

## Klasyfikacja białek.

Implementacja API dla baz Klasyfikujących struktury białkowe:

SCOP2 (<http://scop2.mrc-lmb.cam.ac.uk/graph/restapi.html>)

oraz funkcji umożliwiających pobranie najnowszej wersji bazy

CATH (<http://www.cathdb.info/download>)

### 1. SCOP i SCOP2

Niemal 20 lat temu powstała baza danych strukturalnej klasyfikacji białek (*ang. Structural Classification of Proteins – SCOP*) w Laboratorium Biologii Molekularnej i Centrum Inżynierii Białek w Cambridge. Baza ta gromadzi wiele informacji z badań nad strukturą białek oraz ich ewolucją. SCOP wraz z innymi bazami danych o strukturach białek, jak CATH, stały się cennymi zasobami i narzędziami badania tychże struktur. Pojęcie „ewolucji białka” zawarte w SCOP pozwoliło na dyskretne grupowanie białek w oparciu nie tylko o ich podobieństwo strukturalne, ale także o ich prawdopodobne pochodzenie ewolucyjne. Dyskretne jednostki, domeny, zostały pogrupowane hierarchicznie na podstawie ich wspólnych zależności strukturalnych i ewolucyjnych. Zgodnie ze stopniem ewolucyjnej rozbieżności i strukturalnego podobieństwa, SCOP zorganizował domeny białkowe w rodzinę i nadrodzinę. Zostały one następnie pogrupowane w „strukturalne ufałdowanie” (*ang. folds*), które niekoniecznie wskazywały na wspólne pochodzenie ewolucyjne i klasy odzwierciedlające struktury drugorzędne domen. Każda grupa w klasyfikacji była wynikiem starannej, indywidualnej analizy struktur białkowych i szczegółowej znajomości funkcji i ewolucji białek.

Oryginalna klasyfikacja SCOP opierała się na kilku założeniach:

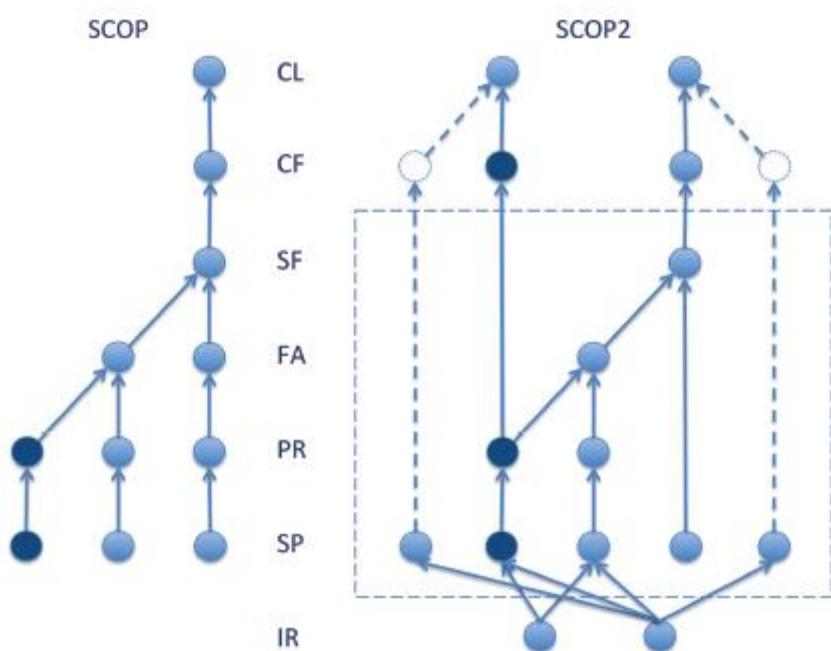
- sekwencje białek pełniących tę samą funkcję molekularną odbiegały od specjalizacji organizmów,
- dana sekwencja białkowa może mieć tylko jedną zwiniętą strukturę „natywną”,
- białka homologiczne składają się na podobne struktury,
- struktury białkowe są ewolucyjnie bardziej konserwatywne niż sekwencje
- białka niezależnych linii ewolucyjnych mogą dzielić wspólną fałdę.

Główym celem bazy SCOP było wsparcie biologów molekularnych w analizie i badaniu strukturalnych podobieństw białek. Prosta hierarchiczna klasyfikacja wspierała rozwój narzędzi i algorytmów i była z powodzeniem stosowana przez wiele aplikacji. Baza znalazła zastosowanie też do innych obszarów badań nad białkami, takich jak przewidywanie struktury białka i analizy genomu na dużą skalę. SCOP został również użyty do przewidywania oddziaływań białko-białko, dopasowania struktury białka z enzymatyczną aktywnością i innych badań.

SCOP2 to nowe podejście do klasyfikacji białek i ma na celu stawienie czoła przeszkodom i niespójnościom wynikającym z licznych przykładów nietrywialnych związków białkowych. SCOP zachowuje najlepsze cechy starej bazy SCOP, ale różni się w kilku kluczowych aspektach i dostarcza nowych danych niedostępnych w starym źródle. Podsumowując, nowy projekt klasyfikacji ma zapewnić nowe postępy w tej dziedzinie i otworzyć nowe kierunki badań.

## 2. Klasyfikacja białek wg SCOP i SCOP2

Podobnie jak w przypadku SCOP, głównym celem SCOP2 jest oparta na wiedzy specjalistyczna analiza i klasyfikacja białek, które są scharakteryzowane strukturalnie i zdeponowane w banku białek (PDB). Białka są zorganizowane zgodnie ze związkami strukturalnymi i ewolucyjnymi, ale w odróżnieniu od SCOP, zamiast prostej hierarchii drzewiastej, związków te tworzą złożoną sieć węzłów (Rys. 1). Klasyfikacja białek jest opisana za pomocą ukierunkowanego wykresu, w którym każdy węzeł definiuje zależność określonego typu i jest zilustrowany przez region struktury i sekwencji białka. Co ważne, w węźle podlegającym może być więcej niż jeden węzeł rodzicielski, który pozwala na wiele tras do określonej relacji



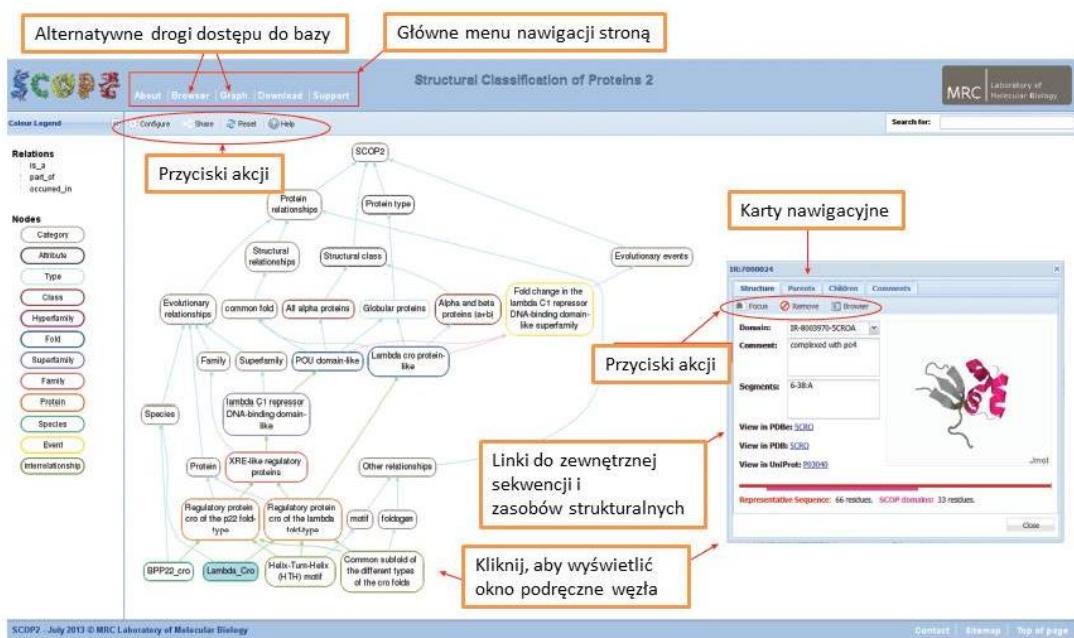
Rys. 1. Wykres zależności klasyfikacji struktur wg SCOP i SCOP2.

W bazie SCOP mamy klasyfikację na sześć poziomów: gatunek białka (*ang. Protein Species - SP*), białko (*ang. Protein - PR*), rodzina (*ang. Family - FA*), nadrodzina (*ang. Superfamily - SF*), fałdowanie (*ang. Fold - CF*) i klasa (*ang. Class - CL*). W SCOP2 relacje strukturalne i ewolucyjne są rozdzielone, co umożliwia klasyfikację homologicznych białek do różnych rodzajów fałdowania i klas strukturalnych, zachowując je w tej samej ewolucyjnej rodzinie i nadrodzinie. Nowa kategoria charakterystyczna dla SCOP2 to „inne” relacje (*ang. Other Relationships – IR*). Relacje te nie są ograniczone do niehierarchicznych związków między homologicznymi i niehomologicznymi białkami, z różnymi fałdami dzielącymi dużą wspólną podstrukturę lub motyw. Zależności w SCOP2 dzielą się na cztery główne kategorie, z których dwie - typy białek i wydarzenia ewolucyjne, nie mają odpowiedników w SCOP. W kategorii typów białek dzielą się na cztery główne typy: rozpuszczalne, membranowe, fibrylarne i wewnętrznie nieuporządkowane, z których każdy w dużym stopniu koreluje z charakterystyczną sekwencją i cechami strukturalnymi. Klasyfikacja białek według ich typów rozwiązuje niespójną klasyfikację białek membranowych i białek typu coiled-coil, wcześniej zorganizowanych w ich własnych klasach i pozwala na umieszczenie ich we właściwych klasach strukturalnych. Białka należące do różnych typów, np. rozpuszczalne i białka membranowe, mogą

teraz dzielić wspólne struktury ufałdowania, a nawet mogą być homologiczne do siebie. Kategoria „zdarzenia ewolucyjne” została wprowadzona w celu ułatwienia adnotacji o różnych rearanżacjach strukturalnych. Trzecia z czterech głównych kategorii SCOP2 - klasy strukturalne, porządkuje ufałdowanie białkowe ściśle według ich drugorzędowej struktury. Czwarta kategoria w SCOP2, „relacje białek”, składa się z trzech podkategorii: relacji strukturalnych, ewolucyjnych i „innych”. „Inne” relacje, kategoria unikatowa dla SCOP2, ma na celu zdefiniowanie i opisanie relacji, takich jak wewnętrzne powtarzania strukturalne, wspólne motywy, które nie były przedmiotem klasyfikacji w bazie danych SCOP.

### 3. Wyszukiwanie w bazie SCOP2

Oprócz związków białkowych, schematycznie pokazanych na Rysunku 1, wykres zawiera dodatkowe kategorie „ontologii”. Wyświetlany wykres można rozwinąć lub zwinąć za pomocą okna wyskakującego. Wyskakujące okno węzła wyświetla również dodatkowe informacje o wybranym węźle, w szczególności powiązanych domenach strukturalnych i sekwencyjnych oraz ich granicach reprezentujących relację. W tym przykładzie (Rys. 2) domena, reprezentująca wspólną strukturę ufałdowania różnych typów Cro, przedstawiona jest w kontekście pełnej długości sekwencji i struktury białka.



Rys. 2. Zrzut ekranu strony internetowej SCOP2 - graf przedstawiający połączony wykres przodków dwóch ortologicznych represorów Cro z bakteriofagów lambda i p22.

### 4. Oprogramowanie do klasyfikacji białek

Opracowany program pozwala na klasyfikację białek na poziomie Protein – Pr. Użytkownik może dokonać analizy białek na poziomie drzewka hierarchii protein. Dla danych pdb\_code (pdb\_id) użytkownik jest w stanie znaleźć klasyfikację białka. Można proces wyszukiwania wykonać także dla wielu białek na raz by potem mieć możliwość porównania ich klasyfikacji.

Możliwy jest także dostęp do wszystkich węzłów hierarchii oraz stworzenie dużo bardziej skomplikowanego podziału białek poczynając od ich formy aż do wszystkich innych zależności między białkami IR – nowa funkcja dodana w SCOP2.

## 5) Szczegółowy opis użytkowania API.

Wszystkie zwracane informacje z bazy danych SCOP2 są zwracane w postaci list słowniów tak by można było bez problemu ich używać do własnych celów takich jak badanie klasyfikacji białka czy poznawanie struktur białka.

Główne funkcje to:

1) Użytkownik może pobrać wszystkie informacje na temat dostępnych id białek

- id\_domain – zapis dostępu do bazy danych SCOP2

- pdb\_code – jest to uniwersalny klucz do danego białka

- pdb\_chain – łańcuch białka

- UnitProt\_id – jest to id za pomocą którego można wyszukać grupę spokrewnionych białek

- Sekwencja – są to sekwencje białek czy najczęściej się powtarzająca sekwencja w danej grupie białek, która świadczy o ich przynależności do grupy czy podobieństwie

A) Użytkownik może wyszukać wszystkie informacje na temat hierarchii IR – interrelationship

Funkcja:

```
scop2_get_interrelationship_id():
```

zwraca listę wszystkich informacji na temat grupy IR w bazie danych SCOP2 dzięki czemu użytkownik może zagłębić się dalej w podgrupy jakie wchodzą w jej skład

B) Użytkownik może wyszukać wszystkie informacje na temat hierarchii SP która jest dzieckiem gałęzi Protein, a ojcem IR.

Funkcja:

```
scop2_get_protein_species_id():
```

zwraca listę wszystkich informacji na temat grupy SP w bazie danych SCOP2 dzięki czemu użytkownik może zagłębić się dalej w podgrupy jakie wchodzą w jej skład

C) Użytkownik może wyszukać wszystkie informacje na temat PR – białek. Dowiedzieć się, które białka zostały dodane do bazy SCOP2 by dalej mieć możliwość ich analizy. PR jest ojcem SP lub IR

Funkcja:

```
scop2_get_proteins_id():
```

zwraca listę wszystkich informacji na temat grupy IR w bazie danych SCOP2 dzięki czemu użytkownik może zagłębić się dalej w podgrupy jakie wchodzą w jej skład

D) Analogicznie dla FA, SF, CF.

Funkcje:

```
scop2_get_families_id()  
scop2_get_super_families_id()  
scop2_get_fold_id()
```

Zwracaja jak powyżej listę słowników informacji na temat danych podgrup białek.

E) Ponadto została dodana funkcja umożliwiająca pobranie wszystkich Node\_ID oraz odpowiadającym im nazwom co umożliwia konstruwanie przez użytkownika dowolnych mu hierarchii, czy klasyfikacji białek.

Funkcja:

```
def scop2_get_nodes_names():
```

Zwraca listę słowników zawierających nazwy poszczególnych Node\_id.

Dostępna jest także funkcja wspomagająca wyszukującą nazwę dla podanego Node\_ID

Funkcja:

```
def scop2_get_nodes_names_by_id(node_id):
```

gdzie node\_id jest to string zawierajacy id danego węzła dzięki czemu jest możliwość sprawdzenia za co on odpowiada. Jest to niezbędne do konstrukcji klasyfikacji białek.

2) Funkcje umożliwiające pobranie wszystkich danych na temat danej wyszukiwanej hierarchii, grupy, podgrupach czy wyspecjalizowanemu białku. Użytkownik ma do dostępu 3 funkcje:

A) Funkcja wyszukiwania według podanej domain\_id, domain\_id jest to unikalny identyfikator dla bazy danych SCOP2. Identyfikatorem domain\_id oznaczone są wszystkie grupy hierarchii opisane na powyższym drzewie. Za pomocą ich identyfikatorów użytkownik może otrzymać wszystkie dane na temat danej grupy. Jedną z najważniejszych rzeczy jest możliwość otrzymania node\_id, które umożliwia głębsze dokładniejsze poszukiwania.

Funkcja:

```
scop2_get_domain_data_by_domain_id(domain_id):
```

Przyjmuje domain\_id jako string i zwraca listę słowników zawierającą wszystkie informacje na temat poszukiwanej domeny. Domain\_id są to unikalne identyfikatory domen z drzewa hierarchii

IR – interrelationships

SP – protein species

PR – protein

FA, SF, CF (opis powyżej)

B) Funkcja pobrania drzewa relacji całej struktury danej grupy, hierarchii białka. Za pomocą term\_id użytkownik może pobrać całą strukturę, z której składa się wyszukiwany obiekt. Funkcja zwraca wszystkie węzły spokrewnione z danym białkiem, które budują jego klasyfikacje.

Funkcja:

```
scop2_get_ontology_tree_by_term_id(term_id):
```

Zwraca listę słowników wszystkich węzłów opisujących poszukiwaną strukturę wg term\_id. Term\_id jest to string.

C) Funkcja umożliwiająca pobranie ojca, dzieci oraz głównego opisu danego węzła. Użytkownik ma możliwość sprawdzenia dokładnie relacji między węzłami za pomocą term\_id.

Funkcja:

```
scop2_get_term_data_by_term_id(term_id):
```

Zwraca listę słowników zawierających relacje dziecko ojciec danego węzła oraz główne informacje na temat wyszukiwanego obiektu

3) Została wprowadzona możliwość wyszukiwania białek i klasyfikacji ich za pomocą uniwersalnego pdb\_code. Użytkownik może wyszukiwać białko oraz listy białek i otrzymuje ich pełną klasyfikację.

A) Jeżeli użytkownik znajdzie jakiś pdb\_code interesującego go białka może otrzymać odowiadające mu node\_id (term\_id) oraz domain\_id, którymi posługuje się baza SCOP2

Funkcja:

```
get_domain_id_list_by_pdb_code(pdb_code):
```

Zwraca listę słowników zawierających domain\_id, term\_id.

4) Klasyfikacja białek za pomocą podanego PDB\_CODE oraz Term\_id

Użytkownik może klasyfikować białka za pomocą obu kluczów dostępu używając funkcji

Funkcje:

```
scop2_get_IR_protein_classification_by_protein_term_id(term_id):
```

term\_id odpowiada id dla grupy białek czyli np. "PR:5000091"

```
scop2_get_IR_protein_classification_by_protein_term_id(pdb_code):
```

```
scop2_get_IR_multiple_protein_classification_by_protein_term_id(pdb_code):
```

Powysze funkcje zwracaj¹ listê sloowników zawieraj¹ca pełn¹ klasyfikacj¹ poszukiwanego bia³ka.  
Ostatnia funkcja mo¿e przyj¹c listê bia³ek i zwróci ich klasyfikacj¹.

## CATHDB

Baza danych CATH jest bezpłatnym, publicznie dostępnym zasobem online, który dostarcza informacji na temat ewolucyjnych związków między domenami białkowymi. Została stworzona w połowie lat dziewięćdziesiątych przez profesor Christine Orengo i jego współpracowników i nadal jest rozwijana przez grupę Orengo w University College London.

Określone doświadczalnie trójwymiarowe struktury białkowe otrzymuje się z banku białek (PDB) i w razie potrzeby rozdziela się na ich kolejne łańcuchy polipeptydowe. Domeny białkowe są identyfikowane w tych łańcuchach przy użyciu mieszaniny metod automatycznych i ręcznego wyznaczania. Domeny te są następnie klasyfikowane w hierarchii strukturalnej CATH:

- na poziomie Klasy (C): domeny są przypisywane zgodnie z ich strukturą drugorzędową, tj. wszystkie alfa, wszystkie beta, mieszanina alfa i beta lub małe struktury drugorzędowe,
- na poziomie Architektury (A): informacje o układzie struktury drugorzędnej w przestrzeni trójwymiarowej są wykorzystywane do przypisania,
- na poziomie Topologii / Złożenia (T): wykorzystuje się informacje o sposobie łączenia i rozmieszczenia drugorzędowych elementów struktur,
- na poziomie nadrodziny homologicznej (H): jeśli istnieją dobre dowody na to, że domeny są powiązane przez ewolucję, tj. są homologiczne.

Dodatkowe dane sekwencji dla domen bez eksperymentalnie określonych struktur są dostarczane przez bazę Gene3D, która jest używana do zapełniania homologicznych nadrodzin. Wykorzystuje się również sekwencje białkowe z UniProtKB i Ensembl w celu przewidywania granic sekwencji domeny i tworzenia homologicznych nadrodzin.

Funkcjonalność:

1) Użytkownik ma możliwość pobierania najnowszej wersji bazy danych CATHDB za pomocą funkcji:

```
cath_download_pdb_id(version="4_1_0", pdb_id="2zjp",
save_file_path="C:/Users/zerg/Downloads/")

cath_download_domain_id(version="4_1_0", domain_id="2zjpS01",
save_file_path="C:/Users/zerg/Downloads/")

cath_download_chain_id(version="4_1_0", chain_id="101mA00",
save_file_path="C:/Users/zerg/Downloads/")
```

Pliki sa zapisywane we wskazanym save\_file\_path. Możliwość pobierania baz danych dla łańcuchów chain\_id, dla domain\_id oraz dla pdb\_id pojedynczych białek.

2) Użytkownik ma także możliwość wyszukiwania wszystkich ID dla danej wersji bazy CATCHDB by mieć listę opcji do pobierania.

A) Wersje bazy CATCHDB

```
get_catch_versions():
```

Zwraca listę dostępnych wersji bazy CATCHDB

B) Funkcja do pobrania listy domain\_id . Wersja podana defaultowo można zmienić.

```
get_protein_domain_id_list(version="4_2_0"):
```

Zwraca listę domain\_id za pomocą, których użytkownik może potem pobierać bazy danych.

C) Funkcja do wyszukania chain\_id. Wersja podana defaultowo można zmienić.

```
get_protein_chain_id_list(version = "4_2_0", chain_character = ""):
```

Zwraca listę chain\_id dla podanego symbolu łańcucha. Dzięki czemu można pobierać bazy danych z CATCH DB za pomocą chain\_id

D) Funkcja do wyszukania pdb\_id. Wersja podana defaultowo można zmienić.

```
D) get_protein_pdb_id_list(version="4_2_0"):
```

Zwraca listę pdb\_id. Dzięki czemu można pobierać bazy danych z CATCH DB za pomocą pdb\_id

Użytkownik ma także możliwość pobrania informacji na temat danej domain\_id.

```
get_id_domain_short_description(version="4_2_0", id_domain=""):
```

Zwraca listę słowników zawierających informacje o danej domain\_id

```
get_id_domain_full_description(version="4_2_0", id_domain=""):
```

Zwraca pełne informacje o danej domain\_id. Uwaga funkcja może działać bardzo długo z powodu bardzo dużej ilości danych. Zwraca STRING.