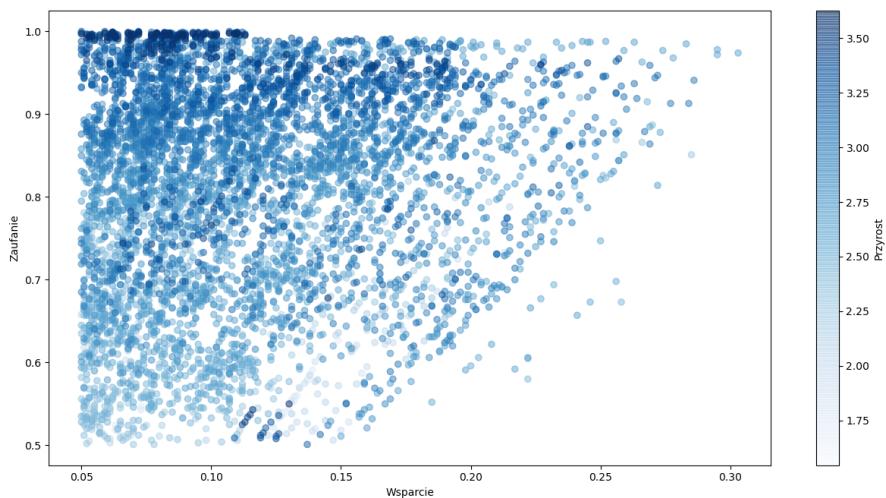


Rysunek 29. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV z wykluczeniem cząsteczek 18 i 28.



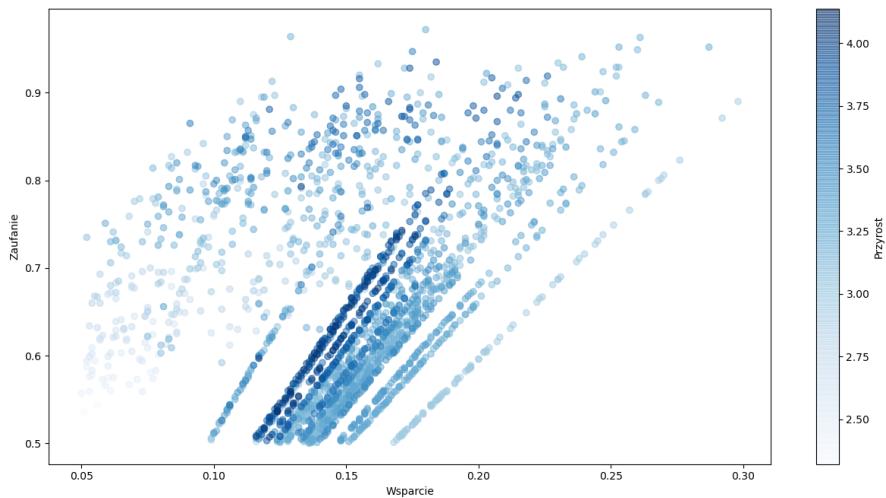
Rysunek 30. Wykres punktowy reguł asocjacyjnych utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV.

takiej energii kolizji cząsteczka 18 już nie występuje tak często, lecz została zastąpiona przez cząsteczkę 28. Niestety tak samo, jak przy widmach o energii kolizji 20 eV ciężko jest zauważać jakieś charakterystyczne punkty inne niż grupa w lewym górnym rogu. Jest to spowodowany tym, że przy tak wysokiej energii kolizji wszystkie większe cząsteczki rozpadają się bardzo szybko i nie są wykrywane na wyjściu. Przez to występuje o wiele



Rysunek 31. Wykres punktowy reguł asocjacyjnych dla ultraconnych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV z wykluczeniem cząsteczek 18 i 28.

więcej małych cząsteczek, które przez swoją ilość są sekwencją wielu reguł i mają bardzo podobną wartość przyrostu.



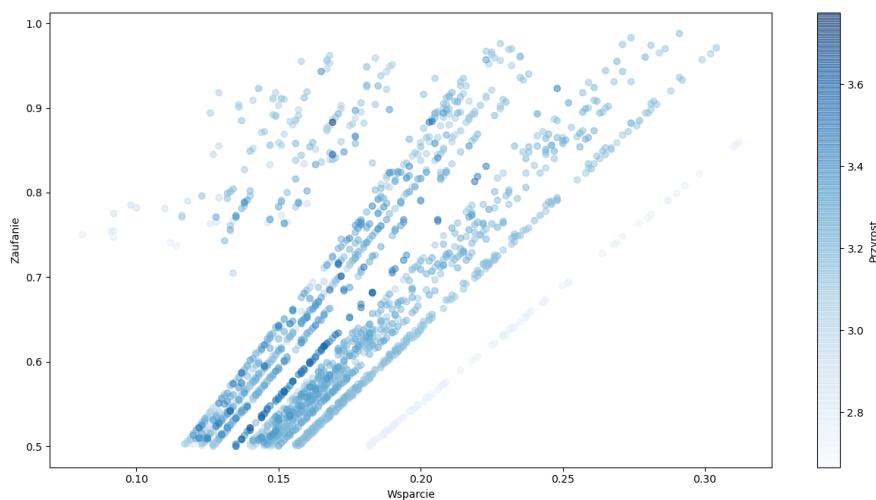
Rysunek 32. Wykres punktowy reguł asocjacyjnych dla ultraconnych cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 10 eV.

Na rysunku 32 znajduje się wykres punktowy reguł asocjacyjnych wykonanych dla widm o energii kolizji 10 eV z grupy lipidów i ich pochodnych. Zaprezentowane na tym wykresie reguły składają się z mniejszych cząsteczek jako poprzedników i większych czą-

7. Wyniki badań

steczek jako sekwencji. Jest to o wiele klarowniejszy wykres w przeciwieństwie do wykresu reguł w kierunku z większej cząsteczki do mniejszej. Można na nim odróżnić poszczególne grupy o wiele wyraźniej. Po lewej stronie wykresu można zaobserwować grupę reguł, które składają się głównie z szumu. Pośrodku tej grupy można zauważać też jedną dość widoczną linię. Jest ona jednak efektem reguł, których poprzednikiem są wartości szumu (wartości dookoła często występujących odległości), a sekwencja jest wartością 252, 292, 154 lub ich sąsiednie wartości. Po prawej stronie wykresu widać pojedynczą linię. Składa się ona z reguł, w których poprzednikiem jest wartość 18 (H_2O). Jako najpowszechniejsza w naturze utracona cząsteczka obojętna jest ona częścią wielu reguł. Zaskakujące jest, że reguły na prawym górnym końcu linii nie zawierają wartości 28 lub 26. Zamiast tego są to wartości 40 i 58, co by się zgadzało, ponieważ cząsteczka o masie 58 może się rozpaść na cząsteczki o masie 40 i 18. Świadczy to też o tym, że cząsteczki 18 i 28 lub 26 występują często razem, ale nie są od siebie zależne. Dalej na lewo można zaobserwować dwie wyraźne linie. Są to reguły, których poprzednikiem jest cząsteczka o masie 40 lub 58. Widać, że są to cząsteczki mocno skorelowane, szczególnie że reguły najbardziej w prawym górnym rogu, która jest widocznie oddzielona od reszty, jest reguła wiążąca cząsteczki 40 i 58. Na lewo od reguł z poprzednikiem 40 lub 58 znajduje się kolejne gęste skupisko reguł. Zaczyna się ono regułami z poprzednikiem 26, 28 i 98. Niestety najsilniejsze reguły łączą cząsteczki 26 i 28 z cząsteczkami 58 i 40. Świadczyłyby to jednak o tym, że cząsteczka 30 (CH_2O) i cząsteczka 32 (CH_4O) powinny być o wiele bardziej popularne. Taka reakcja może zachodzić, jednak ta reguła jest w większym stopniu sformułowana przez sam fakt wspólnego ich istnienia niż zależności pomiędzy nimi. Jedna z najsilniejszych reguł, gdzie cząsteczka 98 jest poprzednikiem, zawierają sekwencję z cząsteczkami 154 lub 252. Jest to oczekiwana reguła, ponieważ 252 może się rozpaść na cząstkę 154 i 98. Jednak jest jeszcze jedna silniejsza reguła łącząca cząsteczkę 98 z cząsteczką 138. Świadczyłyby to o fakcie, że rozpada się ona na cząsteczki 98 i 40. Co też by tłumaczyło, czemu cząsteczka 40 jest silnie powiązana z 98. Po lewej stronie środkowej grupy znajdują się dwie linie z bardzo dużym współczynnikiem przyrostu. Są to reguły z poprzednikami 154 i 252 oraz cząsteczkami z ich otoczenia. Niestety poza regułami łączącymi 154 i 252 ze sobą, większość pozostałych reguł jest stworzona z innymi cząsteczkami z otoczenia. Ostatnim interesującym zjawiskiem jest linia na lewo od najgęstszej grupy. Są to reguły z poprzednikiem 70. Najsilniejszą regułą z tym związkiem okazuje się reguła łącząca związek 70 i 98. Jeśli 70 byłoby częścią rozpadu cząsteczki 98, wyjaśniałoby to popularność cząsteczki 28. Jednak korzystając z pracy Agrawala, taki rozpad nie mógłby zajść przy podanych wzorach, ponieważ cząsteczka 28 (CO) nie jest możliwym wynikiem rozpadu związku 98 (H_3PO_4) niezawierającym atomu węgla. Jednak cząsteczka 98 może pochodzić z rozpadu innej cząsteczki (np. $C_4H_5 - COOH$), lecz bardziej prawdopodobny jest fragment cholesterolu.

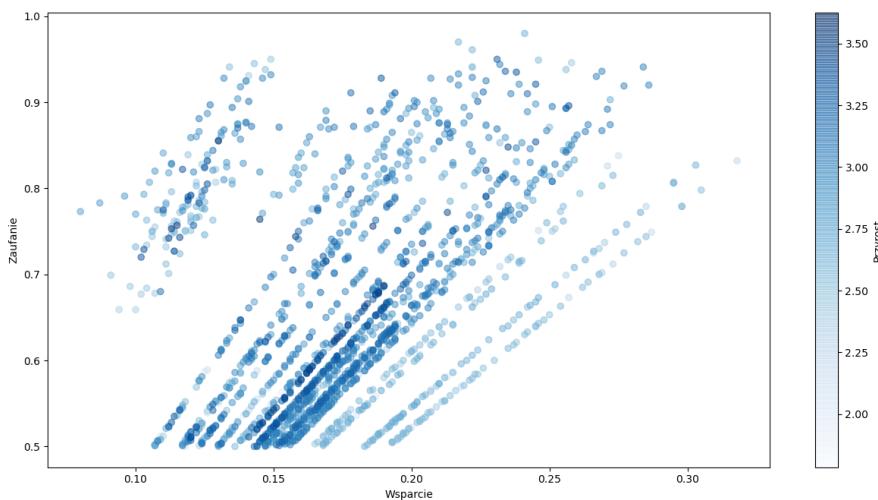
Rysunek 33 zawiera wykres reguł asocjacyjnych dla obliczonych utraconych cząsteczek obojętnych z widm lipidów i ich pochodnych o energii kolizji 20 eV. Jest to wykres bardzo



Rysunek 33. Wykres punktowy reguł asocjacyjnych dla ultracony cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV.

podobny do poprzedniego. Widać na nim jednak różnicę w liczebności cząsteczki 18. Po prawej stronie wykresu wyblakła linia reprezentuje reguły, w których poprzednikiem jest cząsteczka wody (masa 18). Widać, że przyrost tych reguł jest niższy. Dodatkową różnicą jest znaczny mniejszy szum na lewej stronie wykresu. Reszta reguł jest bardzo podobna do reguł opisanych przy wykresie wykonanym dla widm o energii kolizji 10 eV. Zmieniły się tylko wartości przyrostu. Regułami z największą wartością przyrostu są tutaj reguły z poprzednikiem 154.

Wykres punktowy reguł asocjacyjnych dla widm o energii kolizji 40 eV został zaprezentowany na rysunku 34. Można tutaj zaobserwować całkowity zanik reguł, których poprzednikiem była cząsteczka 18. Dodatkowo wiele wartości, które nie były tak popularne w widmach o niższej energii kolizji, stały się teraz jednymi z częściej występujących cząsteczek. Na tym wykresie pierwszą widoczną linię po prawej stronie wykresu stanowią reguły, których poprzednikiem jest najczęściej występująca w tych widmach cząsteczka o masie 28. Najsilniejsze z tych reguł jednak nie składają się z cząsteczek, które są wynikiem reakcji rozpadu. Są one wynikiem spowodowanym samym faktem częstotliwości występowania tych cząsteczek. Kolejna linia składa się z reguł z poprzednikiem o masie 42. Znów najsilniejsze reguły są dziełem przypadku i błędu pomiarowego, lecz można zaobserwować też bardzo silne powiązanie z cząsteczką 56. Kolejne 2 blisko leżące siebie linie to reguły z poprzednikami 56 i 58. Niestety tutaj też powiązane cząsteczki nie są efektem wspólnego rozpadu, ale obie mogą być powiązane z cząsteczką 28 (56 rozpada się na dwie cząsteczki 28, a 58 może się rozpaść na cząsteczkę 28 i 30). Ta zasada powtarza się na całym wykresie. Jej powodem jest fakt, że przy takiej dużej energii kolizji związki rozpadają się prawie



Rysunek 34. Wykres punktowy reguł asocjacyjnych dla ultracony cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV.

od razu do najmniejszych możliwych cząsteczek, więc związki na ścieżce rozpadu nie są wykrywane. Algorytm jedynie może utworzyć reguły na podstawie występowania dwóch produktów rozpadu, a nie na podstawie produktu rozpadu i pierwotnego związku.

7.5. Wnioski

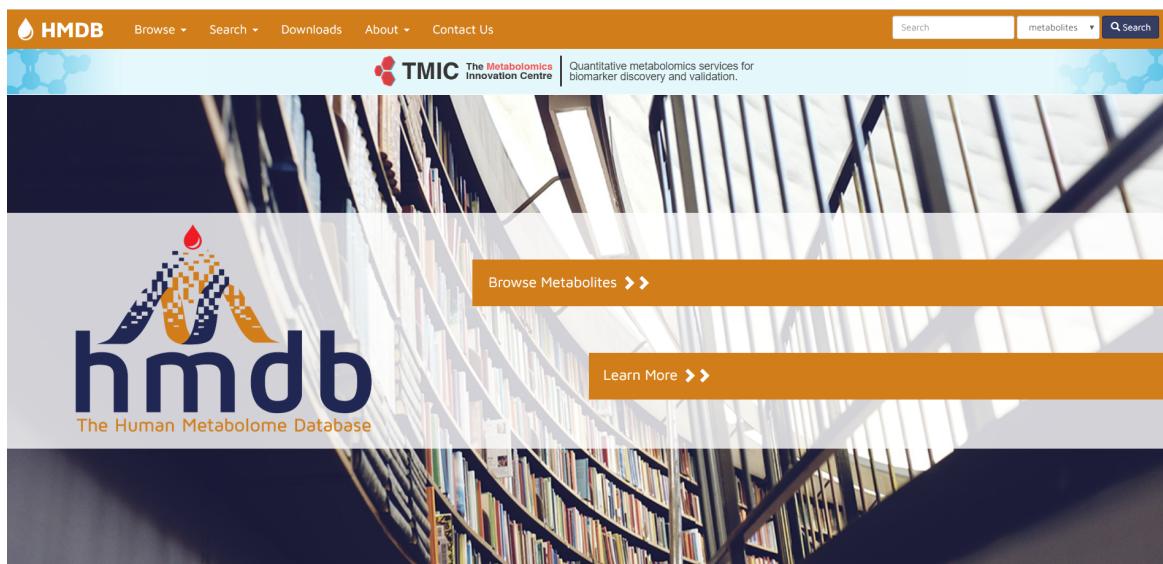
Przeprowadzone badania pozwoliły zaobserwować wiele występujących zależności. Pierwszą z nich jest różnica w liczbie wystąpień grup aminowych (odległość 17 i 27) i liczbie wystąpień grup hydroksylowych (cząsteczki o masach 18, 28 i 44). Związków zasadowych jest średnio około 10 razy mniej niż związków zawierających grupy o charakterze kwaśnym, co świadczy o tym, że znaczna większość ludzkich metabolitów ma charakter kwaśny. Cząsteczka 18 (H_2O) jest najczęstszą wykrytą odlegością, co jest spodziewanym wynikiem. Bardzo dużo związków zawiera atomy wodoru i tlenu i w trakcie rozpadu wydzielana jest cząsteczka wody. Jednak wszystkie widma o energii kolizji 40 eV jej nie zawierają. Przy większej energii kolizji rozpad jest o wiele bardziej gwałtowny i nie następuje tak stopniowo, jak przy mniejszych energiach. Ten efekt widać po porównaniu różnic w wykresach liczebności dla różnych energii kolizji. Wraz ze wzrostem energii kolizji maleje liczebność cząsteczki 18, a wzrasta liczebność cząsteczek 26 i 28. Na wykresach punktowych reprezentujących obliczone reguły asocjacyjne można też zaobserwować charakterystyczne wzory. Pozwoliły one na znalezienie szczególnych miejsc, na które warto było zwrócić uwagę. Charakterystyczne linie są wynikiem reguł asocjacyjnych, których poprzednik jest identyczny. Dzięki temu można w prosty sposób przebadanie i znaleźć zależności związane z tym poprzednikiem. Dla związków ludzkich metabolitów

z grupy lipidów i ich pochodnych szczególnie interesującymi związkami były cząsteczki o masie 154 i 252. Są one szczególnie istotne dla widm o niskiej energii kolizji. Jest to spowodowane łagodniejszym procesem rozpadu, co pozwala na utworzenie reguł z cząsteczek przed i po rozpadzie (część cząsteczek o wysokiej masie się rozpadła na swoje podzespoły, a część została zarejestrowana przez czujnik aparatury). Takie reguły są o wiele bardziej czytelne i łatwe do wychwycenia. W przypadku większych energii kolizji reguły są tworzone w większości pomiędzy wszystkimi wynikami rozpadu i ciężko wywnioskować ich powiązanie. Związki o masach 154 i 252 zostały też mocno powiązane z cząsteczką 98, co jest podstawą, by wierzyć, że cząsteczka 252 rozpadła się właśnie na cząsteczki 154 i 98. Dalsza analiza reguł wykazała, że podobną zależność można znaleźć pomiędzy cząsteczkami 40, 58 i 98. Jednak jeśli taki rozpad cząsteczki 98 był występował, oznaczałoby to, że w związku nie ma grupy fosforylowej, zatem jest to przypadkowa zbieżność. Ważną obserwacją jest też stabilność cząsteczki o masie 154. Nawet przy dużej energii kolizji (40 eV) nie rozpada się ona. Kolejną interesującą zależnością okazały się reguły wiążące związki o masach 98 i 138. Jest to potencjalna kolejna ścieżka rozpadu, która dodatkowo uwalniałaby z siebie cząsteczkę 40. Liczebność wystąpień cząsteczki 18 niestety spowodowała, że bardzo wiele wyników było fałszywych lub była dziełem przypadku (cząsteczka występująca często tworzy regułę z prawie każdym innym związkiem). Jednak mimo to udało się znaleźć regułę wiążącą związek o masie 40 ze związkiem o masie 58. Oba związki występowały wspólnie wielokrotnie. Jednak poprzez fakt, że związków 40 było więcej można wywnioskować, że cząsteczka 58 rozpada się właśnie na cząsteczki 40 i 18. Duża liczebność cząsteczek o masie 70 i reguła łącząca ją ze związkiem 98 też wyjaśniałaby powszechność występowania cząsteczki 28.

Zaprezentowane powyżej wyniki wykazują, że użycie reguł asocjacyjnych do wykrycia nowych i interesujących utraconych cząsteczek obojętnych oraz zależności występujących pomiędzy nimi jest możliwa i daje dobre wyniki dla widm o niskiej energii kolizji. Trzeba jednak pamiętać, że użycie takich reguł nie jest wystarczające, by jednoznacznie określić jakie to są cząsteczki i wymaga przeprowadzenia dodatkowych badań. Wynika to z faktu, że taka sama waga cząsteczki może być uzyskana poprzez kombinacje różnych atomów. Mogą one jednak zostać wykorzystane, aby wstępnie zawęzić obszar badań, znaleźć konkretne wagi utraconych cząsteczek lub grupy związków, co pozwoli na zmniejszenie czasu badań, jak i zużycia potrzebnych do tego procesu odczynników. Jest to szczególnie istotne w takiej dziedzinie jak chemia, gdzie liczba danych jest obszerna, a ręczne szukanie takich zależności bardzo trudne.

8. Eksperymentowanie z utworzoną aplikacją

Poniżej podano kilka przykładów ilustrujących możliwości utworzonego oprogramowania. Każdy przykład składa się ze zdefiniowanych parametrów wyszukiwania, a następnie pokazuje kroki wykonane podczas manualnego wyznaczenia parametrów rejestracji oraz kroki wyznaczenia tych samych parametrów z użyciem utworzonej aplikacji. Dla każdego z nich zmierzono czas wyznaczenia parametrów rejestracji metabolitów, który na końcu jest porównywany. W każdym przykładzie test wykonywany na aplikacji zaczyna się od okna startowego (rysunek 3), a w przypadku manualnego wyznaczania parametrów zaczyna się on od startowej strony HMDB (rysunek 35).



Rysunek 35. Strona startowa HMDB.

8.1. Przykład 1 - wyznaczanie parametrów rejestracji kwasu dikawoiłochinowego

Porównanie czasu wyznaczenia parametrów rejestracji kwasu dikawoiłochinowego metodą manualną i za pomocą utworzonej aplikacji.

Zadanie składa się z następujących czynności:

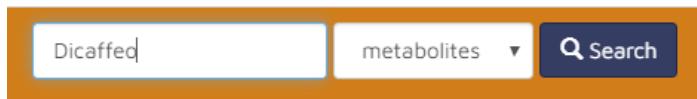
- znalezienia po nazwie kwasu dikawoiłochinowego (ang. Dicaffeoylquinic Acid),
- wyznaczenia i zapisania parametrów rejestracji z intensywnością większą lub równą 20 dla 10 elektronowoltowych widm pozytywnych,
- wyznaczenia i zapisania parametrów rejestracji z intensywnością większą lub równą 20 dla 20 elektronowoltowych widm pozytywnych.

8.1.1. Manualne wyznaczenie parametrów

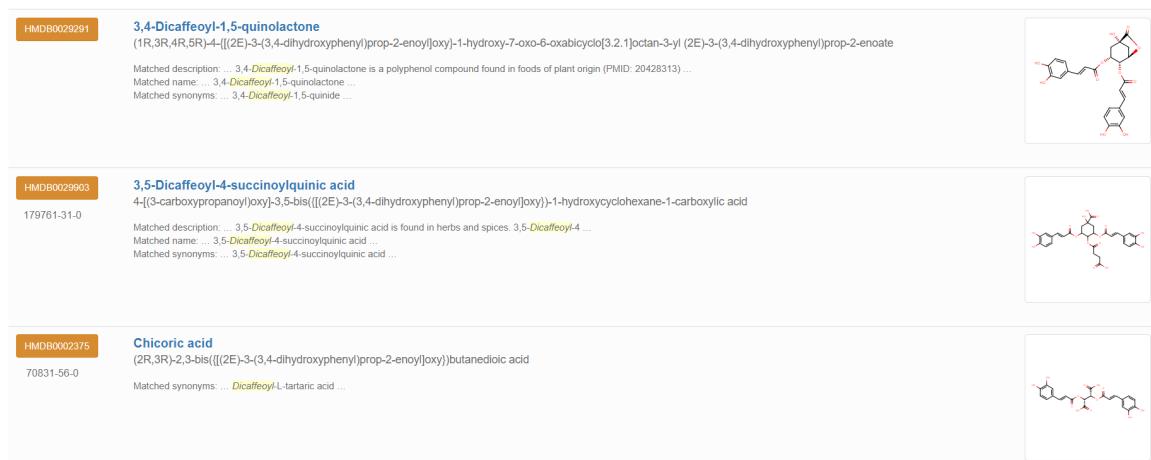
Pierwszym wykonanym krokiem było wpisanie fragmentu nazwy związku w polu wyszukiwania (Dicaffeo). Na rysunku 36 została przedstawiona szukana fraza. Znalezione

8. Eksperimentowanie z utworzoną aplikacją

metabolity (rysunek 37) zawierały szukaną frazę w różnych polach szczegółowych, jednak szukany metabolit nie został znaleziony.

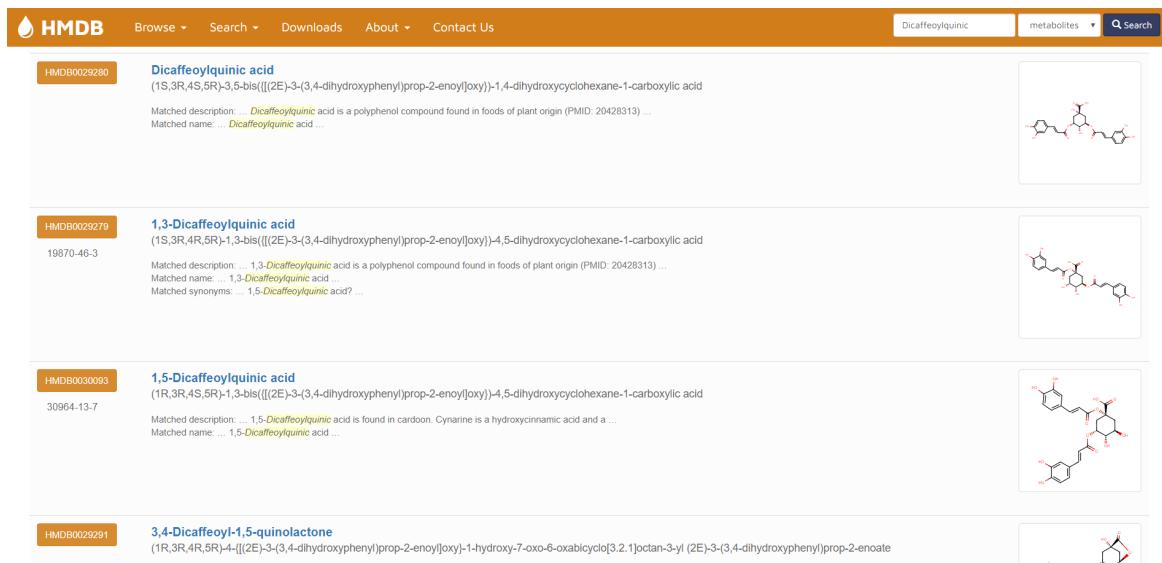


Rysunek 36. Wyszukiwanie kwasu dikawoilochnowego po fragmencie nazwy.



Rysunek 37. Związki znalezione po wyszukianiu fragmentu nazwy kwasu dikawoilochnowego.

Dopiero po wpisaniu pełnej nazwy związku (Dicaffeoylquinic) system zwrócił oczekiwany wynik (rysunek 38). Z listy został wybrany pierwszy związek, co spowodowało przejście na stronę z informacjami o związku oraz granatowym panelem nawigacyjnym (rysunek 39).



Rysunek 38. Wyszukiwanie związku z wykorzystaniem pełnej nazwy kwasu dikawoilochnowego.

8. Eksperymentowanie z utworzoną aplikacją

Showing metabocard for Dicaffeoylquinic acid (HMDB0029280)

Jump To Section: Identification Taxonomy Ontology Physical properties Spectra Biological properties Concentrations Links References XML

Show Metabolites with Similar Structures

Record Information	
Version	4.0
Status	Expected but not Quantified
Creation Date	2012-09-11 17:29:24 UTC
Update Date	2019-07-23 06:03:42 UTC
HMDB ID	HMDB0029280
Secondary Accession Numbers	• HMDB29280

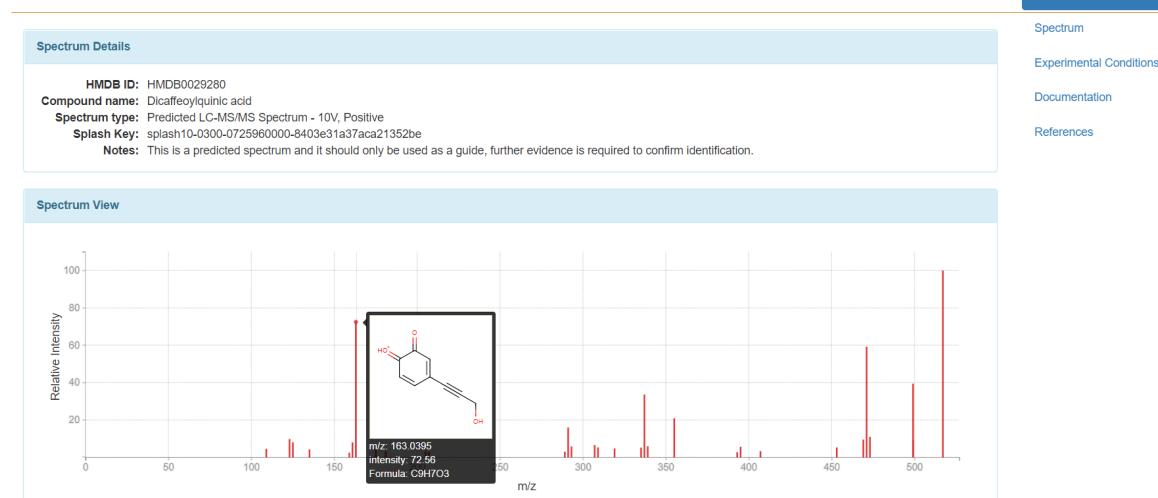
Rysunek 39. Strona HMDB kwasu dikawoilochinowego.

Z panelu nawigacyjnego został wybrany przycisk “Spectra”. Strona została przewinięta i wycentrowana na sekcji dotyczącej widm (rysunek 40).

Spectra			
Spectrum Type	Description	Splash Key	View
Predicted GC-MS	Predicted GC-MS Spectrum - GC-MS (Non-derivatized) - 70eV, Positive	splash10-0079-2553900000-e7b3dfed829c7f0f5984	JSpectraViewer MoNA
Predicted GC-MS	Predicted GC-MS Spectrum - GC-MS (2 TMS) - 70eV, Positive	splash10-01pa-6971405000-e34d11dad7937c86be68	JSpectraViewer MoNA
LC-MS/MS	LC-MS/MS Spectrum - LC-ESI-qToF, Positive	splash10-03di-0900000000-7b6d7228927b3f8dd2ff	JSpectraViewer MoNA
LC-MS/MS	LC-MS/MS Spectrum - LC-ESI-qToF, Negative	splash10-0udi-0609000000-42a893b713ad121e3e8b	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 10V, Positive	splash10-0300-0725960000-8403e31a37aca21352be	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 20V, Positive	splash10-03dr-0905400000-6d80268458b90a9825a4	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 40V, Positive	splash10-08ou-0911100000-5880ea311f7c93f19d1	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 10V, Negative	splash10-01b9-0504980000-756b5a404ac37e609c5a	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 20V, Negative	splash10-0kmi-0637610000-2ea45ea748b397c264c	JSpectraViewer MoNA
Predicted LC-MS/MS	Predicted LC-MS/MS Spectrum - 40V, Negative	splash10-08i9-0914000000-85006063ed7a07854afe	JSpectraViewer MoNA

Rysunek 40. Sekcja zawierająca widma kwasu dikawoilochinowego.

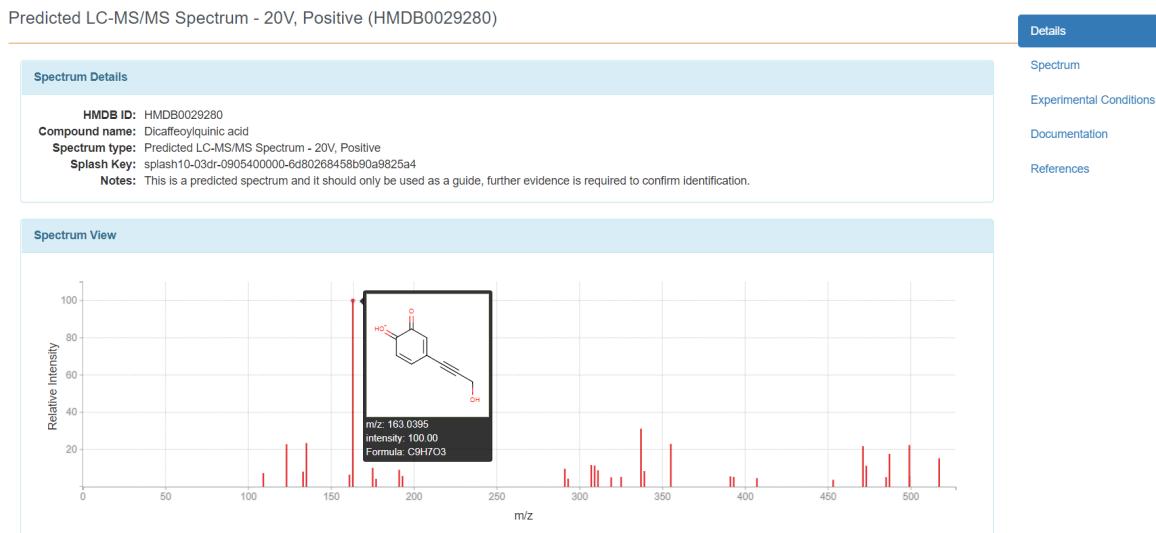
Predicted LC-MS/MS Spectrum - 10V, Positive (HMDB0029280)



Rysunek 41. Widmo kwasu dikawoilochinowego o energii kolizji równej 10 elektronowoltom.

Z listy wszystkich spektr zostały wybrane i otworzone w nowej karcie widma *Predicted LC-MS/MS Spectrum - 10V, Positive* (pokazane na rysunku 41) oraz *Predicted LC-MS/MS Spectrum - 20V, Positive* (pokazane na rysunku 42). Na obu widmach szczegółowe wartości poszczególnych wierzchołków można zobaczyć poprzez najechanie na niego kursorem.

8. Eksperymentowanie z utworzoną aplikacją

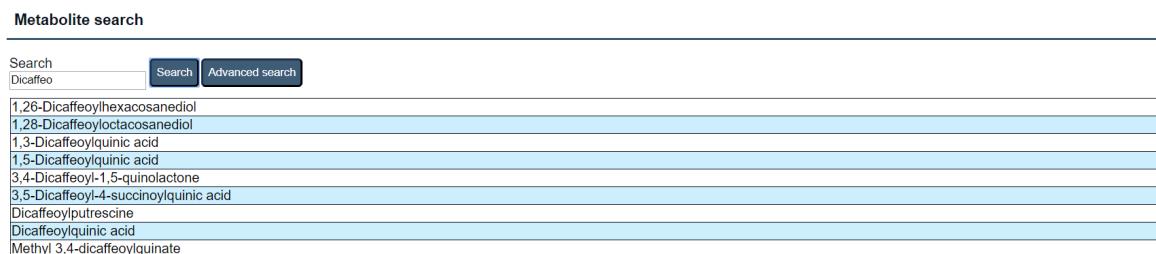


Rysunek 42. Widmo kwasu dikawoilochinowego o energii kolizji równej 20 elektronowoltom.

Ostatnim krokiem było przepisanie wartości wszystkich wierzchołków z obu widm, których wartości intensywności były większe niż 20. Cały proces od pierwszego wyszukania, aż do końca przepisywania szukanych parametrów zajął 3 minuty i 12 sekund.

8.1.2. Wyznaczenie parametrów z wykorzystaniem utworzonego oprogramowania

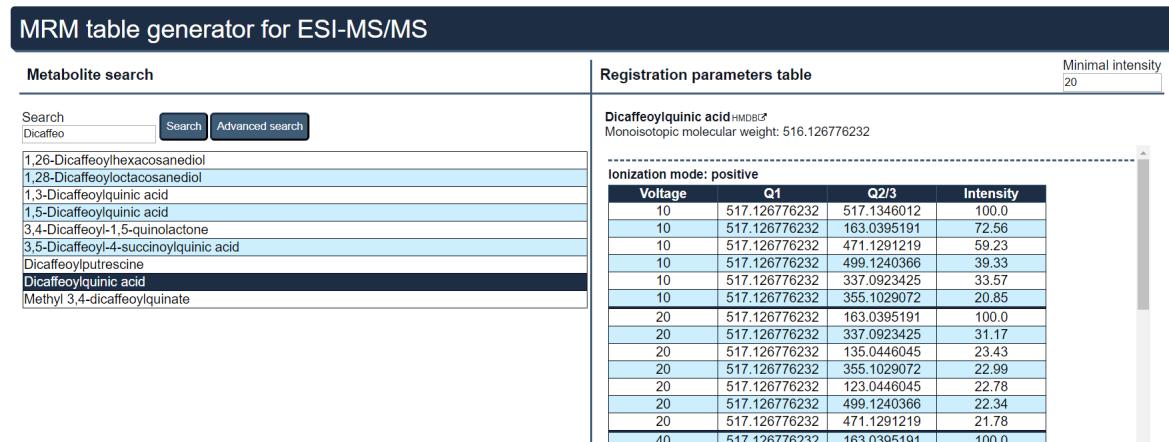
Tak samo jak w powyższym przykładzie z użyciem strony HMDB, pierwszym krokiem jest wyszukanie związku po części jego nazwy (do tego wykorzystano taką samą część całej nazwy: "Dicaffeo"). Po wpisaniu frazy w pole tekstowe i wcisnięciu klawisza enter lub przycisku "Search" pojawiła się lista ze wszystkimi dopasowanymi nazwami (rysunek 43). Lista wyników już w tym przypadku zawiera szukany związek, co pozwala na jego szybszy wybór.



Rysunek 43. Wyszukanie kwasu dikawoilochinowego w aplikacji używając części pełnej nazwy.

Po wybraniu szukanego związku strona ładuje informacje na temat parametrów rejestracji metabolitów i prezentuje je w czytelnej formie tabelki (rysunek 44). W prawym górnym rogu znajduje się pole ograniczające wyświetlane wyniki ustawione domyślnie na wartość 20, co pozwala w łatwy sposób odfiltrować wszystkie nieinteresujące użytkownika dane.

8. Eksperymentowanie z utworzoną aplikacją



Rysunek 44. Strona aplikacji po wybraniu kwasu dikawoiłochinowego.

Ostatnim krokiem jest znalezienie szukanych danych. W tym celu należy przewinąć stronę do sekcji z widmami w trybie jonizacji pozytywnej, a w niej wyszukać wiersze z napięciem 10 i 20 (rysunek 45). Znalezione wiersze można w prosty sposób zaznaczyć i skopiować. Cały proces trwał zaledwie 21 sekund i nie wymagał ręcznego kopiowania, a więc znacznie krócej niż w poprzednim punkcie.

Ionization mode: positive			
Voltage	Q1	Q2/3	Intensity
10	517.126776232	517.1346012	100.0
10	517.126776232	163.0395191	72.56
10	517.126776232	471.1291219	59.23
10	517.126776232	499.1240366	39.33
10	517.126776232	337.0923425	33.57
10	517.126776232	355.1029072	20.85
20	517.126776232	163.0395191	100.0
20	517.126776232	337.0923425	31.17
20	517.126776232	135.0446045	23.43
20	517.126776232	355.1029072	22.99
20	517.126776232	123.0446045	22.78
20	517.126776232	499.1240366	22.34
20	517.126776232	471.1291219	21.78

Rysunek 45. Szukane w przykładzie 1 parametry rejestracji kwasu dikawoiłochinowego.

8.1.3. Porównanie

Powyższy przykład pokazuje, o ile sprawniejsza w działaniu jest utworzona aplikacja. Ręczne wyznaczanie parametrów rejestracji metabolitów zajęło 3 minuty i 12 sekund, podczas gdy wykorzystanie prezentowanego programu przyspieszyło ten proces do 21 sekund. Oznacza to prawie 90% przyspieszenie czasu wyznaczenia. Dodatkowo zmniejsza ono szansę na błędy występujące podczas ręcznego przepisywania wartości. Ważnym elementem jest też sposób wyszukiwania związków. Baza HMDB pozwala na obszerniejsze wyszukiwanie poprzez fakt, że brane pod uwagę jest nie tylko pole nazwy związku, ale też pozostałe pola z informacjami. Jednak taka metoda powoduje utrudnienie, gdy

użytkownik wyszukuje konkretny związek o skomplikowanej nazwie. Aplikacja wyszukuje związki tylko po nazwie i wyświetla je w bardziej zwartej formie, co pozwala na szybsze znalezienie szukanego związku.

8.2. Przykład 2 - wyznaczanie parametrów rejestracji 1-Metylohistaminy

Porównanie czasu wyznaczania parametrów rejestracji 1-Metylohistaminy metodą manualną z wykorzystaniem zapytania tekstowego w bazie HMDB i za pomocą utworzonej aplikacji, wykorzystując zaawansowany tryb wyszukiwania.

Przy wyznaczeniu parametrów rejestracji zostały wzięte pod uwagę następujące znane informacje wstępne:

- szukany związek zawiera histaminę (ang. histamine),
- szukany metabolit znajduje się w ślinie (ang. saliva),
- należy wyznaczyć i zapisać parametry rejestracji metabolitu z intensywnością większą lub równą 10 dla 40 elektronowoltowych widm negatywnych.

Związek użyty w tym przykładzie to 1-Metylohistamina, ale w celu pokazania funkcjonalności programu zostało założone, że użytkownik nie zna pełnej nazwy związku.

8.2.1. Manualne wyznaczenie parametrów

Podczas manualnego wyznaczania szukanych parametrów, należy przejść na stronę wyszukiwania zaawansowanego bazy HMDB, gdzie można wykonywać własne zapytanie tekstowe. W tym celu należy najechać na przycisk “Search” na górnym panelu i wybrać z rozwiniętej listy opcję “Text Query” (rysunek 46).

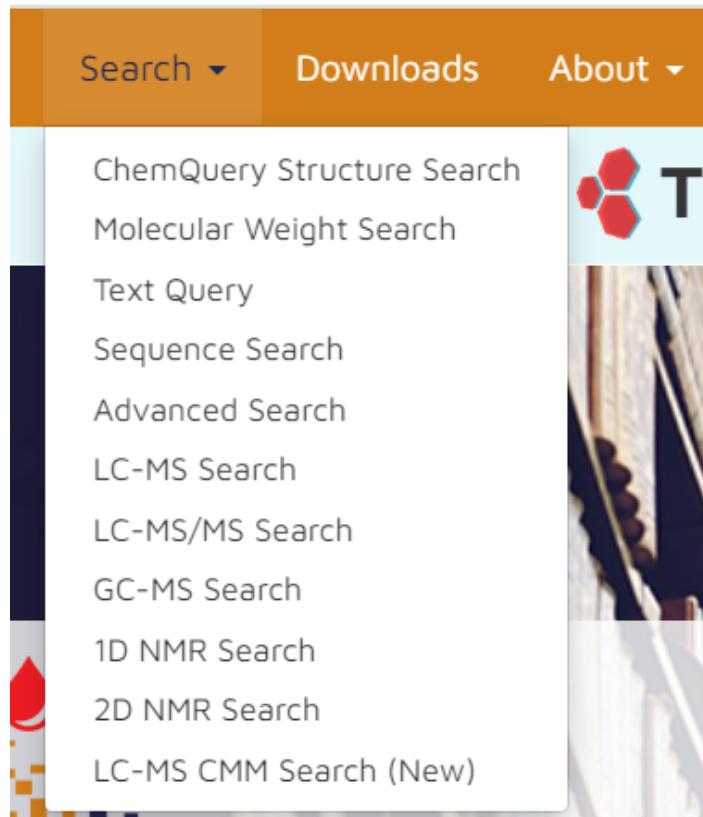
Po przejściu na stronę z polem do wpisania własnego tekstu zapytania zostało ono uzupełnione według podanego schematu (rysunek 47). Wykonane zapytanie miało formę:

```
name:* histamine * AND biofluid_source:saliva
```

Zapytanie to składa się z dwóch warunków oddzielonych operatorem “AND”, co oznacza, że oba warunki muszą zostać spełnione. Każdy warunek składa się z nazwy pola bazy i reguły, które to pole musi spełnić, oddzielonej dwukropkiem. Pierwsza część dotyczy pola nazwy związku “name”. Jedynym warunkiem postawionym na to pole jest to, że musi ono zawierać w swojej nazwie ciąg znaków “histamine”. Zostało to dokonane przy pomocy wyrażenia regularnego. Drugi człon zapytania warunkuje pole “biofluid_source”, który musi być identyczny z podaną wartością “saliva”. Taka opcja wyszukiwania pozwala na bardzo doprecyzowane zapytania, niestety jednak wymaga od użytkownika wiedzy o strukturze bazy danych i nazwie pól, oraz umiejętności posługiwania się wyrażeniami regularnymi.

Po wciśnięciu przycisku “Search” ładowana jest strona z wynikami, zaprezentowana na rysunku 48. Lista wyników zawiera szukaną w tym przykładzie związek. Po wciśnięciu

8. Eksperymentowanie z utworzoną aplikacją



Rysunek 46. Panel wyboru zaawansowanych opcji wyszukiwania na stronie bazy HMDB.

Searching HMDB

A screenshot of the HMDB search bar. It contains a text input field with the query "name:*histamine* AND biofluid_source:saliv" and a dropdown menu next to it labeled "metabolites". To the right of the input field is a blue "Search" button with a magnifying glass icon.

Rysunek 47. Pole do wpisania zapytania tekstowego na stronie bazy HMDB.

nazwy związku użytkownik zostaje przekierowany do głównej strony z informacjami na temat 1-Metylohistaminy (rysunek 49).

Strona składa się ze wszystkich informacji dotyczących danego związku oraz posiada na granatowym polu panel nawigacyjny, dzięki któremu szybko można przejść do wybranej sekcji. Używając go poprzez wcisnięcie przycisku "Spectra" strona została przewinięta i wyśrodkowana na sekcji zawierającej wszystkie widma metabolitu (rysunek 50).

W przypadku rozpatrywanego związku lista ta jest znacznie dłuższa. Przy pomocy przycisku "Show more..." została ona rozwinięta i na niej znalezione interesujące nas widmo (rysunek 51). Aby przejść do widoku szczegółowego widma, należy wcisnąć przycisk oznaczony napisem "JSpectraViewer".

Strona ze szczegółowymi informacjami o widmie zawiera sekcję z parametrami podsta-

8. Eksperymentowanie z utworzoną aplikacją

The screenshot shows two search results from the HMDB database:

- 1-Methylhistamine** (HMDB0000898): A metabolite of histamine, produced by histidine methyltransferase. It has a SMILES string of CNCC1=CC=C1. The page includes a detailed description, metabolite distribution across various body fluids (Blood, CSF, Feces, Saliva, Urine), and a chemical structure diagram.
- Histamine** (HMDB0000870): An amine derived from histidine, acting as a neurotransmitter. It has a SMILES string of CCNc1cc[nH]cn1. The page includes a detailed description, metabolite distribution, and a chemical structure diagram.

Rysunek 48. Wyniki wyszukiwania tekstowego na stronie bazy HMDB.

Showing metabocard for 1-Methylhistamine (HMDB0000898)

This is a detailed view of the HMDB metabocard for 1-Methylhistamine (HMDB0000898). The card includes sections for Record Information, Spectra, and References. Key details include:

- Record Information:** Version 4.0, Status Detected and Quantified, Creation Date 2005-11-16 15:48:42 UTC, Update Date 2019-07-23 05:44:35 UTC, HMDB ID HMDB0000898.
- Spectra:** Includes four GC-MS spectra and one GC-MS/MS spectrum.
- References:** A list of 10 references related to 1-methylhistamine.

Rysunek 49. Strona HMDB 1-Metylohistaminy.

This section displays the spectra for 1-Methylhistamine. It lists four GC-MS spectra and one GC-MS/MS spectrum, each with its corresponding splash key and viewing options (JSpectraViewer or MoNA).

	Spectrum Type	Description	Splash Key	View
	GC-MS	GC-MS Spectrum - GC-EI-TOF (Pegasus III TOF-MS system, Leco; GC 6890, Agilent Technologies) (2 TMS)	splash10-0fk-3900000000-8197b65f233a917f6f5d	JSpectraViewer MoNA
	GC-MS	GC-MS Spectrum - GC-EI-TOF (Pegasus III TOF-MS system, Leco; GC 6890, Agilent Technologies) (Non-derivatized)	splash10-00dr-3900000000-4a05dead99e371226a6c	JSpectraViewer MoNA
	GC-MS	GC-MS Spectrum - GC-EI-TOF (Pegasus III TOF-MS system, Leco; GC 6890, Agilent Technologies) (2 TMS)	splash10-0ddi-9700000000-b8108f1cb31c26ab942f	JSpectraViewer MoNA
	GC-MS	GC-MS Spectrum - GC-MS (2 TMS)	splash10-0fe0-5900000000-	JSpectraViewer

Rysunek 50. Sekcja zawierająca widma 1-Metylohistaminy.

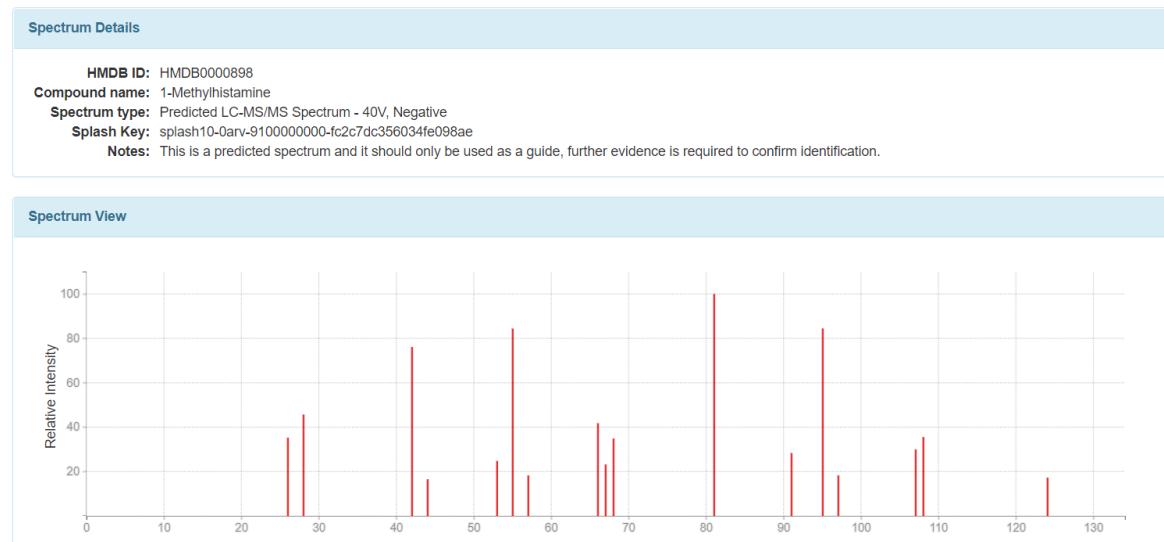
This image shows a specific LC-MS/MS spectrum for 1-Methylhistamine. The x-axis represents retention time, and the y-axis represents intensity. The spectrum shows several peaks, with the most prominent one at approximately 40V.

Rysunek 51. Szukane widmo, z którego zostaną wyznaczone parametry rejestracji 1-Metylohistaminy.

wowymi oraz rysunek widma (rysunek 52). Aby odczytać informacje dotyczące poszczególnych słupków na wykresie, należy najechać na nie kursorem myszy. W pojawiającym się oknie zawarte są dokładne wartości obu osi. Rysunek widma zawiera przedziały poziome oznaczające intensywność w odstępach równych 20. Powoduje to, że wyróżnienie i wybranie słupków, które spełniają postawiony warunek ograniczający intensywność do 10 i wzwyż, jest o wiele trudniejsze do szybkiego oszacowania. Zmniejszenie tego limitu też spowodowało potrzebę przepisania większej liczby parametrów, co dodatkowo

8. Eksperymentowanie z utworzoną aplikacją

wydłużyło czas wykonania i zwiększyło potencjalną liczbę błędów popełnionych podczas przepisywania. Łączny czas przeprowadzenia tego procesu zajął 3 minuty i 45 sekund.



Rysunek 52. Widmo 1-Metylohistaminy o jonizacji negatywnej i energii kolizji równej 40 elektrowniowoltów.

8.2.2. Wyznaczenie parametrów z wykorzystaniem utworzonego oprogramowania

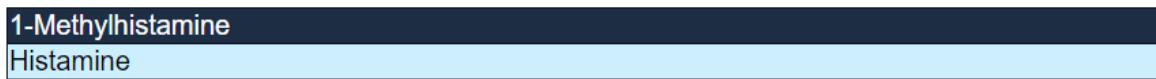
Aby wyszukać ten sam związek, wykorzystując do tego prezentowane oprogramowanie, należy przycisnąć pole odpowiedzialne za otworzenie zaawansowanego panelu wyszukiwania. W tym panelu należy uzupełnić pola dotyczące nazwy oraz "Biospecimen location" (rysunek 53). Pole "Biospecimen location" zawiera listę wszystkich możliwych wyborów, co ułatwia użytkownikowi precyzyjny wybór bez możliwości pomyłki w pisowni.

The screenshot shows a 'Metabolite search' interface. It includes input fields for 'Name' (histamine), 'Super class', 'Main class', 'Sub class', and a dropdown menu for 'Biospecimen location'. The 'Biospecimen location' dropdown is currently open, displaying a list of biological specimens: Blood, Feces, Urine, Cerebrospinal Fluid (CSF), Saliva, Sweat, and Amniotic Fluid. There are also 'Search' and 'Simple search' buttons at the bottom left. The 'Mass' section has two range sliders for 'min' and 'max'.

Rysunek 53. Panel wyszukiwania zaawansowanego w aplikacji.

Po wcisnięciu klawisza "enter" lub przycisku "Search" pojawi się lista wyników (rysunek 54). Znalezione nazwy związków są identyczne z tymi znalezionymi przy użyciu strony bazy danych HMDB.

Po wybraniu szukanego przez użytkownika związku pojawią się szczegółowe informacje dotyczące parametrów rejestracji 1-Metylohistaminy. Na górze tego ekranu (rysunek



Rysunek 54. Wyniki wyszukiwania w aplikacji.

55) znajduje się pole do filtrowania minimalnej wyświetlanej intensywności. Wartość ustaliona jest domyślnie na 20, więc została ona zmieniona na 10. Spowodowało to ponowne odfiltrowanie wszystkich wyników i wyświetlenie tylko takich, które spełniają ten warunek. Następnie wystarczy przewinąć stronę w poszukiwaniu szukanych parametrów rejestracji (rysunek 56), a znalezione wartości skopiować i przenieść do docelowego programu lub narzędzia. Cały wykonywany proces zajął 23 sekundy.



1-Methylhistamine HMDB
Monoisotopic molecular weight: 125.095297367

Rysunek 55. Sekcja z polem do filtrowania intensywności wybranego związku.

Ionization mode: negative			
Voltage	Q1	Q2/3	Intensity
10	124.095297367	124.0880209	100.0
20	124.095297367	124.0880209	100.0
20	124.095297367	95.0614718	23.59
20	124.095297367	97.07712187	17.18
40	124.095297367	81.04582174	100.0
40	124.095297367	95.0614718	84.48
40	124.095297367	55.03017168	84.43
40	124.095297367	42.03492271	76.11
40	124.095297367	28.01927264	45.68
40	124.095297367	66.03492271	41.77
40	124.095297367	108.0567208	35.54
40	124.095297367	26.00362258	35.27
40	124.095297367	68.05057277	34.89
40	124.095297367	107.0614718	29.98
40	124.095297367	91.03017168	28.32
40	124.095297367	53.01452161	24.77
40	124.095297367	67.03017168	23.19
40	124.095297367	97.07712187	18.27
40	124.095297367	57.04582174	18.24
40	124.095297367	124.0880209	17.27
40	124.095297367	44.05057277	16.54

Rysunek 56. Szukane parametry rejestracji 1-Metylohistaminy.

8.2.3. Porównanie

Powyższy przykład pokazuje, o ile łatwiejsza w obsłudze jest utworzona aplikacja. Podczas gdy ręczne wyznaczanie parametrów rejestracji zajęło 3 minuty i 45 sekund, to

8. Eksperymentowanie z utworzoną aplikacją

aplikacja poradziła sobie z tym problemem w czasie 23 sekund. Program znów przyspieszył czas wykonania procesu o prawie 90%. Aplikacja też jest bardziej intuicyjna dla użytkownika, ponieważ nie wymaga od niego wiedzy o strukturze bazy danych, o nazwach pól, ani nie wymaga znajomości wyrażeń regularnych. Tak samo, jak wykazał poprzedni przykład, oprogramowanie zmniejsza szansę na błędy występujące podczas ręcznego przepisywania wartości i znacznie ułatwia szybkie ich użycie.

8.3. Przykład 3 - identyfikacja Karwonu

Porównanie procesu identyfikacji i weryfikacji karwonu przy użyciu strony HMDB i utworzonej aplikacji.

Zdarza się sytuacja, kiedy chemikowi wygodniej jest wyszukać związek wykorzystując do tego jego monoizotopową masę molekularną. Taka potrzeba może się pojawić podczas próby identyfikacji i weryfikacji związku za pomocą spektrometru mas. Taki proces zaprezentowany został na przykładzie identyfikacji i weryfikacji karwonu (ang. Carvone).

Przy identyfikacji zostały wzięte pod uwagę następujące informacje wstępne:

- monoizotopowa masa molekularna związku wynosi 150.104,
- badany metabolit znajdował się w moczu (ang. urine),
- wiadomo, że dany związek należy do lipidów,
- należy porównać otrzymane ze spektrometru wyniki z parametrami rejestracji dla 40 elektronowoltowych widm negatywnych.

W tym przykładzie zadanie jest trudniejsze, ponieważ chemik nie wie jakiego związku szuka. Musi on wyznaczyć parametry rejestracji metabolitu dla wszystkich związków spełniających podane warunki, a następnie porównać z wynikiem analizy metabolitu uzyskanego ze spektrometru mas. W tabelce 8 znajduje się przykładowy wynik analizy karwonu w spektrometrze mas, który został zrobiony w trybie jonizacji negatywnej dla energii kolizji równej 40 elektronowoltom.

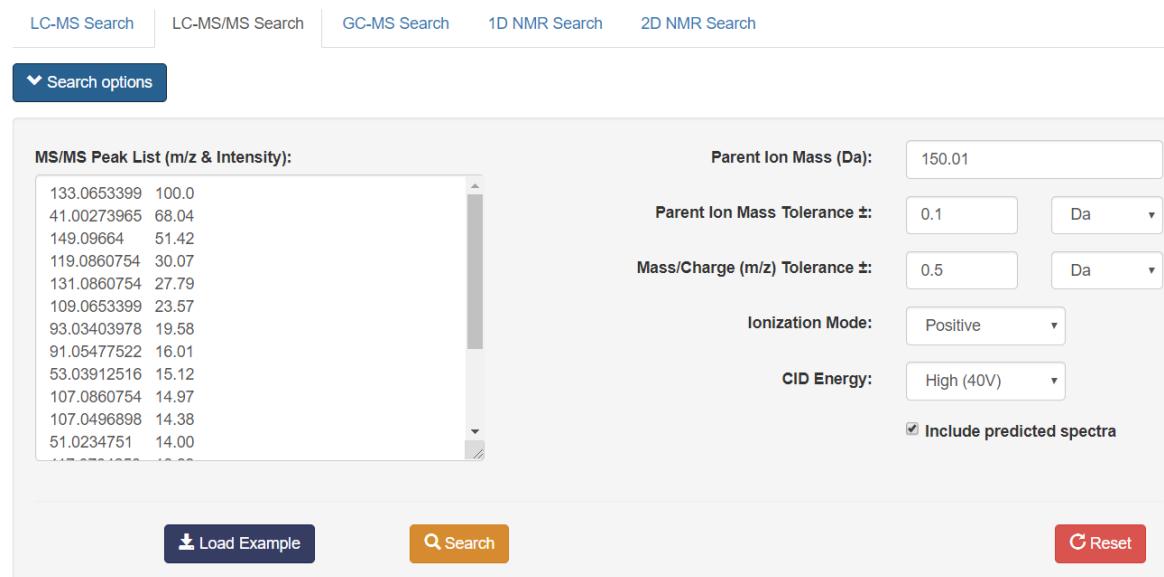
8.3.1. Identyfikacja związku przy pomocy strony HMDB

Do zidentyfikowania związku poprzez porównanie widm, HMDB udostępnia panel wyszukiwania zaawansowanego, który przyjmuje wyniki spektrum i na ich podstawie wyszukuje związek (rysunek 57). Niestety funkcjonalność ta nie pozwala na większe doprecyzowanie i zawężenie wyszukiwania, co powoduje, że wyniki są nieprecyzyjne. W podanym przykładzie szukany związek karwonu był dopiero 30 wynikiem na liście, co oznacza, że użytkownik musiałby ręcznie sprawdzić każdy z nich. Taka mała dokładność jest spowodowana tym, że widma posiadają względную intensywność wobec najsilniejszego sygnału, co powoduje podobieństwo pomiędzy widmami niepowiązanych metabolitów.

Drugą możliwością jest wykorzystanie trybu wyszukiwania zaawansowanego (rysunek 58). W tym trybie można wybrać spośród wyselekcyjowanych pól takie, które pomogą w wybraniu właściwego związku. Niestety nie ma tam możliwości filtrowania metabolitów

Tabela 8. Wynik analizy karwonu w spektrometrze mas na ustawieniu 40 elektronowoltów i jonizacji negatywnej posortowane malejąco po intensywności.

m/z	Intensywność
133.0653399	100.0
41.00273965	68.04
149.09664	51.42
119.0860754	30.07
131.0860754	27.79
109.0653399	23.57
93.03403978	19.58
91.05477522	16.01
53.03912516	15.12
107.0860754	14.97
107.0496898	14.38
51.0234751	14.00
117.0704253	13.89
67.05477522	13.15
105.0704253	12.81
103.0547752	12.56
81.03403978	11.27
39.0234751	10.13



Rysunek 57. Panel zaawansowanego wyszukiwania związku przy użyciu widma na stronie HMDB.

po ich przynależności takonomicznej takiej jak podkласa, główna klasa czy nadklasa. Powoduje to, że pomimo większych możliwości są one nadal niewystarczające, ponieważ liczba wyników jest nadal duża (rysunek 59). W podanym przykładzie liczba wyników wyniosła 32.

8. Eksperymentowanie z utworzoną aplikacją

Match all of the conditions below:

Search Conditions:

Monoisotopic Mass	greater than	150.00	Find metabolites with a given monoisotopic mass
Monoisotopic Mass	less than	151.20	Find metabolites with a given monoisotopic mass
Biofluid	matches	Urine	Find metabolites that have been found in a specific biofluid (e.g. blood)

Display fields:

Rysunek 58. Panel wyszukiwania zaawansowanego na stronie HMDB.

Search Results
Displaying matches 1 - 30 of 32 in total

Guanine

Field	Value
Monoisotopic Mass	151.049409807
Biofluid	Cellular Cytoplasm, Feces, Saliva, Urine

2-Hydroxyadenine

Field	Value
Monoisotopic Mass	151.049409807
Biofluid	Urine

Rysunek 59. Wyniki wyszukiwania zaawansowanego.

8.3.2. Identyfikacja związku z wykorzystaniem utworzonego oprogramowania

Do wyszukania w aplikacji metabolitu spełniającego podane wymagania, należy przejść do panelu wyszukiwania zaawansowanego i wypełnić go wszystkimi znanymi informacjami (rysunek 60). W podanym przykładzie zostało wypełnione pole klasy, ponieważ wiemy, że szukany związek należy do klasy lipidów. Zostało wybrane miejsce występowania związku oraz podany zakres masy związku.

Metabolite search

Name	Super class	Main class	Sub class	Biospecimen location	Mass
		lipid		Urine	150.00 - 150.20
<input type="button" value="Search"/> <input type="button" value="Simple search"/>					

Rysunek 60. Panel wyszukiwania zaawansowanego na stronie utworzonej aplikacji.

Po wcisnięciu przycisku “Search” lub klawisza enter pojawiło się 5 wyników (rysunek 61). Taka liczba pozwala już na ręczne porównanie wyniku spektrometru mas z widmami podanymi w bazie. W przypadku karwonu widmo może się zgadzać dla wszystkich jego kombinacji ((+)-(S)-Carvone, (R)-Carvone, Carvone), ponieważ ich widma mogą być bardzo zbliżone do siebie.

(+)-(S)-Carvone
(R)-Carvone
5-Isopropyl-2-methylphenol
Carvone
Thymol

Rysunek 61. Wyniki wyszukiwania zaawansowanego na stronie utworzonej aplikacji.

8.3.3. Porównanie

W trzecim przykładzie została pokazana możliwość przyspieszenia procesu identyfikacji metabolitu. Na stronie bazy danych HMDB, użycie zaawansowanego wyszukiwania po widmie zwróciło ponad 100 wyników, a dopiero 30 był trafny. Mimo że szukany metabolit był pośród podanych wyników, to użytkownik korzystający z tej funkcjonalności nadal musiałby sprawdzić ręcznie każdy wynik po kolej. Proces ten jednak rozwiązuje potrzebę ręcznego przepisywania parametrów rejestracji metabolitów, ponieważ od razu je porównuje. W przypadku korzystania z wyszukiwania zaawansowanego udało się zawęzić listę metabolitów do 32. Jednak w tym przypadku porównanie nadal musi być wykonane ręcznie poprzez przepisywanie wartości parametrów rejestracji lub utworzenie wykresu z badanego widma i wzrokowym porównywaniu z widmami z bazy.

Opisywana w tej pracy aplikacja pozwoliła zawęzić liczbę wyników do 5, co pozwala na o wiele szybsze sprawdzenie wszystkich możliwości. Dodatkowo parametry rejestracji są posortowane i pogrupowane, co ułatwia porównanie ich z wynikiem wykonanej analizy.

9. Zakończenie

9.1. Zrealizowanie celu pracy

Niniejsza praca realizuje dwa postawione we wstępie cele. Pierwszym z nich jest stworzenie aplikacji, która ma służyć analitykom chemicznym i badaczom podczas procesu wyznaczania parametrów rejestracji metabolitów. Drugim z nich jest analiza utraconych cząsteczek neutralnych w widmach ludzkich metabolitów z grupy lipidów i ich pochodnych.

9.2. Aplikacja webowa

Przedstawiona tutaj praca w pierwszej części skupia się na podstawowych czynnościach wykonywanych podczas badań w dziedzinie metabolomiki. Wynikiem niniejszej pracy jest aplikacja webowa, która została napisana w językach python i javascript. Liczy ona łącznie ponad trzy tysiące linii kodu. Jej głównym zadaniem jest wspomaganie pracy analityków chemicznych i badaczy podczas wyznaczania parametrów rejestracji metabolitów. Zostało wykazane, że prezentowane rozwiązanie znacznie przyspiesza przeprowadzenie tego procesu.

Zaprojektowana aplikacja spełnia wszystkie wymagania funkcjonalne przedstawione na początku pracy. Pozwala ona na łatwe i szybkie wyszukiwanie związków, a następnie oferuje w czytelnej formie parametry rejestracji metabolitów, które w prosty sposób można skopiować i wykorzystać w innych narzędziach. Zaprezentowane rozwiązanie spełnia też wymaganie dostępności, które w przypadku podanej architektury jest globalne. Aplikacja jest napisana w języku angielskim, co pozwoli na wykorzystanie jej przez różnych badaczy. Przeprowadzone testy wykazały również, że interfejs jest bardzo intuicyjny i prosty w nawigacji, co zwiększa jego użytkownictwo.

Aplikacja ta wyróżnia się tym, że jest jedynym publicznie dostępnym narzędziem do wyszukiwania i wyznaczania parametrów rejestracji metabolitów w formie łatwej do dalszej pracy z nimi. Rozwiązania takie jak interfejsy webowe baz danych Metlin i HMDB udostępniają możliwość odczytania tych danych, ale nie oferują widoku, w którym są one efektywnie zebrane i podsumowane. Do najważniejszych osiągnięć tej pracy należy zaliczyć jej intuicyjny interfejs i szybkość działania.

Informatyka od zawsze była narzędziem, którego dużą wartością jest wykorzystanie go w innych dziedzinach nauki. Przedstawiony tu program spełnia dokładnie to zadanie, pomagając badaczom poprzez uproszczenie bardzo czasochłonnego i nieproduktywnego procesu, jaki występuje w badaniach metabolomiki.

Opisywana tutaj aplikacja jest interesującym przykładem rozwiązania informatycznego, które pozwala badaczom na znaczące przyspieszenie (aż o 90%) i polepszenie jakości ich pracy. Wraz z pojawiением się masowych baz danych ze zwiększającą się szybko ilością

informacji, którą należy przeanalizować, takie aplikacje będą coraz bardziej potrzebne i coraz bardziej popularne.

9.3. Analiza utraconych cząsteczek neutralnych

Zaprezentowana w tej pracy analiza utraconych cząsteczek neutralnych skupia się na znalezieniu charakterystycznych wartości oraz zależności między nimi, które pozwolą na następne dokładniejsze ich przebadanie. Taka analiza ma za zadanie wspierać analityków chemicznych i badaczy podczas podejmowania decyzji, na jakich związkach powinni się skupić i jakie związki mogą dać jak najlepsze wyniki. Jej wynikiem są reguły, które opisują zależności pomiędzy często występującymi wspólnie utraconymi cząsteczkami neutralnymi.

Analiza została przeprowadzona z wykorzystaniem reguł asocjacyjnych, które w prosty sposób wizualizują zależności występujące pomiędzy związkami. Dzięki tej metodzie możliwe jest w dość łatwy sposób znalezienie nowych, wartych dalszego zbadania utraconych cząsteczek neutralnych. Główną zaletą jest tu niski nakład środków i zasobów potrzebnych, by przeprowadzić cały proces. Zbadanie w ten sposób związków z grupy lipidów i ich pochodnych pozwoliło na znalezienie reguł w wystąpieniach cząsteczek neutralnych oraz pomogło w wytypowaniu reguł i poszczególnych ścieżek rozpadu.

Opisane rozwiązanie jest dobrym przykładem wykorzystania metod eksploracji danych w dziedzinach nauki, które produkują bardzo duże zbiory danych. Coraz większy dostęp do publicznych baz danych powoduje, że jest coraz więcej informacji, z których można korzystać podczas badań, ale niestety kosztem ich jakości. Dane często są nieznormalizowane lub po prostu błędne. Wykorzystując metody eksploracji danych, można nie tylko przeanalizować cały zbiór, ale też odrzucić zakłócające dane.

9.4. Dalsze kierunki badań

Przestawiona metoda jest jednak tylko początkiem badań, które należałyby wykonać, aby całkowicie przeanalizować utracone cząsteczki neutralne na podstawie widm. Po znalezieniu charakterystycznych zależności dobrym kolejnym krokiem byłoby zbadanie analizowanych związków pod kątem tych cząsteczek.

Zaproponowany algorytm jest uniwersalny, zatem może on też być wykorzystany dla kolejnych podgrup ludzkich metabolitów. Trzeba jednak pamiętać, że przy metodach eksploracji danych liczba reprezentantów grupy musi być wystarczająco duża, aby wyniki były wiarygodne.

Należy też zwrócić uwagę, że architektura opracowanego tu programu, dzięki frameworkowi django, jest dobrze przystosowana na dodanie nowych modułów do już istniejącego rozwiązania. Jednak najważniejszym elementem dalszych prac jest poprawa jakości i atrakcyjności interfejsu. Bardziej rozbudowany interfejs zachęci większą liczbę

9. Zakończenie

użytkowników i spowoduje, że korzystanie z aplikacji będzie znacznie ułatwione, co może być bardzo istotne dla dalszego rozwoju metabolomiki.

Aplikacja może też być rozbudowana o moduł automatycznej weryfikacji i porównania stworzonego przez spektrometr widma z danymi zaczerpniętymi z bazy. Taka funkcjonalność mogłaby wyeliminować potrzebę wykonywania tego ręcznie, a nawet mogłaby być podłączona do odpowiedniego interfejsu spektrometru mas. W takiej konfiguracji spektrometr po wykonaniu pomiaru automatycznie wysyłałby otrzymane rezultaty do porównania. Wyniki, opisane odpowiednimi miarami, byłyby zapisywane i prezentowane dalej w aplikacji.

Bibliografia

- [1] Y. Smith, *History of Metabolomics*, Dostęp zdalny (03.01.2020): <https://www.news-medical.net/life-sciences/History-of-Metabolomics.aspx>, 2019.
- [2] *Metlin - oficjalna strona*, Dostęp zdalny (29.01.2020): https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage.
- [3] F. Bergamin, „Real-time analysis of metabolic products”, Dostęp zdalny (29.01.2020): <https://phys.org/news/2015-09-real-time-analysis-metabolic-products.html>, 2015.
- [4] D. Wishart, D. Tzur, C. Knox i in., „HMDB: the Human Metabolome Database”, *Nucleic Acids Res.*, t. 35, nr. D521-6, 2007.
- [5] D. Wishart, C. Knox, A. Guo i in., „HMDB: a knowledgebase for the human metabolome”, *Nucleic Acids Res.*, t. 37, nr. D603-610, 2009.
- [6] D. Wishart, T. Jewison, A. Guo, M. Wilson, C. Knox i in., „HMDB 3.0 — The Human Metabolome Database in 2013”, *Nucleic Acids Res.*, t. 41, nr. D801-7, 2013.
- [7] D. Wishart, Y. Feunang, A. Marcu, A. Guo, K. Liang i in., „HMDB 4.0 — The Human Metabolome Database for 2018”, *Nucleic Acids Res.*, t. 46, nr. D608-17, 2008.
- [8] *DB-Engines Ranking*, Dostęp zdalny (11.01.2020): <https://db-engines.com/en/ranking>.
- [9] *Oficjalna strona MongoDB*, Dostęp zdalny (10.01.2020): <https://www.mongodb.com/>.
- [10] *Django MongoDB connector*, Dostęp zdalny (15.01.2020): <https://nesdis.github.io/djongo/>.
- [11] *Gunicorn - Python WSGI HTTP Server*, Dostęp zdalny (15.01.2020): <https://gunicorn.org/>.
- [12] *NGINX*, Dostęp zdalny (15.01.2020): <https://www.nginx.com/>.
- [13] *MongoDB Compass*, Dostęp zdalny (29.01.2020): <https://www.mongodb.com/products/compass>.
- [14] R. Agrawal, T. Imielinski i A. Swami, „Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference”, w. czer. 1993, t. 22, s. 207–. DOI: 10.1145/170036.170072.
- [15] *Strona projektu Efficient-Apriori*, Dostęp zdalny (19.05.2020): <https://github.com/tommyod/Efficient-Apriori>.
- [16] Y. Ma, T. Kind, D. Yang, C. Leon i O. Fiehn, „MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra”, *Analytical Chemistry*, t. 86, nr. 21, s. 10724–10731, 2014, PMID: 25263576. DOI: 10.1021/ac502818e. eprint: <https://doi.org/10.1021/ac502818e>. adr.: <https://doi.org/10.1021/ac502818e>.

Wykaz symboli i skrótów

1. **EiT**I – Wydział Elektroniki i Technik Informacyjnych
2. **PW** – Politechnika Warszawska
3. **HMDB** – Human Metabolome Database
4. **METLIN** – Metabolite and Chemical Entity Database

Spis rysunków

1. Przykład widma masowego.	12
2. Diagram architektury aplikacji.	15
3. Początkowy układ strony.	29
4. Podstawowy tryb wyszukiwania.	29
5. Zaawansowany tryb wyszukiwania.	29
6. Lista miejsc występowania związków.	30
7. Przykładowa tabela ze znalezionymi metabolitami.	30
8. Wygląd wiersza po najechaniu na niego.	31
9. Wygląd wcisniętego wiersza.	31
10. Końcowy układ strony.	31
11. Sekcja dotycząca parametrów rejestracji metabolitu.	32
12. Link do zewnętrznej bazy danych HMDB.	32
13. Tabela parametrów rejestracji metabolitów dla jednego trybu jonizacji.	33
14. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 10 eV.	44
15. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 20 eV.	45
16. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o negatywnej jonizacji i energii kolizji 40 eV.	46
17. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 10 eV.	46
18. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 20 eV.	47
19. Liczebność utraconych cząsteczek obojętnych dla wszystkich widm o pozytywnej jonizacji i energii kolizji 40 eV.	47
20. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 10 eV.	49
21. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 20 eV.	49
22. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o negatywnej jonizacji i energii kolizji 40 eV.	50

23. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 10 eV	51
24. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 20 eV	51
25. Liczebność utraconych cząsteczek obojętnych dla widm z grupy lipidów i ich pochodnych o pozytywnej jonizacji i energii kolizji 40 eV	52
26. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 10 eV	54
27. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 10 eV z wykluczeniem cząsteczek 18 i 28.	54
28. Wykres punktowy reguł asocjacyjnych utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV	55
29. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV z wykluczeniem cząsteczek 18 i 28.	56
30. Wykres punktowy reguł asocjacyjnych utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV	56
31. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z większej cząsteczki do mniejszej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV z wykluczeniem cząsteczek 18 i 28.	57
32. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 10 eV.	57
33. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 20 eV.	59
34. Wykres punktowy reguł asocjacyjnych dla utraconych cząsteczek obojętnych w kierunku z mniejszej cząsteczki do większej z grupy lipidów i ich pochodnych w widmach w trybie jonizacji pozytywnej i energii kolizji 40 eV.	60
35. Strona startowa HMDB.	62
36. Wyszukanie kwasu dikowoilochinowego po fragmencie nazwy.	63
37. Związki znalezione po wyszukianiu fragmentu nazwy kwasu dikowoilochinowego.	63
38. Wyszukanie związku z wykorzystaniem pełnej nazwy kwasu dikowoilochinowego.	63
39. Strona HMDB kwasu dikowoilochinowego.	64
40. Sekcja zawierająca widma kwasu dikowoilochinowego.	64
41. Widmo kwasu dikowoilochinowego o energii kolizji równej 10 elektronowoltom.	64
42. Widmo kwasu dikowoilochinowego o energii kolizji równej 20 elektronowoltom.	65
43. Wyszukanie kwasu dikowoilochinowego w aplikacji używając części pełnej nazwy.	65
44. Strona aplikacji po wybraniu kwasu dikowoilochinowego.	66
45. Szukane w przykładzie 1 parametry rejestracji kwasu dikowoilochinowego.	66

46. Panel wyboru zaawansowanych opcji wyszukiwania na stronie bazy HMDB.	68
47. Pole do wpisania zapytania tekstowego na stronie bazy HMDB.	68
48. Wyniki wyszukiwania tekstowego na stronie bazy HMDB.	69
49. Strona HMDB 1-Metylohistaminy.	69
50. Sekcja zawierająca widma 1-Metylohistaminy.	69
51. Szukane widmo, z którego zostaną wyznaczone parametry rejestracji 1-Metylohistaminy.	69
52. Widmo 1-Metylohistaminy o jonizacji negatywnej i energii kolizji równej 40 elektronowoltów.	70
53. Panel wyszukiwania zaawansowanego w aplikacji.	70
54. Wyniki wyszukiwania w aplikacji.	71
55. Sekcja z polem do filtrowania intensywności wybranego związku.	71
56. Szukane parametry rejestracji 1-Metylohistaminy.	71
57. Panel zaawansowanego wyszukiwania związku przy użyciu widma na stronie HMDB. . .	73
58. Panel wyszukiwania zaawansowanego na stronie HMDB.	74
59. Wyniki wyszukiwania zaawansowanego.	74
60. Panel wyszukiwania zaawansowanego na stronie utworzonej aplikacji.	74
61. Wyniki wyszukiwania zaawansowanego na stronie utworzonej aplikacji.	75

Spis tabel

1. Wykorzystane pola metabolitów.	20
2. Wykorzystane pola widm.	21
3. Kolekcje administracyjne.	22
4. Struktura kolekcji hmdb_met_names.	22
5. Struktura dokumentów zawierających dane o utraconych cząsteczkach obojętnych w widmie.	40
6. 20 najczęściej występujących odległości dla każdej kategorii.	43
7. 20 najczęściej występujących odległości dla grupy lipidów i ich pochodnych.	48
8. Wynik analizy karwonu w spektrometrze mas na ustawieniu 40 elektronowoltów i jonizacji negatywnej posortowane malejąco po intensywności.	73