Centrum Zastosowań Matematyki

ISBN 978-83-942947-4-8
PORTAL CZM 2015

NOWE STRATEGIE OBLICZENIOWE PODCZAS BADAŃ ILOŚCIOWYCH ZALEŻNOŚCI STRUKTURA-AKTYWNOŚĆ/STRUKTURA-WŁAŚCIWOŚCI (QSAR/QSPR) LEKÓW

PIOTR KAWCZAK I TOMASZ BĄCZEK

STRESZCZENIE. Projektowanie leków jest procesem napędzanym przełomowymi odkryciami technologicznymi, co oznacza zastosowanie zaawansowanych metod eksperymentalnych i obliczeniowych. Obecnie techniki lub metody projektowania leków odgrywają najważniejsze znaczenie w przewidywaniu profilu biologicznego, identyfikacji potencjalnych struktur, generowania struktur wiodących, a ponadto w celu przyspieszenia optymalizacji struktur wiodących w kierunku znalezienia odpowiednich kandydatów na leki. Ilościowe zależności struktura-aktywność (QSAR), struktura-właściwości (QSPR) lub struktura retencja (QSRR) służą jako cenne narzędzia do przewidywania w projektowaniu leków. Od dziesięcioleci metody QSAR/QSPR/QSRR mają zastosowanie w analizie zależności między właściwościami związków chemicznych i ich biologiczną aktywnością, aby uzyskać wiarygodny model statystyczny do prognozowania aktywności nowo syntezowanych związków.

1. WPROWADZENIE

Modelowanie ilościowych zależności struktura-aktywność/struktura właściwości (QSAR/QSPR) jest jednym z najważniejszych narzędzi obliczeniowych chemii medycznej. Klasyczne badania QSAR obejmują zagadnienia związane z ligandami i ich miejscami wiązania, stałe hamowania, stałe szybkości i inne biologiczne punkty końcowe, a dodatkowo właściwości molekularne w stosunku do innych, takich jak: lipofilowość, polaryzowalność, właściwości elektronowe, przestrzenne oraz określonych cech strukturalnych. Modelowanie QSAR jest powszechnie obecne w nauce, przemyśle oraz instytucjach na całym świecie. Modele QSAR znajdują szerokie zastosowanie do oceny potencjalnego wpływu substancji chemicznych wykazujących aktywność biologiczną na zdrowie człowieka oraz systemy ekologiczne. Obszar aktywnej ekspansji z zastosowaniem QSAR leży w wykorzystaniu modeli predykcyjnych w celach regulacyjnych m.in. przez agencje rządowe, w których wciąż rosnąca liczba specjalistycznych narzędzi regulacyjnych i baz danych jest opracowywana i zatwierdzana [10].

Pracę wykonano w ramach projektu Centrum Zastosowań Matematyki, współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

2. HISTORIA QSAR

Publikacja Hanscha i wsp. [20] stanowi kulminację 15-letnich zmagań ze zrozumieniem podstaw zależności struktura-aktywność (SARs) roślinnych regulatorów wzrostu. Większość czasu spędzono w dążeniu do znalezienia odpowiedniej zależności Hammetta stanowiącej obowiązującą metodologię wyjaśniający efekt obecności podstawników opisujących reaktywność chemiczną. Opierając się na badaniach Veldstra i wsp. [41] zbadany został wpływ lipofilowości na działanie biologiczne, omijając skomplikowana metodologie i zwracając się w kierunku współczynnika podziału n-oktanol-woda, jako wskaźnika lipofilowości [18]. W międzyczasie Fujita zastosował obliczenia kwantowo-chemiczne w celu uwzględnienia zmian aktywności w roślinnych regulatorach wzrostu [17]. Na początku lat 60-tych Fujita zaczął doświadczalnie mierzyć współczynnik podziału n-oktanol-woda ($\log P$). Hansch i Fujita szybko zorientowali się, że $\log P$ było dodatkową właściwością, tj. częściowy wkład podstawnika na podstawie wartości $\log P$ jednej cząsteczki jest często identyczny z wkładem tego podstawnika do wartości $\log P$ innej cząsteczki wprowadzili więc termin π określający wpływ podstawnika na hydrofobowość [10]. Po stwierdzeniu iż zależności pomiędzy $\log P$ i siłą działania biologicznego nie były bardziej precyzyjne niż pomiędzy współczynnikiem Hammetta, σ, i siłą działania Hansch i Fujita połączyli obydwa terminy w nowym równaniu, następnie opublikowano prace [19, 21], które demonstrowały z sukcesem obliczeniowe podejście do modelowania ilościowego wpływu obecności podstawników na siłę działania biologicznego. Zaletami tych prac jest fakt iż wpływ podstawników był oparty na modelach równowagi, współczynnikach podziału oraz pK_a, które są stosunkowo łatwe do zrozumienia, a dodatkowo wartości wpływu podstawników mogą być w dużej mierze przemieszczane pomiędzy poszczególnymi grupami cząsteczek. W powyższych pracach zwrócono uwagę i doceniono zastosowanie komputerów do obliczeń, natomiast znacznie mniejsza uwagę skupiono na różnicującej i umiejętnie objaśniającej roli statystyki. Kiedy podejście Hanscha do metodologii QSAR zostało ustalone bardziej zasadnicze kwantowo-mechaniczne obliczenia stawały się wykonalne umożliwiając zastosowanie alternatywnych sposobów odkrywania elektronowych i przestrzennych wyznaczników aktywności wśród podobnych struktur chemicznych [10]. Tradycyjne zastosowanie QSAR do szeregu związków będących kongenerami wymaga aby każda cząsteczka w zbiorze posiadała mierzalną wartość aktywności biologicznej np. ilościową, niezerową wartość siły działania. Analiza dyskryminacyjna lub logistyczna metoda regresji mogą mieć bardziej ogólne zastosowanie w modelowaniu binarnym lub zastosowaniu odpowiedzi kategorycznych (aktywny/nieaktywnay, mutagenny/niemutagenny) [35]. Chemia kwantowa pozostaje potężnym narzędziem do odkrywania podstawowych wyznaczników reaktywności w badaniach z zastosowaniem QSAR, jak również do obliczeń właściwości ab initio (np. moment dipolowy, μ) lub też wskaźników reaktywności całych cząsteczek tj. energia najwyżej lub najniżej zajętego orbitalu molekularnego ($E_{\text{HOMO}}, E_{\text{LUMO}}$). Zastosowanie na wyższym poziomie teorii *ab initio* jest zwykle ograniczone do pojedynczych systemów (faza gazowa) i stosunkowo niewielkiej liczby cząsteczek. Chemia kwantowa i jej metody półempiryczne oraz mechanika molekularna mają obecnie zastosowanie w metodach 3D-QSAR tj. wirtualnych badaniach przesiewowych ligandów *in silico* czy profilowaniu w odkrywaniu nowych substancji leczniczych. Ponadto takie metody mogą mieć zastosowanie w badaniach względnej stabilności konformerów, które to z kolei mogą wpływać na właściwości 3D molekuł analizowanych metodami QSAR [10].

3. DESKRYPTORY

Deskryptory (indeksy) są stosowane w opisie charakterystyki właściwości mikroskopowych cząsteczek dla celów związanych z modelowaniem molekularnym. Można je zasadniczo podzielić na następujące kategorie:

- konstytucyjne (względne liczby różnych rodzajów atomów);
- topologiczne (opisujące właściwości i połączenia atomów molekuł);
- fizykochemiczne (związane z rozpuszczalnością w wodzie lub lipidach np. moment dipolowy, ładunek całkowity itp.);
- strukturalne (opisujący trójwymiarowo wielkość, kształt i właściwości powierzchniowe cząsteczki);
- kwantowo-chemiczne (np. ładunki cząstkowe, polaryzowalność, energie orbitali, itd. obliczone półempirycznie za pomocą zasad gęstości funkcjonalnej – DFT oraz ab initio z zastosowaniem programów do obliczeń z chemii kwantowej).

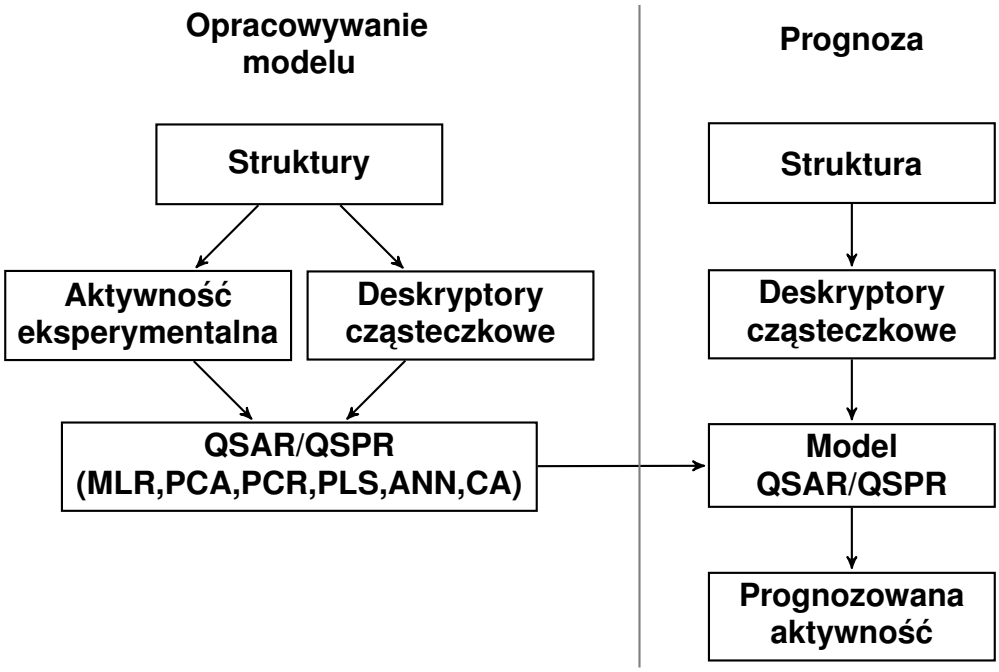
Wiele deskryptorów cząsteczkowych można łatwo obliczyć przy zastosowaniu odpowiedniego oprogramowania tj. Dragon (Talete, Mediolan, Włochy), HyperChem (HyperCube Inc., Gainesville, FL, Stany Zjednoczone).

Ważną grupą deskryptorów strukturalnych stanowią deskryptory trójwymiarowych właściwości pól należące do typu deskryptorów 3D (opisują właściwości cząsteczek dystrybuowane w trzech wymiarach). Te parametry są generowane przez obliczenie energii interakcji atomów pomiarowych w punktach siatki otaczających cząsteczkę. Wartości punktów siatki wokół cząsteczki stanowią deskryptory, które maja za zadanie uwzględnienie faktu dystrybucji danej właściwości w przestrzeni. Najpowszechniej stosowanymi deskryptorami pól cząsteczkowych są obliczone przy zastosowaniu metod CoMFA (*Comparative Molecular Field Analysis*) i CoMSIA (*Comparative Molecular Similarity Indices Analysis*). Metody z zastosowaniem deskryptorów 3D wymagają zgodnego ustawienia cząsteczek w przestrzeni dlatego też, właściwości konformacyjne (elastyczność oraz dystrybucja kształtów 3D) są zwykle istotne.

Odpowiedni dobór deskryptorów ma widoczny wpływ na jakość predykcyjną modelu, a dodatkowo będzie wpływać na zdolność modelu do objaśniania zależności pomiędzy cząsteczkowymi lub innymi mikroskopowymi i fizycznymi parametrami. Jeżeli efektywne deskryptory zostaną znalezione, zwłaszcza jeśli rzadki ich podzbiór zostanie prawidłowo zidentyfikowany wówczas możliwe jest określenie za ich pomocą, wpływu w jaki właściwości mikroskopowe (cząstkowe) mogą zostać zmodyfikowane do poprawy ogólnych właściwości analizowanych struktur [34].

4. METODY BUDOWY MODELI QSAR

Chemia obliczeniowa przedstawia molekularne struktury, jako modele numeryczne i symuluje ich zachowanie zgodnie z zasadami klasycznej fizyki przy pomocy równań kwantowych. Dostępne oprogramowanie pozwala naukowcom łatwo wygenerować i przedstawić dane cząsteczkowe, włączając: geometrię, energię oraz powiązane właściwości elektronowe, masowe oraz spektroskopowe. Najczęściej modelem do prezentacji i dalszych obliczeń na zbiorze tych danych jest tabela, w której analizowane związki chemiczne są opisane za pomocą odpowiednich wierszy i właściwości cząsteczkowych (w postaci deskryptorów) poprzez powiązane kolumny. QSAR próbuje odnaleźć spójną zależność pomiędzy różnicami w wartościach właściwości cząsteczkowych i aktywności biologicznej serii badanych związków tak aby rezultaty mogły posłużyć do ocen nowo



RYSUNEK 1. Modelowanie QSAR/QSPR, na podstawie [36]

syntezowanych struktur (rys. 1). QSAR na ogół przyjmuje postać równania liniowego postaci:

$$\text{Aktywność biologiczna} = \text{Stała} + (C_1 \times P_1) + (C_2 \times P_2) + \ldots + (C_n \times P_n)\,,$$

gdzie parametry od P_1 do P_n są obliczone dla każdej cząsteczki dla serii badanych związków, a współczynniki od C_1 do C_n są obliczone dopasowując wariancję parametrów oraz biologicznej aktywności [33, 44].

Techniki statystyczne lub chemometryczne tworzą matematyczną podstawę do budowy modelu QSAR. Najprostszą metodą wydaje się być regresję liniowa spośród różnych metod statystycznych mających zastosowanie dla QSAR. Regresja ta odzwierciedla bezpośrednią korelację zmiennych niezależnych (x) ze zmienną zależną (y). Ten model zakłada przewidywania wartości y względem zmiennych x. Same wartości mogą należeć do jakościowego lub ilościowego zbioru natomiast wariantami są metody SLR, MLR lub analiza krokowa MLR [36].

4.1. **Prosta regresja liniowa (*Simple Linear Regression* – SLR).** Metoda SLR reprezentuje standardowe obliczenie regresji liniowej w analizach z zastosowaniem modelu QSAR, w postaci równań, które obejmują jeden niezależny deskryptor x oraz y jako zmienna zależna. Ta technika obliczeniowa jest uznawana za bardzo obiecującą dla zależności struktura-aktywność badając niektóre z najważniejszych deskryptorów wpływających na aktywność podczas gdy inne typy oddziaływań mogą zostać pominięte [8, 36]. Prostą regresję liniową można opisać za pomocą następującego równania:

$$y = a + bx,$$

gdzie y jest zmienną zależną, x jest zmienną niezależna, a jest stałą natomiast b jest współczynnikiem regresji.

Regresja liniowa stanowi estymację warunkowej wartości oczekiwanej, w której zakładanym modelem zależności pomiędzy zmiennymi jest funkcja liniowa.

4.2. **Wielokrotna regresja liniowa (*Multiple Linear Regression* – MLR).** Metoda MLR jest rozszerzeniem SLR do więcej niż jednego wymiaru i obliczana jest za jej pomocą standardowa regresja wielu zmiennych. Identyfikacja właściwości leku jest przeprowadzona na zbiorze wszystkich analizowanych deskryptorów. Wystąpienie korelacji jest zobrazowane wartością wielokrotności współczynnika korelacji (r), lub wartością t za pomocą metody *leave-one-out*, natomiast korelacja jest sprawdzana poprzez wartości r^2 lub q^2, które są określane mianem współczynników walidacji krzyżowej [34, 36]. Zależność ta wyrażona w postaci następującego równania:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_mx_m,$$

gdzie y jest zmienną zależną (aktywność), x_i są zmiennymi niezależnymi (deskryptory), b_i określa współczynniki regresji. W zapisie macierzowym równanie wygląda następująco:

$$y = \mathbf{X}b.$$

Całość natomiast może być rozszerzona do regresji wielomianowej postaci:

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + b_4x_2 + b_5x_2^2 + b_6x_2^3 + \ldots,$$

gdzie można stosować wielomiany dowolnej kolejności, a wartości b są dopasowanymi współczynnikami uzyskanymi w oparciu o kryterium najmniejszych kwadratów.

4.3. **Regresja metodą cząstkową najmniejszych kwadratów (*Partial Least Square* – PLS).** Regresja metodą cząstkowych najmniejszych kwadratów jest rozszerzeniem modelu regresji wielorakiej, bez ograniczeń nakładanych przez analizy tj. dyskryminacyjna, korelacja kanoniczna czy regresja metodą głównych składowych. PLS stanowi najprawdopodobniej najmniej restrykcyjną odmianę spośród różnych wielowymiarowych rozszerzeń modelu liniowej regresji wielorakiej, umożliwiając jej zastosowanie w analizie eksploracyjnej oraz identyfikacji obserwacji odstających w klasycznej regresji liniowej. Zastosowanie metody PLS daje statystycznie pewne rozwiązanie nawet wtedy gdy zmienne niezależne są silnie ze sobą powiązane lub liczba zmiennych niezależnych przekracza liczbę obserwacji. PLS jest iteracyjną metoda regresji, która prezentuje swoje rozwiązania bazując na liniowej transformacji z dużej liczby deskryptorów pierwotnych do ich niewielkiej liczby w nowych ortogonalnych warunkach, zwanych zmiennymi ukrytymi, stawiając jednocześnie metodę w szeregu podstawowych metod statystycznych szczególnie przydatną w chemometrii [23, 36, 46].

4.4. **Analiza głównych składowych (*Principal Component Analysis* – PCA).** Analiza głównych składowych stanowi jedną ze statystycznych metod analizy czynnikowej. Podstawowa idea analizy polega na tym, że zmienne ukryte są utworzone poprzez liniowe połączenie zmiennych oryginalnych macierzy danych (X) rozłożone na dwie ortogonalne matryce zwanymi: ładunkami wygenerowanych czynników składowych głównych oraz wynikami. PCA zakłada taki obrót układu współrzędnych aby maksymalizować wariancję najpierw pierwszej współrzędnej a następnie drugiej współrzędnej. W ten sposób konstruowana jest nowa przestrzeń obserwacji, w której najwięcej zmienności wyjaśniają początkowe czynniki. Zasadniczym założeniem więc jest fakt iż w PCA wyniki i ładunki wektorów odpowiadające największym wartościom własnym zawierają najbardziej użyteczne informacje związane z analizowanym problemem. Tak więc za pomocą metody PCA nowy zestaw ortogonalnie położonych deskryptorów zostaje utworzony, zwany głównymi

składowymi (PCs), opisując większość informacji zawartych w zmiennych niezależnych zgodnie ze zmniejszającymi się wartościami wariancji. PCA jest często używana do zmniejszania rozmiaru zbioru danych statystycznych, poprzez odrzucenie ostatnich czynników. PCA może być oparte na macierzy korelacji lub kowariancji pochodzących ze zbioru wyjściowego. W przypadku zastosowania macierzy kowariancji, zmienne w zbiorze wejściowym o największej wariancji mają największy wpływ na wynik, co może być wskazane, jeśli zmienne reprezentują porównywalne wielkości natomiast macierzy korelacji odpowiada wstępna normalizacja zbioru wejściowego tak, aby każda zmienna miała na wejściu identyczną wariancję, co może być wskazane, jeśli wartości zmiennych nie są porównywalne.

Budowa macierzy danych wyjściowych przyjmuje postać:

$$\mathbf{X} = [x_{ji}], \quad j = 1, 2, \ldots, m, i = 1, 2, \ldots, n,$$

gdzie $x_{ji} \geq 0$ – wartość j-tej zmiennej w i-tym obiekcie; po standaryzacji $\mathbf{Z} = [z_{ji}]$.

Podstawowe równanie metody głównych składowych można zapisać w postaci układu równań liniowych:

$$\mathbf{Z}^T = \mathbf{A}\mathbf{G}^T,$$
$$\mathbf{G} = \mathbf{A}^T\mathbf{Z},$$

gdzie $\mathbf{Z}$ – macierz j, standaryzowanych zmiennych pierwotnych, $\mathbf{A}$ – macierz współczynników składowych głównych, $\mathbf{G}$ – macierz składowych głównych.

Podstawą do wyznaczania elementów macierzy ładunków głównych składowych jest macierz korelacji $\mathbf{R}$ postaci:

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}^T\mathbf{Z}.$$

Każdą z głównych składowych G_l można przedstawić jako liniową kombinację pierwotnych zmiennych $\mathbf{Z}$:

$$G_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} a_{i1}Z_j,$$

gdzie m – liczba zmiennych pierwotnych k – liczba składowych głównych, Z_j – j-ta zmienna standaryzowana (pierwotna), G_l – l-ta składowa główna, a_{ij} – ładunki czynnikowe.

Ostatecznie pierwsza główna składowa G_1 jest taką kombinacją zmiennych pierwotnych, dla której wariancja próbkowa opisana jest wzorem:

$$S^2_{G_1} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{i1}a_{j1}S_{ij},$$

gdzie i jest największą wśród wszystkich kombinacji liniowych takich, że: $a_1^T a_1 = 1$. Drugą główną składową można przedstawić w sposób analogiczny. Jest ona kombinacją liniową zmiennych pierwotnych maksymalizującą wariancję przy warunkach: $a_1^T a_1 = 1$ oraz $a_1^T a_2 = 0$. Drugi warunek zapewnia ortogonalność powstałych składowych. Konsekwencją tego jest sumowanie się kolejnych wariancji głównych składowych do wariancji całkowitej. W praktyce ograniczamy się do kilku pierwszych głównych składowych, które wyjaśniają z góry ustaloną część wariancji całkowitej (np.75%).

Za pomocą tej metody model QSAR nie zostaje wygenerowany ale daje ona świadectwo zależności pomiędzy przeciwnymi zmiennymi. PCA zmniejsza wymiarowość pełnego zestawu deskryptorów danych do ich rzeczywistej ilości. Aby wygenerować wielowymiarowe równania liniowe, regresja głównych składowych (PCR) stosuje wyniki rozkładu analizy metodą PCA jako regresory (zmienne niezależne) w modelu QSAR [14, 15, 23, 36, 45].

4.5. **Analiza skupień (*Cluster Analysis* – CA).** Analiza skupień zwana klasteryzacją lub grupowaniem jest metodą klasyfikacji dokonującej grupowania elementów na względnie jednorodne klasy. Następuje tu podział zbioru obserwacji na podzbiory (klastry) w taki sposób aby obiekty w tym samym klastrze były względnie podobne. Algorytmy analizy klasterowej można podzielić na kilka kategorii.

Metody hierarchiczne, do których zaliczają się procedury aglomeracyjne oraz procedury deglomeracyjne. W metodach hierarchicznych algorytm tworzy dla zbioru obiektów hierarchię klasyfikacji, zaczynając od takiego podziału, w którym każdy obiekt stanowi samodzielne skupienie, a kończąc na podziale, w którym wszystkie obiekty należą do jednego skupienia.

Grupa metod k-średnich, w której grupowanie polega na wstępnym podzieleniu populacji na z góry założoną liczbę klas, a następnie uzyskany podział jest poprawiany w ten sposób, że niektóre elementy są przenoszone do innych klas, tak, aby uzyskać minimalną wariancję wewnątrz uzyskanych klas. Metody rozmytej analizy skupień – mogące przydzielać element do więcej niż jednej kategorii, różnią się pod tym względem od metod klasycznej analizy skupień, w których uzyskana klasyfikacja ma charakter grupowania rozłącznego, którego wynikiem jest to, że każdy element należy do jednej i tylko jednej klasy. Analiza skupień jest metodą rozpoznawania wzorców stosowaną do zbadania zależności między obserwacjami związanymi z wieloma innymi właściwościami oraz podziałem zestawu analizowanych na kategorie z podobnymi elementami. Umożliwia ona także określenie która z podgrup charakteryzuje się podobnymi właściwościami fizycznymi [2, 16, 36].

4.6. **Sztuczne sieci neuronowe (*Artificial Neural Networks* – ANNs).** Sieci neuronowe powstały w wyniku badań prowadzonych w dziedzinie sztucznej inteligencji; szczególne znaczenie miały tutaj te prace, które dotyczyły budowy modeli podstawowych struktur występujących w mózgu. Prace te miały na celu naśladowanie zwłaszcza tych cech charakterystycznych dla biologicznych systemów nerwowych, które mogą być szczególnie użyteczne technicznie. Sztuczne sieci neuronowe stanowią użyteczne narzędzia badań QSAR/QSPR, a szczególnie w przypadku, gdy trudno jest określić dokładny model matematyczny opisujący daną zależność struktura-właściwości. Większość prac stosuje sieci neuronowe oparte na algorytmie uczącym propagacji wstecznej, która posiada pewne wady, takie jak minimum lokalne, wolna konwergencja, czasochłonność, nieliniowa optymalizacja iteracyjna oraz trudności w optymalnej konfiguracja sieci [43].

Na podstawie obserwacji mechanizmów zachodzących w naturalnej sieci neuronów, opracowano matematyczne koncepcje sztucznych sieci neuronowych. Podobnie jak w naturalnych sieciach, składają się one z elementarnych komórek – neuronów. Zdając sobie sprawę z braku możliwości dokładnego odwzorowania naturalnych układów i ich budowy, naukowcy opracowali model sztucznego neuronu. Działanie pojedynczego neuronu można opisać następującym wzorem:

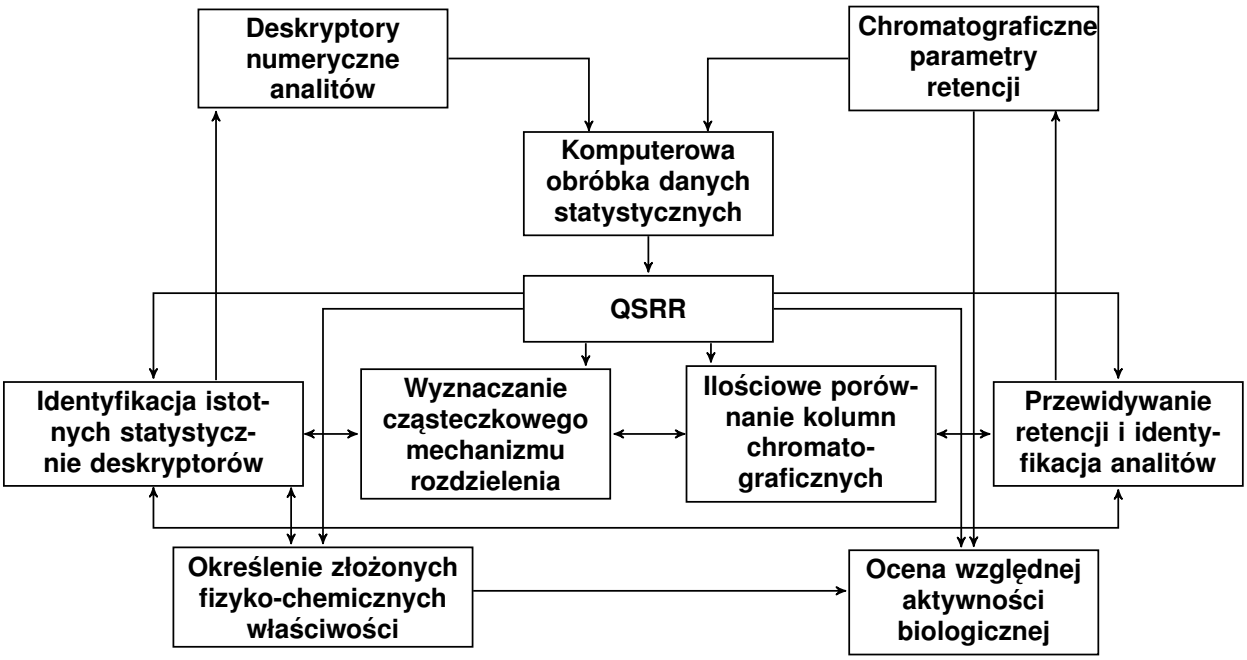
$$y = f\left(\sum_{j=0}^{N} x_j w_j\right).$$

Funkcja aktywacji może mieć różne postaci, każda z nich aby mogła pełnić tę rolę musi być ciągła i łatwo różniczkowalna, wyjątkiem jest perceptron w przypadku którego funkcja aktywacji nie jest poddana tym ograniczeniom. W praktyce stosuje się najczęściej następujące funkcje: liniową, logistyczną, tangens hiperboliczny, sinus oraz signum [40].

Metoda sztucznych sieci neuronowych opiera się na zasadzie działania komórek neuronów obecnych w mózgach zwierząt. Sztuczne sieci neuronowe są równoległym systemem obliczeniowym, składającym się z grup ściśle połączonych elementów przetwarzania zwanych neuronami, które są usytuowane w szeregu warstw. Warstwa wejściowa jako pierwsza i każdy z jej neuronów otrzymuje dane od użytkownika, które odpowiadają jednej ze zmiennych niezależnych wykorzystywane jako wejścia dla QSAR. Za warstwą wejściową, jest wiele warstw neuronów, stanowiących warstwy ukryte. Ostatnia warstwa jest określana jako warstwa wyjściowa, a jej neurony obsługują wyjścia z sieci. Każda warstwa może dokonać obliczeń niezależnych i może przekazać wyniki innej warstwie. Wynik funkcji przenoszenia jest przekazywane neuronom w warstwie wyjściowej, jest to punkt, w którym wyniki są interpretowane i ostatecznie prezentowane [7, 23, 36].

4.7. **Przybliżona funkcja genetyczna (*Genetic Function Approximation* – GFA).** Algorytm przybliżonej funkcji genetycznej (GFA) stanowi technikę generowania statystycznych modeli danych z zastosowaniem procesu ewolucji, tym samym GFA oferuje nowe podejście do problemu konstruowania modeli QSAR i QSPR. Algorytmy genetyczne wywodzą się z analogii do roli jaką odgrywa DNA, gdzie indywidua reprezentowane są za pomocą jednowymiarowego ciągu bitów. Funkcja dopasowania służy do oceny jakości jednostek, w ten sposób, że najistotniejsze z nich, charakteryzować się będą najlepszym wynikiem dopasowania. Algorytmy genetyczne są szczególnie przydatne w rozwiązywaniu problemów o dużej liczbie wymiarów, stanowiąc skuteczne narzędzie analizy. Zastępując analizę regresji algorytmem GFA umożliwiamy budowę modeli konkurencyjnych względem standardowych technik oraz udostępniamy dodatkowe, nie dostarczone informacje. GFA pozwala budować modele przy użyciu nie tylko wielomianów liniowych i nieliniowych ale również wielomianów wyższego rzędu, funkcji sklejanych oraz funkcji Gaussiana. Metoda ta służy jako alternatywa dla standardowej regresji przy wyznaczaniu równania QSAR [37, 38]. Genetyczna metoda cząstkowa najmniejszych kwadratów (G/PLS lub GA-PLS) jest ważnym narzędziem, które ewoluowało, łącząc najlepsze cechy metod GFA i PLS. Metoda ta jest popularna i szeroko wykorzystywane przez naukowców [9, 11–13, 32, 36, 42].

4.8. **Algorytm k-najbliższych sąsiadów (k-*Nearest Neighbors* – k-NN).** Algorytm k-najbliższych sąsiadów jest jednym z algorytmów regresji nieparametrycznej stosowanych w statystyce do prognozowania wartości zmiennej losowej, mający tym samym zastosowanie w klasyfikacji. Podejścia z zastosowaniem algorytmu k-NN są wykonywane poprzez odległości pomiędzy obiektem, który jest nieznany, a wszystkimi obiektami w zestawie treningowym. Na podstawie obliczeń odległości zostają wybrane obiekty zestawu treningowego najbardziej zbliżone do obiektu nieznanego. Ostatecznie, optymalna wartość k jest selekcjonowana za pomocą optymalizacji na podstawie kategoryzacji zestawu testowego. Algorytm k-NN jest przydatny w przypadku złożonej lub nietypowej zależności pomiędzy zmiennymi objaśnianymi i objaśniającymi, tak więc skomplikowanymi przy modelowaniu z zastosowaniem klasycznych metod [1, 36].



RYSUNEK 2. Metodologia i cele QSRR, na podstawie [29]

5. ILOŚCIOWE ZALEŻNOŚCI STRUKTURA-RETENCJA (QSRR)

Ilościowe zależności struktura-retencja chromatograficzna QSRR zależą od retencji chromatograficzną analitu z uwzględnieniem struktury chemicznej i właściwości fizykochemicznych fazy stacjonarnej i ruchomej w chromatografii cieczowej [25–27, 29].

QSRR stanowi statystycznie wyznaczoną zależność pomiędzy parametrami chromatograficznymi i wartościami (deskryptorami) charakteryzującymi strukturę badanych analitów. QSRR jest stosowane przy opisie mechanizmu cząsteczkowego odpowiedzialnego za rozdzielanie w układach chromatograficznych; jednocześnie ocenia kompleksowe fizykochemiczne właściwości analitów oraz chromatograficzne fazy stacjonarnej w przewidywaniu retencji chromatograficznej nowych struktur (rys. 2).

W celu przeprowadzenia analizy QSRR są zbierane dwa zestawy danych: odpowiedni zestaw parametrów opisujących retencję serii analitów oraz zestaw parametrów strukturalnych rozdzielonych analitów (deskryptorów).

W przypadku danych retencyjnych najczęściej stosowanym parametrem jest logarytm ze współczynnika retencji ($\log k$). Alternatywnie, $\log k$, można ekstrapolować do czystej wody ($\log k_w$) jako hipotetycznej fazy ruchomej, celem wyraźniejszego rozróżnienia pomiędzy właściwościami analitów hydrofobowych. Rzadziej stosowany jest bezpośrednio czas retencji (t_R). Deskryptory odzwierciedlające fizyczne i chemiczne właściwości analitów można uzyskać doświadczalnie lub za pomocą metod obliczeniowych.

Najprostszym podejściem jest regresja retencji QSRR (R_t) względem obliczonej wartości współczynnika podziału n-oktanol/woda ($c \log P$) [3, 5, 24, 30]. Równanie przybiera wówczas następującą postać:

$$R_t = k_1 + k_2 c \log P,$$

gdzie k_1 i k_2 są współczynnikami równania regresji.

W następnym modelu QSRR retencja zależy od określonych deskryptorów otrzymanych przy pomocy modelowania molekularnego. Uzyskane równanie umożliwia różnicowanie faz stacjonarnych w chromatografii cieczowej pod względem chemicznych właściwości nośnika oraz ligandów. Równanie pozwala opisać mechanizm interakcji analitu ze złożem badanej kolumny.

Na podstawie prac [4, 6] wykazano, iż stosowanie poniższych deskryptorów zapewnia uzyskanie wiarygodnych wyników: μ – całkowity moment dipolowy, opisujący oddziaływania dipol-dipol pomiędzy analitem i indukowaną chromatograficzną fazą ruchoma oraz fazą stacjonarną; $\rho_{\min}$ – największy ujemny ładunek nadmiarowy w cząsteczce, określa lokalną polarność analitu i zdolność do uczestniczenia w polarnych interakcjach; A_{WAS} – określa powierzchnię dostępną w kontakcie z rozpuszczalnikiem charakteryzując siłę oddziaływań dyspersyjnych (typu Londona) analitu z cząstkami fazy chromatograficznej. Dla takiego modelu równanie przyjmuje następującą postać [31]:

$$\log k_w = k'_1 + k'_2\mu + k'_3\rho_{\min} + k'_4 A_{\mathrm{WAS}},$$

gdzie k'_1, k'_2, k'_3, k'_4 są współczynnikami regresji.

Kolejny model QSRR bazuje na liniowych zależnościach energii swobodnej (LFER) pojęcie to rozszerzono ze sfery właściwości atomowych do sfery interakcji międzycząsteczkowych i określono mianem liniowych zależności energii solwatacji (LSER), stanowiąc termodynamiczny rodzaj zależności, która łączy ze sobą fizyczne i chemiczne modele procesów z pojęciami z termodynamiki [25]. Odpowiednie założenia pozwalają na pominięcie ścisłych praw termodynamiki, zakładając że istniej zależność pomiędzy modelami procesów chemicznych i zasadami termodynamiki z pominięciem derywacji matematycznych zależności [31]. Ten model QSRR zakłada więc istnienie liniowej zależności pomiędzy standardowymi parametrami retencji i zmianą energii swobodnej związanej z rozdzieleniem w procesie chromatograficznym [28]. Zgodnie z tym modelem kolumna chromatograficzna może być postrzegana jako przetwornik energii swobodnej, który przekształca różnice potencjałów chemicznych, wynikających z budowy analitów w ilościowe różnice we właściwościach retencyjnych [25]. Podejście to zostało wprowadzone i rozwinięte przez Abraham i wsp. [39], natomiast ogólne równanie LSER przybiera postać:

$$\log k = \log k_0 + rR_2 + vV_x + s\pi_2^H + a\sum \alpha_2^H + b\sum \beta_2^H,$$

gdzie R_2 oznacza nadmiar refrakcji molowej analitu, V_x jest objętością cząsteczkową algorytmu McGowana, π_2^H jest deskryptorem biopolarności/polaryzowalności, $\sum \alpha_2^H$ jest miarą przekazywania protonu analitowi, $\sum \beta_2^H$ – analogiczna miara odbierania protonu, $\log k_0$ jest wartością stałą, natomiast r, v, s i b są współczynnikami regresji stanowiących zestaw uzupełniających właściwości układu chromatograficznego składającego się z chromatograficznej fazy stacjonarnej i ruchomej.

Obecnie QSRR znajduje szerokie zastosowanie do porównywania kolumn chromatograficznych (faz stacjonarnych) [22].

6. WNIOSKI

Zastosowanie metod prognozowania *in silico* cieszy się wzrastającą popularnością w ostatnich latach. Przewiduje się, że będzie to przedmiot ciągłego rozwoju w przyszłości nie tylko w projektowaniu leków, ale również w obszarze toksykologii prognostycznej. QSAR jest obecna na etapie odkrywania leków, umożliwiając zaprojektowanie bezpiecznego i potencjalnie silnego kandydata

na potencjalny lek. Podczas poszczególnych faz odkrywania leków farmakodynamiczny i farmakokinetyczny profil cząsteczek jest możliwy do uzyskania, stosując różne modele QSAR. Na oceny *in silico* składają się przewidywania różnych właściwości (fizykochemiczne, ADMETox) i aktywności, aby pomóc w optymalizacji i priorytetyzacji kandydatów na leki. Liczne dostępne narzędzia *in silico* w tym modele QSAR/QSPR, drzewa decyzyjne i dokowanie molekularne są proponowane do osiągnięcia wytyczonych celów, natomiast interpretacja deskryptorów uzyskanych przy pomocy metod QSAR pomaga naukowcom projektować coraz ciekawsze struktury.

LITERATURA

[1] S. Ajmani, K. Jadhav, S.A. Kulkarni, *Three-dimensional QSAR using the k-nearest neighbor method and its interpretation*, J. Chem. Inf. Model. 46: 24–31, 2006.

[2] M.S. Aldenderfer, R.K. Blashfield, *A review of clustering methods*, w: M.S. Aldenderfer, R.K. Blashfield (ed.) *Cluster analysis*, SAGE Publications Ltd., London, 33–61, 1984.

[3] M.A. Al-Haj, R. Kaliszan, B. Buszewski, *Quantitative Structure-Retention Relationships with Model Analytes as a Means of an Objective Evaluation of Chromatographic Columns*, J. Chromatogr. Sci. 39: 29–38, 2001.

[4] T. Bączek, *Computer-assisted optimization of liquid chromatography separations of drugs and related substances*, Curr. Pharm. Anal. 4: 151–161, 2008.

[5] T. Bączek, R. Kaliszan, *Combination of linear solvent strength model and quantitative structure-retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography*, J. Chromatogr. A 962: 41–55, 2002.

[6] T. Bączek, R. Kaliszan, K. Novotna, P. Jandera, *Comparative characteristics of HPLC columns based on quantitative structure-retention relationships (QSRR) and hydrophobic-subtraction model*, J. Chromatogr. A 1075: 109–115, 2005.

[7] I.I. Baskin, V.A. Palyulin, N.S. Zefirov, *Neural networks in building QSAR models*, Methods Mol. Biol. 458: 137–158, 2008.

[8] R.A. Berk, *Simple linear regression. Regression analysis: a constructive critique*, SAGE Publications Ltd., London, 21–38, 2003.

[9] B. Breu, K. Silber, H. Gohlke, *Consensus adaptation of fields for molecular comparison (AFMoC) models incorporate ligand and receptor conformational variability into tailor-made scoring functions*, J. Chem. Inf. Model. 47: 2383–2400, 2007.

[10] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *QSAR Modeling: Where Have You Been? Where Are You Going To?*, J. Med. Chem. 57: 4977–5010, 2014.

[11] P.A. Datar, S.A. Khedkar, A.K. Malde, E.C. Coutinho, *Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands*, J. Comput. Aided Mol. Des. 20: 343–360, 2006.

[12] D.K. Dhaked, J. Verma, A. Saran, E.C. Coutinho, *Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA)*, J. Mol. Model. 15: 233–245, 2009.

[13] W.J. Dunn, D. Rogers, *Genetic partial least squares in QSAR*, w: J. Devillers (ed.) *Genetic algorithms in molecular modeling*, Academic Press, London, 109–130, 1996.

[14] G.H. Dunteman, *Basic concepts of principal components analysis*, w: G.H. Dunteman (ed.) *Principal components analysis*, SAGE Publications Ltd., London, 15–22, 1989.

[15] G.H. Dunteman, *Uses of principal components in regression analysis*, w: G.H. Dunteman (ed.) *Principal components analysis*, SAGE Publications Ltd., London, 65–74, 1989.

[16] B.S. Everitt, S. Landau, M. Leese, D. Stahl. *Cluster analysis, Wiley series in probability and statistics, 5th Ed*, John Wiley & Sons., Chichester, UK, 2011.

[17] T. Fujita, S. Imai, K. Koshimizu, T. Mitsui, I. Kato, *Plant Growth Activities of 5- and 8-Halogeno-dihydro- and -Tetrahydro-1-naphthoic Acids*, Nature 184: 1415–1416, 1959.

[18] C. Hansch, *Quantitative Approach to Biochemical Structure?Activity Relationships*, Acc. Chem. Res. 2: 232–239, 1969.
[19] C. Hansch, T. Fujita, $p - \sigma - \pi$ *Analysis. A Method for the Correlation of Biological Activity and Chemical Structure*, J. Am. Chem. Soc. 86: 1616–1626, 1964.
[20] C. Hansch, P. Maloney, T. Fujita, R. Muir, *Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients*, Nature 194: 178–180, 1962.
[21] C. Hansch, R. Muir, T. Fujita, P. Maloney, F. Geiger, M. Streich, *The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients*, J. Am. Chem. Soc. 85: 2817–2824, 1963.
[22] K. Héberger, *Quantitative structure-(chromatographic) retention relationships*, J. Chromatogr. A 1158: 273–305, 2007.
[23] Internetowy Podręcznik Statystyki: `http://www.statsoft.pl/textbook/stathome.html` [dostęp online: 07.07.2015].
[24] J. Jiskra, H.A. Claessens, C.A. Cramers, R. Kaliszan, *Quantitative structure-retention relationships in comparative studies of behavior of stationary phases under high-performance liquid chromatography and capillary electrochromatography conditions*, J. Chromatogr. A 977: 193–206, 2002.
[25] R. Kaliszan, *Quantitive Structure-Chromatographic Retention Relationships*, John Wiley & Sons, New York, 1987.
[26] R. Kaliszan, *Quantitative structure-retention relationships*, Anal. Chem. 64: 619–631, 1992.
[27] R. Kaliszan, *Structure and Retention in Chromatography: A Chemometric Approach*, Harwood Academic Publishers, Amsterdam, 1997.
[28] R. Kaliszan, *Chromatography and capillary electrophoresis in modelling the basic processes of drug action*, Trends Anal. Chem. 18: 400–410, 1999.
[29] R. Kaliszan, *QSRR: Quantitative Structure-(Chromatographic) Retention Relationships*, Chem. Rev. 107: 3212–3246. 2007.
[30] R. Kaliszan, T. Bączek, A. Buciński, B. Buszewski, M. Sztupecka, *Prediction of gradient retention from the linear solvent strength (LSS) model, quantitative structure-retention relationships (QSRR), and artificial neural networks (ANN)*, J. Sep. Sci. 26: 271–282, 2003.
[31] R. Kaliszan, M.A. van Straten, M. Markuszewski, C.A. Cramers, H.A. Claessens, *Molecular mechanism of retention in reversed-phase high-performance liquid chromatography and classification of modern stationary phases by using quantitative structure-retention relationships*, J. Chromatogr. A 855: 455–486, 1999.
[32] S.A. Khedkar, A.K. Malde, E.C. Coutinho, *Design of inhibitors of the MurF enzyme of Streptococcus pneumoniae using docking, 3DQSAR, and de novo design*, J. Chem. Inf. Model. 47: 1839–1846, 2007.
[33] H. Kubinyi, *Quantitative structure-activity relationships. IV. Non-linear dependence of biological activity on hydrophobic character: a new model*, Arzneimittelforschung 26: 1991–1997, 1976.
[34] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, *Quantitative Structure-Property Relationship Modeling of Materials Properties*, Chem. Rev. 112: 2889–2919, 2012.
[35] Y.C. Martin, J.B. Holland, C.H. Jarboe, N. Plotnikoff, *Discriminant Analysis of the Relation between Physical Properties and the Inhibition of Monoamine Oxidase by Aminotetralins and Aminoindans*, J. Med. Chem. 17: 409–413, 1974.
[36] H.M. Patel, M.N. Noolvi, P. Sharma, V. Jaiswal, S. Bansal, S. Lohan, S.S. Kumar, V. Abbot, S. Dhiman, V. Bhardwaj, *Quantitative structure-activity relationship (QSAR) studies as strategic approach in drug Discovery*, Med. Chem. Res. 23: 4991–5007, 2014.
[37] D. Rogers, A.J. Hopfinger, *Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships*, J. Chem. Inf. Comput. Sci. 34: 854–866, 1994.
[38] D. Rogers, *Genetic Function Approximation: Evolutionary Construction of Novel, Interpretable, Nonlinear Models of Experimental Data, Rational Drug Design, The IMA Volumes in Mathematics and its Applications*, 108: 163–189, 1999.
[39] P.C. Sadek, P.W. Carr, R.M. Doherty, M.J. Kamlet, R.W. Taft, M.H. Abraham, *Study of retention processes in reversed-phase high-performance liquid chromatography by the use of the solvatochromic comparison method*, Anal. Chem. 57: 2971–2978, 1985.

[40] Sztuczne sieci neuronowe i algorytmy genetyczne: http://4programmers.net/Z_pogranicza/Sztuczne_sieci_neuronowe_i_algorytmy_genetyczne [dostęp online: 07.07.2015].
[41] H. Veldstra, *The Relation of Chemical Structure to Bio-Logical Activity in Growth Substances*, Annu. Rev. Plant Physiol. 4: 151–198, 1953.
[42] J. Verma, V.M. Khedkar, A.S. Prabhu, S.A. Khedkar, A.K. Malde, E.C. Coutinho, *A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants*, J .Comput. Aided Mol. Des. 22: 91–104, 2008.
[43] B. Walczak, D.L. Massart, *Local modeling with radial basis function networks*, Chemom. Intell. Lab. Syst. 50: 179–198, 2000.
[44] C.S. Walpole, R. Wrigglesworth, S. Bevan, E.A. Campbell, A. Dray, I.F. James, K.J. Masdin, M.N. Perkins, J. Winter, *Analogues of capsaicin with agonist activity as novel analgesic agents; structure-activity studies 3. The hydrophobic side-chain 'C-region'*, J. Med. Chem. 36: 2381–2389, 1993.
[45] S. Wold, K Esbensen, P. Geladi, *Principal component analysis*, Chemometr. Intell. Lab. Syst. 2: 37–52, 1987.
[46] S. Wold, E. Johansson, M. Cocchi, *PLS: partial least squares projections to latent structures*, w: H. Kubinyi (ed.) *3D QSAR in drug design: theory, methods and applications*, ESCOM Science Publishers, Leiden, 523–550, 1993.

PIOTR KAWCZAK
KATEDRA I ZAKŁAD CHEMII FARMACEUTYCZNEJ, WYDZIAŁ FARMACEUTYCZNY Z ODDZIAŁEM MEDYCYNY LABORATORYJNEJ, GDAŃSKI UNIWERSYTET MEDYCZNY, AL. GEN. J. HALLERA 107, 80-416 GDAŃSK
Adres e-mail: p99p@gumed.edu.pl

TOMASZ BĄCZEK
KATEDRA I ZAKŁAD CHEMII FARMACEUTYCZNEJ, WYDZIAŁ FARMACEUTYCZNY Z ODDZIAŁEM MEDYCYNY LABORATORYJNEJ, GDAŃSKI UNIWERSYTET MEDYCZNY, AL. GEN. J. HALLERA 107, 80-416 GDAŃSK
Adres e-mail: tbaczek@gumed.edu.pl