Sabina Podlewska

Development of machine learning-based tools for computer-aided drug design

Jednymi z najbardziej znanych cech procesu projektowania nowych leków są jego stosunkowo długi czas trwania oraz znaczna kosztochłonność. Metody komputerowo wspomaganego projektowania leków są obecnie nieodłącznym elementem każdego etapu opracowywania substancji leczniczych. Zastosowanie metod obliczeniowych rozpoczyna się już na etapie identyfikacji nowych związków, potencjalnie posiadających pożądany profil aktywności biologicznej. Ocena związków *in silico* nie ogranicza się jednak tylko i wyłącznie do ewaluacji aktywności wobec wybranych celów biologicznych, lecz obejmuje także analizę ich własności fizykochemicznych i farmakokinetycznych, a także potencjalnej toksyczności.

Jedną z powszechniej wykorzystywanych metodologii w poszukiwaniu związków wykazujących aktywność biologiczną jest wirtualne przesiewane bibliotek związków chemicznych (ang. *virtual screening*; VS). Polega ono na automatycznym filtrowaniu bibliotek związków chemicznych, dostępnych komercyjnie lub wygenerowanych na drodze kombinatorycznej, w celu identyfikacji potencjalnie aktywnych struktur (tzw. wirtualnych hitów). Technika ta umożliwia ocenę milionów molekuł we względnie krótkim czasie i w związku z tym, pozwala na szybkie wytypowanie związków do wysokowydajnych badań przesiewowych (ang. *high-throughput screening*, HTS).

Wśród szerokiego wachlarza metod stosowanych obszarze obliczeniowej chemii medycznej, największą popularnością cieszą się narzędzia cechujące się jednocześnie wysoką skutecznością, jak i szybkim działaniem. Warunki te spełniane są m.in. przez metody uczenia maszynowego, stanowiące podstawę niniejszej rozprawy. Jej celem było opracowanie zestawu narzędzi opartych o metody uczenia maszynowego do poszukiwania nowych biologicznie aktywnych związków.

Początkowo zoptymalizowano warunki doświadczalne dla wykorzystania tego rodzaju algorytmów, ze zwróceniem szczególnej uwagi na skład zestawu związków nieaktywnych – zarówno jakościowo, jak i ilościowo. W sumie przetestowano 6 metod selekcji związków nieaktywnych (lub o założonej nieaktywności) – losowy i ukierunkowany na maksymalną różnorodność strukturalną wybór z bazy ZINC (biblioteka związków dostępnych komercyjnie), MDDR (baza związków posiadających aktywność biologiczną) oraz tzw. zestawu Directory of Useful Decoys, a liczbę nieaktywnych przykładów uczących optymalizowano w zakresie od 100 do 4000. Wykazano, że do zastosowań w wirtualnym przesiewaniu bibliotek związków chemicznych, największą skuteczność klasyfikacji zapewnia wykorzystanie w procesie uczenia związków nieaktywnych z bazy ZINC, w ok. 9-krotnym nadmiarze w stosunku do związków aktywnych.

Następnie do zastosowań w komputerowo wspomaganym projektowaniu leków wykorzystano nowe algorytmy optymalizacji parametrów maszyny wektorów nośnych oraz

nowe algorytmy uczące oparte na procedurze randomizacji. Wykazano, że nowe metody, którymi się posłużono są nie tylko bardziej skuteczne, lecz również charakteryzują się mniejszą złożonością obliczeniową niż standardowe podejścia używane w zadaniach identyfikacji aktywnych związków.

Przeprowadzono również badania problemu niespójności w danych doświadczalnych, tj. zróżnicowane wyniki aktywności określonych związków wobec wybranych celów biologicznych, które można obserwować w ogólnodostępnych bazach dla stosunkowo wysokiego odsetka przypadków. W celu zwiększenia wiarygodności i skuteczności modeli predykcyjnych konstruowanych na tego rodzaju danych, opracowano protokół wagujący uwzględniający niepewność wyników badań biologicznych, a jego skuteczność przetestowano w eksperymentach klasyfikacyjnych ligandów 25 celów biologicznych. Zastosowana metodologia jest połączeniem wagowania uwzględniającego dysproporcje pomiędzy poszczególnymi klasami oraz wagowania dostarczającego informacji o wariancji wyników powinowactwa wobec rozpatrywanego białka – wagi przypisywane poszczególnym przykładom miały następującą postać:

$$\underbrace{\left(1-\frac{N_{y_i}}{N}\right)}_{\text{liczność klas}} \cdot \underbrace{\left(\frac{1}{1+\text{var}(a_i)}\right)}_{\substack{\text{wariancja} \\ \text{powinowactwa}}} = \frac{N-N_{y_i}}{N(1+\text{var}(a_i))},$$

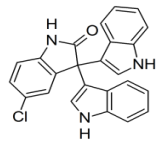
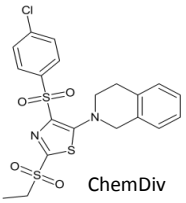
gdzie N jest całkowitą liczbą przykładów, N_y – liczbą przykładów przynależących do określonej klasy, natomiast var(a_i) jest wariancją wyników powinowactwa dostępnych dla danego związku wobec określonego białka.

Metody uczenia maszynowego pozwoliły również na zbudowanie pomostu pomiędzy standardowymi podejściami opartymi o strukturę i właściwości ligandów (ang. *ligand-based*), a metodami opartymi o strukturę celu biologicznego (ang. *structure-based*), dzięki opracowaniu protokołu do automatycznej oceny wyników dokowania opartego o algorytmy uczące. Obejmuje on reprezentację kompleksów ligand-receptor uzyskanych w procedurze dokowania przy pomocy deskryptorów oddziaływań strukturalnych oraz deskryptorów typu Spectrophores, ocenę przez metody uczenia maszynowego oraz wieloetapową analizę wyników uwzględniającą jakość modelu homologicznego białka wykorzystanego w procesie dokowania, skuteczność działania poszczególnych algorytmów klasyfikacyjnych oraz wartość funkcji oceniającej programu dokującego (badania wykonywano w ramach grantu PRELUDIUM 2013/09/N/NZ2/01917).

Ostatecznie, metody uczenia maszynowego zostały zastosowane do poszukiwania nowych ligandów receptora serotoninowego 5-HT_6 w procedurze wirtualnego przesiewania dwóch komercyjnie dostępnych baz związków: Chembridge i ChemDiv. Ocena została przeprowadzona w tzw. podejściu fingerprint-consensus, tj. związek traktowano jako potencjalnie aktywny wtedy i tylko wtedy, gdy został on oceniony jako taki dla wszystkich trzech zastosowanych reprezentacji. Dla dwóch, spośród 22 zamówionych, związków

eksperymentalnie wykazano istotną aktywność wobec receptora 5-HT_6 (119 i 670 nM), przy czym pierwszy posiadał również aktywność do receptora 5-HT_{2A} (296 nM), natomiast drugi wykazywał selektywność powinowactwa wobec pozostałych badanych receptorów (Tabela 1). Z uwagi na usunięcie z rozpatrywanego zbioru związków podobnych do znanych ligandów 5-HT_6R, znalezione nowe aktywne związki posiadają unikalne struktury, a ponadto należą do grupy tzw. ligandów niezasadowych (nie posiadających zasadowego atomu azotu zdolnego do oddziaływania w sprotonowanej formie z resztą kwasu asparaginowego D3.32 poprzez wiązanie wodorowe), co jest nietypowe dla ligandów tego receptora.

Tabela 1. Nowe ligandy receptora 5-HT_6 znalezione w eksperymencie przesiewowym przeprowadzonym przy wykorzystaniu metod uczenia maszynowego.

Struktura/ Baza/ ID związku	K_i [nM]				pK_a	Struktura/ Baza/ ID związku	K_i [nM]				pK_a
	5-HT_6	5-HT_{1A}	5-HT_{2A}	5-HT_7			5-HT_6	5-HT_{1A}	5-HT_{2A}	5-HT_7	
 ChemBridge 7706240	**670**	2450	15830	18080	-6.84	 ChemDiv C848-0334	**119**	12200	296	15770	-8.66

Wyniki uzyskane w ramach przedstawionej rozprawy doktorskiej dostarczają wiedzę i narzędzia do zastosowań metod uczenia maszynowego w poszukiwaniu związków aktywnych biologicznie. Ich użyteczność została potwierdzona nie tylko w obszernych badaniach retrospektywnych, lecz również w badaniach prospektywnych, w których zidentyfikowano dwa nowe ligandy receptora 5-HT_6.