POLSKO-JAPOŃSKA
WYŻSZA SZKOŁA
TECHNIK KOMPUTEROWYCH

Aleksandra Gruca

Bioinformatyczne bazy danych

WYDAWNICTWO
PJWSTK

Notka biograficzna

Aleksandra Gruca jest inżynierem, bioinformatykiem. Od początku swojej pracy naukowej koncentruje się na zagadnieniach związanych z zastosowaniem technik maszynowego uczenia i eksploracji danych dla celów analizy danych medycznych oraz biologicznych. Brała udział w projektach badawczych związanych z analizą danych pochodzących z mikormacierzy DNA, a w szczególności zajmuje się rozwijaniem metod i technik wspomagających proces interpretacji i opisu funkcjonalnego wyników eksperymentów biologicznych. Jest autorką lub współautorką ponad dwudziestu publikacji naukowych.

Streszczenie

Każdego roku w Internecie pojawia się ponad 100 nowych baz danych, które zawierają dane pochodzące z biologicznych i medycznych eksperymentów. Niniejsza książka ma za zadanie przedstawić czytelnikom najważniejsze z tych repozytoriów oraz omówić zagadnienia związane z przetwarzaniem danych w nich zawartych. Pierwsze dwa rozdziały tej książki wprowadzają czytelnika w zagadnienia bioinformatyki, nowej interdyscyplinarnej dziedziny wiedzy oraz w tematykę baz danych. Kolejne rozdziały zawierają przegląd najważniejszych bioinformatycznych baz danych wraz z opisem narzędzi powiązanych z tymi bazami danych. Przedstawiono najpopularniejsze bazy sekwencji nukleotydowych oraz sekwencji białkowych, a także metody przeszukiwania tych baz pod kątem sekwencji podobnych. Książka zawiera również przegląd baz danych rodzin białek oraz struktur białek, a także opis repozytoriów ukierunkowanych na funkcjonalną anotację genów lub białek. Książka jest przeznaczona przede wszystkim dla studentów lub pracowników naukowych kierunków technicznych i przyrodniczych, a w szczególności dla osób zainteresowanych bioinformatyką, które w ramach swojej pracy stykają się z analizą danych medycznych lub biologicznych

Seria: Podręczniki akademickie

Edytor serii: Leonard Bolc
Tom serii: 43

POLSKO-JAPOŃSKA
WYŻSZA SZKOŁA
TECHNIK KOMPUTEROWYCH

Aleksandra Gruca

Bioinformatyczne bazy danych

WYDAWNICTWO
PJWSTK

Edytor
Leonard Bolc

Redaktor techniczny
Ada Jedlińska

Korekta
Anna Bittner

Komputerowy skład tekstu
Grażyna Domańska-Żurek

Projekt okładki
Andrzej Pilich

Wydawnictwo Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych
ul. Koszykowa 86, 02-008 Warszawa
tel. 022 58 44 526, fax 022 58 44 503

Oprawa miękka
ISBN 978-83-89244-90-1

Wersja elektroniczna
ISBN 978-83-63103-51-4

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt „Nowoczesna kadra dla e-gospodarki" – program rozwoju Wydziału Zamiejscowego Informatyki w Bytomiu Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych współfinansowany przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach Podziałania 4.1.1 „Wzmocnienie potencjału dydaktycznego uczelni" Programu Operacyjnego Kapitał Ludzki

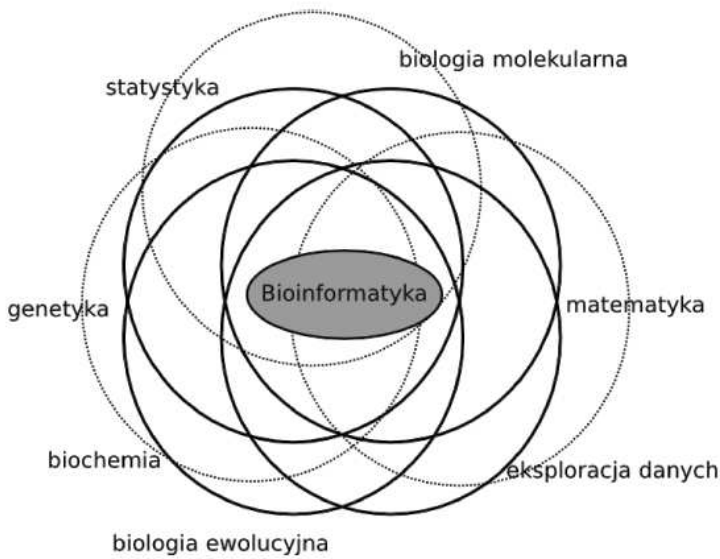
This book should be citated as:
Gruca A. 2010. Bioinformatyczne bazy danych. Warszawa:
Wydawnictwo PJWSTK.

Spis treści

1

Wstęp

Ogromny rozwój technologii badawczych w dziedzinie genomiki i biologii molekularnej, któremu towarzyszył równie dynamiczny rozwój technologii informacyjnych i przyrost mocy obliczeniowej komputerów, zaowocował powstaniem nowej dziedziny nauki, w której do przetwarzania danych biologicznych wykorzystywane są metody obliczeniowe. Bioinformatyka jest stosunkowo nową, interdyscyplinarną dziedziną wiedzy, która powstała na styku różnych odrębnych dziedzin nauki takich jak: biologia i ewolucja molekularna, biologia strukturalna, genetyka, genomika, proteomika, biochemia, statystyka, matematyka, informatyka czy eksploracja danych.



Rysunek 1.1. Bioinformatyka jest nauką integrującą różne dziedziny wiedzy

Na bioinformatykę składają się różne narzędzia i algorytmy pozwalające na badanie, rozwój i zastosowanie komputerowych metod, które wykorzystywane są w biologii do zdobywania, przetwarzania, organizowania, archiwizacji, analizy oraz wizualizacji danych biologicznych. Podstawowym celem tej dziedziny jest dostarczanie narzędzi matematycznych oraz metod komputerowych, w celu dokonywania odkryć biologicznych, które umożliwiają nam głębsze zrozumienie procesów i zależności biologicznych występujących w żywych organizmach. Nie jest nadużyciem stwierdzenie, że bez metod komputerowych dzisiejsza biologia molekularna nie mogłaby się rozwijać w dziedzinach takich jak: mapowanie sekwencji DNA, uliniowianie sekwencji DNA oraz sekwencji białkowych w celu przewidywania właściwości oraz funkcji cząsteczek, poszukiwanie i klasyfikacja rodzin białek, tworzenie przestrzennych modeli biomolekuł, i wiele innych. Tym samym nasza wiedza na temat budowy komórki oraz procesów biologicznych w niej zachodzących byłaby uboższa i znaczenie bardziej ograniczona.

Wyniki eksperymentów biologicznych oraz rezultaty przetwarzania tych wyników gromadzone są od wielu lat w różnych repozytoriach danych, których celem jest umożliwienie dostępu do aktualnej wiedzy biologicznej badaczom z całego świata. Z uwagi na specyficzny rodzaj danych znajdujących się w tych bazach, bazy te określa się pojęciem **bioinformatyczne bazy danych**.

Bioinformatyczne bazy danych to obecnie ogromne, zorganizowane zbiory danych, które pozwalają na przeszukiwanie dostępnych w repozytoriach informacji, przetwarzanie jej, a często również i przesyłanie nowych danych. Tego typu bazy charakteryzują się zazwyczaj prostym i intuicyjnym interfejsem, a także wyposażone są w oprogramowanie, które umożliwia wstępną analizę danych – często przecież korzystają z tych baz osoby, które nie posiadają kierunkowego wykształcenia informatycznego. Liczba dostępnych baz danych wzrasta z każdym rokiem, a ilość danych znajdujących się w najpopularniejszych z nich rośnie wykładniczo. Każdego roku wydawany jest specjalny numer czasopisma *Nucleid Acid Research* poświęcony tylko i wyłącznie bioinformatycznym bazom danych. Na stronach internetowych czasopisma (`http://www.oxfordjournals.org/nar/database/c/`) dostępna jest lista większości popularnych i uznanych bioinformatycznych baz danych, a także ich podział w zależności od typu danych, jaki jest w nich przechowywany. Podział przedstawiono poniżej, na podstawie danych dostępnych na początku 2009 roku [Galperin and Cochrane, 2009]:

- Bazy danych sekwencji nukleotydowych.
  - Sekwencje dostępne w ramach INSDC.
    - DDBJ – DNA Data Bank of Japan.
    - EMBL Nucleotide Sequence Database.
    - GenBank.
  - Bazy kodującego i niekodującego DNA.
  - Struktury genów, intronów, egzonów oraz miejsc splicingu.
  - Miejsca regulatorowe transkrypcji oraz czynników transkrypcji.

- Bazy sekwencji RNA.
- Bazy sekwencji białkowych.
  - Ogólne bazy sekwencji (baza białek NCBI, PIR, UniProt).
  - Bazy własności białek.
  - Bazy lokalizacji białek.
  - Bazy sekwencji, motywów i miejsc aktywnych.
  - Bazy domen i klasyfikacji białek.
  - Bazy rodzin białek.
- Bazy danych struktur (PDB).
- Bazy genomowe (bezkręgowców).
- Bazy ścieżek metabolicznych i ścieżek sygnałowych.
- Bazy genomowe ludzkie oraz kręgowców.
- Bazy ludzkich genów i chorób.
- Bazy danych mikromacierzowych i bazy danych ekspresji genów.
- Bazy zasobów proteomicznych.
- Pozostałe bazy biologii molekularnej.
- Bazy organelli.
- Bazy roślinne.
- Bazy immunologiczne.

Bioinformatyczne bazy danych mogą zawierać różnego rodzaju informacje. Jednakże niezależnie od typu bazy danych, każdy wpis najczęściej składa się z dwóch elementów: części opisowej zawierającej opis danych, anotacje i odnośniki do literatury oraz części głównej zawierającej sekwencję lub wyniki obserwacji. Niektóre bazy danych umożliwiają każdemu użytkownikowi niczym nieograniczone przesłanie wyników swoich eksperymentów biologicznych, inne wymagają, żeby każda informacja przed opublikowaniem została sprawdzona przez tak zwanych kuratorów bazy danych, którzy dbają o to, aby dane znajdujące się w bazie były poprawne, aktualne i zgodne z istniejącą wiedzą biologiczną. Nadzorowane bazy danych są bardziej wiarygodne, jednakże z uwagi na fakt, iż proces sprawdzania poprawności informacji jest czasochłonny, przyrost danych w takich bazach jest dość wolny, z kolei nienadzorowane bazy danych zawierają wyniki najnowszych eksperymentów – nie dając jednak żadnej gwarancji co do jakości danych użytkownikom, którzy takie dane w przyszłości z tej bazy pobierają. Pewnym kompromisem jest stosowanie różnego rodzaju automatycznych procedur kontroli jakości, które przynajmniej częściowo zastąpić mogą nadzór manualny.

Na początku 2009 roku na stronach czasopisma *Nucleid Acid Research* wymieniono 1170 bioinformatycznych baz danych, przy czym z każdym rokiem czasopismo rejestruje około 100 nowych repozytoriów. W niniejszym opracowaniu opisano jedynie niewielką część dostępnych repozytoriów. Skoncentrowano się tutaj na najważniejszych i najpopularniejszych bioinformatycznych bazach sekwencji nukleotydowych oraz sekwencji białkowych, wraz z opisem

metod poszukiwania sekwencji podobnych. W kolejnych rozdziałach przedstawiono przegląd bazy danych rodzin białek oraz bazy struktur białek, a także repozytoria ukierunkowane na funkcjonalną anotację genów lub białek.

2

Wprowadzenie do baz danych

Zbiór informacji, które zawierają ze sobą powiązania i zostały w pewien sposób skatalogowane, określamy mianem bazy danych. W najprostszej postaci bazą danych może być kartka papieru zawierająca listę nazwisk osób przyporządkowanych do ich miejsca zamieszkania. Oczywiście w sytuacji, kiedy posiadanych danych jest coraz więcej, pojawia się potrzeba stworzenia efektywnego systemu, który pozwoli na szybkie wyszukiwanie i przetwarzanie posiadanych przez nas informacji. Stąd też obecnie, kiedy używamy terminu baza danych, mamy na myśli zarówno dane, jak i program komputerowy, który tymi danymi zarządza. System, który przechowuje dane i wyposażony jest w mechanizmy ich udostępniania określamy mianem Systemu Zarządzania Bazą Danych, SZBD (ang. *DataBase Management System, DBMS*).

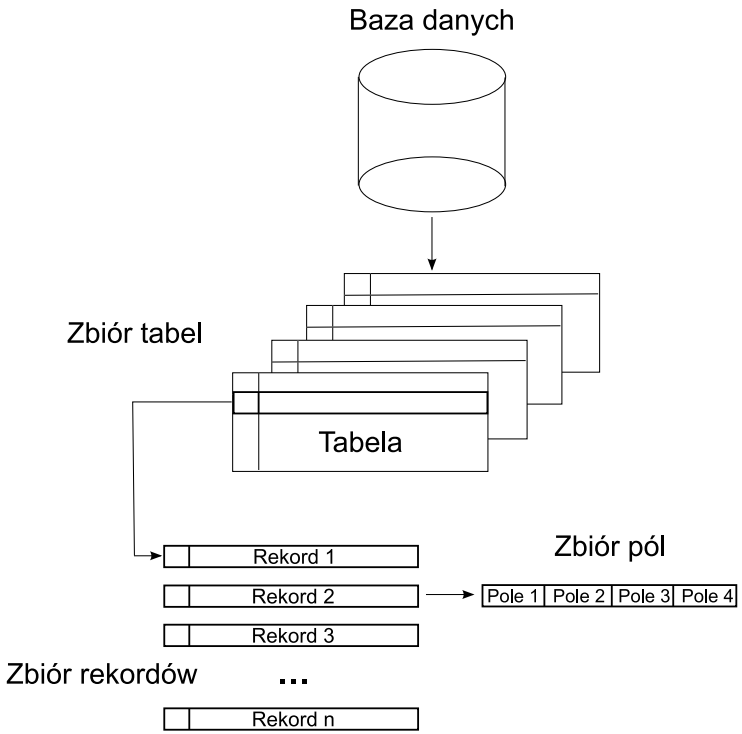
Podstawowe zadania, których oczekuje się od Systemu Zarządzania Bazą Danych są następujące [Garcia-Molina et al., 2006]:

- Umożliwienie użytkownikowi utworzenia nowej bazy danych i zdefiniowania jej struktury.
- Udostępnienie użytkownikowi możliwości pobierania oraz aktualizacji danych za pomocą odpowiedniego języka zapytań (ang. *query language*).
- Umożliwienie przechowywania ogromnych ilości danych oraz zabezpieczenie ich przed niepowołanym dostępem, a także umożliwienie efektywnego dostępu do danych.
- Zabezpieczenie przed utratą spójności danych w przypadku, gdy wielu użytkowników korzysta z bazy danych równocześnie.

Mówiąc o bazach danych oraz o danych w nich zgromadzonych, będziemy używali następujących pojęć:

- Tabela (ang. *table*) – zbiór rekordów tego samego typu.
- Rekord (ang. *record*) – podstawowa jednostka informacji, pojedynczy wpis w tabeli posiadający zdefiniowaną strukturę, będący opisem pewnego konkretnego obiektu.
- Pole (ang. *field*) – najmniejsza część rekordu, która zawiera niepodzielne dane.

Znaczenie powyższych pojęć w odniesieniu do ogólnego schematu bazy danych przedstawiono na rysunku 2.1.



Rysunek 2.1. Ogólny schemat bazy danych. W bazie danych znajdują się tabele przechowujące dane. Każda tabela zawiera rekordy, a każdy rekord składa się z pól.

2.1 Modele danych

W komputerowej bazie danych dane znajdują się w ściśle zdefiniowanych strukturach, które odpowiadają założonemu modelowi danych. Struktury, które znajdują się w bazie danych, pozwalają na tworzenie powiązań pomiędzy danymi. W zależności od sposobu organizacji danych możemy wyróżnić następujące podstawowe typy baz danych:

- Bazy kartotekowe (płaskie pliki) (ang. *flat files*).
- Bazy hierarchiczne (ang. *hierarchical databases*).
- Bazy sieciowe (ang. *network databases*).
- Bazy relacyjne (ang. *relational databases*).
- Bazy obiektowe (ang. *object-oriented databases*).

Pierwsze trzy z wymienionych typów baz danych są na obecną chwilę już rozwiązaniami historycznymi. Najprostszą, a równocześnie historycznie pierwszą formą bazy danych, jest baza kartotekowa, w postaci płaskiego pliku tekstowego, w którym każda linia zawiera odrębny wpis. Cechą charakterystyczną takiego rozwiązania jest brak powiązań pomiędzy poszczególnymi plikami. W związku z tym, jeśli chcemy odszukać konkretny rekord, musimy za każdym razem przeglądać plik od góry. W przypadku baz hierarchicznych model danych reprezentowany jest w postaci struktury drzewiastej. Węzły tego drzewa odpowiadają wpisom. Węzły będące wyżej w hierarchii łączone są z węzłami będącymi niżej relacją jeden-do-wielu (każdy z węzłów-rodziców może mieć wielu potomków, natomiast każdy z potomków może mieć tylko jednego rodzica). Struktura taka umożliwia szybkie wyszukiwanie informacji zgromadzonej w bazie, natomiast do wad tego rozwiązania można zaliczyć problemy związane ze zmianą struktury bazy. Sieciowy model danych zbudowany został na podstawie modelu hierarchicznego, który został rozszerzony o możliwość definiowania relacji wiele-do-wielu pomiędzy węzłami będącymi niżej oraz wyżej w hierarchii. W relacyjnych bazach danych dane reprezentowane są w postaci tabel, pomiędzy którymi istnieją powiązania. Historycznie najnowszym spośród wymienionych modeli jest obiektowy model danych, w którym dane modelowane są za pomocą zbioru powiązanych obiektów, będących pewnymi bytami posiadającymi atrybuty (wartości) oraz metody (funkcje stosowane na obiektach). Obecnie również coraz większą popularność zyskują obiektowo–relacyjne modele danych, które łączą w sobie zalety podejścia obiektowego oraz relacyjnego.

Bazy danych można również podzielić ze względu na zawartość danych, które są w nich przechowywane. Standardowe postacie danych, jakie przechowywane są w bazach danych, to proste dane typu tekstowego (ciągi znaków) lub liczbowego. Istnieją jednak również bazy danych przechowujące dane multimedialne takie jak pliki audio, wideo czy obrazy. Wspólną cechą tego typu danych jest ich rozmiar, stąd też bazy przechowujące tego typu dane muszą być wyposażone w specjalne mechanizmy ich przetwarzania.

2.2 Relacyjne bazy danych

Z uwagi na fakt, iż relacyjne bazy danych są obecnie najpopularniejszym podejściem, model danych w nich występujący zostanie omówiony dokładniej w niniejszym podrozdziale.

Relacyjny model danych został zaproponowany w 1970 roku przez Edgara Codda i praktycznie, od drugiej połowy lat 80-tych, stał się podstawą architektury większości systemów baz danych. Bazuje on na pojęciu *relacji*, czyli pewnej abstrakcji intuicyjnego pojęcia dwuwymiarowej tabeli, która zawiera dane. Przykładem relacji jest na przykład tabela *Pracownik* (relacja *Pracownik*), zawierająca listę osób wraz z ich adresem zamieszkania, przedstawiona

na rysunku 2.2. W dalszej części tego rozdziału terminy relacja oraz tabela używane będą zamiennie.

Imię	Nazwisko	Ulica	Numer	Miasto
Jan	Adamek	Sosnowa	18	Katowice
Adam	Jankowski	Sezamkowa	33	Gliwice
Zofia	Kowalska	Wesoła	15	Warszawa
Joanna	Kwiatkowska	Krótka	20	Zabrze
Piotr	Nowak	Sasanek	36	Katowice

Rysunek 2.2. Relacja *Pracownik*

Nagłówki relacji noszą nazwę *atrybutów*. Atrybuty najczęściej odpowiadają tytułom kolumn relacji i odzwierciedlają w swojej nazwie opis danych, jakie znajdują się w kolumnach. W powyższej relacji atrybuty to: *imię*, *nazwisko*, *ulica*, *numer*, *miasto*. Nazwa relacji oraz zbiór jej atrybutów nazywane są *schematem relacji*. W przypadku podanego przykładu schemat relacji jest następujący:

```
Pracownik(imię, nazwisko, ulica, numer, miasto).
```

Wiersze relacji, czyli rekordy tabeli, nazywane są krotkami. Każdy z atrybutów ma swój odpowiednik w postaci tzw. składowej krotki, tzn. pierwsza krotka przedstawiona na rysunku 2.2 ma pięć składowych: *Jan*, *Adamek*, *Sosnowa*, *18*, *Katowice*, które są kolejnymi wartościami atrybutów *imię*, *nazwisko*, *ulica*, *numer* i *miasto*.

W relacyjnym modelu danych, kolejność atrybutów w relacji ani kolejność krotek nie ma znaczenia. Oznacza to, że ich kolejność może być dowolnie przestawiana, a wszystkie relacje, które są kombinacjami danego zbioru krotek oraz atrybutów, są sobie równoważne. Oczywiście należy pamiętać, że jeśli zmieniamy kolejność atrybutów w schemacie relacji, należy również zmienić kolejność składowych krotek.

Schemat relacji definiowany jest przez nazwę relacji oraz kolejność atrybutów i jest praktycznie niezmienny. Natomiast zbiór istniejących krotek nazywamy *instancją relacji*. Instancja relacji zwykle jest modyfikowana i zmienia się w czasie – na przykład poprzez dodanie nowych krotek, edycję krotek już istniejących lub usunięcie krotek.

Klucze główne

Ponieważ kolejność krotek w relacji nie ma znaczenia, musi istnieć jakiś sposób, który pozwoliłby na identyfikację konkretnego wiersza tabeli. Rolę taką spełniają tak zwane klucze główne (inaczej klucze podstawowe), czyli specjalne atrybuty, które w danej relacji pozwalają na jednoznaczną identyfikację wiersza.

W przykładowej relacji pokazanej na rysunku 2.2, takim atrybutem mogłoby być np. nazwisko. Oczywiście nietrudno wyobrazić sobie, że w tabeli zawierającej listę osób, prędzej czy później pojawi się osoba o nazwisku takim, jakie już w tabeli istnieje. W tej sytuacji można tworzyć klucze złożone – na przykład poprzez wybranie zbioru zawierającego kilka atrybutów.

W praktycznych zastosowaniach wykorzystuje się zazwyczaj klucze pojedyncze (składające się z jednego atrybutu). Przykładem takiego klucza może być na przykład numer PESEL, który pozwoli nam na jednoznaczne zidentyfikowanie konkretnej osoby. Najczęściej jednak do tabeli dołącza się dodatkowy, specjalny atrybut, który pełni rolę klucza głównego. Atrybut taki musi posiadać następujące własności:

- Musi unikalnie identyfikować każdy wiersz.
- Musi posiadać wartość dla każdego z wierszy, w szczególności nie może przyjmować wartości pustej, tak zwanej wartości NULL.
- Wartość tego atrybutu pozostaje niezmienna od momentu utworzenia rekordu i nie może być usunięta, jeśli rekord z nią powiązany istnieje w tabeli.

W większości bioinformatycznych baz danych, które omawiane będą w niniejszym opracowaniu, rekordy posiadają unikalny identyfikator zwany numerem dostępu (ang. *accession number*). Identyfikator taki może w tabeli pełnić rolę klucza głównego. Dodatkową jego zaletą jest fakt, że zawsze pozwala odnaleźć dany wpis (np. sekwencję nukleotydową lub aminokwasową) nawet, jeśli w miarę upływu czasu zawartość rekordu jest aktualizowana.

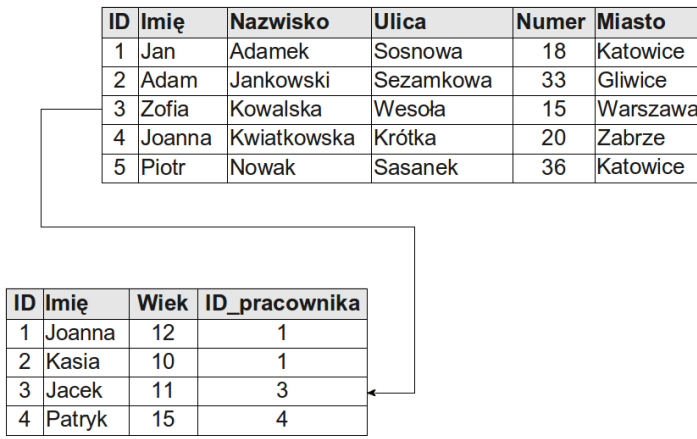
Klucze obce

Klucz główny jednej tabeli może zostać umieszczony w innej tabeli. Tworzy on wówczas, w tej innej tabeli, tak zwany klucz obcy (ang. *foreign key*), który pozwala na łączenie tabel pomiędzy sobą i tworzenie pomiędzy nimi powiązań.

Wyobraźmy sobie, że dla naszej przykładowej relacji *Pracownik* poza danymi o pracowniku, chcemy również w bazie danych umieścić informacje na temat posiadanych przez niego dzieci. Jednym ze sposobów mogłoby być dodanie do tabeli *Pracownik* nowych atrybutów, np. *imię_dziecka* i *wiek_dziecka*. Jednakże okazuje się, że nie jest to dobre rozwiązanie – w przypadku kiedy pracownik posiada więcej niż jedno dziecko, będziemy zmuszeni do powielenia danych znajdujących się w rekordzie tyle razy, ile dany pracownik posiada dzieci. Ponieważ rozwiązanie takie jest bardzo nieefektywne i rodzi wiele problemów (na przykład podczas aktualizacji danych musielibyśmy aktualizować wszystkie powtarzające się rekordy), w praktyce się go nie stosuje. Zamiast tego tworzy się osobną tabelę (na przykład tabelę *Dziecko*), w której znajdują się informacje na temat posiadanych dzieci, a następnie umieszcza się w niej specjalne atrybuty zwane kluczami obcymi, które pozwalają na stworzenia powiązań pomiędzy tabelami.

Schematyczny przykład, pokazujący w jaki sposób klucze główne i obce mogą zostać wykorzystane do utworzenia powiązania pomiędzy relacją *Pracownik*, a relacją *Dziecko*, przedstawiono na rysunku 2.3. Jak widać, do relacji

Pracownik dołożony został nowy atrybut *ID*, będący kluczem głównym tej relacji. Atrybut ten jest równocześnie kluczem obcym relacji *Dziecko* i występuje w niej jako atrybut *ID_pracownika*. Wiedząc, że atrybut *ID_pracownika* w tabeli *Dziecko* odpowiada dokładnie atrybutowi *ID* w tabeli *Pracownik*, w łatwy sposób możemy wzajemnie przyporządkować sobie rekordy z obydwu tabel.



ID	Imię	Nazwisko	Ulica	Numer	Miasto
1	Jan	Adamek	Sosnowa	18	Katowice
2	Adam	Jankowski	Sezamkowa	33	Gliwice
3	Zofia	Kowalska	Wesoła	15	Warszawa
4	Joanna	Kwiatkowska	Krótka	20	Zabrze
5	Piotr	Nowak	Sasanek	36	Katowice

ID	Imię	Wiek	ID_pracownika
1	Joanna	12	1
2	Kasia	10	1
3	Jacek	11	3
4	Patryk	15	4

Rysunek 2.3. Powiązanie typu 1:N pomiędzy tabelami *Pracownik* oraz *Dziecko*. Atrybut ID będący kluczem głównym tabeli *Pracownik* jest kluczem obcym w tabeli *Dziecko*.

Relacje (powiązania) pomiędzy tabelami mogą być następującego typu:

- 1:1 – relacja jeden-do-jeden – jednemu rekordowi w tabeli A odpowiada dokładnie jeden rekord w tabeli B. Tego typu powiązania spotykane są rzadko, gdyż w takiej sytuacji odpowiadające sobie rekordy najczęściej umieszcza się w jednej tabeli.
- 1:N – relacja jeden-do-wielu – rekordowi w tabeli A odpowiada wiele rekordów w tabeli B, zaś jeden rekord w tabeli B ma przyporządkowany dokładnie jeden rekord w tabeli A. Przykładem takiej relacji może być wspomniana wyżej relacja typu rodzic-dzieci.
- M:N – relacja wiele-do-wielu – rekord w tabeli A może być przyporządkowany do wielu rekordów w tabeli B, a równocześnie do jednego rekordu w tabeli B może być przyporządkowanych wiele rekordów z tabeli A. Przykładem takiej relacji może być powiązanie pomiędzy pracownikami oraz zadaniami. Jeden pracownik może być przyporządkowany do wielu zadań, tak samo, jak jedno zadanie może być wykonywane przez wielu pracowników.

Transakcje

Transakcja jest pewnym zbiorem operacji na bazie danych, które stanowią pewną całość. Oznacza to, że wszystkie operacje w ramach transakcji powinny być wykonane od początku do końca, lub nie powinna zostać wykonana żadna z nich. Przykładem transakcji może być przelew pieniędzy z jednego konta na drugie, który składa się z dwóch operacji: pobranie pieniędzy z pierwszego konta i zaksięgowanie pobranej kwoty na drugim koncie. W przypadku jeśli wystąpią problemy z realizacją transakcji, żadna z tych operacji nie powinna zostać wykonana ponieważ wykonanie tylko jednej z nich spowodowałoby wystąpienie nieprawidłowości w bazie – pojawienie się lub zniknięcie pieniędzy.

Transakcje, które wykonywane są w bazie danych powinny spełniać tak zwane warunki ACID: ***a**tomicity* – atomowość, ***c**onsistency* – spójność, ***i**solation* – izolacja, ***d**urability* – trwałość. Atomowość transakcji oznacza jej niepodzielność, czyli że każda transakcja powinna zostać wykonana w całości albo wcale. Spójność oznacza, że po wykonaniu transakcji nie zostaną naruszone zasady integralności – na przykład saldo konta nie powinno być ujemne. Izolacja to cecha, która mówi, że jeśli w danym czasie wykonywanych jest kilka transakcji, to każda z nich powinna się wykonywać oddzielnie, tak jakby była jedyną transakcją wykonywaną w danej chwili w bazie. Trwałość transakcji oznacza, że jeśli dana transakcja zostanie wykonana poprawnie, to efekty jej wykonania zostaną zapisane w bazie.

2.3 Definiowanie zapytań

Język SQL

Język SQL – Strukturalny Język Zapytań (ang. *Structured Query Language*) umożliwia definiowanie zapytań, które pozwalają na dostęp do danych, a w szczególności na ich odczyt oraz modyfikację. Jest on również narzędziem, które pozwala na zarządzanie bazą danych i wykonywanie wszelkich czynności związanych z administrowaniem bazą danych.

Przykładowe polecenie w języku SQL, który w relacji *Pracownik* wyszukuje wszystkie osoby zamieszkałe w Katowicach, jest następujące:

```
SELECT * FROM pracownik WHERE miasto="Katowice"
```

W wyniku wykonania powyższego zapytania zwrócone będą dwie krotki: (*Jan*, *Adamek*, *Sosnowa*, *18*, *Katowice*) oraz (*Piotr*, *Nowak*, *Sasanek*, *38*, *Katowice*).

Przedstawiony przykład instrukcji SELECT zawiera trzy słowa kluczowe: SELECT, FROM oraz WHERE. Słowo kluczowe SELECT określa typ instrukcji. Symbol gwiazdki (∗) oznacza, że zapytanie powinno zwrócić cały rekord spełniający dane warunki. Oczywiście nie zawsze interesują nas wszystkie atrybuty – w takim przypadku należy po słowie kluczowym SELECT umieścić nazwy tych atrybutów, które powinny zostać zwrócone w wyniku zapytania,

rozdzielone przecinkami. Po słowie kluczowym FROM określamy tabelę, z której chcemy pobrać dane, natomiast po słowie kluczowym WHERE definiujemy wyrażenie logiczne, które określa warunki, jakie muszą spełniać atrybuty, aby dany rekord został zwrócony w wyniku zapytania.

Inne typowe instrukcje języka SQL to: INSERT – dodadnie nowego rekordu do istniejącej tabeli, UPDATE – aktualizacja istniejącgo rekordu, DELETE – usunięcie istniejącego rekordu. Oprócz komend związanych z operacjami na tabelach, język SQL umożliwia również wykonywanie takich czynności jak tworzenie i usuwanie baz danych, tworzenie oraz usuwanie tabel w bazie czy też dodawanie użytkowników do bazy i definiowanie ich uprawnień.

Formularze WWW

Z uwagi na kwestie związane z bezpieczeństwem danych, zewnętrzni użytkownicy bioinformatycznych baz danych zazwyczaj nie mają możliwości bezpośredniego połączenia się z bazą danych i formułowania zapytań w języku SQL. Najczęstszym rozwiązaniem jest zapewnienie użytkownikowi dostępu do danych poprzez specjalne serwisy internetowe, zawierające odpowiednio skonstruowane formularze WWW, które pozwalają na wybór atrybutów, których wartości nas interesują oraz definiowanie warunków zapytania. Większość bioinformatycznych baz danych, które omawiane będą w niniejszej pracy, udostępnia swoje zasoby właśnie w taki sposób. Często również, poza możliwością pobierania danych, niektóre z baz udostępniają również dodatkowo formularze, które pozwalają na edycję danych i dodawanie nowych rekordów do tabel już istniejących.

Metody zadawania zapytań oparte o formularze WWW są bardzo wygodne, ale sprawdzają się w przypadku, jeśli ilość informacji, jaką chcemy pobrać z bazy, jest niewielka – na przykład interesują nas informacje na temat pojedynczej sekwencji. Jednak w sytuacji, kiedy ilość przetwarzanych danych jest duża, korzystanie z formularzy WWW okazuje się być nieefektywne i czasochłonne. Dlatego, oprócz formularzy WWW, część bioinformatycznych baz danych udostępnia także specjalne programy i narzędzia pozwalające na definiowanie zaawansowanych zapytań, które pozwalają na równoczesne przetwarzanie większej ilości rekordów. W zależności od potrzeb, stosując takie narzędzia, możemy w jednym zapytaniu pobrać, zaktualizować lub dodać dużą liczbę rekordów do bazy.

Przykładem narzędzia, które pozwala na zdefiniowanie złożonego zapytania, które w wyniku może zwrócić wiele rekordów jest narzędzie *Entrez Programming Utilities (E-utilities)* dostępne w ramach zbioru baz danych Entrez. W tym wypadku użytkownik przesyła odpowiednie zapytanie do bazy poprzez adres internetowy, który tworzony jest według ściśle zdefiniowanych reguł, a w wyniku otrzymuje listę rekordów w formacie XML, które spełniają zadane kryteria. Z kolei przykładem aplikacji, która pozwala na równoczesną edycję lub dodawanie wielu rekordów do bazy danych, jest program *Sequin*, dostępny w ramach bazy sekwencji nukleotydowych GenBank.

3

Bazy danych sekwencji nukleotydowych

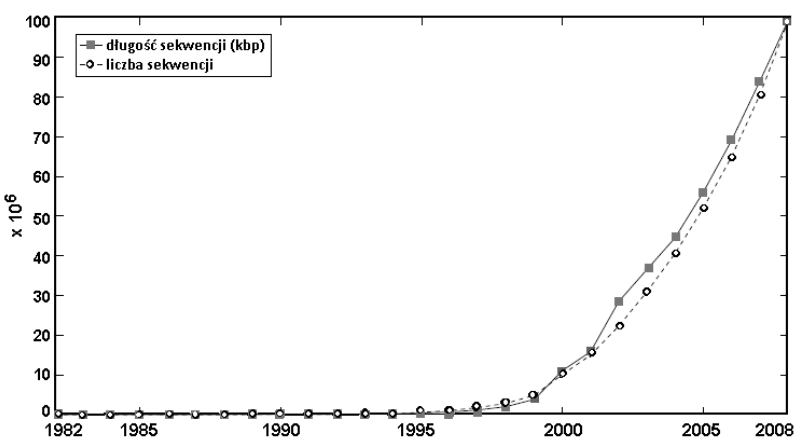
Każde badania, które mają na celu odkrycie funkcji oraz struktury dowolnej biocząsteczki, rozpoczynają się obecnie od określenia jej sekwencji nukleotydowej. Można się spodziewać, że cząsteczki charakteryzujące się podobną sekwencją nukleotydową będą miały podobne właściwości biologiczne i fizykochemiczne. Także porównywanie sekwencji DNA poszczególnych organizmów pozwala nam prześledzić oraz poznać mechanizmy ewolucji gatunków. Stąd też we współczesnej biologii molekularnej ogromny nacisk kładziony jest na gromadzenie oraz udostępnianie odkrytych już sekwencji nukleotydowych tak, aby badacze z całego świata mogli zgromadzoną już wcześniej informację pobierać oraz wykorzystywać w trakcie swoich aktualnych badań. Obecnie na świecie istnieją trzy podstawowe bazy danych, które gromadzą oraz udostępniają niemal wszystkie dotychczas odkryte sekwencje nukleotydowe:

- **EMBL** – baza danych European Molecular Biology Laboratory założona w 1982 roku przez European Bioinformatics Institute (EBI) w Cambridge w Wielkiej Brytanii [Kulkowa et al., 2007].
- **GenBank** – baza utrzymywana przez National Center for Biotechnology Information (NCBI) w US National Institute of Health (NIH) w Bethesda w Stanach Zjednoczonych [Benson et al., 2007].
- **DDBJ** – DNA Databank of Japan, utworzona w 1986 roku baza danych zarządzana w National Institute of Genetics (NIG) [Sugawara et al., 2008].

Każdy z trzech wymienionych powyżej ośrodków działa oddzielnie i dostarcza swoje własne interfejsy, za pomocą których można przesyłać dane. Wspólnie tworzą one International Sequence Database Collaboration (INSDC) i codziennie wymieniają pomiędzy sobą uzyskane informacje, tworząc tym samym spójną bazę danych sekwencji nukleotydowych dostępną dla środowiska naukowego. Dane udostępniane są w postaci plików tekstowych o zdefiniowanym formacie za pomocą FTP lub poprzez dużą liczbę różnych narzędzi i serwisów internetowych, które umożliwiają wyszukiwanie oraz analizę danych dostępnych w bazach.

Bazy sekwencji nukleotydowych tworzone są przez badaczy – zarówno indywidualne laboratoria, jak i wysokoprzepustowe centra analiz danych genomowych deponują odkryte sekwencje w jednej z trzech podstawowych baz danych sekwencji nukleotydowych. Sekwencje przesyłane są bądź za pomocą aplikacji internetowych (na przykład *Webin* bazy EMBL), bądź za pomocą programów komputerowych (na przykład aplikacja *Sequin* bazy GenBank). Przesyłając dane do wybranej bazy danych, badacze dobrowolnie zgadzają się na udostępnianie swoich wyników i w ten sposób je publikują. Większość czasopism naukowych obecnie wymaga, aby publikując nowo odkrytą cząsteczkę, podać odniesienie do jej rekordu w bazie danych. Każdy wpis w bazie danych składa się z sekwencji – pojedynczego, ciągłego odcinka DNA lub RNA, oraz anotacji, czyli opisu tej sekwencji, która zawiera między innymi nazwę organizmu, którego sekwencja dotyczy, odnośniki do literatury oraz opis istotnych cech biologicznych danej sekwencji.

Od czasu powstania bazy GenBank liczba zdeponowanych sekwencji powiększa się dwukrotnie co półtora roku. Na rysunku 3.1 umieszczono liczbę sekwencji oraz ich długość od momentu powstania bazy GenBank. Według danych z grudnia 2007 roku, baza danych GenBank zawierała sekwencje dla ponad 260 tysięcy organizmów (w przeważającej części modelowych), przy czym każdego miesiąca w bazie rejestrowanych jest około 17 tysięcy nowych gatunków. Z uwagi na fakt, iż pomiędzy wszystkimi bazami należącymi do INSDC następuje synchronizacja umieszczonych sekwencji, rysunek 3.1 można traktować jako rysunek poglądowy dla każdego z ośrodków należących do INSDC. Najliczniej reprezentowanym gatunkiem pod względem liczby nukleotydów w tej bazie był człowiek (*Homo sapiens*), następnie mysz domowa (*Mus musculus*), szczur wędrowny (*Rattus norvegicus*), bydło domowe (*Bos taurus*),



Rysunek 3.1. Liczba sekwencji i ich łączna długość (kbps) zdeponowanych w bazie GenBank w latach 1982–2008
źródło: `http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html`

kukurydza zwyczajna (*Zea mays*), danio pręgowany (*Danio rerio*) oraz dzik (*Sus scrofa*). Każda z sekwencji umieszczonych w bazie GenBank należy do pewnej podsekcji, a każda z nich określona jest trzyliterowym skrótem. Obecnie w bazie GenBank istnieje 18 takich podsekcji – ich lista została umieszczona w tabeli 3.1. Podział na podsekcje zgodny jest z organizmem, z którego pochodzi dana sekwencja, lub związany jest z technologią, na podstawie której dana sekwencja została wygenerowana. Obecnie podział na organizmy jest raczej podziałem historycznym i nie odnosi się do aktualnej taksonomii NCBI , a raczej służy jako wygodny system podziału bazy na mniejsze pliki, w których umieszczane są sekwencje należące do tej samej podsekcji.

Tabela 3.1. Podział sekwencji zgodnie z ich typem lub pochodzeniem

symbol sekcji	nazwa sekcji
PRI	sekwencje naczelnych
ROD	sekwencje gryzoni
MAM	sekwencje innych ssaków
VRT	sekwencje innych kręgowców
INV	sekwencje bezkręgowców
PLN	sekwencje roślin, grzybów i glonów
BCT	sekwencje bakterii
VRL	sekwencje wirusów
PHG	sekwencje bakteriofagów
SYN	sekwencje syntetyczne
UNA	sekwencje nieopisane
EST	znaczniki sekwencji ulegających ekspresji
PAT	sekwencje opatentowane
STS	miejsca markerowe sekwencji
GSS	sekwencje przeglądowe genomu
HTG	wysokoprzepustowe sekwencje genomowe
HTC	wysokoprzepustowe sekwencje cDNA
ENV	sekwencje pochodzące ze środowiska o nieznanym pochodzeniu
CON	sekwencje skonstruowane na podstawie innych sekwencji

Sekwencje nukleotydowe w formacie FASTA

Najprostszym, a zarazem najpopularniejszym formatem, który pozwala na reprezentowanie sekwencji nukleotydowych, jest format FASTA. Popularność tego formatu bierze się zapewne stąd, iż jest bardzo prosty i przystępny dla człowieka, a równocześnie te same pliki bez żadnych modyfikacji mogą być przetwarzane przez programy komputerowe.

Sekwencja w formacie FASTA składa się ze znaku początku sekwencji „>” (znak większości), jej nazwy oraz z ciągu znaków małymi lub wielkimi literami – zwyczajowo w jednej linii umieszcza się 60 symboli. Różne bazy danych,

które przechowują daną sekwencję, mogą uzupełniać ją o dodatkowe informacje umieszczone w linii nagłówka, jednakże podstawowy schemat (znak większości, nagłówek, ciąg znaków) zazwyczaj pozostaje niezmieniony. Nie istnieje żadna formalna definicja linii nagłówka, tak więc różne bazy danych po symbolu „>" umieszczają charakterystyczne dla siebie informacje, zachowując przy tym zgodność z formatem.

Poniżej pokazano przykładową sekwencję nukleotydową w formacie FASTA z bazy danych GenBank. W linii nagłówka umieszczono tak zwany numer GI, numer dostępu GenBank, nazwę LOCUS oraz wiersz DEFINITION. Znaczenie poszczególnych elementów zostanie omówione poniżej.

```
>gi|255957385|gb|GQ371214.1| Saccharomyces cerevisiae
strain TCJ154 Ste2p (STE2) gene, partial cds
GCAAGGTTTAGTTAACAGTACTGTTACTCAGGCCATTATGTTTGGTGTCAGATGTGGTGCAGCTGCTTTG
ACTTTGATTGTCATGTGGATGACATCGAGAAGCAGAAAAACGCCGATTTTCATTATCAACCAAGTTTCAT
TGTTTTTAATCATTTTGCATTCTGCACTCTATTTTAAATATTTACTGTCTAATTACTCTTCAGTGACTTA
CGCTCTCACCGGATTTCCTCAGTTCATCAGTAGAGGTGACGTTCATGTTTATGGTGCTACAAATATAATT
CAAGTCCTGCTTGTGGCTTCTATTGAGACTTCACTGGTGTTTCAGATAAAAGTTATTTTCACGGGCGACA
ACTTCAAAAGGATAGGTTTGATGCTGACGTCGATATCTTTCACTTTAGGAATTGCTACAGTTACCATGTA
TTTTGTAAGCGCTGTTAAAGGTATGATTGTGACTTATAATGATGTTAGTGCCACCCAAGGTAAATACTTC
AATGCATCCACAATTTTACTTGCATCCTCAATAAACTTTATGTCATTTGTCCTGGTAGTTAAATTGATTT
TAGCTATTAGATCAAGAAGATTCCTTGGTCTCAAGCAGTTCGATAGTTTCCATATTTTACTTATAATGTC
ATGTCAATCTTTGTTGGTTCCATCGATAATATTCATCCTCGCATACAGTTTGAAACCAAACCAGGGAACA
GATGTCTTAACTACTGTTGCAACATTACTTGCTGTATTGTCTTTACCATTATCATCAATGTGGGCCACGG
CTGCTAATAATGCATCCAAAACAAACACAATTACTTCAGACTTTACAACATCCACAGATAGGTTTTATCC
AGGCACGCTGTCTAGCTTTCAAACTGATAGTATCAACAACGATGCTAAAAGCAGTCTCAGAAGTAGATTG
TATGACCTATATCCTAGAAGGAAGGAAACAACATCGGATAAACATTCGGAAAGAACTTTTGTTTCTGAGA
CTGCAAATGATATAGAGAAAAATCAGTTTTATCAGTTGCCCACACCTACGAGTTCAAAAAATACTAGGAT
```

Za pomocą formatu FASTA definiować można również sekwencje białkowe. W tym przypadku symbole nukleotydów w sekwencji zastąpione są przez symbole aminokwasów:

```
>gi|255957386|gb|ACU43528.1| Ste2p [Saccharomyces cerevisiae]
QGLVNSTVTQAIMFGVRCGAAALTLIVMWMTSRSRKTPIFIINQVSLFLIILHSALYFKYLLSNYSSVTY
ALTGFPQFISRGDVHVYGATNIIQVLLVASIETSLVFQIKVIFTGDNFKRIGLMLTSISFTLGIATVTMY
FVSAVKGMIVTYNDVSATQGKYFNASTILLASSINFMSFVLVVKLILAIRSRRFLGLKQFDSFHILLIMS
CQSLLVPSIIFILAYSLKPNQGTDVLTTVATLLAVLSLPLSSMWATAANNASKTNTITSDFTTSTDRFYP
GTLSSFQTDSINNDAKSSLRSRLYDLYPRRKETTSDKHSERTFVSETANDIEKNQFYQLPTPTSSKNTR
```

Nie istnieje żadne standardowe rozszerzenie pliku zawierającego sekwencje w formacie FASTA, niemniej przyjęło się stosowanie rozszerzenia `.fa` oraz `.fsa`. Z kolei baza danych NCBI stosuje własną konwencję: `.fna` dla plików zawierających geny, `.faa` dla plików zawierających sekwencje kodujące białka, `.ffn` dla genów kodujących białka.

3.1 Baza danych EMBL

Baza danych EMBL jest europejskim oddziałem sieci INSDC. Baza ta utrzymywana jest przez European Bioinformatics Institute (EBI) i wraz z narzędziami, które umożliwiają wyszukiwanie danych oraz ich analizę, dostępna jest na stronie internetowej `http://www.ebi.ac.uk/`. Baza EMBL została założona w 1982 roku i tym samym jest najstarszą europejską bazą sekwencji. Sekwencje umieszczone w tej bazie są publicznie dostępne i pochodzą głównie od indywidualnych badaczy, grup badawczych, European Patent Office (EPO) oraz z wymiany pomiędzy poszczególnymi członkami INSDC. Nowa wersja bazy wydawana jest co cztery miesiące. Sekwencje zdeponowane w bazie dostępne są za pomocą narzędzia Sequence Retrieval System (SRS), FTP, web serwisów oraz narzędzi wyszukiwania sekwencji podobnych.

Każda sekwencja nukleotydowa stanowi odrębny wpis w bazie EMBL, który oprócz samej sekwencji musi zawierać informacje takie jak: identyfikator sekwencji, odnośniki literaturowe oraz anotacje w formie tabeli cech. Tabela cech jest bardzo istotnym elementem rekordu opisującego sekwencje i jej definicja jest wspólna dla wszystkich baz danych należących do INSDC. W tabeli cech znajdują się w zasadzie najważniejsze informacje biologiczne dotyczące danej sekwencji, jakie można uzyskać, analizując dany rekord. Dokumentacja dotycząca tabeli cech opisuje dokładnie, jakie elementy powinny i mogą się w niej znaleźć – między innymi są to istotne biologiczne informacje takie jak regiony kodujące, translacje sekwencji nukleotydów na sekwencję aminokwasów, jednostki transkrypcji, miejsca modyfikacji lub mutacji.

3.1.1 Format rekordu w bazie EMBL

Wpisy w bazie EMBL mają ściśle określoną strukturę i jak w większości tego typu danych w bioinformatyce są tak skonstruowane, aby informacje w nich zawarte bez trudu mogły zostać zinterpretowane przez człowieka, przy równoczesnym zapewnieniu możliwości łatwego ich przetwarzania za pomocą programów komputerowych. Dane reprezentowane są w formie tekstowej, natomiast opisy oraz różnego rodzaju komentarze zapisywane są w języku angielskim. Format rekordu w bazie EMBL różni się nieco od formatu danych baz GenBank i DDBJ.

Pojedyncza sekwencja zdefiniowana w bazie EMBL to wpis, który składa się z różnego typu informacji. Każda linia w pliku opisującym tę sekwencję jest oddzielnym elementem, posiada swój własny format i rozpoczyna się od dwuliterowej etykiety, na postawie której można określić, jakiego typu informację zawiera dana linia.

Poniżej przedstawiono krótki opis większości etykiet linii składających się na jeden rekord w bazie EMBL:

- **ID** – identyfikator sekwencji.
- **PR** – identyfikator projektu INSDC.

- **DT** data utworzenia oraz ostatniej modyfikacji .
- **DE** – ogólne informacje opisujące daną sekwencję. Linia ta może zawierać nazwy genów, których dotyczy sekwencja, lokalizację genomową sekwencji i wszelkie informacje przydatne do identyfikacji sekwencji.
- **KW** – słowo kluczowe, które może być wykorzystywane do identyfikacji danej sekwencji pomiędzy różnymi bazami danych.
- **OS** – gatunek organizmu.
- **OC** – klasyfikacja taksonomiczna organizmu.
- **OG** – typ organelli. Nazwa części komórki z której pochodzi dana sekwencja. Występuje tylko dla sekwencji niepochodzących z jądra komórkowego.
- **Rx** – (RN, RC, itd.) wpisy zawierające informacje na temat publikacji naukowych związanych z daną sekwencją.
- **DR** – odnośniki do innych baz danych, które zawierają informacje związane z tą sekwencją.
- **CC** – komentarze.
- **FH** – nagłówek tabeli cech.
- **FT** – rekordy związane z tabelą cech, która zawiera anotacje danej sekwencji. Wraz z poznaniem właściwości danej sekwencji tabla cech ulega zmianie, tworząc pełniejszy opis zapisanego w bazie ciągu nukleotydów.
- **SEQ** – sekwencja nukleotydowa.
- \\ – zakończenie rekordu.

Nie wszystkie typy linii pojawiają się w każdym rekordzie zawierającym sekwencję. Każdy rekord zawiera takie elementy, jakie wymagane są do jego opisania zgodnie z aktualną wiedzą. Wraz z rosnącą wiedzą na temat danej sekwencji, rekord, który jej dotyczy uzupełniany jest o nowe wpisy. Przykładowy rekord pochodzący z bazy danych EMBL umieszczono w pierwszej części Dodatku.

3.1.2 Dostęp do rekordów bazy EMBL

Przesyłanie nowych sekwencji

Deponowanie sekwencji w publicznie dostępnych bazach stało się standardową praktyką autorów, którzy taką sekwencję chcą opublikować w czasopiśmie naukowym. Każda przesłana sekwencja otrzymuje unikalny numer dostępu (ang. *accession number*), który od tego momentu staje się jej identyfikatorem – niezależnie od zmian jakie w przyszłości zostaną wprowadzone w trakcie edycji rekordu dotyczącego tej sekwencji. Numer dostępu danej sekwencji jest identyczny niezależnie od tego, której z baz danych należących do INSDC będziemy używać do jej przeglądania. Każda z baz danych należących do INSDC udostępnia swój własny interfejs do przesyłania danych. Z punktu widzenia użytkownika, który chce zdeponować daną sekwencję, nie ma znaczenia, do

której bazy danych sekwencja zostanie przesłana – zdeponowane sekwencje codziennie są wymieniane pomiędzy bazami EMBL, GenBank oraz DDBJ.

Baza danych EMBL udostępnia aplikację internetową *Webin* do deponowania nowych sekwencji nukleotydowych. Aplikacja ta pozwala na przesyłanie pojedynczych sekwencji oraz (w przypadku jeżeli liczba przesyłanych sekwencji przekracza 25) udostępnia również procedurę wsadowego przesyłania sekwencji. Procedura wsadowa może być uruchomiana, jeżeli przesyłane sekwencję są ze sobą powiązane (na przykład są to sekwencje tego samego genu, który został zsekwencjonowany dla większej liczby różnych organizmów).

Możliwe jest również przesyłanie danych do bazy za pomocą aplikacji *Sequin*. Jest to odrębna aplikacja, którą należy zainstalować na komputerze użytkownika. Aplikacja ta została stworzona w NCBI i za jej pomocą można przesłać sekwencję nukleotydów do wybranego członka INSDC. Z uwagi na fakt, iż interfejs *Webin* posiada odpowiednie mechanizmy pozwalające na sprawdzenie zgodności przesyłanych danych z formatem EMBL, jest to zalecana przez EMBL metoda przesyłania sekwencji.

Pobieranie sekwencji zdeponowanych

Głównym narzędziem dostępu do sekwencji zdeponowanych w bazie EMBL jest SRS (*Sequence Retrieval System*). Dodatkowo dane udostępniane są za pomocą serwera FTP oraz za pomocą różnego rodzaju narzędzi wyszukiwania sekwencji podobnych. EMBL udostępnia również szereg narzędzi takich jak Dbfetch, Wsdbfetch, netsev, pozwalających na wyszukiwanie konkretnych sekwencji na podstawie numeru dostępu.

3.2 Baza danych GenBank

Baza danych GenBank została założona w 1982 roku, w Los Alamos National Laboratory. Pod koniec lat 80-tych ubiegłego wieku została przeniesiona do National Center for Biotechnology Information (NCBI), gdzie obecnie jest utrzymywana. NCBI jest oddziałem National Library Medicine (NLM) i zlokalizowane jest na terenie US National Health Institute (US NIH) w Bethesda w Stanach Zjednoczonych. Strona internetowa bazy GenBank, jak również i liczne narzędzia do przetwarzania danych zgromadzonych w GenBanku znajduje się pod adresem `http://www.ncbi.nlm.nih.gov/` na stronach NIH. Sekwencje umieszczone w bazie GenBank pochodzą podobnie jak w bazie EMBL głównie od indywidualnych laboratoriów, wysokoprzepustowych centrów sekwencjonowania, US Office Patents and Trademarks Office (USPTO) oraz poprzez wymianę sekwencji pomiędzy członkami INSDC. Nowe wersje bazy wydawane są co dwa miesiące. Tak samo jak w przypadku danych EMBL, sekwencje umieszczone w bazie są publicznie dostępne poprzez różnego rodzaju narzędzia wyszukiwania sekwencji oraz poprzez anonimowe serwery FTP.

3.2.1 Format rekordu w bazie GenBank

Format GenBank jest chyba jednym z najczęściej stosowanych formatów przechowywania sekwencji genomowych. Każdy rekord zawiera zwięzły opis sekwencji, nazwę systematyczną organizmu, z którego pochodzi sekwencja, odnośniki do literatury, tabelę cech oraz oczywiście ciąg nukleotydów składających się na sekwencję. Porównując rekord bazy EMBL z rekordem bazy GenBank, widać, że informacje zawarte w rekordach GenBank są praktycznie takie same jak informacje, które znajdują się w EMBL. Pola rekordu zawierają dane, których typ określony jest poprzez etykiety. Poniżej przedstawiono, jakie informacje zawierają poszczególne etykiety:

- **LOCUS** – pole LOCUS zawiera dane takie jak nazwa locus, czyli kodowe oznaczenie określonego rekordu (obecnie ma znaczenie historyczne), długość sekwencji, typ cząsteczki (zazwyczaj DNA, RNA, mRNA itd.), przynależność sekwencji do podsekcji GenBank (patrz tabela 3.1) oraz datę ostatniej modyfikacji.
- **DEFINITION** – krótki opis sekwencji.
- **ACCESION** – numer odstępu, który jest unikalnym identyfikatorem sekwencji. Identyfikator ten zazwyczaj składa się z jednej litery i pięciu cyfr (np. U12345) lub z 2 liter i sześciu cyfr (np. AF123456).
- **VERSION** – informacja o liczbie zmian w rekordzie, które zostały dokonane od momentu przesłania sekwencji do bazy danych. Wartość umieszczana w polu VERSION tworzona jest na podstawione numeru dostępu (np. dla numeru dostępu podanego powyżej: AF123456.1). Wiersz ten zawiera również identyfikator GI (ang. *geninfo identifier*). Jeżeli w rekordzie zostaną wprowadzone zmiany, wartość VERSION po kropce zwiększana jest o 1 oraz nadawany jest nowy numer GI.
- **KEYWORDS** – słowa kluczowe przypisane do sekwencji przez jej autora. Z uwagi na fakt, że nie istnieją żadne reguły dodawania słów kluczowych do rekordu, obecnie pole to na raczej wartość historyczną.
- **SOURCE** – nazwa organizmu, z którego pochodzi sekwencja. Pole to również zawiera podetykietę **ORGANISM**, w której umieszczana jest formalna, taksonomiczna nazwa organizmu.
- **REFERENCE** – pole zawierające publikacje związane z daną sekwencją.
- **FEATURES** – początek tabeli cech.
- **BASE COUNT** – informacje na temat liczby poszczególnych nukleotydów wchodzących w skład sekwencji.
- **ORIGIN** – pole to może być puste lub zawierać wskazanie lokalizacji genomowej pierwszego nukleotydu sekwencji (w starszych rekordach). Poniżej tego słowa kluczowego podana jest sekwencja nukleotydowa.
- \\ – etykieta końca rekordu.

Przykładowy rekord pochodzący z bazy danych GenBank został umieszczony w drugiej części Dodatku.

3.2.2 Dostęp do rekordów bazy GenBank

Przesyłanie nowych sekwencji

Przesyłanie nowych sekwencji do bazy GenBank odbywa się głównie za pomocą aplikacji internetowej *BankIt* lub za pomocą programu *Sequin*.

Obecnie co trzecia sekwencja przesyłana jest do GenBanku za pomocą aplikacji BankIt, która pozwala na łatwe przesyłanie sekwencji – bez konieczności poznawania reguł formatowania danych czy szczegółowych reguł nazewnictwa. Metoda ta polecana jest jeżeli przesyłane sekwencje nie są długie i ich liczba również nie jest duża. Aplikacja BankIt może być także wykorzystywana do edycji istniejących w bazie sekwencji. W sytuacji, jeżeli liczba sekwencji do przesłania jest większa lub sekwencja jest bardzo długa, polecany jest program *Sequin*, natomiast w sytuacji jeśli przesyłane sekwencje są jeszcze większe (na przykład przesyłany jest cały genom), można wykorzystać do tego celu aplikację linii poleceń Tbl2asn.

Każda przesłana sekwencja podlega walidacji przez pracowników GenBank, a po sprawdzeniu wszystkich błędów nadawany jej jest numer dostępu (ang. *accession number*). Zazwyczaj czas oczekiwania na nadanie numeru wynosi około dwóch dni roboczych (codziennie pracownicy GenBanku nadają około 1600 numerów). Następnie tak sprawdzony rekord przesyłany jest do jego autora w celu zaakceptowania wprowadzonych poprawek. Nadanie numeru dostępu oznacza, że sekwencja została zapisana w bazie i jest dostępna dla jej użytkowników. W przypadku jeżeli autor nie chce jej ujawniać, do czasu pojawienia się publikacji, może ona pozostać utajniona.

Pobieranie sekwencji zdeponowanych

Podstawowym narzędziem dostępu do sekwencji zapisanych w bazie GenBank jest Entrez – internetowa aplikacja podłączona do 35 biologicznych baz danych, która umożliwia wyszukiwanie interesujących informacji (nie tylko sekwencji) zarówno pomiędzy różnymi bazami danych, jak i w pojedynczej, wybranej przez użytkownika bazie danych. Bazy danych dostępne w ramach aplikacji Entrez to bazy zawierające sekwencje nukleotydowe oraz aminokwasowe pochodzące z bazy GenBank i z innych źródeł, mapy genomowe i populacyjne, zbiory sekwencji filogenetycznych oraz środowiskowych, dane pochodzące z ekspresji genów, baza taksonomii NCBI, bazy danych struktur białkowych oraz domen. Każda z baz danych dostępnych w ramach Entrez połączona jest z bazami danych publikacji PubMed oraz PubMed Central.

Innym sposobem dostępu do danych bazy GenBank jest korzystanie z narzędzi służących do porównywania sekwencji. Porównywanie sekwencji jest jednym z najbardziej podstawowych, a zarazem najpopularniejszym sposobem analizy danych dostępnych w ramach bazy GenBank. W tym celu udostępniono szereg narzędzi z godziny BLAST, które pozwalają na wyszukiwanie podobieństw pomiędzy zadaną sekwencją a sekwencjami umieszczonymi w bazie.

Analizy BLAST mogą być wykonywane zarówno poprzez udostępnione internetowe aplikacje, jak i za pomocą aplikacji instalowanych na komputerze użytkownika, a udostępnianych poprzez FTP.

Ostatnim sposobem dostępu do danych zlokalizowanych w bazie GenBank jest pobieranie wersji bazy w postaci plików tekstowych (w formacie GenBank lub ASN.1), które udostępniane są poprzez serwer FTP.

3.3 Baza danych DDBJ

Baza danych DDBJ powstała w 1986 roku w National Institute of Genetics (NIG) w Mishima w Japonii pod nadzorem Japońskiego Ministerstwa Edukacji, Kultury, Sportów, Nauki i Technologii. Dostęp do wszystkich zasobów bazy DDBJ znajduje się pod adresem: `http://www.ddbj.nig.ac.jp/`. Od roku 1987 baza ta stanowi azjatycki oddział INSDC. Format rekordu w bazie DDBJ jest identyczny z formatem rekordu bazy danych GenBank, nie będzie więc on ponownie tutaj opisywany. Nowe wersje bazy danych wydawane są raz na kwartał.

Ponad 90% sekwencji, które przesyłane są z Japonii, umieszczane są w zasobach INSDC poprzez bazę DDBJ, pozostałe sekwencje umieszczane w bazie DDBJ przesyłane są z Korei oraz z Chin. Tak jak w przypadku pozostałych baz danych sekwencje przesyłane są zarówno przez indywidualne laboratoria badawcze, jak i wysokoprzepustowe centra sekwencjonowania genów. Dodatkowo baza danych posiada dział, który zawiera sekwencje opatentowane zbierane i przetwarzane przez Japanese Patent Office, Korean Intellectual Property Office, USPTO oraz EPO.

Deponowanie danych w bazie DDBJ może odbywać się za pomocą internetowej aplikacji *Sakura*. W przypadku konieczności przesyłania dużej liczby sekwencji albo sekwencji opisanych dużą liczbą cech lub bardzo długich (powyżej 500 tys. par zasad), zalecane jest przesyłanie sekwencji bezpośrednio do DDBJ za pomocą procedury MSS (*Massive Submission System*). Tak samo jak w przypadku pozostałych baz należących do INSDC można również korzystać z programu *Sequin*.

3.4 Adresy Internetowe

Adresy internetowych baz danych:

- DDBJ – `http://www.ddbj.nig.ac.jp/index-e.html`
- EMBL – `http://www.ebi.ac.uk/embl/`
- GenBank – `http://www.ncbi.nlm.nih.gov/Genbank/`
- INSDC – `http://www.insdc.org/`
- NCBI – `http://www.ncbi.nlm.nih.gov/`

4

Przeszukiwanie baz danych sekwencji

Podstawowe analizy zawartości publicznych baz danych sekwencji związane są z poszukiwaniem podobieństw (i różnic) pomiędzy sekwencjami, które są dla badacza interesujące. Znajdowanie podobieństw pomiędzy sekwencjami pozwala badaczom na przewidywanie funkcji nowo zsekwencjonowanych genów, przewidywanie nowych członków rodzin genów oraz zrozumienie strukturalnych, funkcjonalnych i ewolucyjnych zależności występujących pomiędzy badanymi sekwencjami. Obecnie gdy w internetowych bazach danych umieszczane są sekwencje całych genomów, poszukiwanie sekwencji podobnych pozwala na przewidywanie lokalizacji oraz funkcji regionów kodujących białka i miejsc regulujących transkrypcję.

Materiał genetyczny, który przekazywany jest z pokolenia na pokolenie, ulega ciągłym zmianom poprzez mutacje, którymi poddawane są sekwencje reprezentujące ten materiał. Najprostsze formy mutacji, jakie mogą się pojawiać na poziomie molekularnym, to: **substytucja** (zamiana jednego nukleotydu na inny), **insercja** (wstawienie dodatkowego nukleotydu) lub **delecja** (usunięcie nukleotydu). Przyjmując założenia, że sekwencje różnicowały się poprzez wymienione proste formy mutacji, porównując sekwencje, badacze mogą określać, czy geny lub białka przez nie reprezentowane są homologami, czyli czy posiadają wspólnego przodka. Zazwyczaj nie jesteśmy w stanie określić, jaka jest sekwencja przodka, ale badając podobieństwo, potrafimy stwierdzić, czy dwie sekwencje mają wspólne pochodzenie ewolucyjne. Dopasowując do siebie dwie sekwencje, możemy znajdować pewne obszary, które pozostają niezmienione lub podlegają zmianie w bardzo niewielkim stopniu – mówimy wtedy, że pewne fragmenty sekwencji są silnie zakonserwowane. Może to świadczyć, że reprezentują one ważne obszary z punktu widzenia funkcjonowania genu czy też białka. Wyszukiwanie podobieństw pomiędzy sekwencjami pełni bardzo istotną rolę nie tylko dla zrozumienia zależności ewolucyjnych pomiędzy sekwencjami, ale także wykorzystywane jest często w procesie automatycznego przewidywania funkcji nowych genów lub białek.

Innym zagadnieniem, które zostanie poruszone w niniejszym rozdziale, jest przeszukiwanie baz danych sekwencji w celu znalezienia grupy sekwencji podo-

bnych do sekwencji, która jest zapytaniem. Jest to zagadnienie o wiele bardziej złożone w porównaniu z zadaniem dokładnego dopasowania dwóch sekwencji. Niemniej obecnie jest to najbardziej podstawowy sposób analizy zawartości baz danych sekwencji. Celem takiego przeszukiwanie jest najczęściej znalezienie grupy sekwencji homologicznych do danej sekwencji, co pozwala określić, które spośród setek tysięcy sekwencji dostępnych w bazie danych mogą być potencjalnie spokrewnione z interesującą nas sekwencją. Biorąc pod uwagę dzisiejszy rozmiar baz danych sekwencji, zadanie to jest nietrywialne i wymaga zastosowania wyspecjalizowanych algorytmów porównywania sekwencji.

4.1 Dopasowywanie dwóch sekwencji

Dopasowywanie (zwane inaczej uliniowieniem) dwóch sekwencji polega na znalezieniu najlepszej relacji pomiędzy sekwencjami, która będzie pokazywać zależność jeden-do-jeden pomiędzy zasadami tworzącymi sekwencję nukleotydowe lub pomiędzy aminokwasami tworzącymi sekwencję białkową. Z uwagi na ogromną liczbę możliwych dopasowań problem, który należy rozwiązać podczas dopasowywania dwóch sekwencji, związany jest z określeniem, jakie zestawienie sekwencji jest najlepsze. Poniżej przedstawiono klika różnych sposobów dopasowania dwóch sekwencji nukleotydowych: `agatccga` oraz `ctagacga`.

Dopasowanie niewykazujące podobieństwa sekwencji:

```
-------ggatccga
ctagacga-------
```

Dopasowanie bez przerw:

```
ggatccga
ctagacga
```

Dopasowanie z przerwami:

```
ggatccga---
c--ta-gacga
```

Inne dopasowanie z przerwami:

```
---ggatccga
ctaga--c-ga
```

Jak widać na powyższym przykładzie, istnieje wiele różnych sposobów dopasowania dwóch sekwencji. Pierwszy problem, jaki się pojawia, związany jest z właściwym wprowadzeniem przerw pomiędzy poszczególnym zasadami. I tak, dopasowując do siebie poszczególne elementy sekwencji, możemy spotkać się z następującymi sytuacjami:

- Dopasowanie (ang. *match*).
- Niedopasowanie (ang. *mismatch*).
- Przerwa (ang. *gap*).

Najprostszym sposobem oceny jakości dopasowania wydaje się stworzenie pewnego systemu punktacji (kar i nagród) za podobieństwo lub jego brak pomiędzy poszczególnymi zasadami lub aminokwasami tworzącymi sekwencję. Dodatkowo każdy system punktacji musi brać pod uwagę nie tylko występowanie substytucji oraz insercji i delecji, ale również długość przerw, które są ich wynikiem. Insercje lub delecje często określane są terminem **indels** z uwagi na fakt, że tak naprawdę nie wiadomo, czy w procesie ewolucji sekwencji pojawiło się wstawienie nukleotydu czy jego usunięcie.

Oceniając jakość dopasowania, możemy zastosować dwa różne podejścia:

- Punktacja za podobieństwo (ang. *similarity scores*) – im bardziej podobne sekwencje, tym wyższa wartość punktacji (ang. *score*).
- Miary odległości (ang. *distance measures*) – im bardziej podobne sekwencje, tym mniejsza wartość miary odległości.

W niniejszym rozdziale przedstawione zostaną sposoby dopasowywania sekwencji bazujące na ich wzajemnym podobieństwie.

Dopasowywanie sekwencji może odbywać się globalnie, kiedy próbujemy dopasować dwie sekwencje na całej ich długości oraz lokalnie, gdy próbujemy znaleźć najlepsze dopasowanie jedynie dla fragmentu sekwencji.

Dopasowywanie sekwencji nukleotydowych różni się od dopasowywania sekwencji białkowych. Podstawowa różnica związana jest z liczbą liter występujących w alfabetach obu rodzajach sekwencji: w sekwencjach nukleotydowych występują tylko 4 litery symbolizujące odpowiednie zasady azotowe, podczas gdy w sekwencjach białkowych mamy 20 liter symbolizujących aminokwasy. Z uwagi na te różnice, obydwa rodzaje sekwencji będą w niniejszym rozdziale rozpatrywane oddzielnie.

4.1.1 Dopasowywanie sekwencji nukleotydowych

Najprostszy schemat oceny dopasowania sekwencji nukleotydowych polega na określeniu jakości dopasowania (ang. *score*), poprzez nagradzanie występowania identycznych par nukleotydów w uliniowieniu oraz karanie sytuacji braku dopasowania lub wystąpienia przerwy. Mógłby on wyglądać następująco:

$$(\textit{ilość dopasowań}) - (\textit{ilość niedopasowań oraz przerw})$$

Powyższy schemat jest bardzo często wykorzystywany przy definiowaniu jakości dopasowywaniu sekwencji. Ważnym elementem takiego dopasowania jest określenie wartości kary za wystąpienie przerwy. Jeśli w danym dopasowaniu mamy przerwę o długości l znaków, to karę za wystąpienie takiej przerwy można przedstawić w postaci pewnej funkcji, której wartość zależna będzie od długości l, i oznaczyć jako $\delta(l)$. Funkcja taka pozwala na wyznaczenie warto-

ści kary za wystąpienie przerwy, przeważnie jest wartością zerową lub ujemną. Najprostszy schemat wyznaczanie kary za wystąpienie przerwy nosi nazwę liniowego modelu przerw (ang. *linear gap model*), gdzie wartość kary wyznacza się następująco: $\delta(l) = -w \cdot l$, przy założeniu, że w jest pewną nieujemną wartością kary (wagą) za wystąpienie przerwy. Funkcja kary za przerwy może pojawić się w dwóch wariantach:

- Nieafiniczny model (ang. *non-affine model*) – każde wystąpienie przerwy traktowane jest tak samo.
- Afiniczny model (ang. *affine model*) – każde utworzenie nowej przerwy jest karane dodatkowo, mamy więc dwie wagi: w_{gap_start} – karę, którą przyznajemy za rozpoczęcie nowej przerwy oraz w_{gap} – wartość kary, którą przyznajemy każdej następnej przerwy w danym ciągu. Stąd też jeżeli w dopasowaniu sekwencji występuje przerwa o długości l, to wartość funkcji kary wyznaczana będzie w sposób następujący: $\delta(l) = w_{gap\_start} + w_{gap}(l)$.

Dodatkowo można określić maksymalną negatywną wartość kary, która będzie przyznawana, jeśli przerwy w dopasowaniu będą zbyt długie.

Poniżej przedstawiono przykłady różnych wartości, jakie można uzyskać dla różnych sposobów dopasowania tych samych sekwencji i dla różnych modeli dopasowania. Wyznaczając wartość dopasowania przyjęto następujące wartości: dopasowanie aminokwasów: +1, brak dopasowania: -1, przerwa: -2, początek przerwy: -4

Sposób I:

```
cgaatcgaacaacatcctca
agattcgac--acc----ca
```

Wartość dopasowania dla nieafinicznego modelu:
-1+1+1-1+1+1+1+1-1-2-2+1+1-1-2-2-2-2+1+1=-6

Wartość dopasowania dla afinicznego modelu:
-1+1+1-1+1+1+1+1-1-6-2+1+1-1-6-2-2-2+1+1=-14

Sposób II:

```
cgaatcaagcaacttctcta
agattcga-c-ac--c--ca
```

Wartość dopasowania dla nieafinicznego modelu:
-1+1+1-1+1+1+1+1-2+1-2+1+1-2-2+1-2-2+1+1=-2

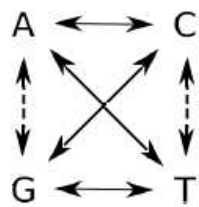
Wartość dopasowania dla afinicznego modelu:
-1+1+1-1+1+1+1+1-6+1-6+1+1-6-2+1-6-2+1+1=-18

Analizując powyższe przykłady, warto porównać wyniki jakości dopasowania dla różnych modeli dopasowania. Można zauważyć, że drugi sposób dopasowa-

nia, który na pierwszy rzut oka wydaje się dopasowywać do siebie poprawnie więcej zasad niż sposób pierwszy, ma o wiele niższą wartość punktacji przy zastosowaniu modelu afinicznego. Bierze się to stąd, iż z punktu widzenia ewolucji nie można założyć, że wystąpienie np. czterech przerw w sekwencji jest tak samo prawdopodobne, jak wystąpienie jednej przerwy o długości cztery. Należy pamiętać, że pojawienie się przerwy jest wynikiem mutacji – samo pojawienie się przerwy jest o wiele mniej prawdopodobne niż jej późniejsze wydłużenie lub skrócenie.

Dopasowywanie dwóch sekwencji w sposób podany powyżej jest kosztowne obliczeniowo, polega na sprawdzeniu wszystkich możliwych kombinacji i wyznaczeniu kombinacji najlepszej. Przy analizie dłuższych sekwencji wykorzystanie takiej metody w praktyce jest niemożliwe – liczba dopasowań dla pary sekwencji o długościach m oraz n wynosi $\binom{m+n}{n}$. Stąd też stosuje się metody przybliżone – tak zwane programowanie dynamiczne (patrz algorytm Needelmana–Wunscha dla dopasowań globalnych [Needleman and Wunsch, 1970] oraz algorytm Smitha–Watermana dla dopasowań lokalnych [Smith and Waterman, 1981]), które pozwalają na znalezienie najlepszego dopasowania sekwencji w czasie wielomianowym.

Przedstawione powyżej podejście nie bierze jednak pod uwagę składu nukleotydowego porównywanych sekwencji. Na przykład znany jest fakt, że tranzycje (zasada purynowa za zasadę purynową, zasada pirymidynowa za zasadę pirymidynową) pomiędzy zasadami o wiele częstsze niż transwersje (zasada purynowa zamiast zasady pirymidynowej i wice wersa). Mechanizm ten przedstawiono na rysunku 4.1.



Rysunek 4.1. Mechanizm tranzycji oraz transwersji pomiędzy zasadami. Przerywanymi strzałkami zaznaczono tranzycje, ciągłymi – transwersje

Jeżeli wszystkie mutacje zdarzałyby się jednakowo często, stosunek tranzycji do transwersji wynosiłby 1/2. Tymczasem analizy dopasowań wykazują, że stosunek ten wynosi około 4. Stąd też wyznaczając kary oraz nagrody dla sekwencji nukleotydowych, często wykorzystuje się różnego rodzaju macierze korekcji tranzycji oraz transwersji. Innego rodzaju macierze korekcji, które również mogą być wykorzystywane przy wyznaczaniu jakości oceny dopasowania dwóch sekwencji tworzone są na podstawie modeli ewolucyjnych, które uwzględniają prawdopodobieństwa mutacji nukleotydów w czasie.

4.1.2 Dopasowywanie sekwencji aminokwasowych

W przypadku dopasowywania sekwencji aminokwasowych liczba możliwych symboli, które mogą się pojawić w sekwencji, wynosi 20. Stąd też dopasowywanie tego rodzaju sekwencji jest trudniejsze i wymaga konstrukcji bardziej złożonych modeli niż w przypadku dopasowywania sekwencji nukleotydowych. Dopasowując sekwencje nukleotydowe, wykorzystuje się wyspecjalizowane macierze substytucji, które pozwalają na uwzględnienie w punktacji prawdopodobieństwa zamiany jednego aminokwasu w drugi. Dwie najpopularniejsze macierze substytucji, które wykorzystywane są do oceny dopasowań sekwencji to macierze **PAM** oraz **BLOSUM**.

Macierze BLOSUM

Macierze BLOSUM (*BLOcks SUbstitution Matrix*) zostały zaproponowane w 1992 roku w pracy [Henikoff and Henikoff, 1992]. Autorzy przeszukali bazę rodzin białek BLOCKS pod kątem występowania silnie zakonserwowanych regionów domen białek. Następnie na podstawie znalezionych „bloków" zakonserwowanych fragmentów (czyli takich zbiorów lokalnych dopasowań sekwencji, w których nie występują przerwy w dopasowaniu), dla każdej pary aminokwasów (w sumie jest 210 takich możliwych par), wyliczono częstości występowania poszczególnych dopasowań oraz częstości oczekiwane. Wartości macierzy BLOSUM wyznaczane są metodą różnic logarytmicznych – odrębnie dla każdej pary, jako logarytm ze stosunku zaobserwowanej częstości (czyli „biologicznie" wyznaczonego prawdopodobieństwa, że dwa aminokwasy zostaną ze sobą zamienione) do częstości oczekiwanej (czyli prawdopodobieństwa przypadkowej zamiany aminokwasów).

Istnieje kilka różnych rodzajów macierzy BLOSUM w zależności od tego, z jaką dokładnością dobierane były sekwencje tworzące bloki, na podstawie których wyliczane są później wartości macierzy. Każda z macierzy oznaczona jest symbolem BLOSUMX, gdzie X określa procent identyczności sekwencji podczas ich grupowania. Tak więc, np. macierz BLOSUM62 oznacza, że sekwencje tworzące blok były co najmniej w 62% identyczne. Im mniej zróżnicowane są sekwencje dopasowywane, tym większy numer powinna zawierać macierz BLOSUM wykorzystywana do ustalenia punktacji dopasowania.

Macierze BLOSUM biorą pod uwagę jedynie wzajemne podobieństwo sekwencji źródłowego dopasowania, autorzy macierzy nie stosują żadnego modelu ewolucyjnego do określenia prawdopodobieństwa przejścia jednego aminokwasu w drugi. Zaletą takiego podejścia jest fakt, że dane do utworzenia macierzy pochodzą bezpośrednio z obserwacji, a nie są budowane na podstawie modelu ewolucji sekwencji.

Macierze PAM

Macierze PAM (*Accepted Point Mutations*) zostały zaproponowane przez grupę M.Dayhoff w 1978 r. [Dayhoff et al., 1978] i oparte są na modelu ewolucyjnym tak zwanych akceptowanych mutacji punktowych.

Akceptowana mutacja punktowa oznacza zastąpienie jednego aminokwasu w sekwencji innym. Takim aminokwasem, który jest „akceptowany" przez ewolucję w tym sensie, że dla danego gatunku mutacja nie tylko powstała, ale również się utrwaliła.

Podobnie jak w przypadku macierzy BLOSUM, konstrukcja macierzy PAM rozpoczyna się od grupowania silnie zakonserwowanych fragmentów domen w bloki, które zawierają sekwencje „odpowiednio podobne" – w oryginalnym podejściu Dayhoff przyjęto założenie, że sekwencje tworzące dany blok nie mogą być od siebie różne bardziej niż w 15%. W przeciwieństwie do macierzy BLOSUM, gdzie dopasowuje się tylko silnie zakonserwowane fragmenty sekwencji (czyli dopasowanie odbywa się lokalnie), w metodzie Dayhoff poszukiwanie sekwencji podobnych ma charakter globalny. Dla każdej grupy sekwencji tworzone są drzewa filogenetyczne (tak konstruowane, aby w całym drzewie liczba podstawień była minimalna), a następnie wykorzystuje się te drzewa do zliczenia wszystkich możliwych przypadków podstawienia aminokwasów. W wyniku otrzymujemy tak zwaną macierz zliczeń, która dla każdej pary aminokwasów zawiera liczbę możliwych podstawień. Na podstawie macierzy zliczeń wyznaczana jest względna mutowalność każdego aminokwasu, czyli prawdopodobieństwo, że w danej jednostce czasu zostanie on zastąpiony innym aminokwasem. 1 PAM oznacza, że w danym czasie sekwencja ta uległa średnio zmianie w 1%, czyli zostało wymienione około 1% wszystkich reszt aminokwasowych (na 100 reszt aminokwasowych jedna uległa zmianie). Oczywiście po okresie 100 PAM nie każdy aminokwas na 100 zostanie wymieniony – niektóre mogły zmutować kilka razy, a inne pozostały niezmienione. Model ewolucyjny dla macierzy PAM zakłada, że zmiany zachodzące w białkach są efektem skumulowanych, nieskorelowanych mutacji. Stąd macierz zawierająca prawdopodobieństwo wystąpienia n podstawień powstaje poprzez n–krotne pomnożenie przez siebie macierzy.

Na podstawie macierzy prawdopodobieństw mutacji, będącej odzwierciedleniem pewnego modelu ewolucyjnego, tworzona jest macierz punktacji prawdopodobieństwa substytucji PAM. Macierz punktacji jest macierzą różnic logarytmicznych i wyznaczamy ją, dzieląc wartości znajdujące się w macierzy prawdopodobieństw mutacji przez wartość prawdopodobieństwa mutacji wynikające tylko z częstości występowania reszt aminokwasowych. W celu dobrania odpowiedniej skali otrzymany wynik dzielenia jest logarytmowany logarytmem o podstawie 10, a następnie mnożony razy 10.

Wartości w macierzy PAM powyżej 1 oznaczają, że dana para aminokwasów ulega mutacji częściej niż wynikałoby to z przypadku. Takie aminokwasy mają zazwyczaj podobne właściwości fizykochemiczne. Wartości równe 1 oznaczają, że prawdopodobieństwo mutacji równe jest prawdopodobieństwu wynikającemu z przypadku, natomiast wartości w macierzy poniżej jedynki oznaczają, że aminokwasy różnią się od siebie własnościami fizykochemicznymi. Podobieństwo (lub jego brak) pomiędzy dwoma aminokwasami może wynikać z ich kształtu, rozmiaru, lokalnych koncentracji ładunku elektrostatycznego,

konformacji powierzchni van der Waalsa czy też ich cech hydrofobowych i hydrofilowych.

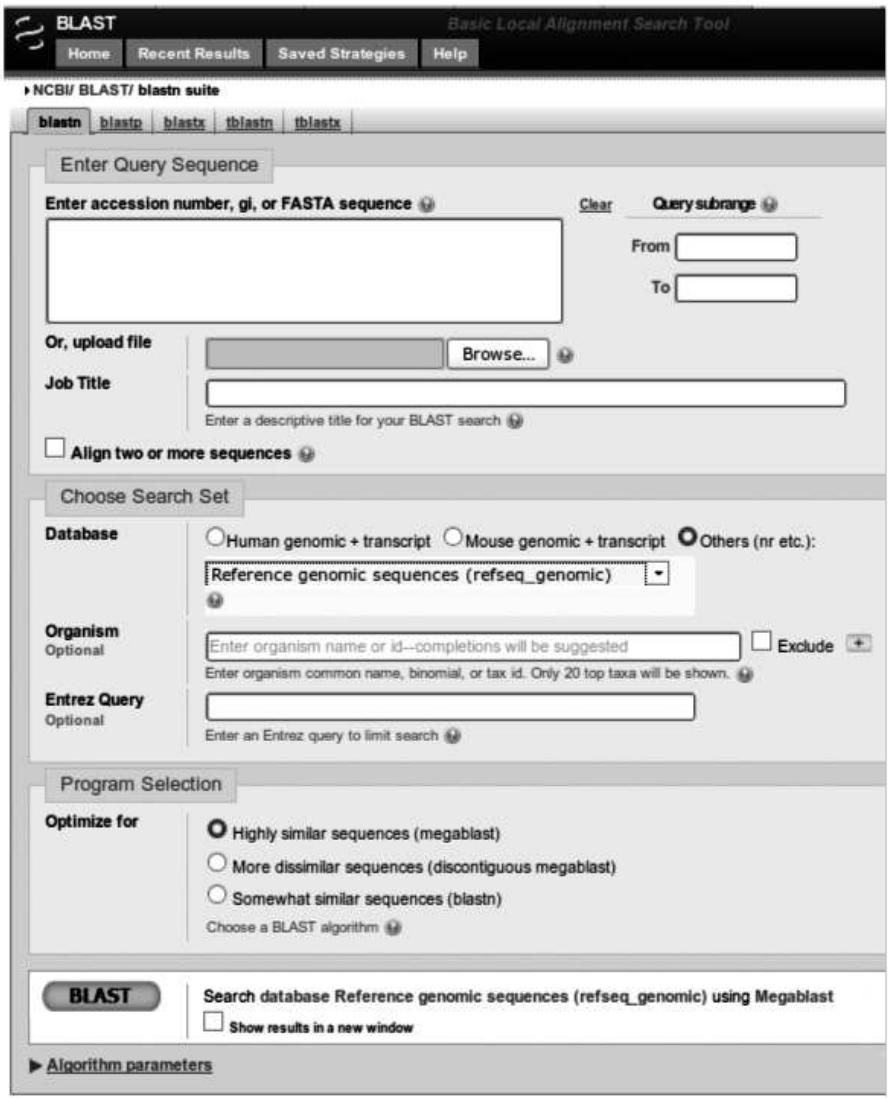
Najpopularniejszą macierzą z rodziny PAM jest macierz PAM250, co wynika z tego, że była to jedyna macierz opublikowana oryginalnie przez grupę Dayhoff. Odległość ewolucyjna 250PAM oznacza, że tylko jeden aminokwas na pięć nie uległ substytucji, czyli sekwencja uległa zmianie w 80% (na każde 100 reszt aminokwasowych przypadło 250 zamian). W zależności od tego, czy analizujemy sekwencje blisko ze sobą spokrewnione, czy dalekie, powinniśmy używać różnych macierzy – dla sekwencji odległych najlepiej sprawdzają się macierze PAM o wysokich numerach (takich jak PAM200 czy PAM250), z kolei dla sekwencji blisko ze sobą spokrewnionych, lepsze wyniki dają macierze o niższych numerach. Warto zwrócić uwagę, że tendencja ta jest odwrotna niż w przypadku macierzy BLOSUM.

4.2 Poszukiwanie sekwencji podobnych w bazach danych - BLAST

W przypadku, gdy przeszukujemy bazę danych względem pewnej konkretnej sekwencji, podstawowym celem, jaki stawiamy przed sobą, jest znalezienie możliwie dużej liczby homologicznych sekwencji, a nie – jak w przypadku uliniowiania sekwencji – znalezienia możliwie najlepszego sposobu na dopasowanie do siebie poszczególnych nukleotydów lub aminokwasów składających się na daną sekwencję. Stąd też do przeszukiwania baz danych wykorzystuje się dedykowane, heurystyczne algorytmy, które pozwalają na możliwe szybkie przeszukiwanie zasobów zawartych w bazach sekwencji i z dużym prawdopodobieństwem umożliwiają znalezienie sekwencji możliwie najbardziej podobnych do sekwencji zadanej. Większość programów poszukiwania sekwencji działa w podobny sposób: w pierwszym korku eliminowane są sekwencje, które są niepodobne do sekwencji będącej zapytaniem, a następnie sekwencje najbardziej podobne są dopasowywane do siebie.

Najczęściej wykorzystywanym narzędziem do wyszukiwania sekwencji homologicznych jest program **BLAST** (ang. *Basic Local Alignment Search Tool*) [Ye et al., 2006] oraz jego liczne rozszerzenia. Program ten dostępny jest w wielu różnych wersjach, między innymi jako aplikacja internetowa. Z punktu widzenia użytkownika obsługa programu jest bardzo prosta: wystarczy bowiem wkleić sekwencję będącą zapytaniem w odpowiednie pole tekstowe, ustawić parametry programu i uruchomić wyszukiwanie, aby po chwili w wyniku otrzymać listę sekwencji najbardziej podobnych do zadanej sekwencji wraz z informacją o jakości tego podobieństwa. Na rysunku 4.2 przedstawiono formatkę wejściową programu BLAST w wersji aplikacji internetowej dostępnej na stronach NCBI.

Zależnie od rodzaju sekwencji, która jest zapytaniem przesłanym przez użytkownika i bazy danych, która przeszukiwana jest względem podobieństwa sekwencji, dostępne są różne wersje programu BLAST:



Rysunek 4.2. Formatka wejściowa dla programu BLAST

- **blastn**, **megablast** – *sekwencja nukleotydowa vs. baza sekwencji nukleotydowych*. Na podstawie sekwencji nukleotydowej program zwraca listę najbardziej podobnych sekwencji nukleotydowych pochodzących z wybranej przez użytkownika bazy sekwencji.

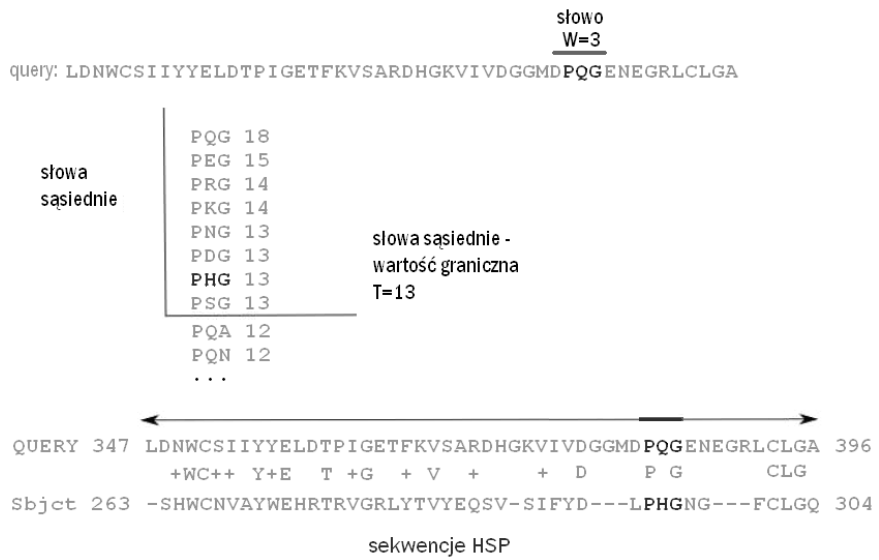
- **blastp**, **psi-blast**, **phi-blast** – *sekwencja białkowa vs. baza sekwencji białkowych.* Na podstawie sekwencji białkowej program zwraca listę najbardziej podobnych sekwencji białkowych pochodzących z bazy sekwencji wybranej przez użytkownika.
- **blastx** – *przetłumaczona sekwencja nukleotydowa vs. baza sekwencji białkowych.* Sekwencja nukleotydowa tłumaczona jest na sekwencję białkową we wszystkich możliwych sześciu ramkach odczytu, a następnie porównywana z sekwencjami białkowymi.
- **tblastx** – *przetłumaczona sekwencja nukleotydowa vs. przetłumaczone sekwencje nukleotydowe.* Sekwencja nukleotydowa tłumaczona jest na sekwencję białkową we wszystkich możliwych sześciu ramkach odczytu, a następnie porównywana z przetłumaczonymi na sekwencje białkowe sekwencjami nukleotydowymi. Celem takiego przeszukiwania jest znalezienie bardzo zależności pomiędzy bardzo odległymi sekwencjami nukleotydowymi.
- **tblastn** – *sekwencja białkowa vs. przetłumaczone sekwencje nukleotydowe.* Sekwencja białkowa porównywana jest z listą sekwencji białkowych pochodzących z tłumaczenia sekwencji nukleotydowych we wszystkich możliwych sześciu ramkach odczytu.

oraz adaptacje programu BLAST takie jak:

- **psi-blast** – program pozwalający na znajdowanie zależności pomiędzy odległymi ewolucyjnie białkami. Na podstawie sekwencji białkowej będącej zapytaniem wyszukiwane są sekwencje podobne, które tworzą „profil" (czyli zestaw cech charakterystycznych) wykorzystywany do przeszukiwania bazy sekwencji białkowych.
- **rps-blast** – wyszukiwanie domen białek.

Działanie programu BLAST oparte jest na metodzie heurystycznej, bazującej na lokalnych dopasowaniach krótkich fragmentów sekwencji. Sekwencja–zapytanie dzielona jest na krótkie, nakładające się słowa o długości W, następnie zaś, baza danych sekwencji przeszukiwana jest w celu znalezienia fragmentów sekwencji o długości W, takich samych jak słowa pochodzące z oryginalnej sekwencji. Dla różnych wersji algorytmu parametr W jest zmienny, np. w blastp domyślnie W=3, w blastn W=11, a w megablast W=28. Dodatkowo, w zależności od wersji programu BLAST, w dopasowywaniu mogą brać udział oryginalne krótkie słowa albo tak zwane „słowa sąsiednie" (ang. *neighbourhood words*), których podobieństwo do słów pochodzących z oryginalnej sekwencji nie przekracza pewnej wartości progowej T. Następnie dla każdego znalezionego identycznego fragmentu sekwencji wyznacza się jego dopasowanie z oryginalną sekwencją, rozszerzając dopasowanie w obie strony i oceniając jakość tego dopasowania zgodnie z macierzami punktacji dla sekwencji białkowych lub nukleotydowych (dla sekwencji białkowych domyślną macierzą punktacji jest macierz BLOSUM62) i założonymi wartościami kar za występowanie przerw w dopasowaniu. Dla każdej pary sekwencji poszukuje się najlepszych dopasowań tworzących pary MSP (ang. *maximal scoring pair*) lub HSP (ang.

high scoring pair). Określona jest pewna wartość progowa S punktacji dopasowania, która musi być spełniona, aby dane uliniowienie zostało przez program uznane za dopasowanie MSP lub HSP. W przypadku, jeżeli dane dopasowanie nie może być poprawione przez dalsze wydłużanie lub skracanie sekwencji dopasowywanej, proces dopasowywania jest przerywany, a dopasowany region zapamiętywany jako wynik działania algorytmu. Schemat poszukiwania par HSP przedstawiono na rysunku 4.3.



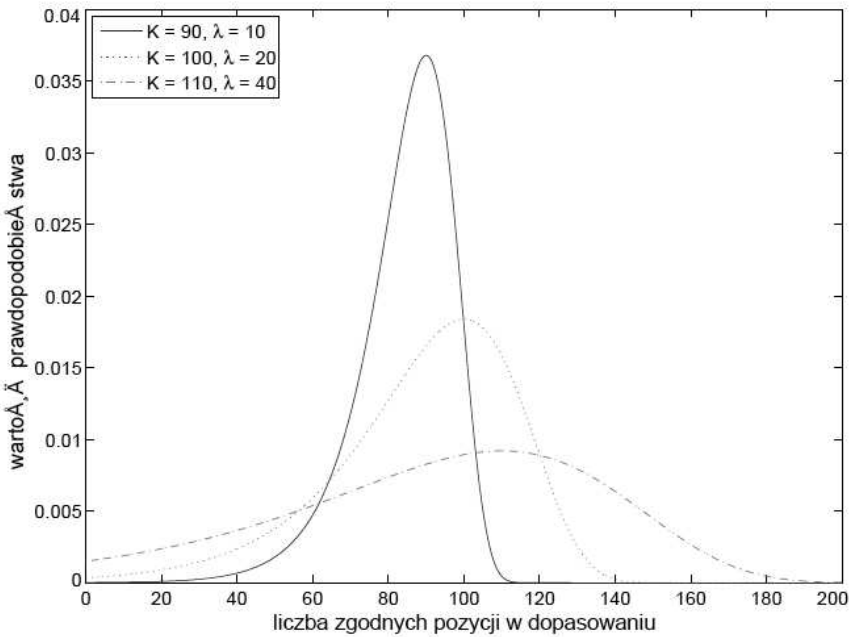
Rysunek 4.3. Schemat wyszukiwania par HSP za pomocą algorytmu BLAST. Na postawie:
`http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html`

Dla każdego zapamiętanego dopasowania wyznaczana jest jego jakość, korzystając z macierzy substytucji. Niemniej nie jest to informacja wystarczająca – potrzebna jest również jakaś metoda, która pozwoli na stwierdzenie, czy dane dopasowanie oznacza, że dwie sekwencje są względem siebie homologiczne. Innymi słowy, potrzebny jest pewien model statystyczny, który pozwoli określić, czy dane dopasowanie jest znamienne statystycznie (a tym samym prawdopodobne ewolucyjnie), czy wynika jedynie z przypadku.

Znamienność statystyczna każdego dopasowania określana jest za pomocą E-wartości (ang. *E-value*, *expected value*), która może być interpretowana jako szansa przypadkowego zaistnienia dopasowania o danej długości i wartości punktacji S. E-wartość jest parametrem, który określa jakiej liczby przypadkowych dopasowań moglibyśmy się spodziewać, gdybyśmy przeszukiwali bazę danych sekwencji o określonej długości. Wraz ze wzrostem wartości S, E-wartość maleje wykładniczo. Przykładowo dopasowanie, dla którego E-wartość

wynosi 1, oznacza że dla bazy danych o aktualnym rozmiarze można się spodziewać jednego przypadkowego dopasowania o wartości punktacji równej S. Im mniejsza E-wartość i im bliższa jest ona zeru, tym większa jest istotność dopasowania.

W celu wyznaczenia E-wartości w programie BLAST wykorzystuje się zaproponowany przez Karlina i Altchula [Karlin and Altschul, 1993] model statystyczny. Model ten może być stosowany dla lokalnych dopasowań sekwencji bez przerw w dopasowaniach. Zgodnie z modelem statystycznym Karlina i Altchula, rozkład wyników lokalnych dopasowań (wartości punktacji) z przypadkowymi sekwencjami dąży do rozkładu wartości ekstremalnej (ang. *extreme value distribution*). Krzywa reprezentująca rozkład wartości ekstremalnej jest niesymetryczna – tempo przyrostu przed maksimum jest mniejsze od tempa zmniejszania się wartości po maksimum. Rozkład ten jest zależny od dwóch parametrów λ oraz K. Parametr K określa wartość maksimum rozkładu, natomiast parametr λ wpływa na szerokość rozkładu. Na rysunku 4.4 pokazano, w jaki sposób parametry λ i K wpływają na kształt krzywej rozkładu.



Rysunek 4.4. Wpływ parametrów rozkładu wartości ekstremalnej na kształt krzywej rozkładu

Przyjmując założenie, że mamy sekwencje–zapytanie o długości m oraz bazę danych sekwencji o długości n, oczekiwana liczba przypadkowych dopasowań

sekwencji o wartości podobieństwa co najmniej S wyznaczana jest wzorem:

$$E = Kmne^{-\lambda S},$$

gdzie wartości λ i K są parametrami związanymi z przestrzenią przeszukiwań i systemem punktacji.

Znamienność statystyczna danego dopasowania związana jest z wielkością przeszukiwanej bazy danych oraz z długością sekwencji–zapytania. Im więcej jest sekwencji w bazie danych, tym wyższa jest ocena S najlepszego dopasowania parą sekwencji, w efekcie czego musi być ona wyższa, by została uznana za znamienną statystycznie. Jest to wynikiem tego, że bazy danych zawierające dużą liczbę sekwencji zwiększają szanse przypadkowego uzyskania takich dopasowań.

Przedstawiony powyżej model statystyczny został skonstruowany dla lokalnych dopasowań niezawierających przerw. W takim przypadku wartości K oraz λ można wyznaczyć w sposób analityczny. Niestety, nie ma obecnie dostępnego modelu dla dopasowań zawierających przerwy. Stąd też dla tego rodzaju dopasowań parametry wyznaczane są w sposób symulacyjny. Dla tego rodzaju dopasowań programy z rodziny BLAST korzystają z gotowych już zestawów parametrów wyznaczonych dla niektórych macierzy substytucji oraz wartości kar za wystąpienie przerwy.

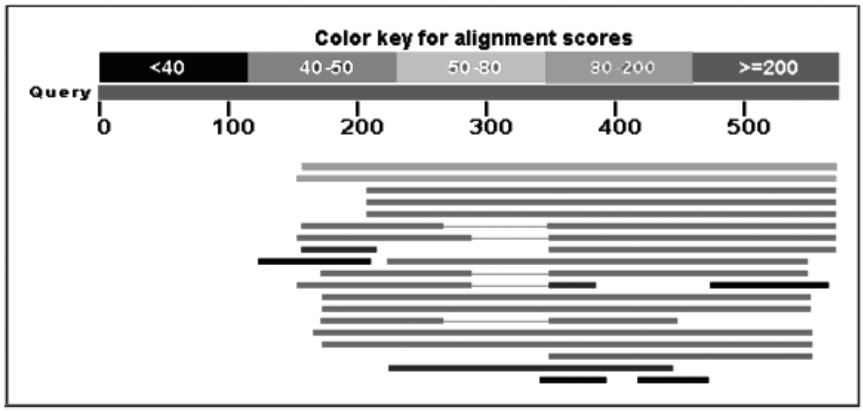
Wyniki wyszukiwania sekwencji za pomocą programów z rodziny BLAST najczęściej przedstawiane mogą być albo w postaci standardowego raportu, w formacie wygodnym do interpretacji przez człowieka, w postaci tabeli trafień (ang. *hit table*), albo w postaci strukturyzowanej w formacie XML lub ASN.1. Poniżej zostanie krótko omówiona postać raportu w formacie standardowym.

Typowy raport rozpoczyna się od części nagłówkowej zawierającej skrótową informacje na temat zapytania, jego identyfikator, typ molekuły, oraz informacje na temat bazy danych, która była przeszukiwana. Właściwe wyniki wyszukiwanie udostępnione są w formie graficznej, która pozwala na szybkie zorientowanie się w rezultatach wyszukiwania. Na rysunku 4.5 przedstawiono przykładowe wyniki przeszukiwania bazy sekwencji białkowych za pomocą narzędzia blastp, dla sekwencji o długości 570 reszt aminokwasowych.

Sekwencja–zapytanie reprezentowana jest za pomocą czerwonego paska na samej górze rysunku. Znalezione w bazie danych sekwencje pasujące są uliniowione w stosunku do sekwencji–zapytania, a kolory reprezentują jakość tego dopasowania. Najgorzej dopasowane sekwencje (poniżej 40 reszt) zaznaczone są kolorem czarnym, a najlepiej (powżej 200 reszt dopasowanych) zaznaczone są kolorem czerwonym. Kliknięcie myszką na dowolne dopasowanie przenosi użytkownika do części wyników zawierających uliniowienie konkretnego dopasowania do sekwencji–zapytania.

Poniżej graficznej reprezentacji wyników znajduje się lista sekwencji, które zostały dopasowane wraz z punktacją dopasowania oraz E-wartością. Przykład takiej listy przedstawiono na rysunku 4.6.

Ostatni fragment raportu zawiera uliniowienie znalezionych dopasowań sekwencji. Pojawiają się tu takie informacje jak: punktacja dopasowania (*score*),



Rysunek 4.5. Graficzna reprezentacja wyników wyszukiwania sekwencji homologicznych za pomocą narzędzia blastp

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|NP_989806.1| mothers against decapentaplegic homolog 3 [G...	184	2e-46	U G
ref|XP_420428.1| PREDICTED: SMAD, mothers against DPP homolog...	180	2e-45	U G
gb|AAF36969.1|AF230190_1 TGF effector Smad2 [Gallus gallus]	170	2e-42	G
ref|NP_989892.1| Sma- and Mad-related protein 2 [Gallus gallu...	170	2e-42	U G
ref|XP_001232181.1| PREDICTED: similar to MADH2 protein, part...	170	2e-42	U G
ref|NP_001019997.1| MAD, mothers against decapentaplegic homo...	125	5e-29	U G
ref|NP_001014968.1| SMAD family member 5 [Gallus gallus] >sp|...	125	6e-29	U G
gb|AAD30150.1|AF143239_1 Smad1 protein [Gallus gallus]	124	2e-28	G
gb|AAF36971.1|AF230191_1 TGF-beta response effector Smad3 [Ga...	122	6e-28	G
gb|AAD30151.1|AF143240_1 Smad5 protein [Gallus gallus]	108	7e-24	G
gb|AAF36983.1|AF233238_1 BMP signal transducer Smad1 [Gallus ...	103	2e-22	G
ref|NP_989579.1| SMAD family member 6 [Gallus gallus] >sp|Q9W...	100	2e-21	U G
gb|AC082015.1| Smad6 [Gallus gallus]	98.2	1e-20	G
gb|AAF36973.1|AF230193_1 TGF-beta signal transducer Smad8 [Ga...	90.9	2e-18	G
gb|AC082014.1| Smad7a [Gallus gallus]	89.4	5e-18	G
ref|NP_001153135.1| TGF-beta signal pathway antagonist Smad7 ...	87.4	2e-17	U G
ref|XP_427238.1| PREDICTED: similar to Smad7 [Gallus gallus]	66.2	4e-11	U G
gb|AAF36972.1|AF230192_1 TGF-beta signal pathway antagonist S...	49.3	6e-06	G
ref|XP_001235820.1| PREDICTED: hypothetical protein, partial ...	47.4	2e-05	U G
ref|XP_417159.2| PREDICTED: similar to Exophilin 5 [Gallus ga...	31.6	1.2	G
ref|XP_419807.1| PREDICTED: similar to Glutaminyl-tRNA syntha...	30.0	3.5	U G
ref|NP_001026173.1| follicular lymphoma variant translocation...	29.6	4.0	U G
ref|NP_001025718.1| adenosine deaminase-like [Gallus gallus] ...	28.9	6.7	U G

Rysunek 4.6. Wyniki wyszukiwania sekwencji za pomocą programu blastp – lista znalezionych sekwencji wraz z punktacją i E-wartością

E-wartość (*expect*), liczba identycznych par w dopasowaniu (*identities*), liczba par, dla których wartości dopasowania w tablicy substytucji są dodatnie (*positives*), oraz liczba przerw w dopasowaniu (*gaps*). Następnie przedstawione jest uliniowienie obydwu sekwencji, ale tylko w tym fragmencie sekwencji-zapytania, który dopasowany jest do wynikowej sekwencji. Dopasowania przedstawione są w wierszach, z których każdy zawiera po 60 reszt. Każdy natomiast wiersz zawiera sekwencję-zapytanie oznaczoną symbolem *Query* (powyżej) i sekwencję znalezioną oznaczoną symbolem *Sbjct* (poniżej), a pomiędzy sekwencjami znajduje się przerwa. Jeśli w uliniowieniu na tym samym miejscu pojawiają się identyczne reszty, to symbol tej reszty zapisany jest w pustej linii pomiędzy sekwencjami, zaś w przypadku gdy reszty są różne, ale mają dodatnie wartości w macierzy substytucji, pomiędzy sekwencjami pojawia się symbol +. Przykład uliniowienia dwóch sekwencji przedstawiono na rys. 4.7.

```
>gb|AAF36983.1|AF233238_1 BMP signal transducer Smad1 [Gallus gallus]
Length=291

 GENE_ID: 395680 SMAD1 | SMAD family member 1 [Gallus gallus]
(10 or fewer PubMed links)

 Score =  103 bits (257),  Expect = 2e-22, Method: Compositional matrix adjust.
 Identities = 58/136 (42%), Positives = 84/136 (61%), Gaps = 3/136 (2%)

Query  155  LQCYQQGGEDSDFVRKAIESLVKKLKDKRIELDALITAVTSNGKQPTGCVTIQRSLDGRL  214
            L  ++QG E+  +  KA+++LVKKLK K+  ++ L  A++  G Q + CVTI RSLDGRL
Sbjct  2    LLGWKQGDEEEKWAEKAVDALVKKLKKKKGAMEELEKALSCPG-QSSNCVTIPRSLDGRL  60

Query  215  QVAGRKGVPHVVYARIWRWPKVSKNELVKLVQCQT--SSDHPDNICINPYHYERVVSNRI  272
            QV+ RKG+PHV+Y R+WRWP + +  +K ++C          +CINPYHY+RV S  +
Sbjct  61   QVSHRKGLPHVIYCRVWRWPDLQSHHELKPLECCEFPFGSKQKEVCINPYHYKRVESPVL  120

Query  273  TSADQSLHVENSPMKS  288
                H E +P  S
Sbjct  121  PPVLVPRHSEYNPQHS  136
```

Rysunek 4.7. Przykład uliniowienia dwóch sekwencji

Domyślnie maksymalna liczba sekwencji dopasowanych przez wersję internetową programu BLAST wynosi 500. Wartość ta może być zmieniona w zaawansowanych opcjach programu. Oprócz punktacji dopasowania oraz uliniowienia sekwencji wiele istotnych informacji związanych ze znalezionymi sekwencjami dostępnych jest dla użytkownika za pomocą odnośników, które ze strony z wynikami wyszukiwania pozwalają przejść do opisu danej sekwencji w systemie Entrez, a co za tym idzie, pozwala na dotarcie do wielu istotnych informacji takich jak opis rodziny białek, do których sekwencja należy, czy też lista publikacji związanych z daną sekwencją.

4.3 Adresy Internetowe

- Macierz BLOSUM62
  `http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt`
- BLAST – `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

5

Bazy danych sekwencji białkowych

Znajomość sekwencji nukleotydowych pozwala na określenie, jakie geny wchodzą w skład DNA badanego organizmu. Tak naprawdę jednak, mimo iż w każdej komórce występuje dokładnie ten sam zestaw genów, komórki mogą pełnić różne funkcje w żywym organizmie. Część genów niezbędnych do pełnienia funkcji życiowych aktywowana jest we wszystkich komórkach, niektóre natomiast aktywowane są tylko w komórkach określonego rodzaju, bądź też aktywują się lub wyciszają pod wpływem specyficznych warunków. Dlatego też prawdziwą wiedzę na temat procesów biologicznych zachodzących w komórkach uzyskujemy dopiero poznając funkcje białek, które powstają w komórce w czasie gdy geny ulegają ekspresji, czyli kiedy zakodowana w DNA informacja zostaje odczytana i przepisana na jego produkty, którymi są m.in.białka. Schemat tego procesu przedstawiono na rysunku 5.1.

Rysunek 5.1. Schemat procesu ekspresji

Obecnie ogromny wysiłek w bioinformatyce położony jest na identyfikację oraz funkcjonalną analizę białek zakodowanych w poznanych genomach licznych organizmów. Początek XXI wieku to rozwój licznych metod identyfikacji białek takich jak spektrometria masowa, która pozwala na szybką identyfikację dużej liczby białek, określanie interakcji występujących pomiędzy nimi, znajdowanie ich lokalizacji w komórce, a także analizę ich biologicznej aktywności. Stąd też białkowe bazy danych pełnią w dzisiejszej biologii i medycynie bardzo istotną rolę jako repozytoria, w których możliwe jest deponowanie odkrytych białek, ich struktury, umieszczanie informacji na temat ich funkcji oraz udostępnianie zgromadzonej wiedzy szerokiemu środowisku naukowemu.

Mówiąc o białkowych bazach danych, istotne jest rozróżnienie pomiędzy nimi, a w szczególności pomiędzy danymi, które są w nich zawarte. Uniwer-

salne bazy danych białek mogą zawierać białka pochodzące ze wszystkich gatunków, podczas gdy specjalizowane bazy mogą zawierać białka z konkretnej rodziny, należące do jednej grupy lub pochodzące z tego samego organizmu. Z kolei uniwersalne bazy danych można podzielić na dwie kategorie: repozytoria sekwencji białkowych, w których zdeponowane sekwencje nie podlegają żadnemu nadzorowi i bazy danych nadzorowane przez grupy eksperckie, gdzie każdy rekord analizowany jest przez kuratorów i w razie potrzeby lub rozwoju wiedzy rozszerzany o dodatkowe informacje zweryfikowane przez ekspertów z dziedziny biologii i medycyny [Apweiler et al., 2004].

5.1 Bazy danych sekwencji białkowych

Niektóre bazy danych funkcjonują jako swego rodzaju ogólnodostępne zbiory sekwencji białkowych. Dane pojawiające się w tych bazach często generowane są w sposób automatyczny (na przykład na podstawie sekwencji nukleotydowych), sekwencje nie są w żaden sposób opisywane lub informacja na ich temat jest bardzo niewielka i zazwyczaj nie prowadzone są żadne prace mające na celu odnalezienie i usunięcie redundantnych sekwencji z baz danych.

5.1.1 Baza GenPept

Najprostszym sposobem uzyskania sekwencji białkowej jest przetłumaczenie jej wprost z sekwencji nukleotydów. Takie tłumaczenie może być robione w sposób automatyczny, na podstawie regionu kodującego sekwencji (CDS – ang. *CoDing Sequence*), kiedy sekwencja nukleotydowa deponowana jest w bazie danych. Taką właśnie bazą danych sekwencji białkowych jest GenPept (*GenBank Gene Products Data Bank*) – baza sekwencji białkowych utrzymywana w ramach zasobów NCBI. Wpisy znajdujące się w tej bazie danych uzyskiwane są na podstawie tłumaczenia sekwencji nukleotydowych deponowanych w bazach należących do konsorcjum INSDC. W efekcie takiego podejścia to samo białko może być w bazie danych reprezentowane przez wiele różnych rekordów, tym bardziej że nie istnieje żadna grupa nadzorująca tę bazę danych, której zadaniem byłoby usuwanie redundantnej informacji. Rekordy znajdujące się w tej bazie charakteryzują się bardzo skromnym opisem właściwości biologicznych sekwencji – znajduje się tam tylko informacja, którą udało się uzyskać na podstawie rekordu zawierającego odpowiadającą sekwencję nukleotydów. Baza danych GenPept nie jest uzupełniana o żadne dodatkowe informacje i nie zawiera białek uzyskanych z sekwencjonowania aminokwasów.

5.1.2 NCBI Entrez Protein

Baza białek NCBI Entrez jest kolejnym repozytorium sekwencji (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein`). Poza sekwencjami białkowymi pochodzącymi z tłumaczeń sekwencji nukleotydowych przesyłanych do

INSDC, baza ta zawiera również sekwencje pochodzące z baz takich jak Protein Information Resource (PIR), bazy SWISS-PROT, Protein Research Foundation (PRF) oraz Protein Data Bank (PDB). W porównaniu z bazą danych GenPept, w bazie tej znajduje się dodatkowa informacja pochodząca z nadzorowanych baz białkowych takich jak SWISS-PROT czy PIR. Podobnie jak w przypadku bazy GenPept sekwencje białkowe umieszczone w bazie NCBI Entrez mogą się wielokrotnie powtarzać.

5.1.3 RefSeq

Baza danych RefSeq (*Reference Sequence*) jest nadzorowaną bazą danych sekwencji (nie tylko białkowych, ale również sekwencji DNA i RNA) utrzymywaną oraz rozwijaną przez NCBI. Głównym założeniem tej bazy jest stworzenie wiarygodnego repozytorium sekwencji, w którym każdy wpis reprezentuje jedną, istniejącą w przyrodzie biomolekułę. Sekwencje umieszczane w bazie RefSeq pochodzą z GenBanku – różnica pomiędzy obiema tymi bazami jest taka, że sekwencja umieszczona w RefSeq jest syntezą większości dostępnej biologicznej wiedzy na jej temat, podczas gdy sekwencja umieszczona w bazie GenBank jest pojedynczą jednostką pochodzącą wprost z badań eksperymentalnych. W przeciwieństwie do bazy GenBank sekwencje umieszczane w bazie RefSeq są sekwencjami pochodzącymi z wybranych, modelowych organizmów – przykładowo w 2007 r. baza RefSeq zawierała sekwencje pochodzące z 4 tys. organizmów, podczas gdy w bazie GenBank w tym czasie umieszczono sekwencje dla 250 tys. różnych organizmów. Dla każdego organizmu opisanego w bazie RefSeq udostępniane są odnośniki do rekordów zawierających DNA tego organizmu, transkrypty genów oraz białka pochodzące z tych transkryptów. Podstawowe cechy tej bazy danych to brak redundancji sekwencji, połączenie każdego białka z odpowiadającą mu sekwencją nukleotydów, aktualizacja rekordów, tak aby informacja w nich zawarta odzwierciedlała aktualną wiedzę biologiczną, walidacja wprowadzanych danych oraz zapewnienie spójności formatu danych, odrębne numery dostępu wskazujące na to, że sekwencja pochodzi z bazy RefSeq. Mimo tego, w połowie 2009 roku, tylko trochę ponad 10% rekordów znajdujących się w tej bazie (1 milion 300 tys.) miało status *Reviewed* lub *Validated* wskazujący, iż rekord ten został sprawdzony przez kuratorów z NCBI. Znacząca większość rekordów (ponad 10 milionów) takiego statusu nie posiadała. Tak więc w dalszym ciągu bazę RefSeq traktować można jako repozytorium sekwencji będących wynikiem eksperymentów, a nie jak nadzorowaną przez kuratorów bazę danych. Takie bazy danych zostaną dokładniej omówione w dalszej części tego rozdziału.

5.1.4 Baza UniProt

Wymienione powyżej bazy sekwencji białkowych pozwalają na szybkie opublikowanie (a tym samym udostępnianie) dowolnej sekwencji białkowej. Nie ulega jednak wątpliwości, że baza danych, która zawiera nie tylko samą sekwencję,

ale również dostarcza użytkownikowi dodatkowych informacji na temat biologicznej funkcji biocząsteczki związanej z daną sekwencją, stanowi o wiele cenniejsze źródło informacji w porównaniu z samą tylko sekwencją aminokwasów. Uniwersalne, nadzorowane bazy danych sekwencji białkowych zawierają nie tylko informacje na temat samej sekwencji, ale także wszelką dodatkową wiedzę biologiczną z nią związaną – przed opublikowaniem rekord jest sprawdzany pod względem poprawności i kompletności informacji przez ekspertów nadzorujących bazę danych. Dodatkowo, w miarę pojawiania się nowej wiedzy na temat danej cząsteczki, informacja o niej na bieżąco uzupełniana jest w bazie. Równocześnie kuratorzy dbają, aby żadna sekwencja w bazie danych nie powtarzała się – informacje pochodzące z wszystkich publikacji dotyczących danej sekwencji białkowej umieszczane są w jednym miejscu.

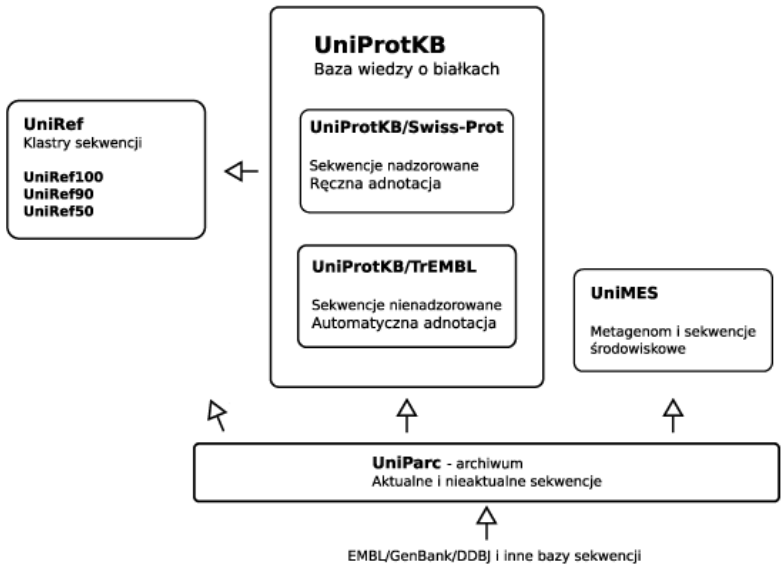
UniProt

Baza UniProt (*Universal Protein Resource*) [Consortium, 2007] jest obecnie największą z nadzorowanych bazą danych sekwencji białkowych. Baza ta powstała w ramach współpracy pomiędzy European Bioinformatics Institute (EBI), Protein Information Resource (PIR) oraz Swiss Institute of Bioinformatics (SIB). Podstawową działalnością tej bazy jest gromadzenie sekwencji białkowych, ręczny nadzór informacji zapisywanych do rekordów bazy wraz z przeprowadzaniem dodatkowych analiz każdej sekwencji, archiwizacja sekwencji, dodawanie odnośników do znajdujących się w innych bazach informacji powiązanych z daną sekwencją, jak również rozwijanie internetowych narzędzi wyszukiwania oraz analizy sekwencji znajdujących się w bazie UniProt. Baza UniProt składa się z czterech podstawowych komponentów: UniProtKB (*UniProt Knowledgebase*), gdzie umieszczone są sekwencje, UniProt Archive (UniParc), archiwum, gdzie znajdują się sekwencje historyczne, UniProt Reference Clusters (UniRef), gdzie znajdują się sekwencje pogrupowane pod względem podobieństwa oraz UniProt Metagenomic and Environmental Sequences (UniMES) – repozytorium przeznaczone dla sekwencji pochodzących z metagenomów oraz danych środowiskowych. Zależność oraz przepływ sekwencji pomiędzy wszystkimi komponentami bazy UniProt przedstawiono na rysunku 5.2.

UniProtKB

Bazę wiedzy o sekwencjach białkowych UniProtKB można podzielić na dwie części: bazę **UniProtKB/Swiss-Prot** oraz **UniProtKB/TrEMBL**.

W bazie UniProtKB/Swiss-Prot znajdują się tylko takie rekordy, których zawartość została sprawdzona i przeanalizowana przez ekspertów. Każda sekwencja znajdująca się w tej bazie uzupełniona jest o odnośniki do literatury oraz wzbogacona o wyniki automatycznych analiz przeprowadzonych pod nadzorem kuratora. Aby zapewnić poprawność zgromadzonej informacji, anotacje sekwencji przeprowadzane są przez kuratorów, którzy są specjalistami w dziedzinie biologii lub medycyny. Na anotację sekwencji białkowej umieszczonej



Rysunek 5.2. Komponenty bazy UniProt
(na podstawie: `http://pir.georgetown.edu/pirwww/dbinfo/uniprot.shtml`)

w bazie UniProt składają się: informacje specyficzne dla konkretnego enzymu, informacja na temat domen oraz innych aktywnych biologicznie miejsc na powierzchni białka, dane na temat modyfikacji posttranslacyjnych, lokalizacja(e) białka w komórce, specyfika tkankowa, struktura białka, interakcje z innymi biocząsteczkami, informacje na temat chorób powiązanych z deficytem lub nieprawidłowościami danego białka itd. Dodatkowo, jeśli w innych bazach danych znajdują są informacje związane z daną sekwencją, w rekordzie umieszczane są odnośniki wiążące sekwencję białkową z informacjami dostępnymi w zewnętrznych bazach danych. W 2008 roku baza UniProtKB powiązana była ze 118 zewnętrznymi bazami danych. Pomiędzy nimi znajdowały się bazy danych sekwencji nukleotydowych i aminokwasowych, bazy specyficzne dla danego organizmu, z którego pochodzi białko, bazy rodzin białek, bazy struktur przestrzennych i inne specjalistyczne bazy danych. Przykładowo przeglądając zawartość rekordu bazy UniProtKB/Swiss-Prot można uzyskać bezpośrednio dostęp do sekwencji nukleotydowej kodującej dane białko, ustalić związane z nim choroby, poznać charakterystykę rodziny, do której białko należy, jak również poznać strukturę przestrzenną tego białka. Z uwagi na fakt, iż każdy rekord w bazie UniProtKB/Swiss-Prot przechodzi szczegółową kontrolę kuratorów, baza ta wyraźnie wyróżnia się pozytywnie na tle innych baz sekwencji białkowych, będąc najpopularniejszą i najbardziej wiarygodną bazą sekwencji wykorzystywaną w badaniach naukowych.

Utworzenie w pełni nadzorowanego wpisu w bazie UniProtKB/Swiss-Prot jest procesem wymagającym dużego nakładu pracy ekspertów, co znacznie ogranicza liczbę nowych sekwencji, które w danym czasie mogą się pojawić w bazie UniProtKB/Swiss-Prot. Dlatego też, aby umożliwić środowisku naukowemu dostęp do większej liczby sekwencji, a przede wszystkim do sekwencji najnowszych, w ramach bazy UniProtKB utrzymywana jest baza UniProtKB/TrEMBL, która zawiera rekordy utworzone automatycznie, anotowane oraz klasyfikowane również w sposób automatyczny. W bazie tej znajdują się sekwencje białkowe odpowiadające wszystkim kodującym sekwencjom nukleotydowym (CDS) umieszczonym w bazach danych sekwencji nukleotydowych należących do INSDC, sekwencje pochodzące z bazy danych struktur białek (PDB), dane pochodzące z sekwencji bezpośrednio przesłanych do bazy UniProtKB oraz sekwencje pochodzące z literatury. Każda sekwencja, która deponowana jest w bazie UniProtKB/TrEMBL, podlega procesowi automatycznego przetwarzania. Jeśli została ona pobrana z zasobów INSDC, jej pierwszy opis pochodzi z rekordu odpowiadającej jej sekwencji nukleotydów. W dalszej części usuwana jest redundancja i jeżeli w bazie istnieje już taka sekwencja, rekordy łączone są ze sobą. W elektronicznym procesie anotacji wykorzystuje się podobieństwo sekwencji anotowanej do sekwencji umieszczonych w bazie Swiss-Prot – informacje zawarte w dobrze opisanych rekordach bazy Swiss-Prot wykorzystywane są do opisania nowych sekwencji deponowanych w bazie TrEMBL. Wybrane rekordy z bazy TrEMBL w przyszłości podlegają weryfikacji przez kuratorów, uzupełniane są o dodatkowe informacje oraz umieszczane w bazie Swiss-Prot przy równoczesnym ich usunięciu z bazy TrEMBL. W sierpniu 2009 roku w bazie Swiss-Prot znajdowało się prawie 500 tysięcy wpisów, podczas gdy baza TrEMBL zawierała ponad 9 milionów rekordów.

UniRef

Baza UniRef [Suzek et al., 2007] zawiera pogrupowane pod względem podobieństwa sekwencje białkowe pochodzące z bazy UniProtKB oraz z wybranych rekordów bazy UniParc. W zależności od wersji bazy UniRef (UniRef100, UniRef90 lub UniRef50) poziom podobieństwa pomiędzy sekwencjami może wynosić 100%, 90% lub 50%. W bazie UniRef100 każda grupa tworzona jest przez identyczne sekwencje lub podsekwencje. Zbiór takich identycznych sekwencji tworzy klaster, który stanowi równocześnie jeden wpis w bazie. Baza UniRef90 zbudowana jest na podstawie bazy UniRef100, zaś baza UniRef50 na podstawie bazy UniRef90, a w celu pogrupowania sekwencji stosowany jest algorytm grupowania hierarchicznego. Ponieważ własności biologiczne białek ściśle powiązane są z sekwencją aminokwasów, które tworzą te białka, podstawowym rodzajem analiz, które wykonywane są w bazach sekwencji białkowych, jest wyszukiwanie sekwencji podobnych. Grupowanie sekwencji pozwala ograniczyć liczbę sekwencji, a tym samym przyspieszyć proces ich wyszukiwania. Stąd też baza UniRef znajduje swoje zastosowanie w takich dziedzinach bioinformatyki jak automatyczne anotacje sekwencji, klasyfikacja rodzin białek,

genomika strukturalna, analizy filogenetyczne czy spektrometria masowa. UniRef100, UniRef90 i UniRef50 pozwalają na zmniejszenie rozmiaru bazy wejściowej odpowiednio o 10%, 40% i 70%. Każdy wpis w bazie UniRef zawiera informację o źródłowych bazach sekwencji, nazwy białek i nazwy taksonomiczne organizmów dla każdej sekwencji tworzącej klaster i reprezentowany jest przez wybraną pojedynczą sekwencję. Dodatkowo rekord zawiera nazwę, liczbę sekwencji w klastrze oraz informacje o najwyższej wspólnej jednostce taksonomicznej wszystkich elementów klastra. Klastry w bazie UniRef są aktualizowane za każdym razem, gdy pojawia się nowa wersja bazy UniProtKB.

UniParc

UniParc jest bazą – archiwum sekwencji, w której odnaleźć można informacje na temat większości kiedykolwiek zarejestrowanych sekwencji białkowych – przechowywane w tej bazie są zarówno sekwencje aktualne, jak i sekwencje nieaktualne, które zostały usunięte ze źródłowych baz danych. Dane umieszczane w bazie UniParc pochodzą nie tylko z bazy UniProtKB, ale również są to translacje regionów kodujących z baz sekwencji nukleotydów GenBank/EMBL/DDBJ oraz sekwencje białkowe pochodzące z takich źródeł jak Ensembl, H-Inv (*H-Invitational Database*), IPI (*International Protein Index*), PDB (*Protein Data Bank*), PRF (*Protein Research Foundation*), RefSeq, sekwencje pochodzące z baz modelowych organizmów (*FlyBase*, SGD, TAIR *Arabidopsis thaliana*, *WormBase*, TROME), a także sekwencje pochodzące z europejskich, amerykańskich i japońskich urzędów patentowych. Każda z sekwencji umieszczonych w bazie UniParc jest sekwencją unikalną, co powoduje, że UniParc jest największym publicznie dostępnym zbiorem niepowtarzających się sekwencji białkowych (w bazie nie istnieje również rozróżnienie pomiędzy sekwencjami pochodzącymi z różnych gatunków). Podstawowe informacje, które powiązane są z każdą sekwencją znajdującą się w UniParc to: identyfikator sekwencji, liczba sekwencji powtarzających się, źródłowa baza danych wraz z numerem dostępu oraz wersji, a także data utworzenia rekordu. Dodatkowo, każdy numer dostępu oznaczony jest jako aktualny lub posiada informację o tym, że został usunięty z bazy źródłowej. Rekordy w bazie UniParc nie są uzupełniane o żadne dodatkowe anotacje, ponieważ anotacje takie mają sens jedynie w przypadku gdy odnoszą się do sekwencji kodującej konkretne białko w konkretnym organizmie – białka o tej samej sekwencji mogą pełnić różne funkcje w zależności od gatunku, tkanki, stadium rozwoju itd.

UniMES

Sekwencje umieszczone w bazie UniProt pochodzą z organizmów, dla których dokładnie określone są właściwe jednostki taksonomiczne. W ostatnich latach rozwinęły się jednakże badania nad sekwencjami, których pochodzenie określić można jako metagenomiczne. Metagenom jest to informacja genetyczna zawarta we wszystkich mikroorganizmach funkcjonujących w danym środowisku, a poprzez badania metagenomiczne rozumiemy prowadzone na dużą skalę

analizy genomów pobranych z mikroorganizmów bezpośrednio z właściwego im środowiska. Taki sposób analizy różni się od tradycyjnego podejścia stosowanego w mikrobiologii, gdzie sekwencje pobierane są z organizmów hodowanych w sterylnych laboratoriach. Obecnie uznaje się, że ogromna grupa mikroorganizmów pochodząca z różnych środowisk nie może zostać wyhodowana w warunkach laboratoryjnych (a tym samym nie można zsekwencjonować ich DNA). Aby poznać te organizmy, konieczne jest pobieranie próbek pochodzących bezpośrednio ze środowiska, w którym one funkcjonują. Baza danych UniMES – Metagenomic and Environmental Sequences – zawiera dane pochodzące z projektu Global Ocean Sampling Expedition (GOS). 25 milionów sekwencji nukleotydowych mikroorganizmów, które zostały zebrane w wyniku tej ekspedycji, zostały zapisane w bazach sekwencji nukleotydowych należących do INSDC i na ich podstawie przewidziano istnienie około 6 milionów białek. Sekwencje umieszczane w UniMES poddawane są procesowi automatycznej anotacji oraz klasyfikacji, dzięki czemu baza UniMES stanowi obecnie unikatowe źródło sekwencji białkowych pochodzących ze środowiska. Sekwencje zdeponowane w bazie UniMES nie są umieszczane w bazie UniProtKB oraz UniRef, natomiast zapisywane są również w bazie UniParc.

5.1.5 PIR

Historycznie pierwszym opublikowanym zbiorem sekwencji białkowych był wydany przez Margaret Dayhoff w 1965 roku *Atlas of Protein Sequence and Structure*. Atlas ten wydawany był w formie książki do 1978 roku i można powiedzieć, że był on pewnego rodzaju „bazą danych" w formie papierowej. Grupa badawcza związana z Margaret Dayhoff pierwszy raz opublikowała w formie elektronicznej sekwencje w 1984 roku pod nazwą PIR PSD (*Protein Sequence Database of Protein Information Resource*). Przez wiele lat baza danych PIR PSD (a później PIR PSD International) była uznanym zbiorem sekwencji białkowych, ale z uwagi na fakt, iż jej zawartość zaczęła się pokrywać z zawartością bazy UniProt, w 2004 roku została wydana ostatnia wersja PIR PSD, a zasoby bazy zostały przepisane do bazy UniParc oraz do innych odpowiednich sekcji bazy UniProt.

Obecnie działalność PIR [Barker et al., 2000] koncertuje się wokół następujących zagadnień: rozwój i utrzymywanie bazy UniProt, integracja danych – bazy iProClass, iProLINK oraz rozwój bazy rodzin białek PIRSF. Dokładniejsze omówienie baz danych iProClass, iProLINK oraz PIRSF znajduje się w niniejszym rozdziale.

5.2 Bazy rodzin białek

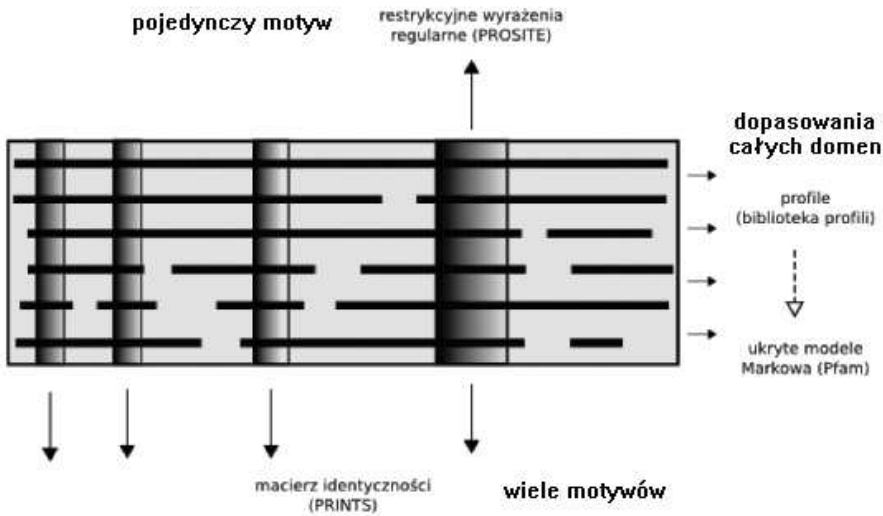
Mając pewną sekwencję białkową i przeszukując internetowe zasoby biologicznych baz danych, poszukujemy najczęściej odpowiedzi, jakie cechy posiada oraz jakie funkcje biologiczne może pełnić białko reprezentowane przez intere-

sującą nas sekwencję. W przypadku jeżeli taka sekwencja istnieje już w bazie danych, uzyskanie takiej informacji jest bardzo łatwe – możemy wprost skorzystać ze zgromadzonej wcześniej informacji. Co jednak w przypadku, gdy próbujemy określić własności całkowicie nowej sekwencji białkowej? Dawno już zauważono, iż białka o podobnych sekwencjach charakteryzują się podobieństwem pod względem funkcjonalnym, stąd też mając nieznaną wcześniej sekwencję białkową i wiedząc, jakie funkcje pełnią białka o sekwencji podobnej, możemy przewidywać funkcje naszego nowego białka. Im mamy więcej białek o sekwencji podobnej pełniących podobne funkcje biologiczne, tym większe jest prawdopodobieństwo, że nasze przewidywania będą poprawne. Stąd też w dzisiejszej bioinformatyce ogromnie ważna rola przypada tak zwanym bazom rodzin białek (nazywanych również bazami danych wzorców sekwencji lub wtórnymi bazami danych).

Tworząc wtórne bazy danych, wykorzystuje się fakt, iż ogromna liczba sekwencji umieszczonych w bazach danych takich jak UniProtKB może, na podstawie podobieństw pomiędzy sekwencjami, zostać podzielona na grupy nazywane rodzinami białek. Okazuje się, że białka należące do poszczególnych rodzin zazwyczaj pełnią podobne funkcje biologiczne i często pochodzą od wspólnego przodka. Badając rodziny białek zaobserwowano, że pewne fragmenty sekwencji zostały w trakcie ewolucji lepiej zakonserwowane od innych elementów. Zachowane fragmenty zazwyczaj okazują się ważne z punktu widzenia funkcjonowania białka, a także często mają wpływ na kształt przestrzennej struktury białka. Analizując stałe i zmienne fragmenty sekwencji białek należących do danej rodziny, można określić tak zwaną sygnaturę rodziny lub domeny, która pozwala na odróżnienie wszystkich członków należących do tej rodziny od pozostałych białek. Inne określenie takiego powtarzającego się fragmentu sekwencji to motyw. Często stosowaną analogią jest tu porównywanie sygnatury do odcisku linii papilarnych. Tak jak odcisk linii papilarnych może być wykorzystany w celu identyfikacji konkretnej osoby, tak samo sygnatura białkowa może zostać wykorzystana, aby przyporządkować nowo zsekwencjonowane białka do właściwej im rodziny białek, co pozwala na formułowanie hipotez na temat biologicznych funkcji pełnionych przez dane białko. Z uwagi na fakt, że bazy rodzin białek tworzone są na podstawie dopasowań wielosekwencyjnych, wykrywanie odległego pokrewieństwa pomiędzy sekwencjami jest zazwyczaj skuteczniejsze niż bezpośrednie przeszukiwanie baz danych sekwencji.

Podział sekwencji na rodziny białek może się odbywać według różnych kryteriów, stąd też istnieje wiele różnych baz rodzin białek. Niniejszy rozdział stanowi próbę przedstawienia czytelnikowi kilku najważniejszych baz danych rodzin białek wraz z krótki opisem metod, które każda z baz danych wykorzystuje do znajdowania rodzin białek. Istnieje wiele odmiennych podejść, które mogą być wykorzystywane do wyszukiwania takich sekwencji – na rysunku 5.3 przedstawiono różne schematy, które wykorzystywane są do tworzenia wzorców charakteryzujących poszczególne typy rodzin białek. Każda z metod wyszukiwania rodzin białek ma swoje wady oraz zalety i każda z nich posiada

swój obszar zastosowań, w którym sprawdza się najlepiej. Żadnego z podejść nie można traktować jako podejścia najlepszego bądź najgorszego, a raczej należy uznać, że metody te uzupełniają się wzajemnie i oferują zróżnicowane możliwości tworzenia dopasowań wielosekwencyjnych.



Rysunek 5.3. Schemat obrazujący trzy główne podejścia do tworzenia baz rodzin białek. Klasyfikacja sekwencji może być prowadzona na podstawie: (1) istnienia pojedynczego motywu, (2) jednoczesnego występowania wielu motywów lub (3) wyznaczonego dopasowania sekwencji całych domen.
Na podstawie: [Higgs and Attwood, 2008]

5.2.1 PROSITE

Baza danych PROSITE [Hulo et al., 2008] powstała w 1988 roku i tym samym historycznie jest najstarszą bazą danych rodzin białek. Rekordy umieszczone w bazie PROSITE można podzielić na trzy grupy: dokumenty, które zawierają opis rodzin białek, domen i miejsc funkcyjnych (odcinków sekwencji białka o ustalonej funkcji biologicznej), oraz wzorce i profile, które pozwalają na ich identyfikację. Uzupełnieniem bazy PROSITE jest ProRule – zbiór reguł zbudowanych na podstawie wzorców i profili, które zwiększają zdolności dyskryminacyjne wzorców oraz profile poprzez dostarczanie dodatkowej informacji o funkcjonalności iłub strukturze najważniejszych aminokwasów. Reguły ProRule wykorzystywane są w procesie anotacji i pozwalają na automatyczne wygenerowanie opisu białka w formacie UniProtKB/Swiss-Prot. W momencie powstania w bazie PROSITE dostępnych było 58 wzorców, podczas gdy w sierpniu 2009 roku w bazie znajdowało się 1308 wzorców, 862 profile i 868 reguł ProRule.

PROSITE Patterns

Twórcy bazy wzorców PROSITE uznali, że każdą rodzinę białek można scharakteryzować za pomocą krótkich, dobrze zakonserwowanych fragmentów sekwencji aminokwasów. Zazwyczaj, poszukiwany wzorzec zawiera około 10-20 aminokwasów i jest bardzo istotnym fragmentem sekwencji z punktu widzenia właściwości, jakie posiada dane białko. Istnienie takiego krótkiego, dobrze zakonserwowanego motywu najczęściej związane jest z istnieniem aktywnego miejsca enzymu, miejsca wiązania jonu metalu albo ligandu, czy też miejscem powstawania wiązań disulfidowyuch. Motyw sekwencyjny w bazie PROSITE reprezentowany jest w formie wyrażenia regularnego (ang. *regular expression*) lub inaczej wzorca. Na każdej pozycji takiego wzorca znajdować się może dowolny aminokwas lub pewien akceptowalny podzbiór aminokwasów, może również występować pewna ilość powtórzeń. Istnieją również takie miejsca wzorca, do których pasuje tylko jeden aminokwas lub można określić, jakie aminokwasy nie powinny się na danej pozycji znajdować. W szczególnych przypadkach niektóre rodziny białek są charakteryzowane przez współwystępowanie kilku różnych wzorców.

Zalety stosowania wzorców to przede wszystkim łatwość ich zrozumienia przez użytkownika oraz fakt, że wzorce zorientowane są na najbardziej zakonserwowane fragmenty sekwencji, co ma swoje uzasadnienie ewolucyjne z uwagi na znaczenie zachowanych motywów dla własności biologicznych białek. Ponieważ zakonserwowane sekwencje są krótkie, przeszukiwanie bazy danych pod kątem wybranego wzorca jest operacją, którą można wykonać bardzo szybko. Natomiast podstawową wadą wzorców jest to, że są one wyrażeniami jakościowymi, to znaczy, że dany fragment sekwencji może pasować do wzorca lub jest przez ten wzorzec odrzucany i nie istnieje żaden sposób oceny, który pozwoliłby określić, w jakim stopniu analizowana sekwencja podobna jest do wzorca. Taka wrażliwość na zmianę nawet pojedynczego aminokwasu w sekwencji może rodzić problemy, szczególnie wtedy, gdy pokrewieństwo pomiędzy przedstawicielami niektórych rodzin jest zbyt odległe.

PROSITE Profiles

W celu zmniejszenia ograniczeń, jakie niesie ze sobą stosowanie wzorców, w bazie PROSITE stworzono profile, które pozwalają na szacowanie dopasowania analizowanej sekwencji do wzorca nie tylko pod względem jakościowym, ale również ilościowym. Profil (lub zamiennie – tablica wag) jest tablicą zawierającą punktację, którą przyznaje się za występowanie zmian na kolejnych pozycjach dopasowania. Każde wystąpienie w sekwencji określonego typu aminokwasu, zastąpienie go przez inny aminokwas, usunięcie lub wstawienie jest wartościowane. Na podstawie sumy punktów przyznanych dopasowywanej sekwencji można wyznaczyć wartość podobieństwa dla dopasowania pomiędzy profilem a sekwencją. Możemy określić również pewną graniczną wartość dopasowania i uznać, że dopasowanie, dla którego wyznaczona wartość podobieństwa przekracza wartość graniczną, oznacza wystąpienie motywu. Profile

pozwalają na wykrywanie odległych relacji ewolucyjnych pomiędzy sekwencjami, w których jedynie nieliczne fragmenty sekwencji są zakonserwowane. Różnią się również od wzorców tym, że charakteryzują rodziny domeny białek nie tylko na niewielkim, najbardziej zakonserwowanym fragmencie sekwencji, ale na całej jej długości.

Wzorce oraz profile PROSITE są dwoma narzędziami, które wzajemnie się uzupełniają. Wzorce ograniczone do niewielkich, zakonserwowanych ewolucyjnie fragmentów sekwencji sprawdzają się znakomicie w zadaniach przewidywania biologicznych funkcji nowych białek – np. ich aktywności enzymatycznej. Z drugiej strony, profile, które wyznaczają podobieństwo w obszarze całej domeny białka, z powodzeniem wykorzystywane do przewidywania struktury białka.

5.2.2 PRINTS

Analizując dopasowania sekwencji, zaobserwowano, że większość rodzin białek charakteryzuje się występowaniem nie jednego, ale kilku silnie konserwatywnych motywów sekwencyjnych. Zbiór takich motywów (albo ich większości) występujących w rodzinie białek określa się mianem **śladu rodziny białek** (ang. *fingerprint*). Ślad jest zestawieniem charakterystycznych cech sekwencji definiujących ich przynależność do pewnej rodziny białek. Klasyfikowanie sekwencji na podstawie kilku różnych motywów charakteryzuje się lepszą zdolnością rozpoznawczą, gdyż zwykle nie wymaga się, aby w dopasowaniu brały udział wszystkie motywy tworzące ślad danej rodziny białek. Przykładowo sekwencja, która dopasowuje cztery z siedmiu motywów definiujących rodzinę białek, może wciąż zostać uznana za sekwencję należącą do danej rodziny, jeżeli pasujące motywy ułożone są w odpowiedniej kolejności, a odległości pomiędzy pasującymi motywami są zgodne z oczekiwanymi odległościami, jakie powinny występować pomiędzy motywami sekwencji należącej do danej rodziny białek.

Baza danych zawierająca ślady rodzin białek nazywa się PRINTS, a jej pierwsza elektroniczna wersja została wydana w 1993 roku [Attwood, 2002]. Obecnie baza danych zarządzana jest przez uniwersytet w Manchesterze (Wielka Brytania), a wersja bazy z lutego 2009 liczyła 1950 rekordów zawierających 11 625 indywidualnych motywów. W każdym kolejnym wydaniu bazy pojawia się 50 nowych rekordów ze śladami rodzin białek. Z uwagi na fakt, iż ślady rodzin białek wyszukiwane są ręcznie, przyrost nowych rekordów w bazie jest dość powolny. Do 2003 roku nowe wydania bazy pojawiały się regularnie co kwartał, natomiast w ostatnich latach tempo dodawania nowych rekordów do bazy wyraźnie zmalało – pomiędzy 2005 a 2009 rokiem w bazie pojawiło się tylko 100 nowych rekordów. Dlatego, aby przyspieszyć proces wyszukiwania nowych sekwencji, opracowano dodatek do bazy PRINTS – narzędzie **prePRINTS** służące do automatycznego wyszukiwania śladów rodzin białek. Znalezione przez prePRINTS ślady rodzin białek są potencjalnymi kan-

dydatami do umieszczenia w bazie PRINTS – po manualnym sprawdzeniu i uzupełnieniu przez kuratorów.

5.2.3 Pfam

Baza danych Pfam [Bateman et al., 2004] jest kolejną bazą domen oraz rodzin białek. Identyfikacja rodziny, do której należy białko w bazie Pfam, odbywa się na podstawie całej sekwencji (a nie jednego czy kilku wybranych motywów), a dopasowania reprezentowane są przez profile i przez ukryte modele Markowa (HMM, ang. *Hidden Markov Models*). Ukryte modele Markowa są statystyczną metodą klasyfikacji sekwencji zdarzeń. Łańcuch Markowa może być traktowany jako pewien proces stochastyczny, którego ewolucja zależy od jego aktualnego stanu. Stan reprezentuje zaistnienie pewnego zdarzenia (np. wystąpienie danego znaku, brak wystąpienia znaku) oraz istnieje pewien zestaw dopuszczalnych przejść pomiędzy stanami. W zastosowaniu do modelowania konserwatywnych odcinków sekwencji model HMM przybiera postać liniowego łańcucha trzech typów stanów: stanu dopasowującego element sekwencji (ang. *match state*), stanu wstawiającego element sekwencji (ang. *insert*) oraz stanu usuwającego element sekwencji (ang. *delete state*).

Rodziny białek znajdujące się w bazie Pfam podzielone są na dwie kategorie: **Pfam-A** oraz **Pfam-B**. Każda rodzina należąca do bazy Pfam-A reprezentowana jest za pomocą trzech odrębnych elementów: (1) zbioru dopasowań zalążkowych (ang. *seed alignment*) – ręcznie zweryfikowanego przez kuratora zbioru sekwencji reprezentatywnych dla danej rodziny, (2) profili HMM zbudowanych na podstawie dopasowań zalążkowych oraz (3) automatycznie wygenerowanego pełnego dopasowania, które zawiera wszystkich członków rodziny wykrytych podczas przeszukiwania podstawowych baz sekwencji. Różnica pomiędzy zbiorem dopasowań zalążkowych a zbiorem pełnych dopasowań ułatwia uaktualnianie bazy danych: dopasowania zalążkowe są stałe, podczas gdy pełne dopasowania oraz profile HMM mogą być generowane automatycznie dla każdego nowego wpisu pojawiającego się w bazie sekwencji. Wpisy umieszczone w bazie Pfam-B są automatycznie generowane na podstawie zasobów bazy ProDom i reprezentowane są poprzez pojedyncze dopasowanie. Tak więc z jednej strony użytkownik ma do wyboru bazę Pfam-A, która nadzorowana jest ręcznie przez kuratorów, co uwiarygadnia otrzymany w wyniku analizy zbiór dopasowań i zapewnia wysoką jakość anotacji, a z drugiej strony użytkownik może skorzystać z automatycznej bazy Pfam-B, dzięki czemu liczba pojawiających się nowych rekordów w bazie Pfam zwiększa się wraz z przyrostem danych w bazach sekwencji.

Każda rodzina określona jest za pomocą nazwy, stałego numeru dostępu i zawiera opis parametrów modelu, który został wykorzystany do zidentyfikowania jej członków. Dołączony jest również krótki opis funkcjonalny, informacje na temat interakcji z innymi rodzinami oraz struktura domeny. Bardzo często umieszczone są również odnośniki do dokumentacji z innych źródeł takich jak bazy PROSITE czy PRINTS.

Większość rodzin znajdujących się w bazie Pfam powstała na bazie odpowiadającym tym rodzinom zbiorów motywów zdefiniowanych w bazie PROSTIE lub w bazie PRINTS. W wielu przypadkach jednakże przyporządkowania sekwencji do poszczególnych rodzin mogą się różnić pomiędzy bazą Pfam a bazami PROSTIE i PRINTS. Jest to oczywiście wynik różnych metod stosowanych do przyporządkowywania sekwencji do poszczególnych rodzin. Najczęstsze różnice polegają na tym, że wzorce PROSTIE lub ślady PRINTS rozpoznają silnie zakonserwowany motyw, który dzielony jest pomiędzy członków nadrodziny białek, i traktują jej członków jako należących do jednej rodziny, podczas gdy w bazie Pfam wyróżnionych zostanie kilka rodzin białek. Z drugiej strony zdarzają się sytuacje, gdy baza Pfam rozpozna nadrodzinę białek, podczas gdy w bazach PROSTIE lub PRINTS sekwencje należące do tej nadrodziny zostaną przyporządkowane do oddzielnych rodzin charakteryzowanych przez odrębne motywy.

5.2.4 ProDom

Zasoby bazy danych ProDom [Bru et al., 2005], utrzymywanej oraz rozwijanej na Uniwersytecie Claude Bernard we Francji, generowane są w sposób automatyczny. Do konstruowania rodzin białek w bazie ProDom wykorzystuje się programu MKDOM2, który iteracyjnie przegląda bazy danych sekwencji białkowych za pomocą algorytmu PSI-BALAST (*Position specific iterative BLAST*) w poszukiwaniu homologicznych domen. Sekwencje źródłowe wykorzystywane do zbudowania bazy ProDom pochodzą z bazy UniProtKB i są to tylko sekwencje ciągłe (tj. pozbawione przerw), natomiast do inicjalizacji procedury grupowania wykorzystano domeny pochodzące z bazy danych SCOP. Znalezione sekwencje tworzące rodzinę są do siebie dopasowywane.

Każdy wpis w bazie ProDom charakteryzowany jest przez unikalny numer dostępu. Z uwagi na fakt, że każde wydanie bazy budowane jest od nowa, konieczne jest stworzenie takiego sposobu, który pozwoliłby w różnych wersjach bazy danych nadawać te same numery dostępu odpowiadającym sobie rodzinom białek. W tym celu napisany został program *MatchDom*, który na podstawie podobieństwa pomiędzy rekordami różnych wersji bazy ProDom przepisuje numery dostępu odpowiadającym sobie rodzinom białek. Zdefiniowano również zasady nadawania numeru dostępu w przypadku, gdy struktura rodziny ulega zmianie. Jeżeli rodzina zostanie podzielona na dwie części, wówczas numer dostępu z poprzedniej wersji przypisywany jest jednej z nowo tworzonych rodzin oraz tworzony jest nowy numer dostępu dla drugiej rodziny. Jeśli natomiast kilka wpisów zostanie połączonych w jedną rodzinę, nowy rekord otrzymuje poprzedni numer dostępu jednej z rodzin tworzących nowy wpis. Numery dostępu pozostałych rodzin również wskazują na nowy rekord, niemniej jednak oznaczone są komentarzem mówiących o ich dezaktualizacji (ang. *obsolete*).

Wersja bazy ProDom z listopada 2008 roku zawierała 574656 rodzin domen posiadających co najmniej dwie sekwencje.

5.2.5 PIRSF

Baza danych PIRSF – (*PIR SuperFamily*) [Wu et al., 2004] jest bazą danych rodzin białek rozwijaną od 1993 roku przez grupę badawczą Protein Information Resources. Rodziny PIRSF uporządkowane są w sposób hierarchiczny wokół koncepcji rodzin i nadrodzin obejmujących białka o podobnej sekwencji. System klasyfikacji bazuje na poszukiwaniu podobieństwa między białkami poprzez analizę całych sekwencji, a nie poprzez porównywanie domen czy motywów. Zależności pomiędzy rodzinami reprezentowane są jako sieć powiązań zbudowana na podstawie zależności ewolucyjnych występujących pomiędzy białkami. W sieci reprezentowanej w formie acyklicznego grafu skierowanego można wyróżnić trzy poziomy: poziom rodzin homeomorficznych, poziom nadrodzin oraz poziom podrodzin. Każdy węzeł grafu reprezentowany jest przez rodzinę, nadrodzinę lub podrodzinę PIRSF oraz unikalny identyfikator (UID), w postaci PIRSFxxxxxx (gdzie x oznacza cyfrę).

Podstawowe węzły tworzone są przez **rodziny homeomorficzne**, które zawierają białka będące równocześnie homologami (pochodzące od jednego przodka), jak i homomorfami (charakteryzującymi się podobieństwem na długości całej sekwencji oraz podobną architekturą domenową). Każde białko może być przypisane tylko i wyłącznie do jednej rodziny, która może posiadać jednego lub więcej rodziców oraz jednego lub więcej potomków. Każdy węzeł musi zwierać opis, na który składają się: nazwa rodziny, typ relacji rodzic/potomek, listę białek wchodzących w skład rodziny oraz sygnatura architektury domenowej białka. Opis może być rozszerzony o informacje na temat rodziny, odnośniki do literatury oraz słowa kluczowe/terminy GO opisujące daną rodzinę. Dodatkowo dla każdego węzła automatycznie generowane są dopasowania wielosekwencyjne, drzewo filogenetyczne rodziny oraz modele HMM rodziny.

Ponad węzłami podstawowymi zdefiniowana została sieć węzłów **nadrodzin** (ang. *superfamily nodes*), które stanowią połączenia pomiędzy odległymi rodzinami białek i pojedynczymi białkami nie należącymi do żadnej z rodzin. Nadrodziny mogą być tworzone poprzez sekwencje homeomorficzne – rodziny homeomorficzne o wspólnej architekturze domenowej wraz z podobieństwem na całej długości sekwencji, chociaż częściej tworzone są na podstawie podobieństwa pomiędzy domenami – rodziny domenowe o wspólnej architekturze domenowej z częściowym podobieństwem na całej długości sekwencji. Opis węzłów poziomu nadrodzin zawiera nazwę nadrodziny, typ relacji rodzic/potomek, listę białek wchodzących w skład nadrodziny oraz listę wspólnych domen nadrodziny. Dodatkowo opis może być rozszerzony o informacje na temat rodziny, odnośniki do literatury oraz słowa kluczowe/terminy Ontologii Genowych.

Z kolei poniżej węzłów podstawowych znajdują się węzły **podrodzin**, które reprezentowane są przez homeomorficzne i homologiczne grupy białek i tworzą podział białek zgodny z pełnioną przez nie funkcją biologiczną iłub posiadających zmienną architekturę domenową. Podobnie jak w przypadku węzłów rodzin homeomorficznych, anotacje węzłów podrodzin zawierają nazwę

podrodziny, typ relacji rodzic/potomek, listę członków podrodziny i dodatkowo mogą zawierać opis, listę publikacji oraz słowa kluczowe. Dodatkowo dla każdej podrodziny generowane są dopasowania wielosekwencyjne, drzewa filogenetyczne, profile HMM, a także dostępne są dopasowania wielosekwencyjne ręcznie nadzorowane przez kuratorów.

5.3 Integracja zasobów pochodzących z odrębnych baz danych

Projektując bioinformatyczną bazę danych, której zadaniem jest integracja danych pochodzących z różnych źródeł, można przyjąć kilka różnych podejść. Podstawowe z nich to: połączenie różnych baz danych za pomocą hiperłączy, utworzenie hurtowni danych, do której fizycznie zostaną skopiowane dane z innych źródeł, zadawanie bezpośrednich zapytań pomiędzy różnymi bazami danych oraz stworzenie federacji, która w ramach jednego projektu udostępniać będzie dane pochodzące z różnych źródeł. Stworzenie hurtowni danych to podejście „danych ściśle pozwiązanych" – dane pochodzące z wielu różnych baz konwertowane są do jednej zunifikowanej bazy. Zaletą takiego rozwiązania jest uzyskanie kontroli nad lokalnymi danymi, wadą zaś problem aktualizacji danych pochodzących z wielu źródeł. Przeciwieństwem do podejścia „danych ściśle pozwiązanych" jest podejście „danych luźno powiązanych", gdzie stosuje się model łączenia różnych źródeł za pomocą hiperłączy. Takie podejście pozwala użytkownikowi uzyskać mnóstwo informacji – o ile będzie na tyle cierpliwy, aby w celu ich uzyskania podążać za kolejnymi odnośnikami.

5.3.1 InterPro

W 1999 r. powstało konsorcjum InterPro [Mulder et al., 2002] założone przez grupę SWISS-PROT z European Bioinformatics Institute, Swiss Institute of Bioinformatics oraz założycielskie bazy danych Prints, PROSITE, Pfam oraz ProDom. Wynikiem tej inicjatywy było powstanie bazy InterPro, której pierwsze wydanie miało miejsce w 2000 roku. W późniejszym czasie do konsorcjum dołączyły kolejne bazy danych: SMARTSMART oraz TIGRFAMs, PIRSF, GENE3D, SUPERFAMILY, a ostatnio HAMAP.

Każda z baz danych wchodząca w skład InterPro rozwija metody, które mogą zostać wykorzystane w celu wyznaczenia jakości dopasowania sekwencji białkowej do zadanej sygnatury. Dla niektórych metod klasyfikacja może być binarna (czyli pasuje – nie pasuje), w innych przypadkach otrzymujemy pewną wartość liczbową i sami możemy określić, jaka wartość progowa określa dopasowanie sekwencji do sygnatury. W tabeli 5.1 przedstawiono metody wyszukiwania rodzin białek wykorzystywane przez bazy danych będące członkami InterPro wraz z informacją, skąd pochodzą źródłowe sekwencje wykorzystywane przez każdą z tych baz.

Tabela 5.1. Bazy danych rodzin białek będące członkami InterPro

Baza danych	Źródło danych	Informacja
PROSITE patterns	UniProtKB/SwissProt	proste wyrażenia regularne
PROSTIE profiles	UniProtKB/SwissProt	tablice wag
HAMAP	sekwencje mikrobów pochodzące z UniProtKB/SwissProt	tablice wag
PRINTS	UniProtKB	ślady sekwencji białkowych
PANTHER	UniProt	modele HMM
PIRSF	UniProtKB	modele HMM
Pfam	UniProtKB, GenPept, dane metagenomiczne	modele HMM
SMART	UniProtKB, ENSEMBL	modele HMM
TIGRFAMs	UniProtKB, GenPept, dane metagenomiczne	modele HMM
Gene3D	GenBank	modele HMM
SUPERFAMILY	UniProtKB, PDB, kompletne genomy organizmów	modele HMM
ProDom	UniProtKB	grupowanie sekwencji za pomocą PSI-BLAST

Każda z metod generowania rodzin białek posiada swoje zalety i każda z nich najlepiej sprawdza się na swoim własnym polu zastosowań, dlatego nie należy spośród nich wyróżniać lepszych i gorszych rozwiązań, a raczej traktować je jako wzajemnie uzupełniające się.

Wzorce, które są proste do skonstruowania i bardzo dobrze sprawdzają się do wykrywania krótkich sekwencji istotnych dla biologicznej funkcji białka, zazwyczaj nie radzą sobie w sytuacjach dopasowywania odległych członków rodzin – w tym zastosowaniu najlepiej sprawdzają się profile oraz modele HMM, które dopasowują sekwencję na większym jej obszarze, a z uwagi na fakt, że nie muszą przestrzegać ścisłych zasad dotyczących tego, jakie aminokwasy są w określonej części sekwencji akceptowalne, potrafią odległe sekwencje dopasować do danej rodziny. Ślady rodzin białek również mają problemy przy analizie odległych rodzin, w przypadkach dopasowań wielosekwencyjnych, jeżeli liczba sekwencji ciągłych (ang. *ungapped*) w dopasowaniu jest zbyt mała, natomiast znakomicie sprawdzają się w problemach klasyfikacji podrodzin białek – do czego z kolei nie bardzo nadają się profile oraz modele HMM. Bazy danych grupowania sekwencji takie jak ProDom znajdują swoje zastosowanie na polu identyfikacji domen. Automatyzacja procesu analizy w bazie ProDom pozwala uzyskać wysokie pokrycie źródłowych baz danych sekwencji, ale też z uwagi na

brak nadzoru kuratorów, pojawiają się wątpliwości dotyczące wiarygodności informacji umieszczonej w bazie.

Każda z baz rodzin białek rozwijana jest w celu stworzenia jak najlepszej klasyfikacji białek, ale każdą z nich również charakteryzuje inne podejście do analizy sekwencji, czego wynikiem jest powstawanie różnych i w większej części niezależnych źródeł informacji o białkach. Z jednej strony bazy te opisują (pokrywają – ang. *coverage*) podobne zbiory sekwencji źródłowych, a bazy wzorców mają nawet podobny rozmiar, z drugiej – zawartość baz jest inna. Niektóre z nich koncentrują się na domenach, inne na miejscach aktywnych, a jeszcze inne na rodzinach białek. Stąd też, zajmując się analizą białka, warto przeszukiwać wszystkie dostępne zasoby baz rodzin białek, co pozwala uzyskać możliwie obszerny i szczegółowy opis analizowanej sekwencji.

Powyższe różnice w podejściach definiowania rodzin białek reprezentowane przez różne bazy danych przy równoczesnej uzasadnionej potrzebie analizy informacji pochodzącej ze wszystkich tych źródeł zaowocowały inicjatywą powstania bazy InterPro, której celem jest zintegrowanie danych pochodzących z różnych źródeł, a tym samym uproszczenie całego procesu pozyskiwania informacji na temat rodzin białek, domen i miejsc aktywnych. Obecnie w bazie InterPro znaleźć można informacje pochodzące z baz PROSITE, PRINTS, Pfam, ProDom, SMART oraz TIGRFAMs. Sygnatury pochodzące z tych baz danych, które opisują tę samą rodzinę białek, domenę, powtórzenia oraz modyfikacje posttranslacyjne zostały zintegrowane, tworząc pojedynczy wpis.

Proces integracji danych pochodzących z tak wielu różnych baz danych, z których każda stosuje własne kryteria podziału białek oraz definicje domen, jest skomplikowany i wymaga zdefiniowania reguł dotyczących tego, kiedy daną rodzinę pochodzącą z jednej bazy można uznać za odpowiadającą rodzinie pochodzącej z innej bazy. Dwie sygnatury pochodzące z różnych baz danych mogą zostać zintegrowane, jeżeli przynajmniej częściowo zachodzą na siebie na tej samej pozycji w sekwencji białkowej. Dodatkowo lista białek pokrywana przez te sygnatury musi się zgadzać co najmniej w 75% i muszą one opisywać tę samą jednostkę biologiczną (rodzinę, domenę itp.). Nowe sygnatury pochodzące z baz danych członkowskich są ręcznie dodawane do bazy InterPro przez kuratorów – w takim przypadku mogą one zostać dopisane do już istniejących rekordów albo utworzyć nowy wpis w bazie. Anotacje tych samych sygnatur pochodzące z różnych rodzin białek są scalane w ramach jednego wpisu.

Scalanie utrudnione jest poprzez istnienie złożonych relacji pomiędzy rekordami w poszczególnych bazach danych. Problemy pojawiają się wtedy, gdy sygnatura (lub sygnatury) znajdująca się w jednej z baz danych dopasowuje pewien zbiór białek, który jest podzbiorem większej rodziny i dopasowany jest przez inną sygnaturę, która równocześnie nakłada się z sygnaturą (lub sygnaturami) rozpoznającą mniejszą grupę. W takiej sytuacji każda z sygnatur otrzymuje unikalny numer dostępu InterPro i ustanawiana jest pomiędzy nimi relacja. W bazie InterPro istnieją dwa typy relacji: rodzic/potomek (ang. *parent/child*) oraz zawiera/znaleziony (ang. *contains/found in*). W przypadku relacji rodzic/potomek sygnatura–potomek powinna dopasowywać podzbiór

sekwencji rozpoznawanych przez sygnaturę–rodzica. Przykładem takiej relacji jest rodzina tubulin (wpis w bazie InterPro o numerze dostępu IPR000217), która określa wszystkie białka tubulin. Rodzina ta może zostać podzielona na podrodziny zawierające specyficzne tubuliny:

aplha (IPR002452), beta (IPR002453), gamma (IPR002454),
delta (IPR002967), epsilon (IPR004057) i zeta (IPR004058).

Każde białko znajdujące się na liście białek dopasowanych przez dowolną podrodzinę tubulin znajduje się również na liście białek dopasowywanych przez wpis rodzica (IPR000217).

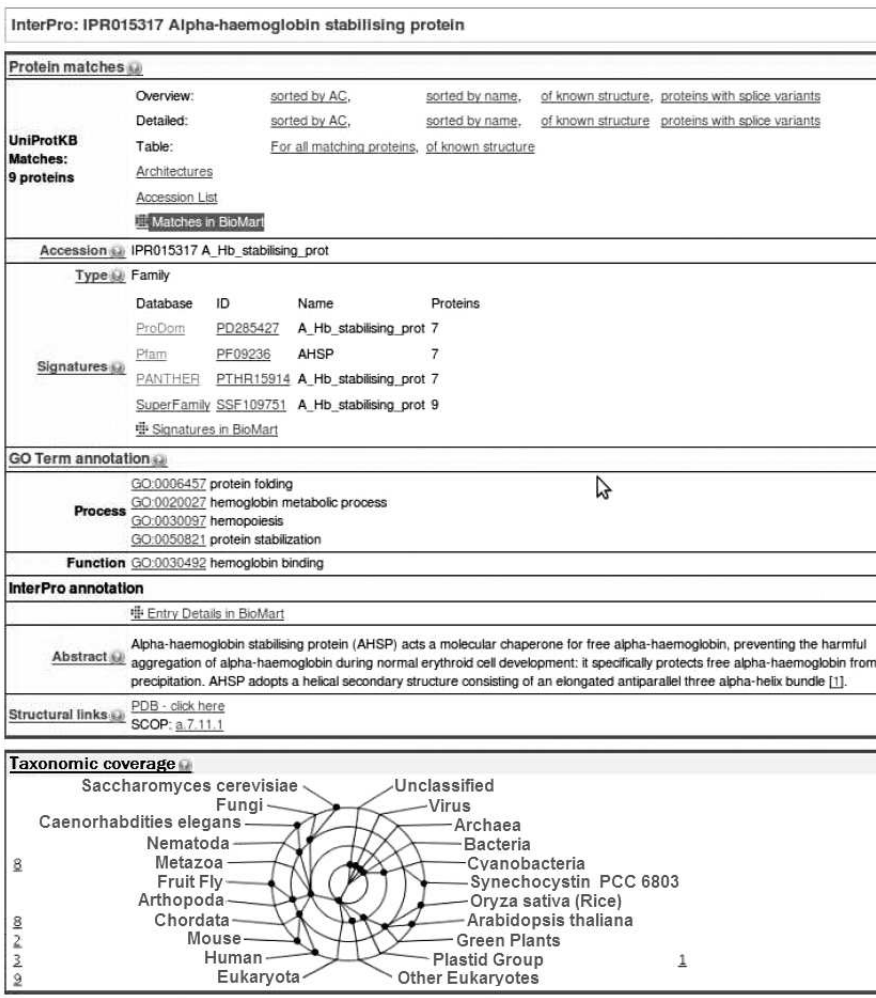
Drugi rodzaj relacji pomiędzy wpisami w bazie InterPro, relacja zawiera/znaleziony wykorzystywana jest do wskazywania zależności pomiędzy domenami, które występują w rodzinach białek odmiennych zarówno pod względem strukturalnym, jak i funkcjonalnym. Pojedyncza domena jest oddzielną, ruchomą jednostką – ta sama domena często może być obserwowana w kilku różnych białkach w konfiguracjach z innymi domenami. Przykładem może być tu domena C2 (IPR000008), którą można znaleźć w różnych rodzinach białek takich jak fosfolipaza D (IPR011402) czy synaptogamina (IPR001565).

Dopasowania w bazie InterPro wyznaczane są dla sekwencji białkowych pochodzących z baz UniProtKB i UniParc. Wersja bazy danych z lipca 2009 roku zawierała 18843 rekordy reprezentujące 5428 domen, 11379 rodzin, 79 miejsc aktywnych, 52 miejsca wiązania ligandu, 506 fragmentów zakonserwowanych, 1123 regiony, 23 posttranslacyjne modyfikacje oraz 253 powtórzenia. Nowe wersje bazy wydawane są co 2–3 miesiące.

Struktura rekordu w bazie InterPro

Każdy rekord w bazie InterPro posiada unikalny numer dostępu – zbudowany według schematu: IPRxxxxxx (gdzie x oznacza cyfrę). Numer dostępu znajduje się w polu `Accession`, które zawiera też zwięzły opis – unikalny w całej bazie InterPro. Pole `Type` określa typ wpisu – może to być rodzina, domena, region, powtórzenie lub miejsce aktywne. Pole `Signature` zawiera lista sygnatur białkowych związanych z danym wpisem. Dla każdej sygnatury znajduje się tu nazwa bazy, z której ona pochodzi, nazwa sygnatury oraz liczba sekwencji przez nią dopasowywanych. Pole `InterPro Relationships` zawiera numery dostępu rekordów, które z danym rekordem są w relacji rodzic/potomek lub zawiera/znaleziony. Pole `GO Term annotation` zawiera listę terminów Ontologii Genowych (patrz rozdział 6) związanych z danym wpisem. Streszczenie informacji na temat sygnatury pobrane z innych baz danych wraz z odnośnikami do literatury znajduje się w polu `Abstract`. Pole `Structural links` zawiera odnośniki do baz danych struktur białkowych związanych z danym wpisem, podczas gdy pole `Database links` odnośniki do zasobów innych baz danych. Mogą to być odnośniki do rekordów bazy danych członkowskich, na przykład do rekordów dokumentacji rodzin białek w bazie danych PROSITE, a także do zasobów dowolnych baz danych zawierających informacje

związane z danym wpisem, takich jak baza Enzyme Commission (EC) zawierająca klasyfikację enzymów czy inne specjalistyczne baz danych. W polu `Publications` znajduje się lista publikacji, które zostały wykorzystane do zredagowania streszczenia. Dodatkowo w polu `Additional Reading` znajduje się lista publikacji pochodzących z anotacji członkowskich baz danych, które nie zostały umieszczone w polu `Publications`. Przykładowy rekord pochodzący z bazy InterPro przedstawiono na rysunku 5.4.



Rysunek 5.4. Przykładowy rekord pochodzący z bazy InterPro

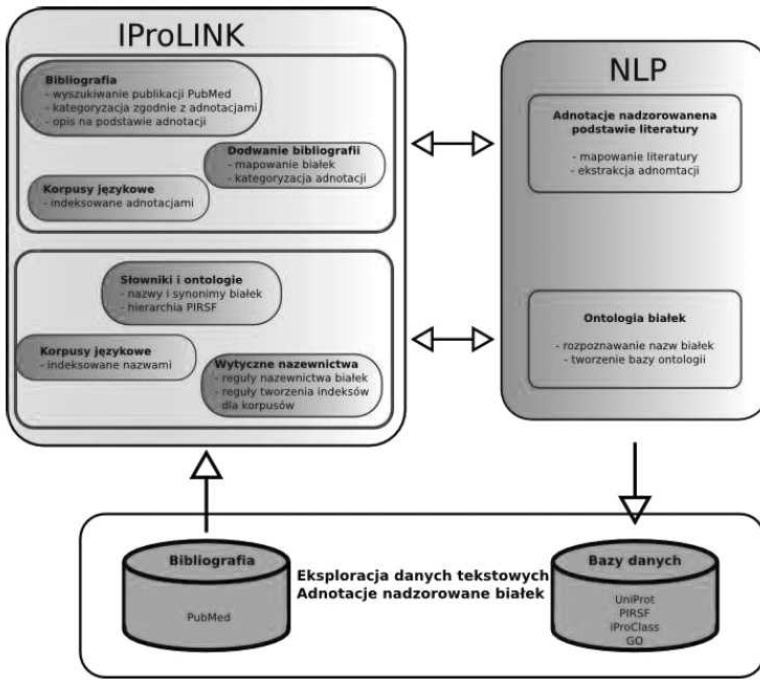
5.3.2 iProClass

Inną bazą danych, która powstała w celu zintegrowania informacji pochodzących z różnych baz danych dotyczących białek, jest baza *iProClass* – (*Integrated Protein Knowledgebase*) [Huang et al., 2003]. W bazie iProClass znajdują się informacje na temat sekwencji pochodzących z bazy UniProt oraz rodzin białek z bazy PIRSF. Projektując tę bazę danych, z powodzeniem połączono ze sobą modele ściśle i luźno powiązanych danych. W fizycznych zasobach bazy iProClass zgromadzone są podstawowe dane na temat białek pobrane z różnych źródeł, co pozwala na ich szybką analizę oraz anotację. Dodatkowo każdy rekord uzupełniony jest dużym zbiorem odnośników do innych źródeł, dzięki czemu udaje się uniknąć problemów z aktualizacją pomiędzy różnymi bazami danych rozproszonymi w Internecie. Obecnie za pomocą bazy iProClass można uzyskać informacje na temat sekwencji białkowych pochodzących z bazy UniProtKB oraz wybranych sekwencji znajdujących się w bazie UniParc, a dane, które użytkownik może uzyskać, pochodzą z ponad 90 różnych zewnętrznych źródeł. Są to m.in.: bazy rodzin białek, interakcji, funkcjonalne, ścieżek sygnałowych, struktur, klasyfikacji strukturalnej, genowe i genomowe, ontologii, literaturowe oraz taksonomiczne. Rekordy bazy iProClass pozwalają ich użytkownikom szybko i w jednym miejscu zebrać aktualne informacje dotyczące biocząsteczki, pochodzące z bioinformatycznych baz danych o różnym profilu. Pozwala to na uzyskanie o wiele pełniejszej anotacji cząsteczki niż można by uzyskać w dowolnej, „pojedynczej" bazie danych.

5.3.3 iProLINK

Integrated Protein Literature INformation and Knowledge [Hu et al., 2004] nie jest typową bazą danych, lecz raczej narzędziem, które pozwala na stworzenie swego rodzaju połączenia pomiędzy bazami danych sekwencji białkowych a internetowymi zasobami publikacji naukowych z dziedziny biologii i medycyny. Narzędzie iProLink rozwijane jest w ramach grupy badawczej Protein Information Resource i powstało w odpowiedzi na widoczny w ostatnich latach wzrost znaczenia metod eksploracji danych z tekstu (ang. *text mining*) w zastosowaniu do automatyzacji procesu anotacji danych genomowych i proteomowych. W ciągu ostatnich lat zanotowano lawinowy wzrost liczby publikacji z zakresu biologii i medycyny oraz towarzyszący mu wzrost zainteresowania metodami przetwarzania języka naturalnego (ang. *NLP – natural language processing*) pozwalającymi na automatyczną ekstrakcję wiedzy na podstawie informacji zawartych w internetowych bazach danych publikacji. Stąd też główny cel, jaki postawili przed sobą twórcy iProLink, to stworzenie nadzorowanego repozytorium danych tekstowych, które będzie można wykorzystywać w celu przyporządkowywania bibliografii, ekstrakcji anotacji, wyszukiwania i rozpoznawania nazw białek oraz rozwoju Ontologii Białkowej. Źródła danych dostępne w bazie iProLink można podzielić na dwie główne grupy z uwagi na ich przydatność dla procedur eksploracji tekstu:

- Mapowanie literatury i ekstrakcja anotacji.
- Rozpoznawanie nazw białek i rozwój Ontologii Białkowej.



Rysunek 5.5. Baza danych iProLink
(na podstawie: `http://pir.georgetown.edu/iprolink/`)

W pierwszej grupie mieszczą się narzędzia do wyszukiwanie odpowiednich źródeł literaturowych oraz znajdowania anotacji biologicznych, które mogą później zostać wykorzystane do rozszerzenia opisu biologicznego białka znajdującego się w bazie danych. Źródła danych dostępne w ramach tej części systemu iProLink to: system bibliograficzny – lista publikacji zawierających informacje opisujące i charakteryzujące dane białko wraz z identyfikatorami bazy PubMed – PMID oraz korpusy językowe – kilkaset abstraktów i pełnych artykułów indeksowanych za pomocą eksperymentalnie zweryfikowanych modyfikacji posttranslacyjnych (acetylowanie, glikozylacja, metylacja, fosforylacja, hydroksylowanie) anotowanych w bazie sekwencji PIR. Dodatkowo w ramach tej części systemu użytkownik ma możliwość dodawania kolejnych publikacji do listy publikacji powiązanych z danym białkiem. Druga grupa zastosowań metod eksploracji tekstu obejmuje rozpoznawanie nazw białek oraz rozwój Ontologii Białkowej. Narzędzia dostępne w ramach iProLink powiązane z tymi zastosowaniami to m.in.: BioThesaurus – system wyszukiwania

rekordów w bazie UniProt na podstawie nazw białek; słowniki białek zawierające zasady nazewnictwa, synonimy, akronimy; słowniki symboliczne zawierające terminy biomedyczne, terminy chemiczne, makromolekuły itd.; Ontologie Białkowe wraz z podziałem na rodziny białek pochodzącym z bazy PIRSF; zbiór zasad nadawania nazw białkom oraz rodzinom białkowym; korpusy językowe indeksowane nazwami białek.

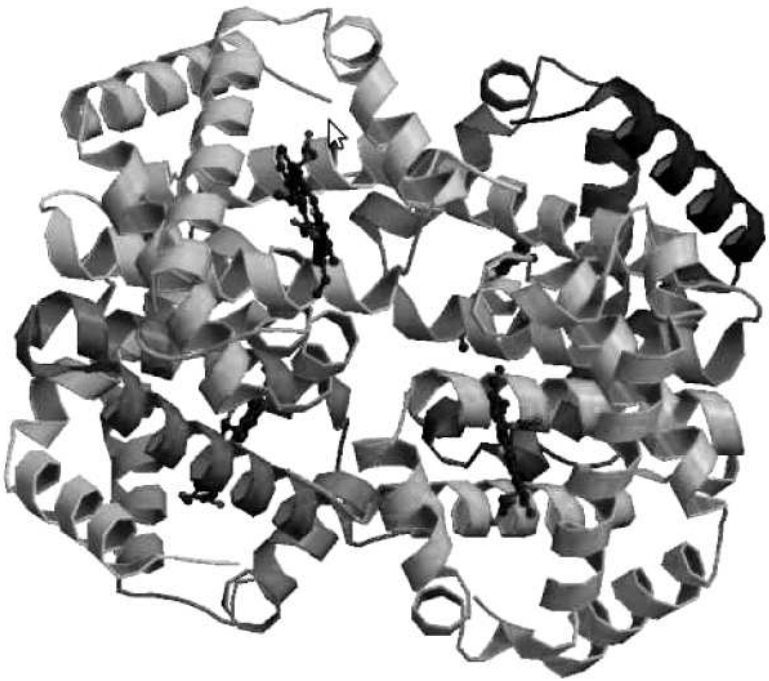
5.4 Bazy danych struktur białek

Informacja na temat kolejności aminokwasów w łańcuchu jest pierwszą i podstawową informacją, jaką możemy uzyskać na temat każdego białka. Jednakże sama znajomość sekwencji aminokwasów, zwana również strukturą pierwszorzędową lub pierwotną białka, niewiele jeszcze mówi na temat biologicznych i fizykochemicznych właściwości białka. Równie istotna wiedza na temat właściwości białka zawarta jest w informacji na temat jego struktury drugo– i trzeciorzędowej. Struktura drugorzędowa zawiera informacje o lokalnych strukturach białka powstających w wyniku tworzenia się wiązań wodorowych pomiędzy nieodległymi od siebie aminokwasami. Struktury drugorzędowe występują w postaci helis alfa (ang. α *helix*), beta kartek (ang. β *sheet*) jak również beta pętli (ang. *turn*). Struktura trzeciorzędowa to inaczej struktura przestrzenna białka zawierająca informacje o tym, w jaki sposób względem siebie położone są elementy struktury drugorzędowej.

Informacje na temat struktur białek można nabywać, porównując sekwencje aminokwasów białka, o którym chcemy się czegoś dowiedzieć, z sekwencją aminokwasów białka już rozwiązanego, czyli takiego, którego struktura trzeciorzędowa jest znana i została wyznaczona w sposób eksperymentalny – za pomocą metod krystalografii, spektroskopii jądrowego rezonansu magnetycznego (ang. *Nuclear Magnetic Resonance* - NMR) lub mikroskopii elektronowej. Im więcej białek o takiej samej sekwencji aminokwasów ma podobną strukturę przestrzenną, tym bardziej wiarygodne mogą być przewidywania struktury białkowej nowej cząsteczki, dla której znana jest sekwencja aminokwasów. Stąd też bardzo istotną rolę w biadaniach nad własnościami białek pełnią bazy przestrzennych struktur białek. Ogólnie bazy danych zawierające struktury białek można podzielić na dwie grupy: bazy dane zawierające informacje na temat struktur przestrzennych białek (współrzędne przestrzenne atomów w strukturze) oraz bazy danych zawierające klasyfikację i charakterystykę znanych struktur białek.

5.4.1 PDB

Baza danych PDB (*Protein Data Bank*) [Berman, 2008] została założona w roku 1971 w Brookhaven National Laboratory w Stanach Zjednoczonych przez krystalografów jako odpowiedź na konieczność stworzenia repozytorium, które



Rysunek 5.6. Struktura przestrzenna białka hemoglobiny ludzkiej

pozwoliłoby środowisku krystalografów na swobodną wzajemną wymianę struktur białek pomiędzy różnymi laboratoriami. W 1999 r. opiekę nad bazą przejęło konsorcjum Research Collaboratory of Structural Bioinformatics (RCSB PDB), którego założycielami byli Uniwersytet Stanowy w New Jersey Rutgers, San Diego Supercomputer Center na Uniwersytecie Kalifornia San Diego oraz National Institute of Standards and Technology. W 2003 r. baza PDB połączyła się z europejską bazą danych struktur Macromolecular Structure Database w European Bioinformatics Institute (MSD-EBI) oraz japońską PDB Japan (PDBj) w Institute for Protein Research na Uniwersytecie Osaka. W wyniku tego kroku baza PDB stała się największym światowym ogólnodostępnym zbiorem przestrzennych struktur białek.

W bazie PDB znajdują się struktury białek rozwiązane za pomocą metod eksperymentalnych takich jak krystalografia rentgenowska, spektroskopia NMR czy mikroskopia elektronowa. W tabeli 5.2 przedstawiono liczbę struktur różnego rodzaju biocząsteczek, które znajdują się w bazie PDB z podziałem na eksperymentalne metody, które zostały wykorzystane do ich znalezienia. Struktury białek przesyłane do PDB przez laboratoria badawcze są przed opublikowaniem sprawdzane pod względem poprawności oraz anotowane przez

kuratorów. Do przesłania nowej struktury do bazy PDB wykorzystać można aplikację internetową ADIT, która zapewnia kontrolę formatu danych i pozwala na automatyczne stworzenie końcowego raportu. Po przesłaniu danych sekwencja białkowa i odniesienia do literatury przesłane wraz ze strukturą są porównywane z publikacjami w czasopismach naukowych i informacjami znajdującymi się w bioinformatycznych bazach danych. Sprawdzana jest jakość wprowadzonej struktury wraz z jej długościami i kątami wiązań, kątami torsyjnymi, itd. Każda struktura otrzymuje unikalny identyfikator, nazwę oraz synonimy, określana jest również formalna nazwa organizmu, z którego pochodzi białko oraz dodawana jest informacja biologiczna na jego temat. Po sprawdzeniu i opisaniu przez kuratora struktura przesyłana jest do jej autora w celu akceptacji wprowadzonych zmian. Po ich zaakceptowaniu struktura może zostać opublikowana w bazie PDB – jeżeli autorzy nie chcą upubliczniać struktury, dane zdeponowane w PDB mogą pozostać przez pewien czas (nie dłużej niż rok) ukryte.

Tabela 5.2. Zawartość bazy danych PDB we wrześniu 2009 roku

Metoda	Białka	Kwasy nukleinowe	Kompleksy białko/ kwas nukleinowy	Inne	Suma
Krystalografia	48225	1168	2216	17	51625
NMR	6993	869	150	6	8018
Mikroskop elektronowy	171	16	65	0	252
Pozostałe	130	5	5	10	150
Suma	**55519**	**2058**	**2436**	**33**	**60046**

Format danych dostępny w bazie PDB

Informacja na temat struktury białek reprezentowana może być w bazie PDB za pomocą trzech różnych formatów plików: PDB, mmCIF (*macromolecular Crystallographic Information File*) i PDBML (*Protein Data Bank Markup Language*) – reprezentacja danych PDB w formacie XML . Podstawowa informacja, która zawarta jest w plikach PDB reprezentujących strukturę białka, dotyczy współrzędnych atomów, które składają się na opisywaną biomolekułę. Każdy taki plik zawiera listę atomów oraz ich lokalizację w przestrzeni trójwymiarowej.

Pliki PDB wykorzystują do zapisu informacji tzw. metodę **reguł chemicznych**. Wykorzystuje ona znajomość praw fizyki przy generowaniu połączeń pomiędzy atomami. Np. wykorzystanie prawa „średnia długość stabilnego wiązania C–C równa jest około 1.5 Å" spowoduje, że jeśli mamy dwa atomy węgla oddalone od siebie od 1.5 Å, to wówczas zawsze będą tworzyć pojedyncze wiązanie. Założenie jest więc takie, że do rekonstrukcji struktury wymagana jest jedynie tablica długości oraz typów wiązań dla każdej możliwej oddziałującej

ze sobą pary atomów. Zastosowanie takiego rozwiązania powoduje zmniejszenie liczby informacji, które mogą być przechowywane w plikach zawierających struktury, ale wymaga równocześnie zastosowania skomplikowanych algorytmów programistycznych do odtworzenia cząsteczki oraz powoduje problemy w przypadku gdy pojawiają się odstępstwa od przyjętych reguł.

Informacja o strukturze cząsteczki jest w plikach PDB zakodowana w dwóch postaciach: w postaci sekwencji jawnej (ang. *explicit sequence*) oraz w postaci sekwencji ukrytej (ang. *implicit sequence*). Sekwencje jawne znajdują się w wierszach oznaczonych symbolem `SEQRES`. Przykładowa linia `SEQRES` przedstawiona jest poniżej i zawiera: kolejny numer rekordu `SEQRES` (17), identyfikator łańcucha biopolimerowego reprezentowany przez duże litery A–Z (tutaj A), liczbę aminokwasów w łańcuchu (321) oraz nazwy kolejnych aminokwasów (`MET` – metioniona, `ILE` – izoleucyna, `GLU` – glutamina itd.).

```
SEQRES 1 A 321 MET ILE GLU ARG ARG LYS ILE ALA VAL ILE GLY SER GLY
```

Informacja znajdująca się w części jawnej nie wystarcza do rekonstrukcji cząsteczki. Programy służące do wizualizacji struktur korzystają również z informacji o strukturze zakodowanej w postaci sekwencji ukrytej. Często zdarza się, że programy do wizualizacji korzystają jedynie z tej części informacji zawartej w plikach PDB. W tej części każdy atom reprezentowany jest osobno, zaś jego pozycję określają współrzędne (x, y, z), które reprezentują odległości wzdłuż każdej osi względem ustalonego punktu odniesienia w przestrzeni. Pojedyncza linia opisująca każdy atom rozpoczyna się od słowa kluczowego `ATOM` lub `HETATM`. Słowo kluczowe `ATOM` zazwyczaj identyfikuje atomy białek lub kwasów nukleinowych, podczas gdy słowo kluczowe `HETATM` oznacza atomy mniejszych molekuł. Następnie pojawia się szereg informacji dotyczących danego atomu takich jak: jego nazwa, numer w pliku, nazwa i numer reszty, do której on należy, jedna litera określająca łańcuch, współczynnik temperatury oraz informacja na temat liczby różnych konformacji, w jakich występuje dany atom. Poniżej przedstawiono pojedynczą linię `ATOM` pochodzącą z pliku PDB:

```
ATOM  3  C MET A  17  11.464  26.829  -4.734  1.00 69.67
```

Linia rozpoczyna się słowem kluczowym `ATOM`, następnie pojawia się cyfra 3 oznaczająca kolejny numer atomu w pliku. `C` określa nazwę atomu (w tym przypadku węgiel), `MET` oznacza nazwę aminokwasu, do której należy atom (w tym przypadku metionina), następnie znajduje się identyfikator łańcucha (`A`) oraz numer sekwencji aminokwasowej (17). Wartości `11.464 26.829 -4.734` oznaczają odpowiednio współrzędne x,y,z atomu. Wartość `1.00` oznacza, że atom występuje tylko w jednej konformacji, natomiast wartość `69.6` określa jego gęstość elektronową.

5.4.2 MMDB

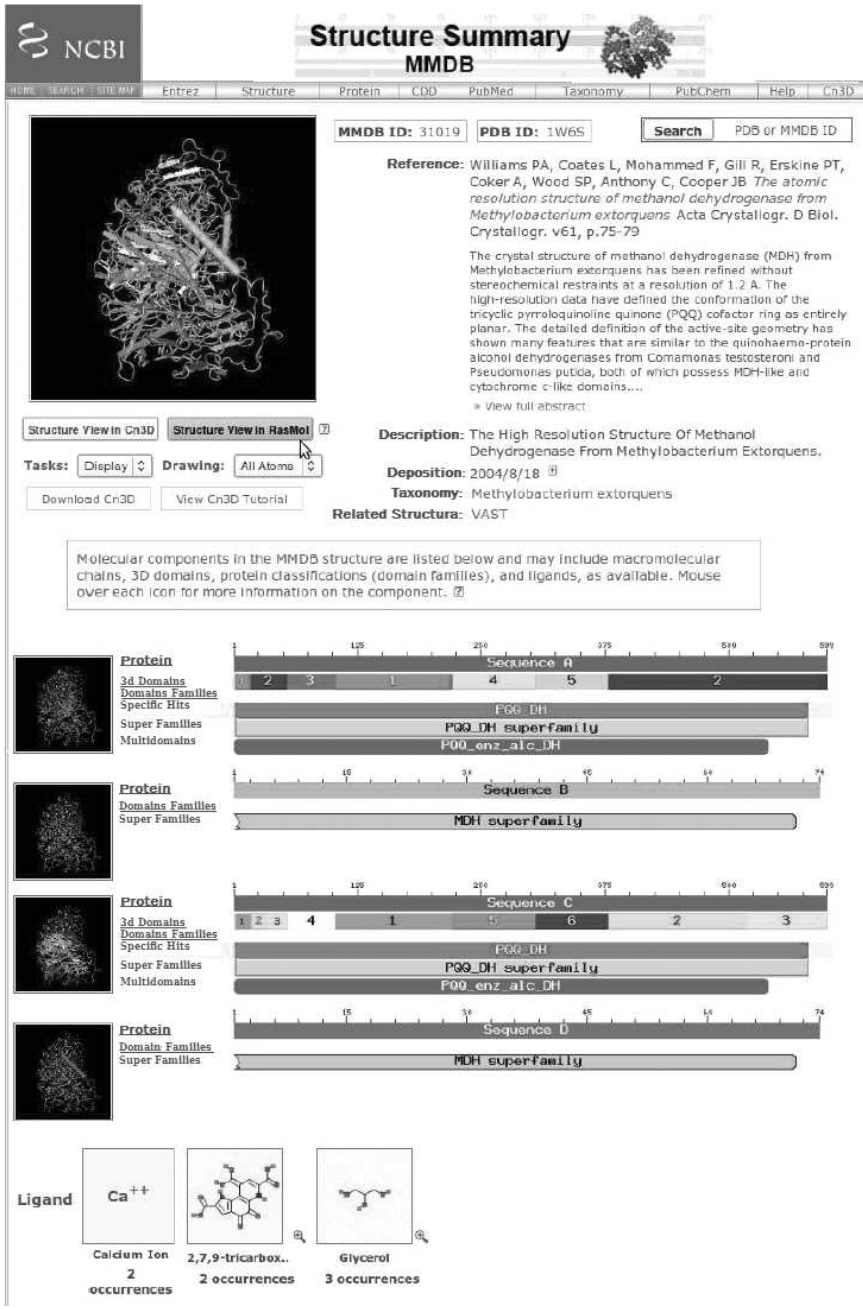
Baza danych modelowania molekularnego MMDB (*Molecular Modeling Database*) [Wang et al., 2002] w NCBI zawiera struktury molekularne pochodzące

z PDB. Dane na temat struktur zintegrowane są z systemem wyszukiwania danych Entrez, co pozwala użytkownikom na szybki dostęp nie tylko do informacji o strukturze, ale także do wszelkich biologicznych anotacji związanych z daną cząsteczką. W momencie gdy cząsteczka dodawana jest do bazy MMDB, do bazy sekwencji nukleotydowych dodawana jest odpowiadająca jej sekwencja, a informacja zawarta w plikach PDB przepisywana jest do formatu ASN.1 (*Abstract Syntax Notation*). Weryfikowana jest zgodność danych zawartych w części opisującej strukturę białka oraz części zawierającej współrzędne atomów – w przypadku ewentualnych niezgodności, do utworzenia struktury wykorzystuje się dane zapisane w postaci sekwencji ukrytej. Format ASN.1 pozwala na przechowywanie dodatkowych informacji na temat cząsteczki takich jak ujednolicona definicja struktury drugorzędowej oraz dane dotyczące przestrzennej budowy domenowej – informacje te wykorzystywane są podczas poszukiwania podobnych struktur białkowych. Użytkownik, który przegląda strukturę w bazie MMDB, posiada listę odnośników do publikacji w bazie MEDLINE związanych z cząsteczką, odnośniki do bazy danych taksonomii, listę pokrewnych sekwencji białkowych i nukleotydowych dla każdego łańcucha polipeptydowego w strukturze, odnośniki do bazy domen białkowych (CDD – *Conserved Domain Database*) zidentyfikowanych za pomocą algorytmu RPS-BLAST oraz odnośniki do bazy porównawczej struktur VAST. Wizualizacja struktury cząsteczki może odbywać się za pomocą programu Cn3D lub za pomocą programu RasMol. Na rysunku 5.7 przedstawiono przykładowe podsumowanie pochodzące z bazy MMDB dla struktury 1W6S.

Pliki MMDB wykorzystują do opisu cząsteczki tzw. metodę **wiązań jednoznacznych** (ang. *explicit bonding approach*). Oznacza to, że każdy plik zawierający strukturę posiada również informacje na temat wiązań, jakie istnieją pomiędzy atomami cząsteczki. Struktury cząsteczek zawarte w pliku MMDB reprezentowane są w postaci hierarchicznej jako grafy połączeń pomiędzy atomami, resztami i molekułami. Odczytanie tej informacji wymaga skorzystania ze standardowego słownika reszt. Słownik reszt dostarczany wraz z bazą MMDB zawiera informacje na temat atomów oraz wiązań 20 aminokwasów budujących białka oraz 8 grup rybonukleotydowych i deoksyrybonukleotydowych występujących w DNA i RNA. Oprogramowanie, które wykorzystuje dane zawarte w pliku MMDB do odtworzenia struktury cząsteczki i wygenerowania wiązań pomiędzy atomami w cząsteczce, korzysta ze słownika reszt.

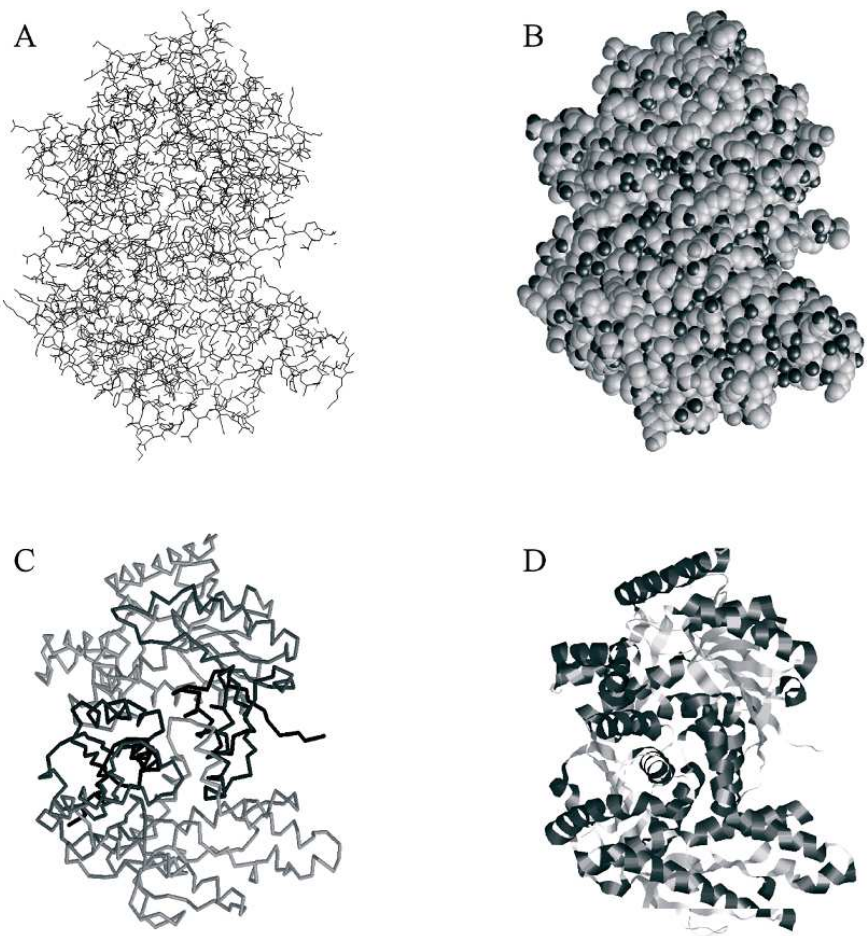
5.4.3 Wizualizacja struktur białek

Pliki zawierające struktury białek wykorzystywane są przez różnego rodzaju aplikacje, które potrafią interpretować ich zawartość i na jej podstawie generować przestrzenne struktury białek. Graficzna reprezentacja białka ułatwia poznawanie i odkrywanie własności cząsteczki. Większość programów umożliwia wczytanie pliku struktury, wyświetlenie struktury z podkreśleniem tych aspektów, które najbardziej interesują użytkownika, a także udostępnia róż-



Rysunek 5.7. Przykładowe podsumowanie pochodzące z bazy MMDB dla struktury 1W6S – dehydrogenazy metylowej

nego rodzaju narzędzia pozwalające na mierzenie odległości pomiędzy atomami, kątów pomiędzy wiązaniami czy identyfikację pewnych cech struktury. W zależności od informacji, jaką zainteresowany jest użytkownik, różna może być forma prezentacji struktury. Na rysunku 5.8 przedstawiono cztery różne, najbardziej typowe formaty wyświetlania struktur białkowych, na podstawie cząsteczki dehydrogenazy mleczanowej (identyfikator struktury 2FN7). Struktury zostały wygenerowane za pomocą aplikacji RasMol.



Rysunek 5.8. Przykładowa wizualizacja struktury dehydrogenazy mleczanowej (2FN7) wykonana za pomocą programu RasMol.
A – struktura szkieletowa; B – struktura czaszowa; C – struktura łańcucha głównego, D – wizualizacja struktury drugorzędowej

Przykład A na rysunku 5.8 przedstawia cząsteczkę w postaci szkieletowej (ang. *wireframe*). W przypadku takiej reprezentacji linia rysowana jest pomiędzy każdym wiązaniem kowalencyjnym pomiędzy dwoma atomami. Przykład B reprezentuje wizualizację w formie schematu czaszowego (ang. *spacefill*), gdzie wokół każdego atomu rysowana jest kula odzwierciedlająca relatywny kształt atomu. Na przykładzie C przedstawiono wizualizację białka w postaci głównego łańcucha (ang. *backbone*), który wyznaczany jest za pomocą wiązań łączących sąsiadujące ze sobą atomy węgla *alpha*. Z kolei przykład D przedstawiono schemat cząsteczki podkreślający jej strukturę drugorzędową – alfa helisy zaznaczone są tu w postaci skręconych wstążek, natomiast beta kartki w postaci strzałek.

Różnego rodzaju reprezentacje wykorzystywane mogą być w różnych celach. Oglądając strukturę białka, zawsze warto sprawdzić różne wizualizacje i różne schematy kolorów w celu znalezienia najlepszej reprezentacji dla aktualnie wykonywanego zadania. Postać szkieletowa jest wygodna, jeśli interesują nas szczegóły struktury, takie jak aktywne miejsca lub sposób połączenia pomiędzy jonami metalu. Struktury czaszowe pokazują ogólny kształt białek oraz ich rozmiar, i sprawdzają się wtedy, gdy chcemy zrozumieć, w jaki sposób różnego rodzaju białka oddziałują ze sobą. Z kolei schematy łańcucha głównego i schematy podkreślające drugorzędową strukturę cząsteczki przydatne są w sytuacjach, kiedy analizujemy sposób, w jaki białko się zwija (ang. *fold*) lub porównujemy zwijanie się pomiędzy różnymi białkami. Warto również eksperymentować z różnymi sposobami kolorowania cząsteczek. Do najczęściej wykorzystywanych schematów kolorów należą: *monochrome* – całość cząsteczki w jednym kolorze, *CPK* – każdy pierwiastek w innym kolorze (węgiel - jasnoszary, tlen - czerwony, wodór - biały, azot - niebieski itd.), *shapely* – każdy aminokwas i nukleotyd ma inną barwę, *group* – każdy łańcuch ma inny kolor, kolorowane są one stopniowo od niebieskiego, poprzez zielony, żółty i pomarańczowy aż do niebieskiego, N-koniec polipeptydów oraz 5' koniec kwasów nukleinowych jest niebieski, podczas gdy C-koniec polipeptydów oraz 3' koniec kwasów nukleinowych jest czerwony, *chain* – każdy łańcuch polipeptydowy (podjednostka) ma inny kolor, *temperature* – określa zakres swobody drgań poszczególnych atomów (im cieplejszy, tym większa możliwa oscylacja, im zimniejszy – tym mniejsza), *structure* – barwa podkreśla strukturę drugorzędową białka (różowe alfa helisy, żółte beta kartki).

Istnieje duża liczba programów, które potrafią odczytywać pliki w formacie PDB, mmCIF czy MMDB. We wrześniu 2009 roku na stronach PDB umieszczono listę 35 aplikacji zdolnych przetwarzać dane w formacie PDB. Niektóre z nich oferują najprostsze możliwości takie jak wyświetlanie samej struktury, inne pozwalają na bardzo złożone analizy. Część oprogramowania udostępniana jest w ramach licencji open source, z kolei inne są płatne lub dostępne w ramach ograniczonej licencji akademickiej. Listę dostępnego oprogramowania można znaleźć pod następującym adresem: `http://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html`. Z po-

śród wielu dostępnych programów warto wymienić narzędzia takie jak RasMol – jeden z najpopularniejszych programów wizualizacji struktur oparty na języku skryptowym, napisany w języku Java program *JMol*, który stosuje podobne schematy wizualizacji jak program RasMol i dostępny jest zarówno jako odrębna aplikacja oraz jako aplet przeglądarki internetowej, program Cn3D, który stosowany jest jako przeglądarka struktur umieszczonych w bazie MMDB czy zaawansowany program *SwissPDB Viewer* , który pozwala nie tylko na wizualizacje molekuł, ale umożliwia również przeprowadzanie bardziej złożonego modelowania strukturalnego.

5.4.4 SCOP

Dla prawie każdej struktury białkowej określić można jej podobieństwo do innych struktur – wiedza na temat podobieństwa strukturalnego białek jest istotna dla odkrywania funkcji nowych białek (na podstawie strukturalnego podobieństwa do cząsteczek, na temat których posiadamy już biologiczną wiedzę), a klasyfikacja molekuł względem podobieństwa pozwala na lepsze zrozumienie procesów ewolucji żywych organizmów – stąd tak wielkie znaczenie w biologii molekularnej baz zawierających klasyfikację oraz charakterystykę struktur białkowych.

Baza danych SCOP (*Structural Classification of Protein Database*) [Andreeva et al., 2008], utrzymywana przez laboratorium biologii molekularnej przy MRC (*Medical Research Council*), jest bazą danych przestrzennych struktur białkowych uporządkowanych zgodnie z zależnościami ewolucyjnymi i strukturalnymi białek. Stopień podobieństwa pomiędzy strukturami białek wyznaczać można na podstawie wzajemnego położenia elementów tworzących strukturę drugorzędową białka i topologii łańcuchów polipeptydowych. To podobieństwo ściśle związane jest z podstawowymi fizycznymi i chemicznymi własnościami białek. Metody klasyfikacji nowych białek dołączanych do bazy obejmują – w zależności od podobieństwa analizowanej struktury do struktur już w bazie istniejących – w pełni automatyczną klasyfikację, automatyczną klasyfikację i ręczną walidację wyników lub całkowicie ręczną klasyfikację.

SCOP jest bazą hierarchiczną. Podstawowe jednostki klasyfikacji w bazie to domeny, które zostały wyznaczone w trakcie biologicznych eksperymentów rozwiązywania struktur białek. Niewielkie lub średnie cząsteczki zazwyczaj składają się z jednej domeny, stąd też takie białka traktowane są w procesie klasyfikacji całościowo. Z kolei większe struktury muszą zostać rozdzielone na domeny, które później klasyfikowane są oddzielnie. Klasyfikacja domen odbywa się na różnych poziomach hierarchii zależnie od struktury, sekwencji oraz wzajemnych funkcjonalnych i strukturalnych relacji pomiędzy domenami. Przy przechodzeniu hierarchii SCOP od liścia do korzenia reguły grupowania białek na poszczególnych poziomach zmieniają się od strukturalno–funkcjonalnych do czysto strukturalnych.

Możemy wyróżnić następujące poziomy hierarchii:

- **Species** (gatunki). Ten poziom hierarchii reprezentowany jest przez pojedyncze sekwencje białkowe – zarówno w formie istniejącej naturalnie w środowisku, jak i w ich sztucznych odmianach.
- **Protein** (białko). Grupuje podobne sekwencje, które pełnią zasadniczo te same funkcje biologiczne – pochodzące bądź z różnych gatunków, bądź będące różnymi izoformami w tym samym organizmie.
- **Family** (rodzina). Zawiera białka o podobnej sekwencji, ale pełniące różne funkcje biologiczne.
- **SuperFamily** (nadrodzina). Na tym poziomie hierarchii połączone są ze sobą rodziny białek o podobnych cechach funkcjonalnych i strukturalnych pochodzących od wspólnego przodka.
- **Folds** (zwoje). Ten poziom zawiera podobne strukturalnie rodziny białek (czyli takie, gdzie główne elementy struktury drugorzędowej są uporządkowane w ten sam sposób i tak samo połączone) niezależnie od tego, czy białka są ze sobą spokrewnione ewolucyjnie czy też nie.
- **Class** (klasy). Na ostatnim etapie hierarchii SCOP zwoje powiązane są ze sobą, tworząc klasy hierarchii. Podstawą podziału na klasy jest drugorzędowa struktura białka. Na jej podstawie w bazie SCOP wyodrębniono 11 klas, w tym 7 podstawowych: (1) *all alpha* – białka zawierające w swojej strukturze głównie alfa helisy, (2) *all beta* – białka zawierające w swojej strukturze głównie beta kartki, (3) *alpha and beta* – białka zawierające struktury alfa i beta przemieszane ze sobą, (4) *alpha plus beta* – białka, w których struktury alfa i beta występują oddzielnie, (5) *multi domain* – zawierające dwie lub więcej domen należących do różnych klas, (6) *membrane and cell surface proteins and peptides* – białka błonowe i związane z powierzchnią komórki, (7) *small proteins* – małe białka. Pozostałe klasy zawierają struktury kwasów nukleinowych oraz modele teoretyczne. W tabeli 5.3 przedstawiono siedem głównych klas bazy struktur SCOP . Przedstawione struktury odpowiadają 38221 strukturom w bazie PDB.

5.4.5 CATH

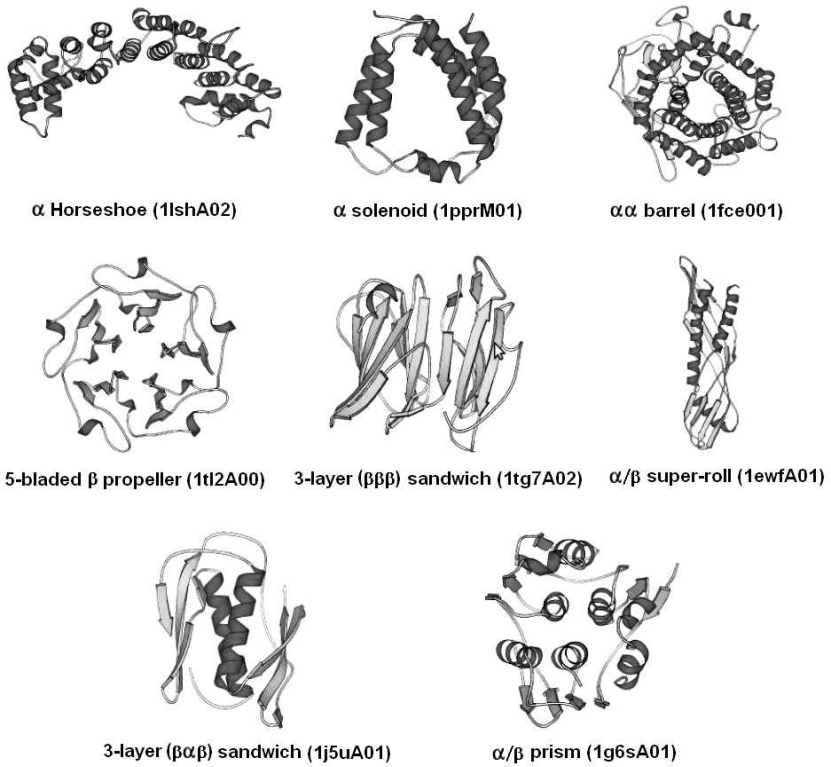
Baza danych CATH (*Class, Architecture, Topology, Homology*) jest hierarchiczną bazą danych przestrzennych domen białek, firmowaną przez Instytut Biologii Strukturalnej i Molekularnej w University College London [Cuff et al., 2009]. Klasyfikacji podlegają tu pojedyncze domeny białek , tak więc każda struktura białkowa składająca się z większej liczby domen musi być podzielona na składowe domeny. Taki podział może odbywać się automatycznie, jeśli białko wykazuje wysokie podobieństwo sekwencyjne i strukturalne do cząsteczki znajdującej się już w bazie CATH lub – jeśli takiego podobieństwa nie ma – ręcznie, przy pomocy odpowiednich narzędzi. Tak otrzymane domeny są następnie klasyfikowane – w sposób automatyczny, jeśli w bazie CATH znajduje się już domena o sekwencji i strukturze podobnej (wówczas

Tabela 5.3. Podział na klasy białek w bazie danych SCOP. Dane pochodzą z lutego 2009 roku. W zestawieniu pominięto struktury kwasów nukleinowych oraz modele teoretyczne.
Źródło: `http://scop.mrc-lmb.cam.ac.uk/scop/count.html# scop-1.75`

Klasa białek	Liczba		
	zwojów	nadrodzin	rodzin
α	284	507	871
β	174	354	742
α/β	147	244	244
$\alpha + \beta$	376	552	1055
multi domain	66	66	89
membrane and cell surface	58	110	123
small proteins	90	129	219
suma	**1195**	**1962**	**3902**

nowa domena otrzymuje przynależność do tej samej klasy) lub ręcznie, jeżeli takiego podobieństwa nie ma. Proces ręcznej klasyfikacji wspomagany jest przez liczne narzędzia automatycznego porównywania sekwencji oraz analizy treści publikacji naukowych powiązanych z daną cząsteczką. Hierarchia bazy CATH podzielona jest na cztery kategorie:

- **Class** lub **C–level** (klasa). Na tym poziomie domeny klasyfikowane są na podstawie struktury drugorzędowej. Istnieją trzy podstawowe klasy: klasa 1, która zawiera domeny składające się głównie z helis alfa, klasa 2 zawierająca domeny o strukturze beta kartki oraz klasa 3 zawierająca domeny, których struktura składa się zarówno z helis alfa, jak też z beta kartek (zarówno struktury $\alpha + \beta$, jak i α/β). Dodatkowo istnieje klasa 4 zawierająca domeny o słabo zdefiniowanej strukturze drugorzędowej.
- **Architecure** lub **A–level** (architektura). Domeny zgrupowane na poziomie klasy w hierarchii bazy CATH są następnie grupowane względem architektury określonej przez podobieństwo uporządkowania struktur drugorzędowych w przestrzeni 3D, przy czym nie bierze się tu pod uwagę sposobu połączeń pomiędzy strukturami. Proces przypisywania domen do elementów tej kategorii przeprowadzany jest ręcznie, poprzez obrazowe opisywanie kształtów, jakie tworzą w przestrzeni struktury drugorzędowe; np. beczka (ang. *barrel*), podkowa (ang. horseshoe). Na rysunku 5.9 przedstawiono przykładowe architektury występujące w bazie danych CATH.
- **Topology (Fold family)** lub **T-level** (topologia). Na tym poziomie hierarchii brane są pod uwagę zarówno kształty, jak i sposoby połączenia struktur drugorzędowych w domenach. Grupy tworzone są przez domeny, których podstawowe struktury zwijają się w ten sam sposób, tworząc tzw. rodziny zwojów (ang. *fold family*).



Rysunek 5.9. Przykładowe architektury domen w bazie CATH.
Źródło: Cuff A.L. et al. (2008) The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies, *Nucleic Acids Res.*, 37:D310–D314

- **Homologous Superfamily** lub **H-level** (nadrodziny homologiczne). Na tym poziomie podział uwzględnia ewolucyjne zależności pomiędzy domenami. Grupy tworzone są tu przez homologi, czyli domeny, które posiadają wspólnego ewolucyjnie przodka. Określone są ścisłe zasady, na podstawie których domeny uznawane są za podobne, a porównanie obejmuje zarówno podobieństwo sekwencji i struktury, jak również funkcje biologiczne białek.
- **Sequence Family Levels: (S,O,L,I,D)** lub **S–level** (rodziny sekwencji). Grupowanie nadrodzin zdefiniowanych na poziomie S pozwala na zdefiniowanie rodzin, które charakteryzują się podobieństwem sekwencji. W zależności od stopnia podobieństwa sekwencji można wyróżnić różne poziomy podobieństwa sekwencyjnego: S35 – rodzina sekwencji (ang. *sequence family*) – podobieństwo sekwencji większe lub równe 35%; S60 – rodzina ortologów (ang. *orthologous family*) – podobieństwo sekwencji większe lub

równe 60%; S95 – prawie jak domena (ang. *like domain*) – podobieństwo sekwencji większe lub równe 90%; S100 – identyczne domeny (ang. *identical domain*) – podobieństwo 100%. Dodatkowo istnieje tak zwany poziom D – licznik domen (*ang. domain counter*), który dodawany jest do hierarchii w celu weryfikacji unikalności domen w bazie – za jego pomocą można sprawdzić, czy każda domena ma unikalny identyfikator (CATHSOLID).

5.5 Adresy Internetowe

Adresy internetowych baz danych:

- CATH – `http://www.cathdb.info/`
- Gene3D – `http://gene3d.biochem.ucl.ac.uk/`
- HAMAP – `http://www.expasy.org/sprot/hamap/`
- InterPro – `http://www.ebi.ac.uk/interpro/`
- iProLINK – `http://pir.georgetown.edu/iprolink/`
- MMDB – `http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure`
- PANTHER – `http://www.pantherdb.org/`
- PDB – `http://www.wwpdb.org/`
- Pfam – `http://pfam.sanger.ac.uk/`
- PIRSF – `http://pir.georgetown.edu/pirsf/`
- PRINTS – `http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/`
- ProDom – `http://prodom.prabi.fr/`
- PROSITE patterns – `http://www.expasy.ch/prosite/`
- PROSITE profiles – `http://www.expasy.ch/prosite/`
- SCOP – `http://scop.mrc-lmb.cam.ac.uk/scop/`
- SMART – `http://smart.embl.de/`
- SUPERFAMILY – `http://supfam.org/`
- TIGRFAMs –`http://www.tigr.org/TIGRFAMs/`

Adresy stron wybranych programów do prezentacji struktur białek:

- Cn3D – `http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml`
- JMol – `http://jmol.sourceforge.net/`
- RasMol – `http://rasmol.org/`
- SwissPDB Viewer – `http://spdbv.vital-it.ch/`

6

Bazy danych anotacji funkcjonalnych

W poprzednich rozdziałach opisano biologiczne bazy danych, które zawierają reprezentację pewnych, fizyczne istniejących molekuł – ich sekwencje nukleotydowe, sekwencje białkowe czy też struktury przestrzenne. Ogólnie można powiedzieć, że zawartość tych baz danych powstaje wprost w wyniku zapisywania do nich rezultatów eksperymentów biologicznych, bezpośrednio przez grupy badawcze, które przeprowadzają te eksperymenty lub poprzez klasyfikację zawartości tych baz danych (w przypadku baz rodzin lub struktur białek).

Jednym z większych wyzwań dzisiejszej bioinformatyki jest stworzenie zasobów, które stanowić będą kompletną reprezentację naszej wiedzy na temat procesów biologicznych zarówno na poziomie komórki, jak i całego organizmu. Stąd też obecnie intensywnie rozwijane są bazy danych, które koncertują się na opisie procesów biologicznych i biochemicznych zachodzących w komórkach żywych organizmów, próbując opisać te procesy oraz relacje istniejące pomiędzy nimi. Zawartość tych baz danych tworzona jest najczęściej na podstawie informacji znajdujących się w publikacjach naukowych, które przeglądane są przez kuratorów baz danych. Zgromadzona w jednym miejscu informacja na temat szlaków metabolicznych, sieci sygnałowych genów czy innych oddziaływań międzycząsteczkowych pozwala na szybki dostęp do wiedzy, całościowe spojrzenie na problem, co pozwala na lepsze zrozumienie zależności pomiędzy procesami biologicznymi. Informacja zawarta w takich bazach danych często wykorzystywana jest w różnego rodzaju automatycznych procedurach wspomagających proces interpretacji biologicznych funkcji pełnionych przez geny lub białka albo w procesach automatycznego przewidywania funkcji genów, przebiegu procesów komórkowych, a nawet zachowania organizmów na podstawie genomicznej i molekularnej informacji.

6.1 KEGG

Baza danych KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [Kanehisa et al., 2008] jest zbiorem baz danych, które powstały w celu integracji geno-

micznych, chemicznych i systematycznych informacji na temat funkcjonowania systemów biologicznych oraz interakcji pomiędzy nimi. Od 1995 roku baza ta rozwijana jest w Kanehisa Laboratories w Kyoto University Bioinformatics Center i Human Genome Center na University of Tokyo. W ramach KEGG dostępnych jest kilkanaście różnych baz danych – każda z nich reprezentuje inny aspekt funkcjonowania systemów biologicznych. Przykładowo w bazie KEGG Genes podstawową jednostką informacji są geny oraz białka. KEGG Ligand zawiera informacje na temat substancji chemicznych oraz reakcji istotnych dla prawidłowego funkcjonowania komórek. KEGG Pathway jest bazą danych, która zawiera diagramy reprezentujące szlaki metaboliczne, ścieżki interakcji oraz reakcji związane z procesami komórkowymi, natomiast w hierarchicznej bazie KEGG Brite znajdują się, reprezentowane za pomocą kontrolowanego słownictwa, informacje na temat relacji i zależności pomiędzy różnego rodzaju procesami biologicznymi, cząsteczkami, związkami chemicznymi i innymi, szeroko pojętymi, systemami biologicznymi.

Poniżej przedstawiono informacje na temat zawartości oraz źródeł danych w poszczególnych bazach tworzących bazę KEGG:

- **Pathway** – metaboliczne ścieżki reakcji i interakcji na poziomie molekularnym, procesy komórkowe oraz choroby ludzkie. Zawartość bazy danych powstaje poprzez ręczne jej tworzenie na podstawie publikacji naukowych.
- **Brite** – hierarchiczna, funkcjonalna klasyfikacja białek, związków chemicznych i innych elementów dostępnych w ramach baz danych KEGG. Źródłem są tu publikacje naukowe analizowane przez kuratorów, którzy ręcznie tworzą zawartość tej bazy.
- **Genes** – baza Genes stanowi zbiór wszystkich kompletnych genomów (oraz niektórych częściowych genomów) uzyskanych na podstawie informacji dostępnych w publicznych bazach sekwencji. Genomy znajdujące się w tej bazie są ręcznie anotowane do bazy KEGG Orthology, a także podlegają analizie SSDB. Zasoby bazy Genes generowane są głównie na podstawie informacji z bazy NCBI RefSeq lub (w mniejszym stopniu) z innych publicznych baz danych. Podział komponentów, na które składa się ta baza, jest następujący:
  - **Genome** – sekwencje genomowe organizmów.
  - **Genes** – geny pochodzące z kompletnych, dobrze poznanych genomów, zawartość tej części bazy anotowana jest w sposób manualny.
  - **EGenes** – zbiory sekwencji genów w postaci kontigów EST.
  - **DGenes** – baza zawierająca informacje na temat genów pochodzących ze słabo poznanych, szkicowych (ang. *draft genome*).
  - **VGenes** – geny pochodzące z wirusów.
  - **OGenes** – geny pochodzące z organelli komórkowych.
- **KEGG Orthology (KO)** – grupy ortologów wyznaczone ręcznie na podstawie baz Pathway oraz Brite. Elementy bazy KO odpowiadają węzłom ścieżek KEGG Pathway oraz węzłom hierarchii KEGG Brite.

- **SSDB** – baza zawierająca punktację podobieństwa sekwencji oraz informacje o najlepszych trafieniach dla każdej pary sekwencji genów – na podstawie tej informacji możliwe jest wyszukiwanie ortologów, paralogów jak również zakonserwowanych grup genów. Zawartość bazy powstaje na podstawie bazy Genes poprzez porównywanie parami wszystkich sekwencji kodujących białka.
- **Ligand** – baza odzwierciedlająca aktualną wiedzę na temat chemicznych cząsteczek (ligandów), które wchodzą w reakcje z innym molekułami. Jest to również złożona baza danych, która dalej dzieli się następująco: **Compound** (C) – baza struktur metabolitów i niewielkich molekuł, **Glycan** (G) – struktury glikanów, **Reaction** (R) – reakcje biochemiczne, **RPair** (RP) – baza zawierająca porównania chemicznych struktur substrat–produkt dla reakcji chemicznych znajdujących się w bazie Reaction oraz **Enzyme** (EC) – baza nomenklatury enzymów. Zawartość katalogu Enzyme powstaje na podstawie bazy ExplorEnz, natomiast zawartości pozostałych katalogów tworzone są ręcznie, na podstawie analizy dostępnej literatury.
- **Disease** – baza danych chorób – głównie dotycząca chorób człowieka. Każdy wpis w bazie powiązany jest z odpowiednią ścieżką sygnałową w bazie KEGG Pathway w części *Human Diseases* i opisany zestawem genów związanych z daną chorobą.
- **Drug** – baza danych zawierająca struktury związków chemicznych lub komponentów reprezentujących znane leki dostępne na rynkach w Japonii (wszystkie dostępne), USA (większość dostępnych) oraz w Europie. Baza zawiera nie tylko opis samego leku, ale również obiektu, na który ukierunkowane jest leczenie. W postaci ścieżek sygnałowych KEGG zawartych w części *Drug Development* udostępniona jest także informacja na temat struktury związków chemicznych potrzebnych do wytworzenia danego leku. W przyszłości dostępny będzie tu również opis enzymów oraz transporterów biorących udział w metabolizmie leku, a także interakcje danego leku z innymi lekami.

Każdy obiekt (poza wpisami reprezentującymi geny) w bazie KEGG jest identyfikowany za pomocą pięciu cyfr poprzedzonych dużą literą (np. K05032 lub D00336) z wyjątkiem baz KEGG Pathway oraz KEGG Brite, gdzie cyfry poprzedzone są 2-4 literowym kodem (np. map00250, hsa04930). W tabeli 6.1 przedstawiono prefiksy obiektów KEGG dla poszczególny baz.

Mimo ogromnej liczby różnych biologicznych danych, które dostępne są w bazie KEGG, najbardziej charakterystyczną (a także unikalną) informacją, z którą najczęściej kojarzona jest baza KEGG, jest jej część związana ze ścieżkami sygnałowymi (KEGG Pathway). Baza danych ścieżek sygnałowych jest to zbiór ręcznie namalowanych diagramów, reprezentujących istniejącą wiedzę na temat szlaków metabolicznych w postaci sieci interakcji oraz reakcji, które konieczne są do prawidłowego funkcjonowania komórek.

Podstawowy podział informacji zawartej w tej bazie obejmuje następujące aspekty:

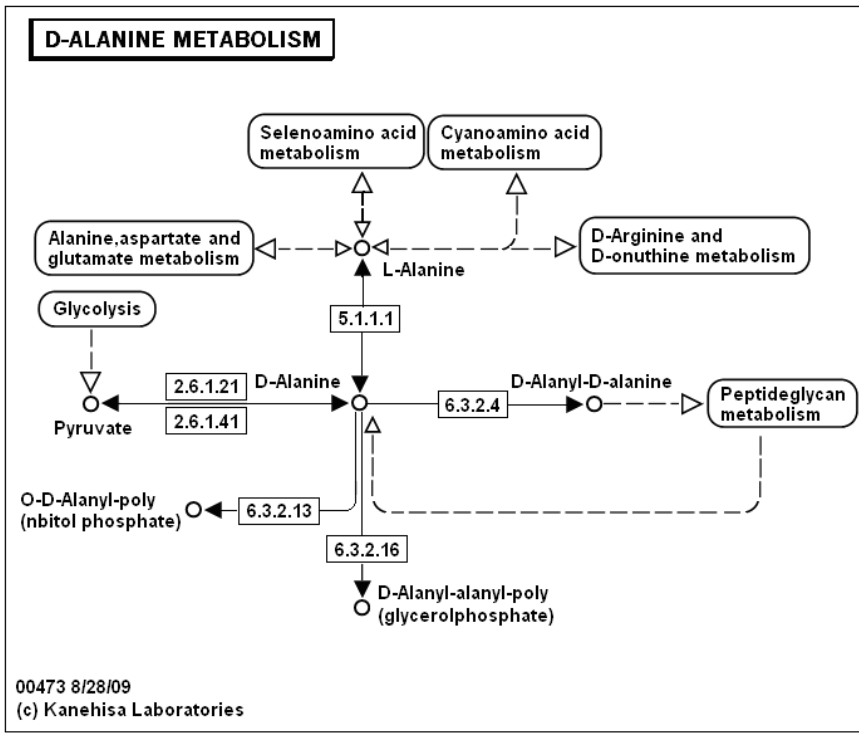
Tabela 6.1. Prefiksy obiektów KEGG

Prefiks	Baza danych
K	KEGG Orthology
C	KEGG Ligand/Compound
D	KEGG Drug
G	KEGG Ligand/Glycan
R	KEGG Ligand/Reaction
RP	KEGG Ligand/RPair
map/ko/ec/rn/(org)	KEGG Pathway
br/ko/(org)	KEGG Brite
M	KEGG Module
H	KEGG Disease
T	KEGG Genome

- Metabolism.
- Genetic Information Processing.
- Environmental Information Processing.
- Cellular Processes.
- Human Diseases.
- Drug Development – struktury związków chemicznych reprezentujące znane leki.

Każda ścieżka sygnałowa znajdująca się w bazie reprezentowana jest przez plik graficzny w formacie PNG. Na rysunku 6.1 przedstawiono przykładowy szlak metaboliczny – metabolizm jednego z aminokwasów: D–alaniny. Zaokrąglone prostokąty oznaczają inne szlaki metaboliczne powiązane z danym szlakiem, prostokąty oznaczone są numerem identyfikującym konkretny enzym z bazy Ligand/Enzyme i są równocześnie odnośnikami do terminów bazy KEGG Orthology, natomiast okręgi reprezentują związki chemiczne znajdujące się w bazie Ligand/Compound. Na przykład E5.1.1.1 to enzym *alanine racemase* – wpis K01775 z bazy KEGG Orthology będący równocześnie wpisem bazy Ligand/Enzyme EC:5.1.1.1; L-Alanine to z kolei związek o symbolu C00041 bazy Ligand/Compound.

Graficzne przedstawienie przebiegu procesów metabolicznych i złożonych zależności pomiędzy nimi jest naturalne i najlepsze z punktu widzenia interpretacji przez człowieka, natomiast wadą takiego rozwiązania jest brak możliwości komputerowego przetwarzanie informacji zapisanej w takiej formie. Stąd też każdy diagram reprezentowany jest również w postaci dodatkowego pliku w formacie KGML (*KEGG Markup Language*), który zawiera informację przedstawioną w obiekcie graficznym. Pliki KGML umożliwiają komputerowe przetwarzanie informacji, która przedstawiona jest w formacie graficznym, a także pozwalają na automatyczne rysowanie diagramów.



Rysunek 6.1. Szlak metabolizy D-alaniny

Pliki KEGG dla ścieżek metabolicznych zawierają informacje na temat dwóch rodzajów obiektów: jakie są wzajemne relacje pomiędzy enzymami (przedstawionymi w postaci prostokątów) oraz reakcje pomiędzy związkami chemicznymi (przedstawionymi w postaci okręgów). Pliki KGML dla ścieżek regulatorowych zawierają jedynie informacje, jakie są relacje pomiędzy białkami (przedstawionymi za pomocą prostokątów). W bazie ścieżek metabolicznych prostokąty identyfikowane są za pomocą terminów Ontologii KEGG bazy KEGG Orthology, ale z uwagi na zaszłości historyczne oznaczone są za pomocą identyfikatorów EC bazy Ligand/Enzyme.

6.2 Gene Ontology

Baza danych Ontologii Genowych (*Gene Ontology*) jest uniwersalną, hierarchiczną bazą danych zawierającą opisy genów oraz produktów genowych [Ashburner et al., 2000]. Baza Ontologii Genowych stworzona w roku 1998, jest utrzymywana przez konsorcjum a jej celem jest standaryzacja i ujednoli-

cenie informacji na temat genów oraz ich produktów. Idea stworzenia takiej bazy informacji powstała w momencie, gdy okazało się, że biolodzy z całego świata, odkrywając nowe geny w różnych organizmach i poznając ich funkcje, tworzyli swoje własne nazewnictwo. W efekcie nie byli w stanie w żaden sposób porównać wyników swoich odkryć pomiędzy organizmami ani stwierdzić, czy w innym organizmie występuje gen o podobnej funkcji. W chwili obecnej setki osób zaangażowanych w projekt *Gene Ontology* przeglądają literaturę naukową w celu odnalezienia informacji na temat nowo odkrytych genów i opisania ich za pomocą terminów Ontologii Genowych, a także w celu weryfikacji podanych informacji na temat istniejących genów.

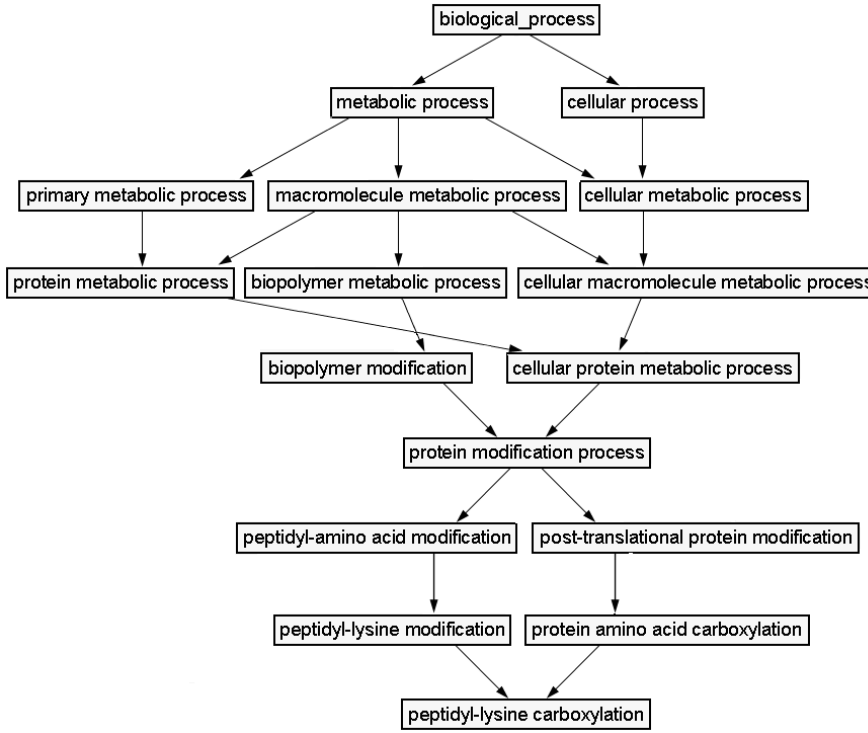
Baza Ontologii Genowych podzielona jest na trzy „podbazy" (trzy główne ontologie):

- Proces Biologiczny (ang. *Biological Process* – BP).
- Funkcja Molekularna (ang. *Molecular Function* – MF).
- Komponent Komórkowy (ang. *Cellular Component* – CC).

Każda z tych trzech ontologii dostarcza informacji na temat produktów genowych w kontekście procesów biologicznych, funkcji molekularnej oraz komponentu komórkowego. Produkt genowy może być związany lub zlokalizowany w określonej części komórki, aktywny w różnych procesach biologicznych, w ramach których pełni specyficzne funkcje molekularne. Przykładowo białko *cytochrom c* może być opisane za pomocą funkcji molekularnej aktywność oksydoreduktazy (ang. *oxidoreductase activity*), procesu biologicznego fosforylacja oksydacyjna (ang. *oxidative phosphorylation*) i wywołanie śmierci komórki (ang. *induction of cell death*), i komponentu komórkowego macierz mitochondrialna (ang. *mitochondrial matrix*) oraz wewnętrzna membrana mitochondrium (ang. *mitochondrial inner membrane*).

Proces Biologiczny, Funkcja Molekularna oraz Komponent Komórkowy są korzeniami trzech, oddzielnych, skierowanych acyklicznych grafów zawierających terminy Ontologii Genowych zorganizowanych w sposób hierarchiczny. Każdy termin określony jest przez nazwę, unikalny identyfikator (7 cyfr poprzedzonych przedrostkiem GO – np. GO:0018235) oraz miejsce w hierarchii drzewa ontologii. Im głębiej w dół grafu, tym terminy są bardziej specyficzne i dokładniej określają funkcje genów. Terminy reprezentowane są jako węzły grafu. Jeden termin może posiadać kilku rodziców, a relacje pomiędzy terminem a jego rodzicami określane są za pomocą zwrotów *is_ a*, *part_ of*, *regulates*, *negatively regulates* oraz *positively regulates*. Na rysunku 6.2 przedstawiono niewielki fragment struktury ontologii Proces Biologiczny, zawierający wszystkich rodziców terminu peptidyl-lysine carboxylation (GO:0018235).

Wszystkie terminy Ontologii Genowych tworzące graf muszą spełniać *regułę prawdziwości ścieżki*, co oznacza, że relacje, które występują pomiędzy terminami, jeśli przechodzimy od wybranego węzła do korzenia poprzez wszystkie terminy będące jego rodzicami, muszą zawsze opisywać prawdziwe zależności biologiczne występujące w żywych organizmach.



Rysunek 6.2. Fragment grafu ontologii *Proces Biologiczny* – na rysunku przedstawiono wszystkie terminy będące rodzicami terminu peptidyl-lysine carboxylation (GO:0018235)

Podstawowym formatem, w jakim rozprowadzana jest baza Ontologii Genowych, jest plik tekstowy w formacie OBO (*Open Biomedical Ontologies*) o nazwie `gene_ontology.1.2.obo`. Baza dostępna jest również w postaci relacyjnej za pomocą plików zawierających obraz zawartości bazy w formacie MySQL oraz SQL, a także w postaci XML. Podstawowe cechy, jakie charakteryzują dane zapisane w postaci OBO, to: możliwość łatwej interpretacji przez człowieka, łatwość przetwarzania za pomocą komputerowych metod, możliwość rozszerzania informacji oraz możliwie niska redundancja danych.

Format pliku OBO jest następujący:

```
<header>

<stanza>
<stanza>
...
```

Puste linie są ignorowane. Wpisy w pliku mają charakter klucz–wartość. W pliku występuję nagłówek (ang. *header*), który kończy się wraz z rozpoczęciem pierwszej strofy (ang. *stanza*). Każda strofa jest oznakowaną sekcją dokumentu będącą właściwym opisem obiektu. Strofy składają się z nazwy w nawiasach kwadratowych ([Term]) oraz z listy atrybutów obiektu przedstawionych w postaci klucz–wartość. Poniżej przedstawiono przykładową definicję terminu Ontologii Genowych pochodzącą z pliku OBO:

```
[Term]
id: GO:0048505
name: regulation of timing of cell differentiation
namespace: biological_process
def: "The process controlling the activation andór rate at which
relatively unspecialized cells acquire specialized features."
[GOC:bf, GOC:jic]
synonym: "timing of cell differentiation" RELATED []
is_a: GO:0040034 ! regulation of development, heterochronic
is_a: GO:0045595 ! regulation of cell differentiation
```

Znaczenie poszczególnych kluczy jest następujące: *id* – identyfikator terminu GO, *name* - nazwa terminu, *namespace* – jeden z trzech typów ontologii, do których należy termin, *def* – definicja terminu zakończona umieszczoną w nawiasach kwadratowych informacją referencyjną o jej pochodzeniu (może to być odnośnik do publikacji, wpisu w innej bazie danych itd.; w powyższym przypadku [GOC:bf, GOC:jic] identyfikują konkretnych kuratorów bazy, którzy są autorami definicji), *synonym* – inna nazwa, która może być wykorzystywana dla określenia tego samego terminu, *is_ a* – określa rodzica danego terminu oraz rodzaj relacji.

6.2.1 Anotacje genów za pomocą terminów GO

Anotacja genu za pomocą Ontologii Genowych polega na przyporządkowaniu do danego genu bądź też produktu genowego terminu Ontologii Genowej, który według istniejącej wiedzy biologicznej najlepiej go opisuje. Przykładem produktu genowego może być kwas rybonukleinowy lub białko. Ponieważ jeden gen może opisywać wiele różnych produktów genowych, konsorcjum *Gene Ontology* zaleca, aby terminy Ontologii Genowych opisywały produkty genowe, a nie poszczególne geny. W sytuacji gdy produkt genowy nie posiada swojej odrębnej nazwy, dopuszcza się anotację genu. W wielu przypadkach grupy nadzorujące proces anotacji dla danego gatunku nie posiadają odrębnej bazy produktów genowych. Na przykład grupa SGD (*Saccharomyces Genome Database*) zajmująca się anotacją drożdży piekarskich (*Saccharomyces cerevisiae*) opisuje terminami ontologii jedynie geny, przyjmując za podstawę opisu produkty tych genów, traktując gen oraz jego produkt równoważnie.

Gen lub produkt genowy może być przyporządkowany do jednego lub więcej węzłów grafu ontologii na dowolnym jej poziomie. Poszczególne opisy za

pomocą terminów Ontologii Genowych są od siebie niezależne. Każda anotacja musi zawierać odniesienie do źródła (publikacji), z którego pochodzi informacja, na podstawie której została ona zdefiniowana oraz informacje, jakiego typu analizy zostały przeprowadzone przez autora publikacji. Informacja na temat typu analiz określona jest przez kod ewidencji anotacji (ang. *evidence code*).

Podstawowy podział kodów ewidencji jest następujący:

- Eksperymentalne.
- Analizy obliczeniowe.
- Oświadczenie autora.
- Oświadczenie kuratora.
- Wnioskowane na podstawie elektronicznych anotacji.

Kody ewidencji typu eksperymentalnego odnoszą się do anotacji, które powstały na podstawie źródła opisującego wyniki przeprowadzonego w laboratorium eksperymentu biologicznego.

Analizy obliczeniowego dotyczą sytuacji, kiedy gen lub produkt genowy został przyporządkowany do terminu Ontologii Genowej na podstawie analiz *in slico* sekwencji genowych, a następnie wyniki te zostały potwierdzone poprzez analizę literatury w danej dziedzinie.

Oświadczenie autora określa, że anotacja została utworzona na podstawie stwierdzenia autora znajdującego się w publikacji odniesienia.

Oświadczenie kuratora znajduje swoje zastosowanie w sytuacji, kiedy trudno zastosować jeden z dostępnych typów ewidencji, jednakże na podstawie innych, istniejących już wcześniej anotacji można przyporządkować dany gen lub produkt genowy do danego terminu Ontologii Genowej.

Wnioskowanie na podstawie elektronicznych anotacji dotyczy wszystkich sytuacji, w których gen został opisany terminem Ontologii Genowej na bazie obliczeń komputerowych (m.in. analizy sekwencji genowych), jednakże nie zostały one potwierdzone poprzez przegląd dostępnej literatury. Z tego powodu przyporządkowania produktów genowych od terminów, które powstały na bazie anotacji elektronicznych, niezweryfikowanych w żaden sposób, traktuje się jako mniej wiarygodne informacje.

6.3 Anotacje funkcjonalne grup genów

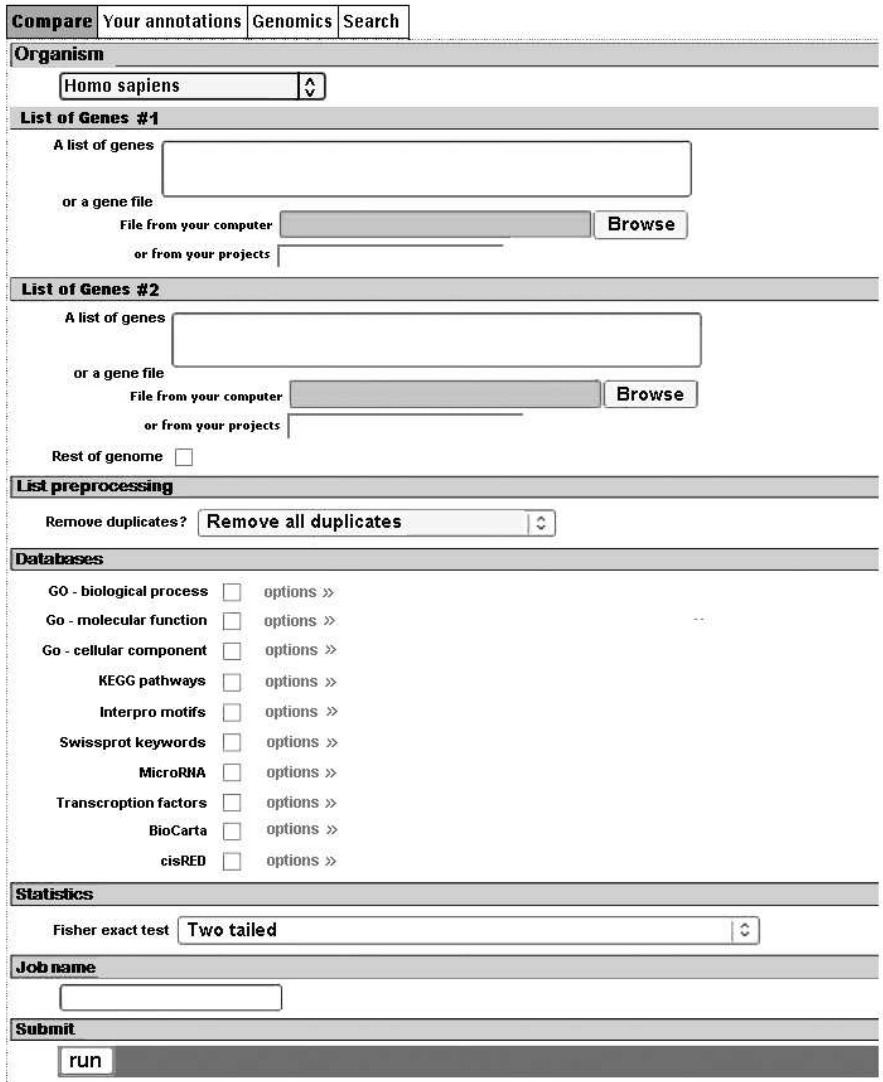
Często w wyniku różnego rodzaju eksperymentów biologicznych badacze otrzymują pewne zbiory genów, które są dla nich interesujące – może to być na przykład grupa genów, która w warunkach przeprowadzanego eksperymentu biologicznego wykazuje się podobnym zachowaniem. W takich sytuacjach badacze zainteresowani są określeniem, jakie funkcje pełnią geny należące do takiej grupy. Jedną z popularniejszych i szeroko stosowanych metod poszukiwania tego rodzaju zależności jest statystyczna analiza występowania częstości pojęć

charakteryzujących geny (takich jak np. Ontologie Genowe czy szlaki metaboliczne KEGG) w interesującej badacza grupie oraz pośród pozostałych genów. Jeśli w wyniku analiz statystycznych obserwuje się w analizowanej grupie częstsze niż wynikałoby to z przypadku, zagęszczenie pewnych słów kluczowych (w porównaniu z grupą odniesienia), wówczas naukowcy mają podstawę do wyciągania wniosków na temat zależności pomiędzy zachowaniem genów wynikającym z pomiarów eksperymentalnych, a ich funkcją lub procesem biologicznym opisywanym przez dane pojęcie.

Dostępnych jest wiele niekomercyjnych narzędzi zarówno w postaci samodzielnych aplikacji, jak i serwisów internetowych, które umożliwiają przeprowadzanie tego rodzaju analiz. Znajdując liczbę genów opisanych danym terminem ontologii w analizowanej grupie oraz pośród pozostałych genów, a następnie stosując odpowiednie testy statystyczne, można uzyskać listę terminów ontologii nadreprezentowanych lub niedoreprezentowanych w interesującej badacza grupie genów. Analizę taką przeprowadza się oddzielnie dla każdego pojęcia opisującego geny w badanej grupie, przeprowadzając oddzielnie dla każdego z nich test statystyczny. Przyjmuje się tu hipotezę zerową mówiącą o tym, że występowanie danego słowa kluczowego w badanej grupie genów jest przypadkowe i szuka się statystycznie istotnych dowodów pozwalających na odrzucenie hipotezy zerowej. Najczęściej stosowane do tego celu testy statystyczne to: dokładny test hipergeometryczny (zwany też schematem urnowym), test χ^2 lub test dwumianowy.

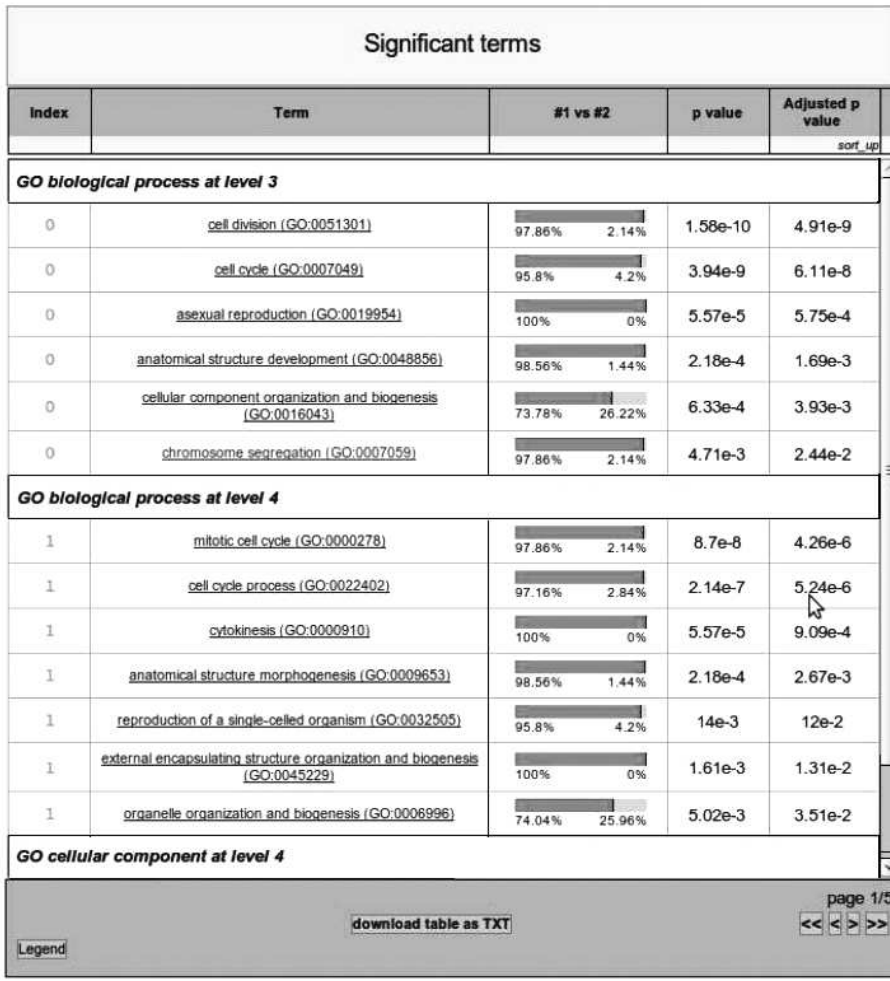
6.3.1 FatiGO – funkcjonalna anotacja grup genów

Przykładem jednego z najpopularniejszych programów wykorzystywanych do funkcjonalnych anotacji grup genów jest internetowa aplikacja *FatiGO* dostępna w ramach systemu *Babelomics* [Al-Shahrour et al., 2005]. Portal Babelomics rozwijany jest od 2005 roku w Departamencie Bioinformatyki w Centro de Investigacion Principe Felipe w Valencji w Hiszpanii i jest to dostępny w jednym miejscu zbiór wielu powiązanych ze sobą narzędzi, które pozwalają na funkcjonalną anotację wyników eksperymentów biologicznych przeprowadzanych na skalę genomową. FatiGO umożliwia przeprowadzanie analiz funkcjonalnych grup genów za pomocą pojęć takich jak Ontologie Genowe, szlaki metaboliczne KEGG, motywy InterPro (ang. *InterPro motifs*), słowa kluczowe Swissprot (ang. *Swissprot keywords*), microRNA, czynniki transkrypcyjne (ang. *transcription factors*), BioCarta, cisRed. Przeprowadzanie anotacji funkcjonalnych genów za pomocą programów takich jak FatiGO odbywa się zazwyczaj według bardzo podobnego schematu. Użytkownik wybiera gatunek, którego geny będą analizowane i przesyła dwie listy genów: zbiór genów oraz interesującą go listę referencyjną. Istnieje również możliwość wyboru pojęć, którymi mogą być anotowane geny, możliwość wyboru testu statystycznego oraz możliwość określenia poziomu znamienności statystycznej, na którym odrzucana będzie hipoteza zerowa. Na rysunku 6.3 przedstawiono formatkę programu FatiGO.



Rysunek 6.3. FatiGO – formatka wprowadzania danych do anotacji funkcjonalnej grup genów

W wyniku przeprowadzonej analizy statystycznej użytkownik otrzymuje listę pojęć znamiennych statystycznie, które opisują interesującą go grupę genów. Na rysunku 6.4 przedstawiono przykładowe wyniki funkcjonalnej anotacji grupy genów za pomocą terminów Ontologii Genowych. Widoczna jest lista terminów znamiennych statystycznie – dla każdego terminu podano tu jego na-



Significant terms

Index	Term	#1 vs #2	p value	Adjusted p value (sort_up)
GO biological process at level 3				
0	cell division (GO:0051301)	97.86% 2.14%	1.58e-10	4.91e-9
0	cell cycle (GO:0007049)	95.8% 4.2%	3.94e-9	6.11e-8
0	asexual reproduction (GO:0019954)	100% 0%	5.57e-5	5.75e-4
0	anatomical structure development (GO:0048856)	98.56% 1.44%	2.18e-4	1.69e-3
0	cellular component organization and biogenesis (GO:0016043)	73.78% 26.22%	6.33e-4	3.93e-3
0	chromosome segregation (GO:0007059)	97.86% 2.14%	4.71e-3	2.44e-2
GO biological process at level 4				
1	mitotic cell cycle (GO:0000278)	97.86% 2.14%	8.7e-8	4.26e-6
1	cell cycle process (GO:0022402)	97.16% 2.84%	2.14e-7	5.24e-6
1	cytokinesis (GO:0000910)	100% 0%	5.57e-5	9.09e-4
1	anatomical structure morphogenesis (GO:0009653)	98.56% 1.44%	2.18e-4	2.67e-3
1	reproduction of a single-celled organism (GO:0032505)	95.8% 4.2%	14e-3	12e-2
1	external encapsulating structure organization and biogenesis (GO:0045229)	100% 0%	1.61e-3	1.31e-2
1	organelle organization and biogenesis (GO:0006996)	74.04% 25.96%	5.02e-3	3.51e-2
GO cellular component at level 4				

page 1/5 << < > >>

download table as TXT

Legend

Rysunek 6.4. FatiGO – formatka zawierająca przykładowe wyniki anotacji funkcjonalnej grupy genów

zwę i symbol, procentowe porównanie występowania danego terminu w analizowanej grupie i w grupie odniesienia, poziom istotności testu (p-wartość) oraz poziom istotności testu po korekcji związanej z testowaniem wielokrotnym.

FatigGO jest przykładem jednej z wielu aplikacji, które umożliwiają przeprowadzanie funkcjonalnej anotacji grup genów. Inne aplikacje, które umożliwiają przeprowadzanie podobnych analiz, to np. DAVID, GOTM (*Gene Ontology Tree Machine*), Onto–Express czy Pathway–Express.

6.4 Adresy Internetowe

- DAVID – `http://david.abcc.ncifcrf.gov/`
- FatiGO – `http://babelomics.bioinfo.cipf.es/`
- Gene Ontology – `http://www.geneontology.org/`
- GOTM – `http://bioinfo.vanderbilt.edu/gotm/`
- KEGG – `http://www.genome.jp/kegg/`
- Onto–Express – `http://vortex.cs.wayne.edu/ontoexpress/`
- Pathway–Express – `http://vortex.cs.wayne.edu/ontoexpress/`

Literatura

[Al-Shahrour et al., 2005] Al-Shahrour et al. (2005). Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Research*, 33(Web Server issue):W460–W464.

[Andreeva et al., 2008] Andreeva, A. et al. (2008). Data growth and its impact on the scop database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–D425.

[Apweiler et al., 2004] Apweiler, R., Bairoch, A., and Wu, C. (2004). Protein sequence databases. *Curr Opin Chem Biol*, 8(1):76–80.

[Ashburner et al., 2000] Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25:25–29.

[Attwood, 2002] Attwood, T. (2002). The prints database: a resource for identification of protein families. *Brief Bioinform*, 3(3):252–263.

[Barker et al., 2000] Barker, W. et al. (2000). The protein information resource (pir). *Nucleic Acids Res*, 28(1):41–44.

[Bateman et al., 2004] Bateman, A. et al. (2004). The pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–D141.

[Baxevanis and Ouellette, 2004] Baxevanis, A. and Ouellette, B. (2004). *Bioinformatyka. PodrÄtcznik do analizy genÃłw i biaÅ'ek*. Wydawnictwo Naukowe PWN, Warszawa.

[Benson et al., 2007] Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D. (2007). Genbank. *Nucleic Acids Res*, 35(Database issue):D21–D25.

[Berman, 2008] Berman, H. (2008). The protein data bank: a historical perspective. *Acta Crystallogr A*, 64(Pt 1):88–95.

[Bru et al., 2005] Bru, C. et al. (2005). The prodom database of protein domain families: more emphasis on 3d. *Nucleic Acids Res*, 33(Database issue):D212–D215.

[Consortium, 2007] Consortium, U. (2007). The universal protein resource (uniprot). *Nucleic Acids Res*, 35(Database issue):D193–D197.

[Cuff et al., 2009] Cuff, A. et al. (2009). The cath classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*, 37(Database issue):D310–D314.

[Dayhoff et al., 1978] Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:345–351.

[Ewens and Grant, 2006] Ewens, W. and Grant, G. (2006). *Statistical Methods in Bioinformatics : An Introduction (Statistics for Biology and Health)*. Springer.

[Galperin and Cochrane, 2009] Galperin, M. and Cochrane, G. (2009). Nucleic acids research annual database issue and the nar online molecular biology database collection in 2009. *Nucleic Acids Res*, 37(Database issue):D1–D4.

[Garcia-Molina et al., 2006] Garcia-Molina, H., Ullman, J., and Widom, J. (2006). *Systemy baz danych*. Wydawnictwa Naukowo Techniczne.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

[Higgs and Attwood, 2008] Higgs, P. and Attwood, T. (2008). *Bioinformatyka i ewolucja molekularna*. Wydawnictwo Naukowe PWN, Warszawa.

[Hu et al., 2004] Hu, Z. et al. (2004). iprolink: an integrated protein resource for literature mining. *Comput Biol Chem*, 28(5-6):409–416.

[Huang et al., 2003] Huang, H. et al. (2003). iproclass: an integrated database of protein family, function and structure information. *Nucleic Acids Res*, 31(1):390–392.

[Hulo et al., 2008] Hulo, N. et al. (2008). The 20 years of prosite. *Nucleic Acids Res*, 36(Database issue):D245–D249.

[Kanehisa et al., 2008] Kanehisa, M. et al. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484.

[Karlin and Altschul, 1993] Karlin, S. and Altschul, S. (1993). Application and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA*, 90:5873–5877.

[Kulkowa et al., 2007] Kulkowa, T. et al. (2007). EMBL nucleotide sequence database in 2006. *Nucleic Acids Res*, 35(Database issue):D16–D20.

[Mulder et al., 2002] Mulder, N. et al. (2002). Interpro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*, 3(3):225–235.

[Needleman and Wunsch, 1970] Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

[Sugawara et al., 2008] Sugawara, H. et al. (2008). Ddbj with new system and face. *Nucleic Acids Res*, 36(Database issue):D22–D24.

[Suzek et al., 2007] Suzek, B. et al. (2007). Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288.

[Wang et al., 2002] Wang, Y. et al. (2002). Mmdb: Entrez's 3d-structure database. *Nucleic Acids Res*, 30(1):249–252.

[Wu et al., 2004] Wu, C. et al. (2004). Pirsf: family classification system at the protein information resource. *Nucleic Acids Res*, 32(Database issue):D112–D114.

[Ye et al., 2006] Ye, J., S., M., and Madden, T. (2006). Blast: improvements for better sequence analysis. *Nucleic Acids Res*, 34(Web Server issue):W6–W9.

Dodatek

1 Przykład rekordu pochodzącego z bazy sekwencji EMBL

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
```

```
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="HSSP:P26205"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
...
agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac      1800
tttgaattaa aagtcttttt ttattttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa       1859
/
```

2 Przykład rekordu pochodzącego z bazy sekwencji GenBank

```
LOCUS       X56734 1859 bp mRNA linear PLN 25-NOV-2005
DEFINITION  Trifolium repens mRNA for non-cyanogenic beta-glucosidase.
ACCESSION   X56734 S46826
VERSION     X56734.1  GI:21954
KEYWORDS    beta-glucosidase.
SOURCE      Trifolium repens (white clover)
  ORGANISM  Trifolium repens
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae;
            Trifolium.
REFERENCE   1  (bases 1 to 1859)
  AUTHORS   Oxtoby,E., Dunn,M.A., Pancoro,A. and Hughes,M.A.
  TITLE     Nucleotide and derived amino acid sequence of the cyanogenic
            beta-glucosidase (linamarase) from white clover (Trifolium repens
            L.)
  JOURNAL   Plant Mol. Biol. 17 (2), 209-219 (1991)
   PUBMED   1907511
REFERENCE   2  (bases 1 to 1859)
  AUTHORS   Hughes,M.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-NOV-1990) Hughes M.A., University of Newcastle Upon
            Tyne, Medical School, Newcastle Upon Tyne, NE2 4HH, UK
COMMENT     On Jun 10, 2005 this sequence version replaced gi:233395.
FEATURES             Location/Qualifiers
     source          1..1859
                     /organism="Trifolium repens"
                     /mol_type="mRNA"
                     /db_xref="taxon:3899"
                     /clone="TRE361"
                     /tissue_type="leaves"
                     /clone_lib="lambda gt10"
     mRNA            1..1859
                     /experiment="experimental evidence, no additional details
                     recorded"
     CDS             14..1495
                     /EC_number="3.2.1.21"
                     /note="non-cyanogenic"
                     /codon_start=1
                     /product="beta-glucosidase"
                     /protein_id="CAA40058.1"
                     /db_xref="GI:21955"
                     /db_xref="GOA:P26204"
                     /db_xref="InterPro:IPR001360"
                     /db_xref="UniProtKB/Swiss-Prot:P26204"
                     /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGF
                     IFGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGI
                     MKDQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWD
                     LPQVLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTN
                     APGRCSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNW
                     LMPLDDNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLV
                     NGSFDFIGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYV
                     YPYMFIQEDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYR
                     HLYYIRSAIRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
```

```
ORIGIN
        1 aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt
       61 cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag
      121 tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga
      181 aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata
      241 tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta
          ...
     1741 agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac
     1801 tttgaattaa aagtcttttt ttatttttct aaaaaaaaaa aaaaaaaaaa aaaaaaaaa
/
```

Indeks

ポーランド日本情報工科大学

PJWSTK ポーランド日本情報工科大学

POLSKO-JAPOŃSKA WYŻSZA SZKOŁA TECHNIK KOMPUTEROWYCH

Jedna z najlepszych uczelni w Polsce – wyróżniana przez pracodawców, studentów i media.

Akredytacja Państwowej Komisji Akredytacyjnej

Informatyka

Studia inżynierskie, magisterskie uzupełniające, podyplomowe, studia doktoranckie, uprawnienia habilitacyjne, studia przez internet.

Specjalizacje:

animacja 3D, bazy danych, eksploracja www, inteligentne systemy przetwarzania danych, inżynieria oprogramowania i baz danych, multimedia, programowanie aplikacji biznesowych, programowanie gier, programowanie systemowe i sieciowe, robotyka, sieci urządzeń mobilnych, systemy rozproszone i równoległe.

Architektura Wnętrz

Studia licencjackie

Kultura Japonii

Studia licencjackie

Sztuka Nowych Mediów (Grafika)

Studia licencjackie, magisterskie uzupełniające

Zarządzanie Informacją

Studia inżynierskie

Kursy:

Akademia Sieciowa CISCO; LPI Linux; Microsoft

Akademickie Liceum Ogólnokształcące przy PJWSTK

02-008 Warszawa, ul. Koszykowa 86
tel.: 22 58 44 500, fax: 22 58 44 501
e-mail: pjwstk@pjwstk.edu.pl
www.pjwstk.edu.pl

PJWSTK w Bytomiu
Informatyka
Sztuka Nowych Mediów, Grafika
Studia inżynierskie, licencjackie
41-902 Bytom, Aleja Legionów 2
tel.: 32 387 16 60
e-mail: bytom@pjwstk.edu.pl
www.bytom.pjwstk.edu.pl

PJWSTK w Gdańsku
Informatyka
Studia inżynierskie
80-045 Gdańsk, ul. Brzegi 55
tel. 58 683 59 75
e-mail: gdansk@pjwstk.edu.pl
www.gdansk.pjwstk.edu.pl

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt **"Nowoczesna kadra dla e-gospodarki"** – program rozwoju Wydziału Zamiejscowego Informatyki w Bytomiu Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych, współfinansowany przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach Podziałania 4.1.1. "Wzmocnienie potencjału dydaktycznego uczelni" Programu Operacyjnego Kapitał Ludzki

ISBN 978-83-89244-90-1

Egzemplarz bezpłatny