Analiza interakcji związków farmakologicznych na podstawie bazy danych **Drugbank**

Marlena Osipowicz, Stanisław Massalski

15 stycznia 2019

1 Wstęp

Projektowanie leków to dynamicznie rozwijająca się dziedzina wiedzy. Liczba leków stale rośnie. Analiza sieci powiązań pomiędzy istniejącymi lekami może nietylko pomóc w tworzeniu nowych substancji leczniczych, ale również w wykorzystaniu już istniejących w innym celu. W poniższym raporcie został przedstawiony proces analizy bazy danych zawierającej informacje zarówno o aktualnych lekach, jak i o tych wycofanych oraz eksperymentalnych. Na podstawie bazy danych stworzono sieć interakcji, którą badano metodami biologii systemów.

2 Materiały

W projekcie korzystano z bazy **Drugbank** (*www.drugbank.ca*) zawierającej informacje o lekach i związkach z którymi oddziałują (ich targetach). Baza ta pozwala stworzyć sieć powiązań typu lek-cel, ale również dostarcza wiele informacji, między innymi na temat struktury leku (atrybut *classification* informujący o typie związku chemicznego) jak również o dacie uzyskania patentu. Inną pomocną informacją jest ta mówiąca o statusie leku, czyli o tym, czy dany lek został zatwierdzony, wycofany, zabroniony bądź czy jest lekiem eksperymentalnym. Dostpęna jest również baza targetów leków - ich lokalizacji komórkowej bądź procesów w jakich uczestniczą.
Do analizy informacji z bazy **Drugbank** korzystano głównie z języka programowania Python oraz z programu OpenBabel (wyliczanie podobieństwa pomiędzy cząsteczkami). Biblioteki wykorzystane na różnych etapach analizy to: `xml` (parsowanie pliku xml), `Networkx` (obliczanie parametrów sieci), `numpy`, `matplotlib`, `seaborn` (tworzenie wykresów), .

3 Metody

3.1 Parsowanie danych pobranych z Drugbank

Parsowanie danych z pliku typu xml nie było zadaniem oczywistym, gdyż wielkość bazy **Drugbank** przekroczyła możliwości używanych komputerów.

W celu wczytania dannych z pliku xml do obiektów łatwych do analizy w Pythonie wykorzystano skrypt udostępniony na stronie *github.com/dhimmel/drugbank/blob/gh-pages/parse.ipynb*. Dokonano jednak kilku jego modyfikacji. Najważniejszą z nich była zmiana funkcji `parse` z pakietu `xml.etree.ElementTree` na funkcję `iterparse`, która pozwala na wczytywanie pliku fragmentami i usuwanie niepotzebnych danych na bieżąco. W ten sposób rozwiązano problem wielkości pliku xml, który okazał się zbyt duży by móc zostać w szybkim czasie wczytany pierwotnie wykorzystaną metodą. Dodatkową zmianą był zapis leków oraz ich targetów do obiektów typu Drug i Target, posiadających atrybuty potrzebne w dalszej analizie. Obiektami wyjściowymi procesu parsowania były dwa słowniki: jeden zawierający obiekty Drug przypisane do klucza będącego ich ID, drugi zawierający obiekty Target przypisane do klucza będącego ich numerem z bazy Uniprot. Skrypt wykorzystany w tej części zadania znajduje się w */skrypty/parse.py*.

3.2 Tworzenie sieci interakcji

Sieć interakcji tworzona jest na podstawie obiektów wczytanych z pliku xml (leki z listą krawędzi, targety) o interesujących własnościach bądź atrybutach. W przypadku analizy tylko podgrupy węzłów bądź celów dodatkowo filtruje się zbiór leków, na podstawie którego przeprowadzana jest dalsza analiza. W celu wyliczenia podstawowych parametrów sieci, za pomocą biblioteki `networkx` tworzony jest graf. Skrypt wykorzystany do tworzenia sieci oraz wyliczania jej parametrów można znaleźć w pliku */skrypty/network_ analysis.py*.

3.3 Klasyfikacja leków

W bazie **Drugbank** do każdego leku przypisano informacje na temat jego taksonomii - klasyfikacji opartej na rodzaju związku chemicznego. Na podstawie tej informacji każdy z leków został przypisany do klasy związków (ustalonej na podstawie atrybutu *superclass* w bazie). Następnie zbadano cechy każdej z podgrup leków, należących do tej samej klasy. Sprawdzono czy któraś z klas preferuje jakiś rodzaj celów (target, enzym, transporter, nośnik) oraz jaki jest średni stopień węzłów dla każdej z nich (informacja ta mówi o poziomie specyficzności leków z danej klasy w porównaniu do średniej dla całej sieci). Skrypt wykorzystany w tej części analizy znajduje się w */skrypty/class.py*.

3.4 Wspólne cele leków, wspólne leki celów

Ciekawym zagadnieniem wartym zbadania są wspólne cele różnych leków, jak również wspólne leki dla różnych celów. Rozważono dwie sieci zmodyfikowane pod kątem połączeń między węzłami.

W pierwszej z nich, oznaczanej jako *drug-interactions*, krawędzie pomiędzy węzłami występowały tylko pomiędzy lekami, posiadającymi wspólny cel. Dodatkowo do każdej krawędzi przypisano wagi. Im więcej wspólnych celów występuje pomiędzy dwoma danymi lekami, tym waga łączącej je krawędź jest wyższa.

W drugiej siec, oznaczanej jako *target-interactions*, krawędzie pomiędzy węzłami dodano tylko pomiędzy celami, które oddziałują z tym samym lekiem. Tutaj również do krawędzi przypisano wagi, zależne od ilości wspólnych leków łączących daną parę celów (im więcej wspólnych leków tym wyższa waga krawędzi).

Dla obu sieci zbudowanych w opisany powyżej sposób obliczono podstawowe parametry. Celem było wyróżnienie cech wspólnych oraz specyficznych dla każdej z nich. Skrypt wykorzystany w tej części projektu można znaleźć w pliku */skrypty/interactions.py*.

3.5 Podobieństwo leków

W bazie **Drugbank** do każdego leku przypisany jest zapis jego struktury chemicznej w formacie `smiles`. Na tej podstawie można policzyć podobieństwo poszczególnych cząsteczek leków do siebie nawzajem. Umożliwia to przeprowadzenie ciekawej analizy starającej się odpowiedzieć na pytanie, czy podobne leki oddziałują na te same cele biologiczne. W tym celu wykorzystano podobieństwo FP2, liczone za pomocą programu `openbabel`. Skrypt wykorzystany w tym celu znajduje się w */skrypty/similarity.py*.

3.6 Leki wielocelowe

Leki wielocelowe (*multitarget drugs*) to leki oddziałujące z więcej niż jednym celem. Analiza tego typu węzłów może dostarczyć ciekawych informacji na temat procesów na które oddziałują. Kolejnym nasuwającym się pytaniem jest to, czy rok zatwierdzenia leku ma związek z jego wielocelowością (czy leki "nowocześniejsze" są bardziej specyficzne czy bardziej uniwersalne?). Skrypt wykorzystany w tej części projektu można znaleźć w pliku */skrypty/multitarget_ drugs.py*.

3.7 Analiza lokalizacji komórkowej celów leków

Targety leków są zlokalizowane w różnych częściach komórki, bądź poza nią. Obecności informacji na ten temat w bazie **Drugbank** umożliwiła analizę zależności lokalizacji celów w zależności od parametrów leku. Skupiono się na dacie patentu leku, na tym czy istnieje jakaś zależność pomiędzy czasem wprowadzenia preparatu na rynek a miejscem jego działania w komórce. W tym celu przeanalizowano stworzoną sieć lek-cel. Pierwszym krokiem był podział komórki na interesujące podgrupy, tj. plasma membrane, cytosol, extracellular space, mitochondrion, nucleus, unknown/other. Następnie do każdej z tych grup przypisano odpowiednie cele. Rozważając leki wprowadzone na rynek w poszczególnych latach zliczono ich cele zlokalizowane w różnych miejscach w komórce. Skrypt wykorzystany w tej części zadania znajduje się w pliku */skrypty/date_ cell.py*.

4 Wyniki i dyskusja

4.1 Podstawowe parametry sieci lek-cel

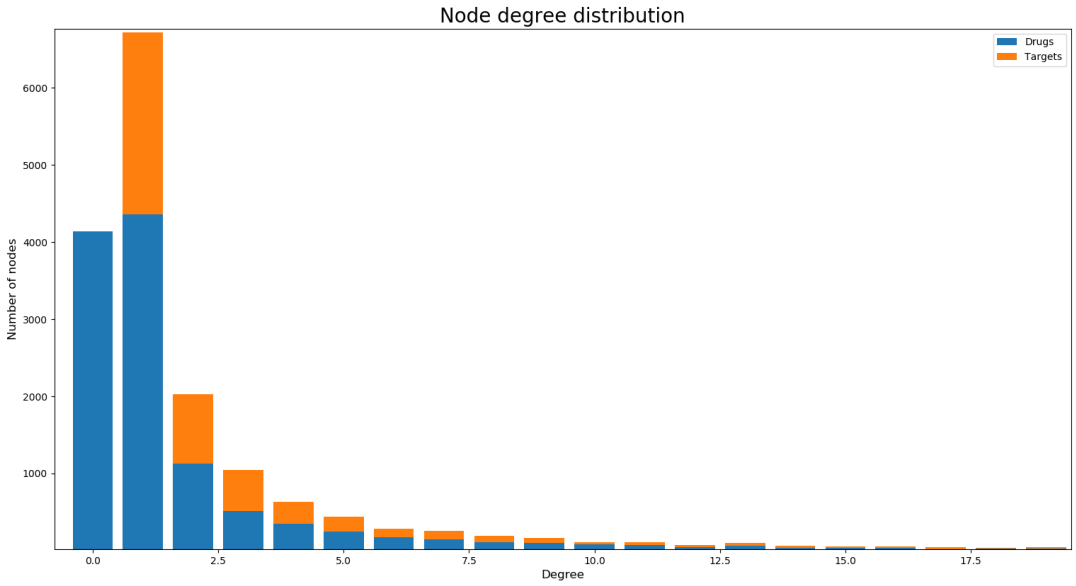
Sieć interakcji stworzona na podstawie bazy Drugbank składa się z 16903 wierzchołków (na które składa się 11922 leków oraz 4981 celów na które leki oddziałują) oraz z 28401 krawędzi reprezentujących oddziaływania pomiędzy związkami. Podstawowe statystyki sieci zostały przedstawione w tabeli 1.

Tabela 1: Podstawowe parametry sieci lek-cel dla wszystkich węzłów.

Liczba wszystkich węzłów:	16903
Liczba leków:	11922
Liczba celów:	4981
Izolowane węzły:	4137
Średni stopień węzłów:	3.44
Średni stopień leków:	2.38
Średni stopień celów:	6.21
"Betweeness centrality":	0.0003
Liczba węzłów w GC:	11466
Średnia długość ścieżki:	5.88
Współczynnik klasteryzacji:	0.0
Gęstość sieci:	0.00034
Entropia sieci:	1.898

Rozważając tylko węzły reprezentujące leki można zauważyć, że ich średni stopień jest niższy niż dla całej sieci. Wydaje się więc, że leki są średnio dosyć specyficzne. Potwierdza to ilość leków powiązanych z jednym tylko celem - jest ich 4365. Leków o stopniu mniejszym bądź równym 3 jest aż 10140 - stanowi to aż 85% wszystkich z nich.

Wykres 1 to histogram przedstawiający rozkład stopni węzłów w stworzonej sieci. Kolorami zaznaczono odpowiednio węzły reprezentujące leki i cele. Wyraźnie widać, że im większy stopień, tym mniejsza ilość węzłów. Liczba targetów o danym stopniu wydaje się zmniejszać nieco wolniej od leków. Potwierdza to wartość średniego stopnia węzła dla celów w porównaniu do leków (porównaj tabelę 1).



Wykres 1: Rozkład stopni węzłów dla sieci, kolorami zaznaczono jaka część węzłów o danym stopniu jest lekami, a jaka celami.

4.2 Sieci dla subtypów celów

Cele leków można podzielić na podgrupy zależnie od funkcji jaką pełnią względem cząsteczki leku. W bazie **Drugbank** można wyróżnić cztery podgrupy: cele (*targets*), nośniki (*carriers*), transportery (*transporters*) oraz enzymy (*enzymes*). Na tej podstawie stworzono różne podsieci oparte na wybranych grupach leków, które oddziałują z celami z danych podgrup. Wyniki podstawowej analizy tego typu sieci zostały przedstawione w tabeli 2.

Pierwszą cechą rzucającą się w oczy jest wyższy średni stopień węzłów dla podsieci transporterów oraz enzymów, oraz niższy średni stopień węzłów dla nośników. Jeszcze wyraźniejszą różnicę widać w średnim stopniu celów - można z tego wywnioskować, że cząsteczki oddziałujące z lekami jako transportery i enzymy są średnio mniej specyficzne, w odróżnieniu od nośników, które wykazują przeciwną tendencję.

Tabela 2: Podstawowe parametry sieci lek-cel dla poszczególnych podtypów celów.

Statystyka	**Cała sieć**	**Carriers**	**Transporters**	**Enzymes**
Liczba wszystkich węzłów	16903	490	1032	1973
Liczba leków	11922	404	832	1600
Liczba celów	4981	86	200	373
Izolowane węzły	4137	0	0	0
Średni stopień węzłów	3.44	2.39	4.99	4.98
Średni stopień dla leków	2.38	1.45	3.09	3.07
Średni stopień dla celów	6.21	6.82	12.87	13.17
"Betweeness centrality"	0.0003	0.004	0.002	0.001
Liczba węzłów w GC	11466	447	921	1817
Średnia długość ścieżki	5.88	3.50	3.63	3.68
Współczynnik klasteryzacji	0.0	0.0	0.0	0.0
Gęstość sieci	0.00034	0.0049	0.0048	0.002
Entropia sieci	1.898	1.00	2.05	1.94

4.3 Sieci dla subtypów leków

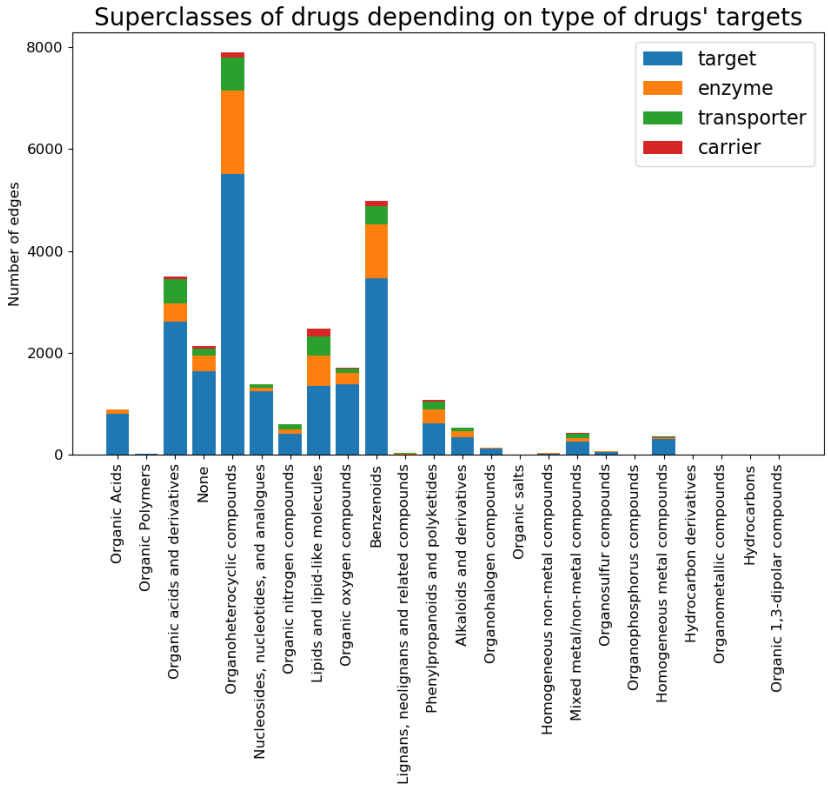
Leki można podzielić na podgrupy ze względu na ich stan prawny, tj. czy zostały dopuszczone do użycia (*Approved*), są na etapie eksperymentalnym (*Experimental*), zostały zakazane (*Illicit*), wycofane (*Withdrawn*) bądź też rozpoczął się już proces zmierzający do ich zatwierdzenia (*Investigational*). Na podstawie każdej z tych podgrup stworzono sieć, których podstawowe parametry przedstawiono w tabeli 3.

Tabela 3: Podstawowe parametry sieci lek-cel dla poszczególnych podtypów leków.

Statystyka	**Cała sieć**	**Experim.**	**Appr.**	**Invest.**	**Illicit**	**Withdrawn**
Liczba wszystkich węz.	16903	8394	6690	6604	347	606
Liczba leków	11922	5764	3729	3920	205	252
Liczba celów	4981	2630	2961	2684	142	354
Izolowane	4137	1107	1447	1926	104	79
Średni stopień węzłów	3.44	2.12	5.27	3.40	5.62	3.12
Średni stopień leków	2.38	1.54	4.72	2.87	4.76	3.75
Średni stopień celów	6.21	3.38	5.95	4.19	6.87	2.67
"Betweeness centr."	0.0003	0.0005	0.0001	0.0007	0.0125	0.0049
Liczba węzłów w GC	11466	5731	5101	4328	240	420
Średnia dł. ścieżki	5.88	7.60	4.82	5.16	4.10	5.15
Współczynnik klast.	0.0	0.0	0.0	0.0	0.0	0.0
Gęstość sieci	0.00034	0.0003	0.0012	0.001	0.03	0.0065
Entropia sieci	1.898	1.29	2.42	2.02	2.64	1.91

4.4 Klasyfikacja leków

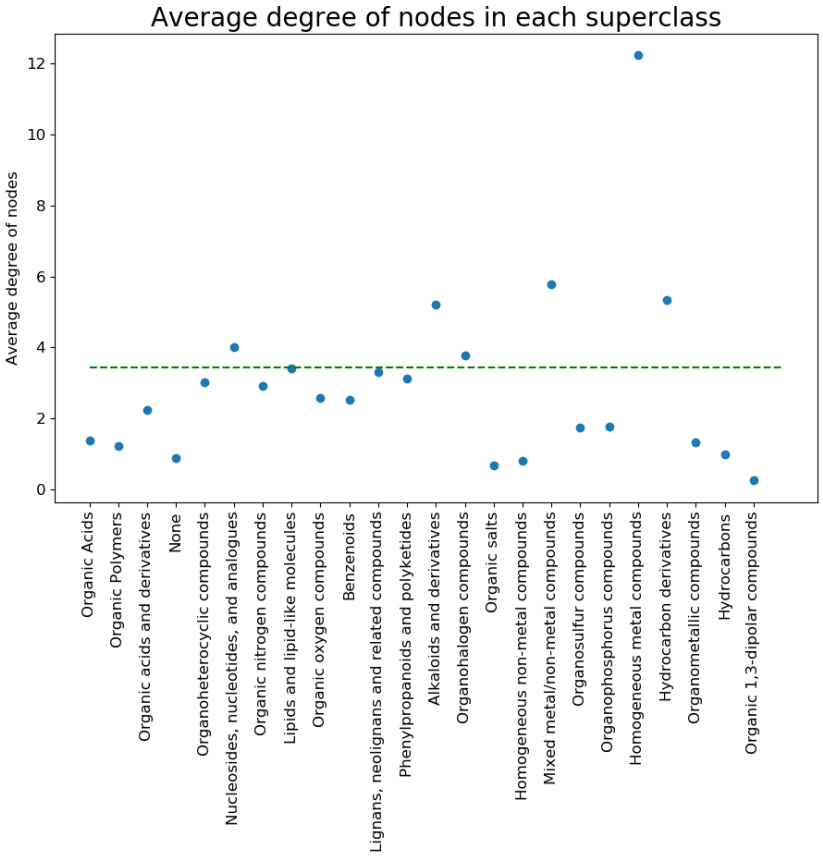
Pierwszym krokiem w analizie leków pod kątem ich klasyfikacji była kontrola, czy któraś z klas jest preferowana przez jakiś subtyp targetów (zobacz wykres 2).



Wykres 2: Wykres przedstawia ilość powiązań z celami różnego typu (w zależności od koloru słupka) dla leków należących do różnych klas (każdy słupek to osobna klasa).

Na podstawie stworzonego histogramu ciężko wysnuć jednoznaczne wnioski na temat poszczególnych klas leków. Najwięcej leków należy do klasy organicznych związków heterocyklicznych oraz do benzenoidów (czyli wielopierścieniowych węglowodorów aromatycznych). Leki z tych dwóch klas względnie często (w porównaniu do innych klas) oddziałują z enzymami.

Następnie zbadano średni stopień węzłów w każdej z grup (zobacz wykres 3). Na podstawie wykresu można stwierdzić, że leki sklasyfikowane jako jednorodne związki metali (*homogeneous metal compounds*) są znacznie bardziej promiskuitywne niż średnia dla całej sieci. Z kolei najbardziej specyficzną grupą leków są organiczne 1,3-dipole.



Wykres 3: Wykres przedstawia średni stopień węzłów z każdej podgrupy (w zależności od klasy do której należą). Przerywana linia symbolizuje średni stopień węzłów dla sieci zbudowanej na podstawie wszystkich węzłów.

4.5 Huby sieci

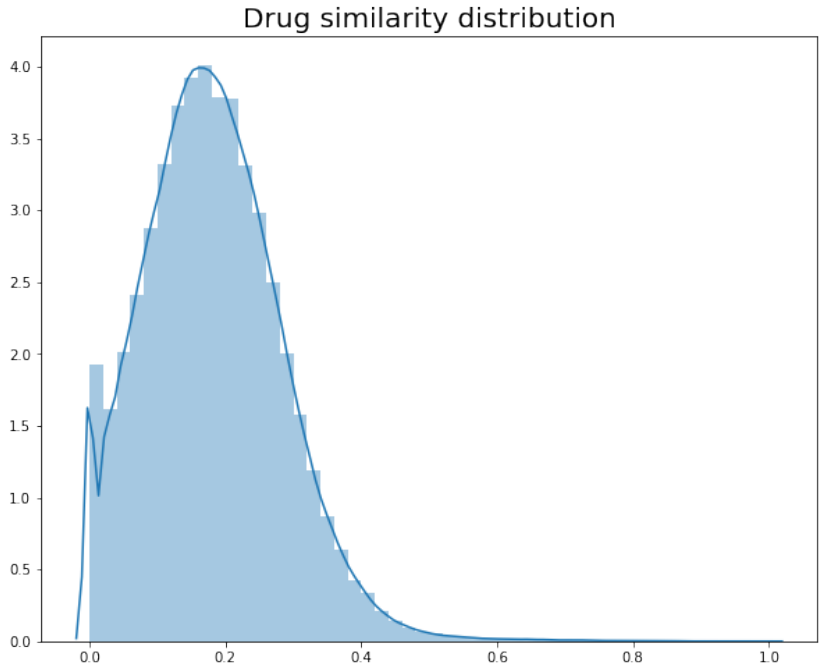
Huby to takie węzły sieci, których stopień jest równy bądź wyższy od stopnia 95% wszystkich węzłów. W rozważanej sieci lek-cel takich węzłów jest 855 (zobacz tabelę 4). Z informacji zawartej w tabeli widać, jak bardzo niski jest średni stopień węzłów w rozważanej sieci - wystarczy, że węzeł jest stopnia 13 bądź wyżej, aby został uznany za huba. Przypuszczenie to potwierdza rozkład stopni węzłów z wykresu 1. Największym hubem jest *Cytochrom P450 3A4*, będący bardzo powszechnie występującym białkiem związanym z procesami metabolicznymi. Jego powszechność w organiźmie człowieka z pewnością wpływa na wysoki stopień węzłą go reprezentującego.

Tabela 4: Parametry hubów sieci.

Liczba wszystkich hubów:	855
Liczba leków-hubów:	462
Liczba celów-hubów:	393
Maksymalny stopień huba:	919
Nazwa największego huba:	*Cytochrome P450 3A4*
Minimalny stopień huba:	13

4.6 Podobieństwo leków

Pierwszym krokiem badania podobieństwa leków było wyliczenie podobieństwa typu FP2 dla każdej pary leków. Rozkład uzyskanych wartości przedstawiono na wykresie 4. W następnym kroku z wszystkich par leków wybrano te, pomiędzy którymi wskaźnik podobieństwa jest wyższy niż 0.5. Uzyskano 6182 pary leków, z których każdy ma średnio 66 % wspólnych targetów ze swoją parą. Jest to wynik dużo wyższy niż można byłoby oczekiwać po losowo wybranych węzłach. Na tej podstawie można stwierdzić, że podobne leki oddziaływują na podobne cele biologiczne.



Wykres 4: Histogram wartości podobieństw dla każdej pary leków w sieci.

4.7 Wspólne cele leków, wspólne leki celów

Na podstawie bazy **Drugbank** zbudowano sieci, w której krawędzie występują jedynie pomiędzy lekami posiadającymi wspólny cel bądź pomiędzy celami oddziałującymi z tym samym lekiem. Waga krawędzi zależy od liczby wspólnych celów (zobacz sekcję 3.4). Parametry otrzymanych sieci przedstawiono w tabeli 3 oraz 5.

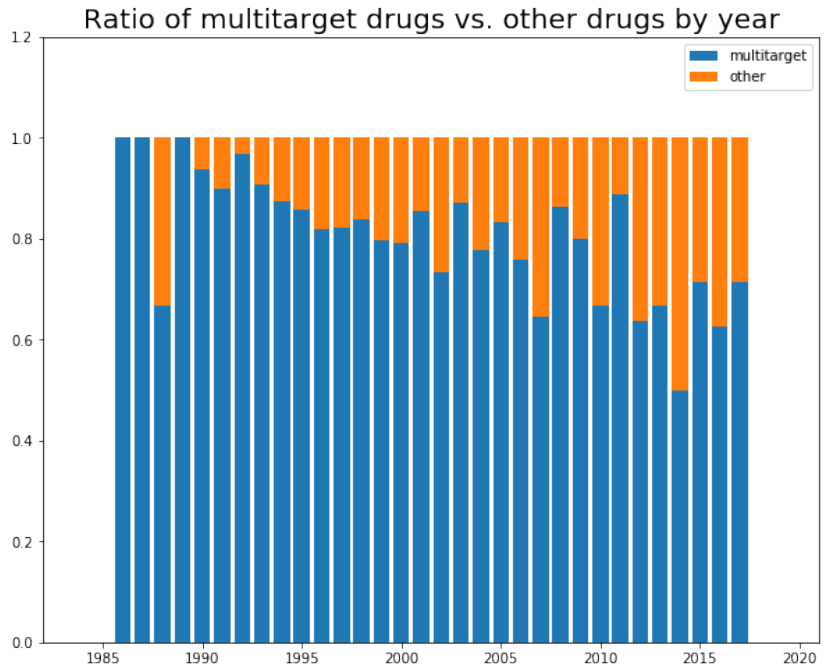
Tabela 5: Podstawowe parametry sieci lek-cel dla sieci zmodyfikowanej pod względem połączeń pomiędzy targetami.

Statystyka	**Zwykła sieć**	**Drug-interact.**	**Target-interact.**
Liczba wszystkich węzłów	16903	7446	4593
Liczba leków	11922	7446	0
Liczba celów	4981	0	4593
Liczba krawędzi	28401	750630	181778
Średni stopień węzłów	3.44	79.15	201.62
Liczba hubów	855	373	382
Minimalny stopień huba	13	1138	308
Maksymalny stopień huba	919	1671	1251
Nazwa maks. huba	*Cytochrome P450*	*Cytochrome P450*	*Amitriptyline*
"Betweeness centrality"	0.0003	0.0002	0.0004
Liczba węzłów w GC	11466	6950	4416
Średnia długość ścieżki	5.88	2.95	2.9
Współczynnik klasteryzacji	0.0	0.83	0.78
Gęstość sieci	0.00034	0.027	0.017
Entropia sieci	1.898	5.178	4.715

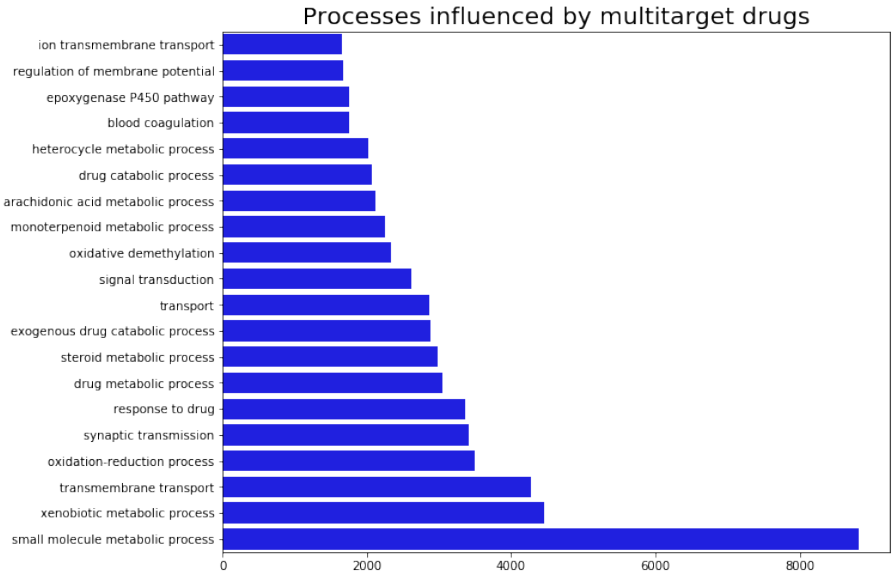
4.8 Leki wielocelowe

Ze wszystkich leków zawartych w bazie **Drugbank** wybrano te, które oddziałują na więcej niż jeden target. Na wykresie 5 przedstawiono histogram zliczający liczbę wielocelowych leków wprowadzonych na rynek w poszczególnych latach, znormalizowany co do ogólnej liczby leków z danego roku. Widoczny jest wyraźny trend zmniejszania ilości leków wielocelowych w czasie. Najnowsze leki, cechują się większą specyficznością w porównaniu do leków starszych. Prawdopodobnie dzięki temu powodują mniej efektów ubocznych, wpływają tylko na procesy wymagające leczenia.

Na wykresie 6 przedstawiono 20 najczęstszych procesów biologicznych z którymi związane są targety leków wielocelowych. Widać, że zdecydowanie najczęstszym procesem tego typu są metaboliczne procesy małych cząsteczek.



Wykres 5: Wykres przedstawia jaką część wszystkich leków wprowadzonych w danym roku stanowią leki wielocelowe.

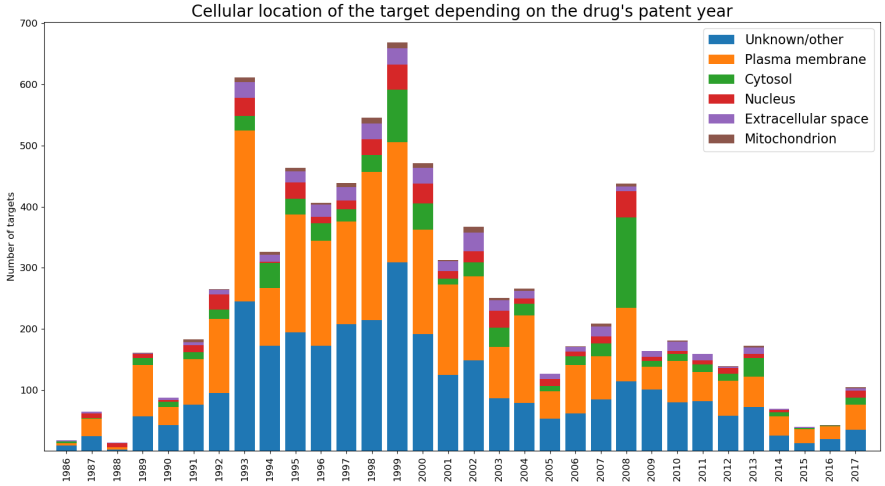


Wykres 6: Wykres przedstawia 20 najczęstszych procesów biologicznych z którymi związane są targety leków wielocelowych.

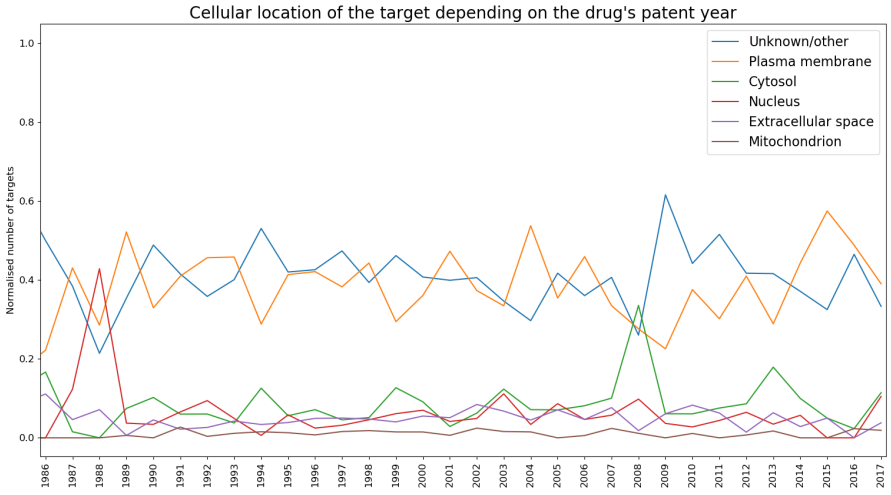
4.9 Analiza lokalizacji komórkowej celów leków

Na pierwszy rzut oka ciężko dopatrzyć się wyraźnych trendów związanych z lokalizacją komórkową celów leków oraz daty wprowadzenia leku na rynek. Na wykresie 7 przedstawiono jaka część białek (wysokość słupków - oś Y) będących targetami leków opatentowanych w danym roku (oś X) jest zlokalizowana w poszczególnych miejscach w komórce (kolory słupków). Tą samą informację, tylko znormalizowaną pod wzlędem liczby targetów w każdym roku, zawierają: wykres 8 (wersja liniowa) oraz wykres 9.

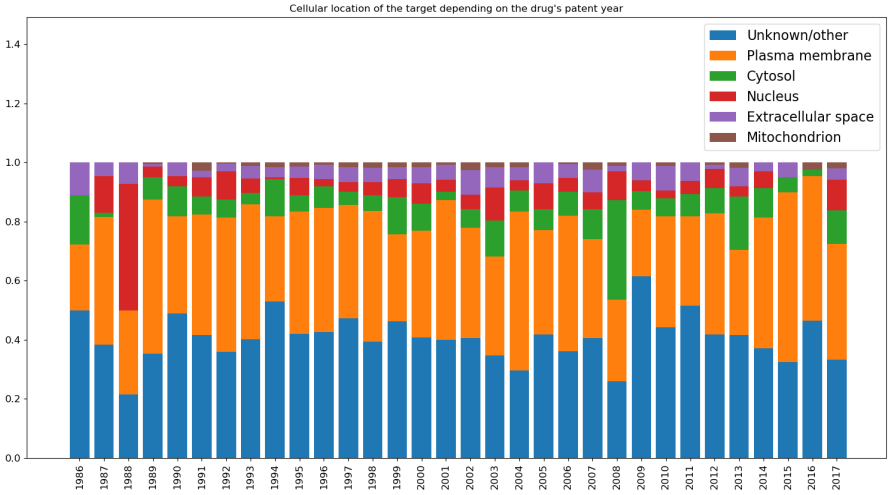
Na wykresie 8 można zauważyć nieznaczny wzrost w ostatnich latach leków oddziałujących na cele zlokalizowane w cytozolu.



Wykres 7: Wykres przedstawia liczbę targetów leku w danej lokalizacji komórkowej w zależności od roku wprowadzenia leku na rynek.



Wykres 8: Wykres przedstawia znormalizowaną liczbę targetów leku w danej lokalizacji komórkowej w zależności od roku wprowadzenia leku na rynek.



Wykres 9: Wykres przedstawia znormalizowaną liczbę targetów leku w danej lokalizacji komórkowej w zależności od roku wprowadzenia leku na rynek.

5 Podsumowanie

Analiza sieci interakcji lek-cel umożliwia uzyskanie bardzo wielu danych na temat leków, białek z którymi oddziałują i ich wzajmnych interakcji. Jednak wyciągnięcie wniosków z otrzymanego ogromu informacji jest zadaniem bardzo złożonym. Podczas przedstawionej w tym raporcie analizy wiele rzeczy nie było widocznych na pierwszy rzut oka. Wiadomo, że zbiór targetów jest znacznie mniejszy od ilości samych leków, wiele z nich działa na te same cele. Jest to również zwiazane z podobieństwem strukturalnym cząsteczek leków do siebie nawzajem. Im jest ono wyższe, tym więcej spodziewanych wspólnch celów dla danej pary leków. Rozważając leki wielocelowe warto wspomnieć o fakcie, że część tego typu leków wprowadzanych na rynek w ostatniej dekadzie znacząco się zmniejszyła. Może to świadczyć o postępie i projektowaniu coraz bardziej specyficznych cząsteczek, powodujących mniej efektów nieporządanych. Można spodziewać się kontunacji tego trendu w kolejnych latach, gdyż nauka cały czas się rozwija, postęp jest bardzo szybki i zauważalny w naszym codziennym życiu.