

Rozszerzenie banku pseudoatomów UBDB w oparciu o najczęściej występujące typy atomów w Cambridge Structural Database

Autor: Aleksandra Sypko, promotor: prof. dr hab. Paulina Maria Dominiak

Celem przeprowadzonych badań była poprawa statystyki rozpoznawalności atomów i całych molekuł w banku UBDB

Badania podzielone były na 3 etapy:

- Przeszukanie bazy CSD
- Odfiltrowanie struktur niepoprawnych chemicznie
- Przygotowanie nowych wpisów do bazy UBDB

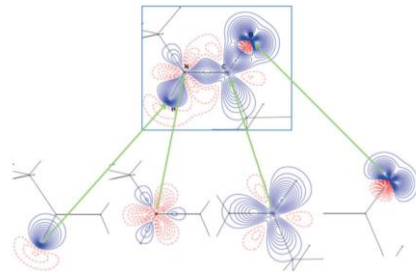
1. Przeszukanie CSD pod kątem struktur zawierających węgiel i dowolną kombinację atomów H, N, O, P, S, F, Cl

Baza CSD została przeszukana z wykorzystaniem wszystkich filtrów dostępnych w programie ConQuest (dotyczących braku nieporządku, wielkości czynnika R itp.). Za pomocą definiowania odległości pomiędzy węglem i wodorem odfiltrowana została większość struktur całkowicie pozbawionych wodorów w miejscach, w których powinny być obecne.

University at Buffalo Databank (UBDB)

Bank UBDB składa się z typów atomów o zbliżonym otoczeniu chemicznym. Może służyć do:

- poprawy wyników procesu udokładniania modelu strukturalnego względem rentgenowskich danych dyfrakcyjnych metodą TAAM,
- szacowania elektrostatycznych właściwości cząsteczek.



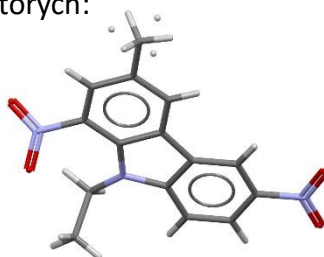
Rys. 1. Gęstość deformacyjna przykładowych typów atomowych w UBDB (dolna część rysunku) i zrekonstruowana z niej gęstość deformacyjna fragmentu cząsteczki (wiązanie peptydowe; górna część rysunku).

2. Odfiltrowanie struktur niepoprawnych chemicznie za pomocą filtrów zaprogramowanych w języku Ruby

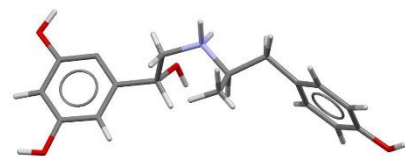
Pomimo zastosowania dostępnych w ConQuest filtrów wiele spośród wyselekcjonowanych struktur nadal była niepoprawnych chemicznie. Aby usunąć je z analizowanej bazy danych napisane zostały skrypty w języku Ruby. Filtry usuwały struktury, w których:

- Wodory były w nieporządku (Rys. 2)
- Struktura była pozbawiona części wodorów (Rys. 3)

Za pomocą zaprogramowanych filtrów usunięto 2,539% plików strukturalnych, pozostawiając tym samym 139175 struktur w bazie.



Rys. 2. Wodory w nieporządku



Rys. 3. Brak wodoru w strukturze

3. Przygotowanie nowych wpisów do bazy UBDB

Przygotowanie nowych wpisów do bazy UBDB odbywało się w kilku etapach:

- znalezienie najliczniejszej grupy nierozpoznanych typów atomów za pomocą programu TypeQuest autorstwa dr Michała Chodkiewicza;
- wybranie 10-15 plików strukturalnych zawierających typ atomu, który ma zostać opisany i przygotowanie wpisu do bazy zgodnie ze wzorem wpisu w formacie UBDB2021;
- sprawdzenie poprawności wpisu za pomocą testu z użyciem programu TypeQuest;
- przesłanie wpisu uzupełnionego wartościami domyślnymi parametrów multipolowych do członka zespołu odpowiedzialnego za ich dokładne obliczenie.

Na skutek wielokrotnego powtórzenia tej procedury statystyki rozpoznawalności dla pojedynczych atomów oraz całych molekuł w banku UBDB wzrosły (Tabela 1.).

Tabela 1. Poprawa statystyk UBDB na skutek wprowadzenia przygotowanych opisów typów atomowych do banku

	UBDB2018	UBDB2018+AS	UBDB2021-AS	UBDB2021
Rozpoznawalność pojedynczych atomów[%]	96,59	97,49 (▲0,9)	97,28 (▲0,69)	98,17 (▲1,58)
Rozpoznawalność całych molekuł[%]	48,31	56,87(▲8,56)	56,10 (▲7,79)	66,76 (▲18,45)
Liczba typów atomów w UBDB	463	552 (▲89)	562 (▲99)	651 (▲188)

Wnioski z przeprowadzonych badań:

- zidentyfikowano najczęściej występujące w CSD typy atomów, których wciąż brakowało w banku UBDB;
- dodano 188 nowych typów atomów, podnosząc statystyki rozpoznawalności banku o 1,58 punktu procentowego dla pojedynczych atomów oraz o 18,45 punktu procentowego dla całych molekuł;
- Zaprojektowano powtarzalną procedurę, która może być używana do dalszej poprawy wydajności banku UBDB.