STUDIA INFORMATICA 2011
Volume 32 Number 3B (99)

Artur SIĄŻNIK, Bożena MAŁYSIAK-MROZEK, Dariusz MROZEK
Politechnika Śląska, Instytut Informatyki

EKSPLORACJA DANYCH GENETYCZNYCH BAZY GENBANK Z WYKORZYSTANIEM USŁUG SIECIOWYCH

Streszczenie. National Center for Biotechnology Information (NCBI) gromadzi ogromne liczby danych opisujących różne organizmy biologiczne na wiele różnych sposobów. Dane te są przechowywane we właściwych bazach danych, zarządzanych przez NCBI. Baza danych GenBank jest jedną z najbardziej znanych na świecie baz NCBI przechowujących dziesiątki milionów sekwencji nukleotydowych DNA i RNA. W niniejszym artykule przedstawiono autorski system eksploracji danych genetycznych bazy GenBank. System *search GenBank* pozwala nie tylko wyszukiwać i przeglądać dane biologiczne bazy GenBank, ale także łączyć znalezione wpisy bazy GenBank z danymi w innych bazach danych NCBI, dając w ten sposób możliwość międzybazowej eksploracji danych.

Słowa kluczowe: bioinformatyka, DNA, RNA, bazy danych, usługi sieciowe

EXPLORATION OF GENETIC DATA FROM GENBANK USING WEB SERVICES

Summary. National Center for Biotechnology Information (NCBI) collects huge amounts of data describing various biological organisms in several ways. These data are stored in appropriate databases, managed by the NCBI. GenBank is one of the world's most famous NCBI database storing tens of millions of nucleotide sequences of DNA and RNA. In this article, we present a new system designed to explore genetic data in the GenBank database. The *search GenBank* system not only allows to search and browse biological data in the GenBank, but also combine the GenBank database entries with items in other NCBI databases. Therefore, the *search GenBank* provides the cross-database exploration possibilities.

Keywords: bioinformatics, DNA, RNA, databases, web services

1. Wprowadzenie

W związku z gwałtownym rozwojem informatyki, który mogliśmy obserwować w przeciągu minionych lat, wiele innych dziedzin nauk może dziś korzystać z rozwiązań i technologii, których ów rozwój jest niezaprzeczalną przyczyną.

Połączenie nauk medycznych z informatyką śniło się niegdyś tylko pisarzom fantastyki naukowej, a jednak dziś wspólna praca naukowców z całkowicie różnych dziedzin nauki nie jest niczym nadzwyczajnym. Co więcej, proces współpracy naukowców o innych kręgach zainteresowań zapoczątkował tworzenie się pośrednich dziedzin nauki, takich które potrafią w harmonii połączyć wiedzę i doświadczenie z, wydawać by się mogło, odległych światów. Zdaje się, że bez współpracy ponad podziałami, wynikającymi z różnic pomiędzy różnymi dziedzinami nauki, dzisiejszy obraz pracy naukowej wyglądałby zupełnie inaczej. Na szczęście, nie jest nam dane się przekonać, co by było, gdyby historia potoczyła się zupełnie inną ścieżką.

Medycyna i nauki przyrodnicze kilka lat temu zaczęły być gałęziami nauki, których badania generują ogromną liczbę danych. Ta niewyobrażalnie duża liczba danych musiała zostać w jakiś sposób przetworzona i zachowana. Zadaniem tym zajęli się informatycy i bioinformatycy, którzy, posiadając odpowiednie technologie baz danych, a także dobre podstawy teoretyczne związane z naukami przyrodniczymi, potrafili rozwiązać problem z przechowywaniem i przetwarzaniem dużej ilości informacji natury biologicznej.

Ze względu na nieustannie rosnący zbiór informacji, rozwiązania i techniki skupiające się na zadaniu przechowywania danych medycznych w bazach danych są coraz to bardziej dopracowywane. Jedną z głównych przyczyn powstania bioinformatycznych baz danych [1, 2] jest bardzo duża ilość informacji genetycznych, w tym sekwencji nukleotydowych, dla których nie ma lepszej metody przechowywania niż systemy baz danych. Praktycznie od 1981 roku, kiedy została wynaleziona metoda sekwencjonowania Sangera, problem przechowywania informacji genetycznych jest cały czas aktualny. Jedną z najbardziej znanych na świecie baz danych przechowujących informacje genetyczne jest baza GenBank [3] utrzymywana przez National Center for Biotechnology Information (NCBI)[1].

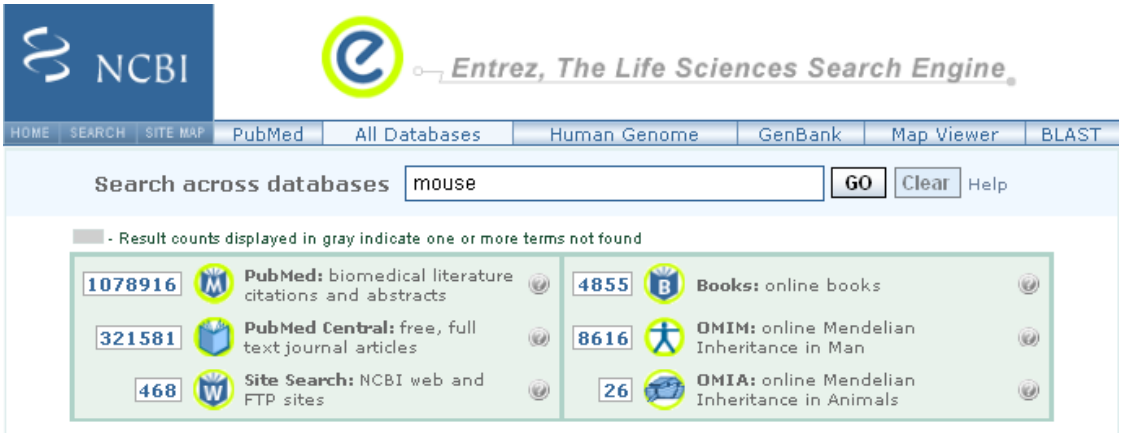
W niniejszym artykule przedstawiono narzędzie do eksploracji danych genetycznych bazy GenBank. Specjalnie zaprojektowany portal internetowy, wykorzystując usługi sieciowe, pozwala na wyszukiwanie informacji w tejże bazie danych, a także w bazach powiązanych z bazą GenBank.

[1] National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov

2. Pobieranie informacji z baz danych NCBI

Do wyszukiwania i pobierania informacji z baz danych, utrzymywanych przez NCBI, wykorzystuje się system Entrez [4], który jest zarówno narzędziem do indeksowania rekordów baz danych. Pierwsza wersja systemu była rozprowadzana na płycie CD-ROM (1991 rok). W tym czasie Entrez (rysunek 1) obsługiwał bazę danych GenBank z zamieszczonymi tam sekwencjami nukleotydowymi, a także bazę sekwencji aminokwasowych Proteins [5], która to przechowywała sekwencje białek odpowiadających sekwencjom nukleotydowym w bazie GenBank, a także bazę abstraktów prac naukowych PubMed [6].

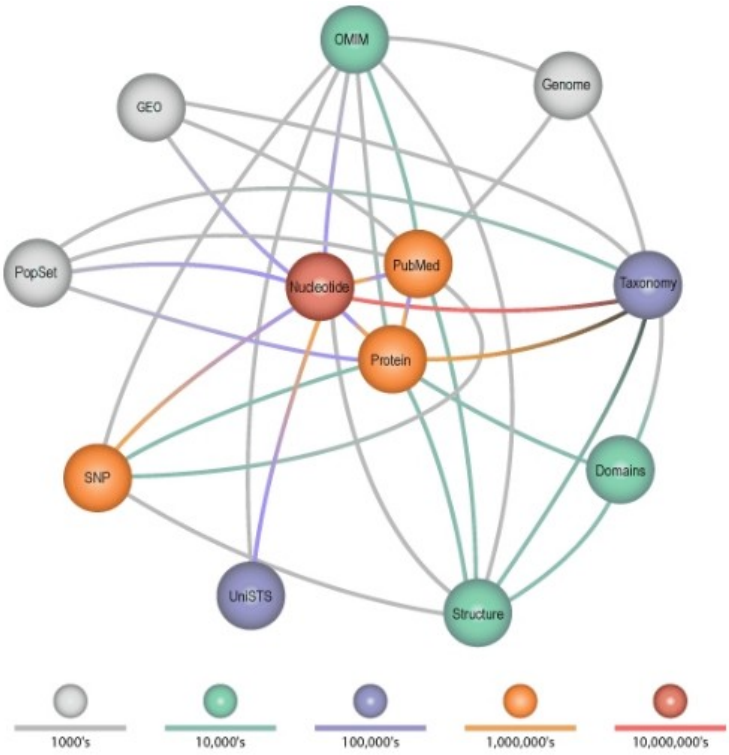
Działanie systemu Entrez opiera się na połączeniach pomiędzy węzłami, które odpowiadają konkretnym bazom danych. Na rysunku 2 została przedstawiona struktura i połączenia węzłów systemu Entrez.



Rys. 1. Przykładowy wynik globalnego zapytania wysłanego do systemu Entrez
Fig. 1. Result of sample, global query submitted to Entrez system

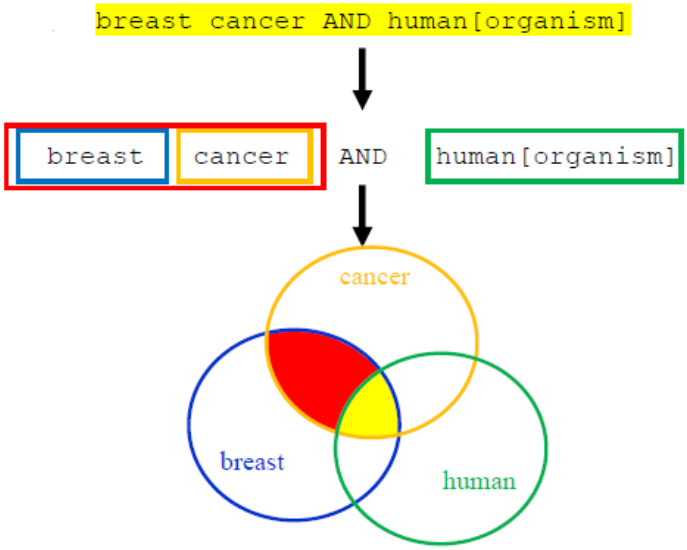
System cały czas jest rozwijany i udoskonalany. Liczba węzłów obsługiwanych przez Entrez cały czas rośnie. Oryginalny trójwęzłowy system, tj. GenBank (Nucleotide), PDB (Protein) i PubMed, wyewoluował w przeciągu ostatnich lat, dodając węzły, takie jak [7]:

- Taxonomy, zorganizowany wokół nazw i powiązań filogenetycznych pomiędzy organizmami;
- Structure, zorganizowany wokół struktur trójwymiarowych białek i kwasów nukleinowych;
- Genomes, gdzie każdy rekord reprezentuje chromosom danego organizmu;
- PopSet, który jest kolekcją sekwencji z pojedynczego studium populacyjnego;
- OMIM, który jest bazą wszystkich znanych chorób o podłożu genetycznym;
- SNP, zorganizowany wokół zjawiska poliformizmu pojedynczego nukleotydu;
- inne.



Rys. 2. Struktura powiązań pomiędzy węzłami systemu Entrez
Fig. 2. Relationships between database nodes in Entrez

3. Budowa i składnia zapytań



Rys. 3. Graficzna reprezentacja wyszukiwania informacji przez system Entrez
Fig. 3. Graphical representation of information searching in Entrez

Przeszukując bazy danych NCBI, użytkownicy wprowadzają zwykle do wyszukiwarki NCBI Entrez słowa kluczowe lub frazy, które mają zostać wyszukane. Ciągi znaków wprowadzanych do systemu Entrez są konwertowane na zapytania o następującym formacie:

```
Term1[field1]OpTerm2[field2]OpTerm3[field3]Op...
```

gdzie: *Term*, określa frazę zapytania, *field* jest polem wyszukiwania (zapisywany zawsze w kwadratowych nawiasach), natomiast *Op* jest jednym z dostępnych operatorów logicznych: AND, OR lub NOT (operatory muszą być zapisane dużymi literami).

Przykład: `human[organism] AND topoisomerase[protein name]`

System Entrez dzieli wpisane zapytanie na serie elementów, które w oryginalnym zapytaniu były rozdzielone spacją. Jeśli w zapytaniu znajdują się operatory logiczne, to system podzieli zapytanie na serie, najpierw względem operatorów logicznych, a następnie względem spacji. Każdy element zapytania jest przetwarzany osobno, a wyniki wyszukiwania są następnie łączone, zgodnie z operatorami użytymi w zapytaniu (rysunek 3). Domyślnym operatorem logicznym jest operator AND.

W przypadku gdy zapytanie składa się tylko z listy UID[2] lub numerów dostępu, system Entrez zwróci tylko te rekordy, do których podane identyfikatory się odwołują. Nie zachodzi wtedy żadne dodatkowe przetwarzanie zapytania.

4. Automatyczne mapowanie fraz w zapytaniu do systemu Entrez

System Entrez podczas przetwarzania zapytania automatycznie przeszukuje bazę danych dla każdej z frazy zapytania względem kryteriów:

1. Węzeł taksonomiczny – każda z fraz jest limitowana do pola `[organism]` lub `[All Fields]`. Na przykład dla frazy `mouse`, system automatycznie mapuje frazę na: `„mus musculus”[organism] OR mouse[All Fields].`
2. Nazwa czasopism – baza danych jest przeszukiwana względem nazw czasopism, np.: `science` ⟶ `science[Journal].`
3. Nazwa autora – rezultat wykonania zapytania jest zawężany do pola `[Author]`. Nie każda fraza może zostać mapowana na pole nazwy autora, ponieważ za prawidłową nazwę autora uważa się tylko słowo, po którym występuje jedna lub dwie litery. Na przykład: `Siaznik A` ⟶ `„Siaznik A”[Author]`

Jeśli system nie zwrócił wyników po procesie automatycznego mapowania, wtedy to usuwana jest fraza, która jest wysunięta najbardziej na prawo w zapytaniu i proces mapowania zostaje powtórzony dopóki, dopóty system zwróci wyniki. Jeśli mimo to system dalej nie zwraca wyników, to wszystkie frazy zapytania są limitowane do pola `All Fields`, oraz są połączone operatorem logicznym AND.

W tabeli 1 przedstawiono przykłady zapytań przed i po procesie automatycznego mapowania [7].

[2] UID – (ang. *Unique Identifier*) nazwa unikalnego identyfikatora rekordu w bazach danych NCBI.

Tabela 1

Przykłady automatycznego mapowania fraz w zapytaniu do systemu Entrez

Zapytanie oryginalne	**Zapytanie po procesie automatycznego mapowania**
`cancer cell receptor`	`("Cancer Cell"[Journal] OR ("cancer"[All Fields] AND "cell"[All Fields]) OR "cancer cell"[All Fields]) AND receptor[All Fields]`
`cell receptor cancer`	`cell[All Fields] AND receptor[All Fields] AND ("Cancer"[Organism] OR cancer[All Fields])`
`mouse p53`	`("Mus musculus"[Organism] OR mouse[All Fields]) AND p53[All Fields]`
`wheat nuclear protein`	`("Triticum aestivum"[Organism] OR wheat[All Fields]) AND nuclear protein[All Fields]`
`wheat w nuclear protein`	`wheat w[Author] AND ("nuclear proteins"[MeSH Terms] OR ("nuclear"[All Fields] AND "proteins"[All Fields]) OR "nuclear proteins"[All Fields] OR ("nuclear"[All Fields] AND "protein"[All Fields]) OR "nuclear protein"[All Fields])`

5. Narzędzia Entrez Programming Utilities

Tabela 2

Dostępne narzędzia eUtils

Nazwa narzędzia	**Opis**
Finfo	Uzyskiwanie informacji na temat dostępnych baz danych lub konkretnej bazy danych, takie jak: liczba rekordów zaindeksowanych dla każdego pola wyszukiwania, data ostatniej aktualizacji, dostępne powiązania z innymi bazami danych.
EGQuery	Odpowiada na zapytanie, zwracając liczbę rekordów pasujących podanym frazom wyszukiwania w każdej bazie obsługiwanej przez system Entrez.
ESearch	Odpowiada na zapytanie, zwracając listę unikalnych identyfikatorów (UID) rekordów, które pasują do zadanego zapytania.
ESummary	Zwraca dla podanej listy UID podsumowania rekordów z konkretnej bazy danych.
EPost	Akceptuje listę UID i wysyła ją na serwer historii, zwracając odpowiedni adres w postaci parametrów: WebEnv i query_key.
EFetch	Odpowiada na listę UID, zwracając kompletne rekordy z danej bazy danych.
ELink	Zwraca listę UID rekordów z podanej bazy danych powiązanych z oryginalną listą UID rekordów z bazy wejściowej.
ESpell	Zwraca sugestie poprawnej pisowni wprowadzonego przez użytkownika zapytania.

Pod nazwą Entrez Programming Utilities [8, 9] kryje się zestaw ośmiu programów, działających po stronie serwera, które świadczą stabilny interfejs systemu Entrez w NCBI.

Z eUtils, bo tak skrótowo określa się ów zestaw narzędzi, można skorzystać w dwojaki sposób. Pierwszym sposobem jest przesłanie przez aplikację odpowiednio złożonego adresu URL do serwera, na którym znajdują się narzędzia, a następnie odebranie odpowiedzi danego narzędzia w formacie XML. Drugim, bardziej dystyngowanym, sposobem jest wykorzystanie w tym celu protokołu SOAP [10]. NCBI udostępnia na swoich stronach internetowych odnośniki do plików WSDL z opisem usług sieciowych, które stanowią narzędzia eUtils.

W tabeli 2 został przedstawiony spis wszystkich narzędzi eUtils wraz z krótkim opisem ich funkcji [9].

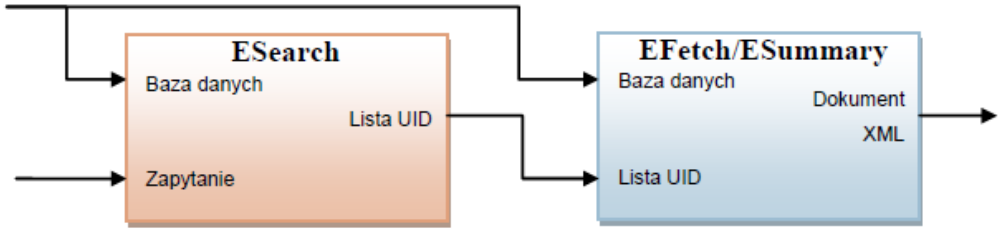
6. Łączenie narzędzi Entrez Programming Utilities

W celu zbudowania aplikacji zdolnej do przeszukiwania i pobierania informacji z baz danych Entrez, ważne jest umiejętnie połączenie dostępnych narzędzi. Podstawowym połączeniem programów, zawartych w pakiecie eUtils, jest:

ESearch ⟶ EFetch/ESummary

Połączenie to pozwala na przeszukanie bazy danych i pobranie odpowiednich rekordów, które spełniają warunki zawarte w podanym zapytaniu.

Program ESearch jest odpowiedzialny za wygenerowanie listy identyfikatorów rekordów w podanej bazie danych, zgodnych z wpisanym zapytaniem, natomiast program ESummary lub EFetch pobiera z bazy danych rekordy o identyfikatorach zgodnych z podaną listą UID (rysunek 4).



Rys. 4. Schemat blokowy, obrazujący działanie połączenia narzędzia ESearch z EFetch/ESummary
Fig. 4. Diagram showing colaborative functioning of ESearch tool and EFetch/ESummary tool

Jeśli zadaniem aplikacji jest wyszukanie rekordów w pewnej bazie danych, a następnie znalezienie rekordów powiązanych z nimi w innej bazie danych, to narzędziami, jakie muszą zostać użyte do wykonania tego zadania, są:

ESearch ⟶ ELink ⟶ EFetch/ESummary

W powyższym połączeniu ELink jest odpowiedzialny za wygenerowanie listy UID, która odpowiada rekordom, konkretnej bazy danych, powiązanym z rekordami bazy danych, dla której oryginalnie przeprowadzono przeszukiwania.

Jeżeli interesuje nas szersze spektrum powiązań, dla przykładu naszym zadaniem jest znalezienie sekwencji aminokwasowych powiązanych z genami, które są w pewien sposób połączone z sekwencjami nukleotydowymi ze zbioru populacyjnego sekwencji, należących do myszy, wtedy to liczba wywołań programu ELink wynosi 3:

ESearch ⟶ ELink ⟶ ELink ⟶ ELink ⟶ EFetch

Dla takiego przypadku ESearch zwraca nam liste UID wszystkich zbiorów populacyjnych sekwencji myszy z bazy PopSet, pierwsze wywołanie ELink wyszukuje nam listę UID sekwencji nukleotydowych w bazie Nucleotide, które zawierały się w znalezionych zbiorach, drugie wywołanie ELink generuje listę genów z bazy Gene, powiązanych z sekwencjami nukleotydowymi, a wynikiem ostatniego wywołania programu ELink jest lista UID białek z bazy danych Protein, które są powiązane z wcześniej uzyskaną listą genów.

7. Aplikacja search GenBank

search GenBank jest programem napisanym w formie portalu internetowego, który pozwala użytkownikom na przeszukiwanie i pobieranie informacji z bazy danych GenBank. Aplikacja, oprócz obsługi bazy sekwencji nukleotydowych, pozwala także użytkownikom na przeszukiwanie innych baz danych.

Program został przetestowany dla następujących baz danych (wytłuszczone nazwy baz danych wskazują na bazy danych, dla których przygotowano formę reprezentacyjną rekordów):

- **Nucleotide** – główna baza sekwencji nukleotydowych (GenBank),
- **dbEST** – baza sekwencji EST[3],
- **dbGSS** – baza sekwencji GSS[4],
- Genome – baza genomów różnych organizmów,
- **PopSet** – baza sekwencji z pojedynczego studium populacyjnego,
- Taxonomy – baza taksonomii,
- Gene – baza znanych genów,
- OMIM – baza wszystkich znanych chorób o podłożu genetycznym,
- SNP – baza poliformizmów pojedynczego nukleotydu,
- PubMed – baza abstraktów publikacji naukowych,
- PMC – baza dostępnych publikacji naukowych,

[3] Expressed Sequence Tag – krótki odcinek sekwnecji z sekwencji cDNA (mRNA). Może być użyty jako identyfikator transkryptów genów.

[4] Genome survey sequence – sekwencje podobne do sekwencji EST, z wyjątkiem że większość z nich jest sekwencjami genomowymi, a nie uzyskiwanymi z mRNA/cDNA.

- Journals – baza czasopism naukowych,
- **Protein** – baza sekwencji aminokwasowych.

Rys. 5. Strona główna aplikacji *search GenBank*
Fig. 5. Main page of *search GenBank*

Aplikacja pozwala na przeszukiwanie zasobów, wymienionych baz danych w taki sam sposób, jak jest to rozwiązane na stronach NCBI[5]. Oprócz standardowej metody wyszukiwania, w której użytkownik wpisuje zapytanie ręcznie w pole wyszukiwania, udostępniono także moduł wyszukiwania zaawansowanego, który pozwala na zdefiniowanie odpowiednich limitów ograniczających wyniki zapytania.

[5] http://www.ncbi.nlm.nih.gov/guide/

Portal internetowy *search GenBank* został także wyposażony w moduł budowania makr, które służą do zautomatyzowania wyszukiwania rekordów w innych bazach danych, powiązanych z rekordami, które są wynikiem zapytania oryginalnego. Ów moduł jest nowością i nie znaleziono żadnych odpowiedników, które spełniałyby takie same funkcje.

Interfejs aplikacji został zaprojektowany z wykorzystaniem najnowszych trendów w świecie szablonów stron internetowych; został on skomponowany, mając na uwadze wymagania nowoczesnych użytkowników serwisów internetowych i jest zgodny z przyjętymi zasadami przejrzystości prezentowania informacji na stronach WWW.

Program pozwala zalogowanym użytkownikom na skorzystanie z prostego systemu zapisywania wprowadzonych zapytań i zbudowanych makr. Udogodnienie to wprowadza do serwisu możliwość ponownego wykorzystania zapisanych elementów, bez przypominania sobie konfiguracji zbudowanego makra czy pisania od nowa skomplikowanego zapytania.

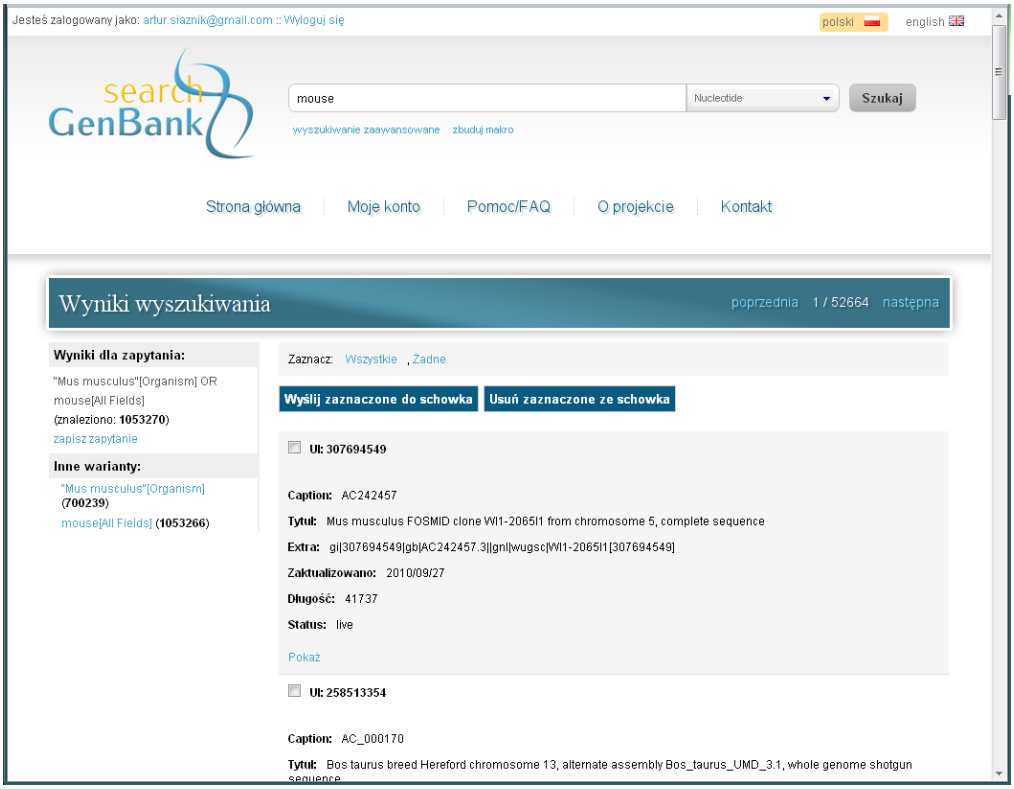
Aplikacja jest dostępna pod adresem: **http://sgb.bioaut.pl**

7.1. Proste wyszukiwanie danych

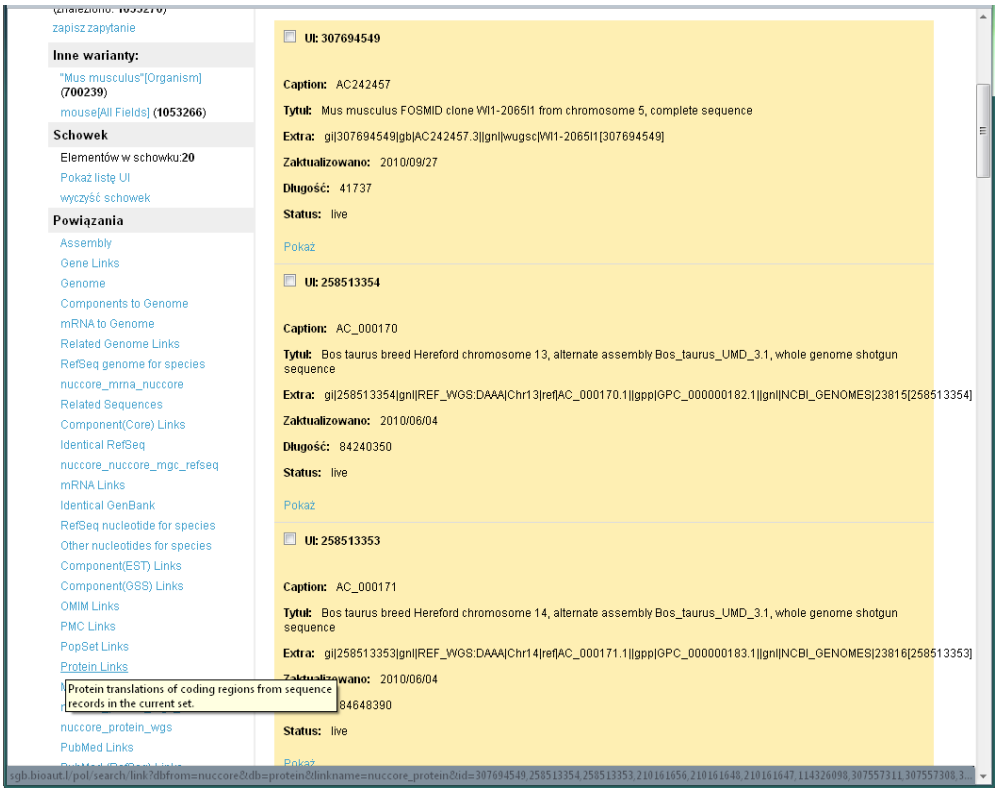
Proste wyszukiwanie jest udostępnione na każdej ze stron portalu eksploracji *search GenBank*. W górnej części każdej strony znajduje się pole, do którego można wprowadzić pojedyncze słowo lub frazę, która ma zostać wyszukana. Dodatkowo, z listy rozwijanej po prawej stronie należy wybrać bazę danych, którą należy przeszukać. Wyniki wyszukiwania danych genetycznych w bazie danych *Nucleotide* dla przykładowej frazy *mouse* zostały przedstawione na rysunku 6.

Warto zwrócić uwagę, do jakiej formy została przekształcona zadana fraza *mouse*. Po lewej stronie strony wyników mogą pojawić się bloki:

- *Wyniki dla zapytania* – pokazuje zapytanie wraz z liczbą znalezionych rekordów, daje również możliwość zapisania zapytania (*zapisz zapytanie*);
- *Inne warianty* – pokazuje listę sugerowanych, alternatywnych zapytań wraz z oczekiwaną liczbą wyników;
- *Czy chodziło Ci o?* – wskazuje na ewentualne błędy w pisowni we wprowadzonym zapytaniu.



Rys. 6. Wyniki wyszukiwania w systemie *search GenBank*
Fig. 6. Results of sample query in *search GenBank*



Rys. 7. Zawartość schowka i powiązania z innymi bazami danych dostępne przez system *search GenBank*
Fig. 7. Content of clipboard and links to other data sources available through *search GenBank*

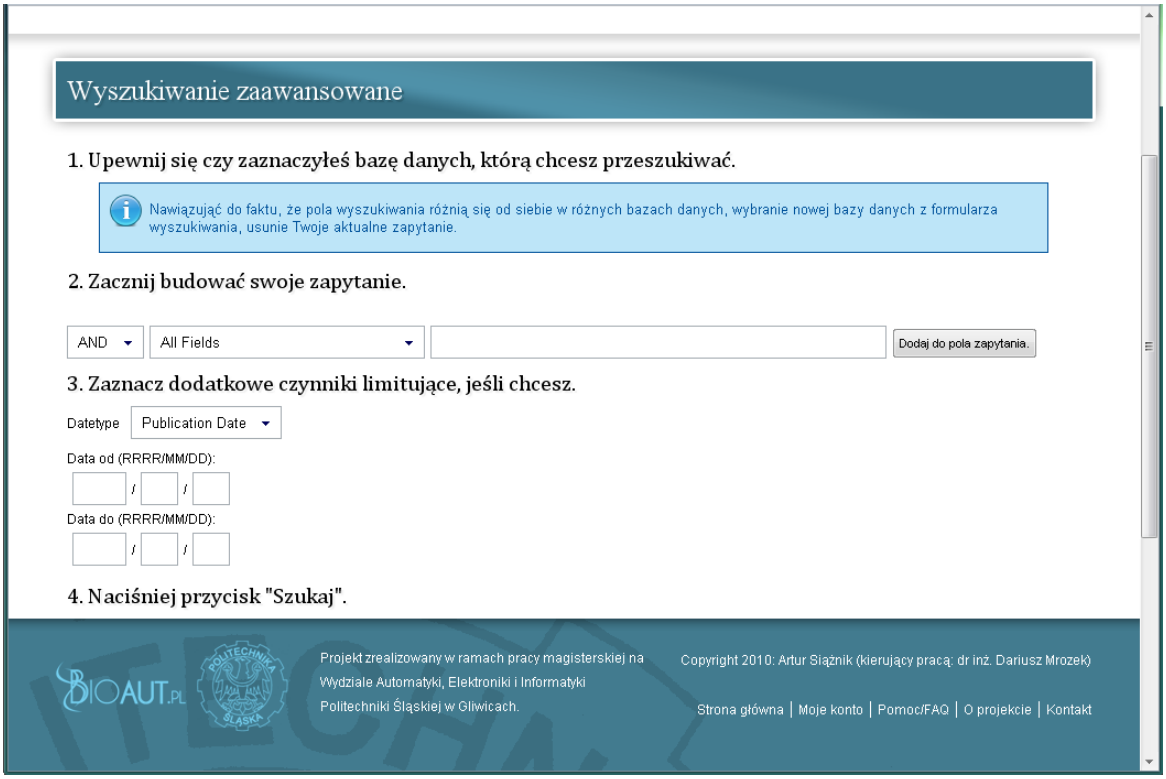
Rekordy, które zostały zwrócone w wyniku wykonania zapytania, można przeglądać oraz dodawać do podręcznego schowka. Podczas przeglądania pojedynczego rekordu lub przeglądania zawartości schowka jest również możliwe powiązanie danego rekordu z danymi innej bazy. Na przykład, szukając danych genetycznych w bazie *Nucleotide*, można je powiązać z sekwencjami białkowymi bazy *Proteins* lub danymi bibliograficznymi bazy *PubMed*. Odbywa się to poprzez szereg powiązań, które są udostępniane użytkownikowi w oknie aplikacji po lewej stronie (rysunek 7, sekcja *Powiązania*).

7.2. Zaawansowane wyszukiwanie danych

Wyszukiwanie zaawansowane w portalu *search GenBank* pozwala na dokładne złożenie zapytania, wykorzystując do tego odpowiednie dla wybranej bazy danych pola wyszukiwania oraz dodatkowe czynniki ograniczające. Zapytania buduje się zgodnie z regułami przedstawionymi w rozdziale 3, zwykle łącząc operatorami logicznymi wiele prostych warunków wyszukiwania. Na przykład:

```
"oxygen"[TITL] AND "hemoglobin"[GENE] AND "Arabidopsis thaliana"[ORGN]
```

W portalu *search GenBank* udostępniono odpowiedni kreator zapytań dla budowania zapytań złożonych przedstawiony na rysunku 8.



Rys. 8. Strona wyszukiwania zaawansowanego w systemie *search GenBank*
Fig. 8. Advanced search web page in *search GenBank* system

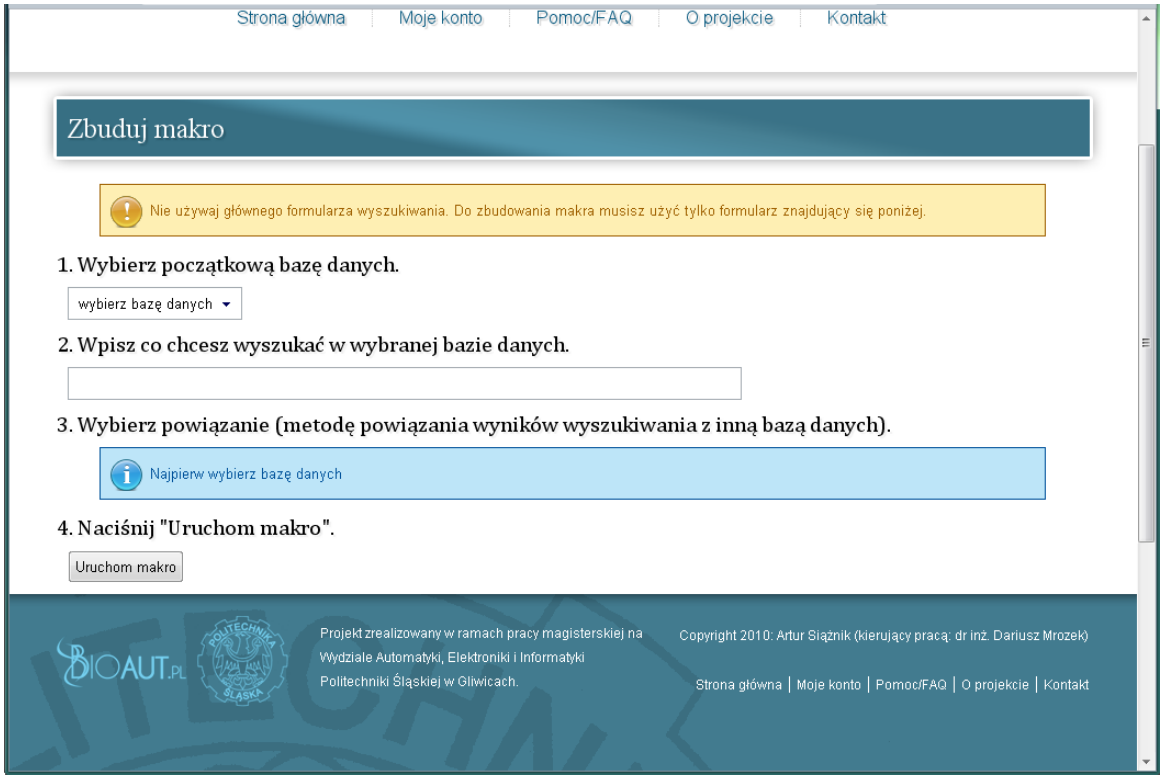
Skrótowo proces wyszukiwania zaawansowanego można opisać w punktach:

1. Wybierz bazę danych z głównego formularza zapytania.
2. Zbuduj swoje zapytanie:
   a) wybierz operator logiczny łączący frazy zapytania,
   b) wybierz pole wyszukiwania,
   c) wpisz frazę wyszukiwania,
   d) naciśnij przycisk Dodaj do pola zapytania,
   e) powtórz punkty od a) do d), jeżeli jest taka potrzeba.
3. Naciśnij przycisk *Szukaj*, który znajduje się obok pola wyboru bazy danych na górze strony WWW.

Moduł wyszukiwania zaawansowanego pozwala także na wprowadzenie dodatkowych czynników ograniczających zakres wyników wyszukiwania. Wspomnianymi czynnikami jest zakres daty publikacji lub daty modyfikacji rekordów w określonej bazie danych. Aby skorzystać z tej funkcji, należy wypełnić odpowiednie pola na stronie z formularzem wyszukiwania.

Prezentacja wyników zbudowanego zapytania jest taka sama jak prezentacja wyników w przypadku wyszukiwania podstawowego. Opcje schowka i powiązań są dostępne i działają tak samo dla wyników zapytań zbudowanych przez formularz wyszukiwania zaawansowanego.

7.3. Tworzenie makr



Rys. 9. Budowanie makr na stronie *search GenBank*
Fig. 9. Construction of macros on *search GenBank*

Makra pozwalają na automatyzację procesu wyszukiwania rekordów, w jakiś sposób powiązanych z rekordami w innej lub w tej samej bazie danych. Do zbudowania makra jest niezbędne określenie bazy danych, w której wyszuka się rekordów pasujących do wpisanego zapytania. Następnie użytkownik wybiera z listy dodatkowe powiązania z innymi bazami danych. Teoretycznie liczba powiązań wprowadzonych do makra jest nieskończona, jednakże należy mieć na uwadze fakt, iż nie wszystkie rekordy w bazach danych NCBI posiadają adnotacje, wskazujące na powiązane elementy w innych bazach. Z biegiem czasu liczba powiązań między bazami danych NCBI wzrasta, dlatego też można być dobrej myśli, iż makra okażą się w przyszłości świetną alternatywą dla żmudnego procesu eksploracji danych pomiędzy bazami. Budowanie makr odbywa się poprzez odpowiedni formularz dostępny w serwisie *search GenBank*, przedstawiony na rysunku 9.

Po skonstruowaniu makra można je wykonać lub zapisać do słownika makr, o ile jest się zalogowanym użytkownikiem. W tabeli 3 przedstawiono kilka przykładów makr.

Tabela 3

Przykłady makr

<table>
<tr><td colspan="3">Problem: Znajdź wszystkie dostępne w bazie danych rekordy genów dla rekordów bazy sekwencji aminokwasowych, odpowiadające białku o nazwie: topoisomerase</td></tr>
<tr><td>Zapytanie:
topoisomerase[protein name]

Baza:
Protein</td><td colspan="2">Powiązanie:
Gene Links

Znaleziono: 19</td></tr>
<tr><td colspan="3">Problem: Znajdź sekwencje nukleotydowe dla myszy, a następnie wszystkie dostępne dla nich artykuły z bazy PubMed</td></tr>
<tr><td>Zapytanie:
mouse

Baza:
Nucleotide</td><td colspan="2">Powiązanie:
Pubmed Links

Znaleziono: 264</td></tr>
<tr><td colspan="3">Problem: Znajdź wszystkie możliwe rekordy z bazy PopSet, odpowiadające zapytaniu o raka piersi, następnie wyszukaj powiązane z nimi sekwencje nukleotydowe. Powiąż znalezione sekwencje nukleotydowe z sekwencjami białek</td></tr>
<tr><td>Zapytanie:
Breast cancer

Baza:
PopSet</td><td>Powiązanie:
Nucleotide Links</td><td>Powiązanie:
Protein Links

Znaleziono: 881</td></tr>
</table>

8. Podsumowanie

Portal internetowy *search GenBank* daje duże możliwości prostego i zaawansowanego przeszukiwania bazy GenBank oraz innych baz danych utrzymywanych w Stanach Zjedno-

czonych przez National Center for Biotechnology Information. Ponadto, możliwość tworzenia makr pozwala na międzybazową eksplorację powiązanych ze sobą danych. Jest to cecha unikalna systemu *search GenBank*. Siły tego rozwiązania nie można aktualnie wykorzystać w pełni ze względu na fakt, iż powiązania pomiędzy rekordami różnych baz nie są obecnie tak bardzo rozbudowane. Jednakże potencjał idei, jaki drzemie właśnie w rozwiązaniach automatyzujących proces wyszukiwania informacji w bioinformatycznych bazach danych, może być w niedalekiej przyszłości całkowicie wykorzystany.

Portal internetowy *search GenBank* został zaprojektowany dla osób zajmujących się analizą danych biologicznych, m.in. biochemików, biologów molekularnych, lekarzy medycyny, pracowników laboratoriów genetycznych, patologów molekularnych. Zarejestrowani i zalogowani użytkownicy systemu mogą zapisywać raz już skonstruowane zapytania i makra w specjalnych słownikach po to, by w przyszłości, prowadząc podobne badania, móc do nich powrócić.

Portal internetowy *search GenBank* koncentruje się wprawdzie w dużej mierze na danych genetycznych, wychodząc z założenia, że dane genetyczne są obecnie najczęściej wykorzystywanymi danymi w naukach o życiu (ang. *life sciences*), jednakże umożliwia również przeszukiwania i przeglądania innych baz danych NCBI.

BIBLIOGRAFIA

1. Mrozek D.: Bioinformatyczne bazy danych – rola, miejsce i klasyfikacja. [w] Bazy danych: Struktury, Algorytmy, Metody. Wydawnictwa Komunikacji i Łączności, Warszawa 2006, s. 117÷128.
2. Mrozek D., Małysiak B.: Bioinformatyczne bazy danych – poziomy opisu funkcjonowania organizmów. [w] Bazy danych: Struktury, Algorytmy, Metody. WKŁ, Warszawa 2006, s. 107÷116.
3. Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Wheeler D. L.: GenBank: update. Nucleic Acids Res., Vol. 32, 2004, s. 23÷26.
4. Hogue C., Ohkawa H., Bryant S.: A dynamic look at structures: WWW-Entrez and the Molecular Modelling Database. Trends Biochem. Sci. 21, 1996, s. 226÷229.
5. Wheeler D. L., Chappey C., Lash A. E., Leipe D. D., et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 28(1), 2000, s. 10÷14.
6. McEntyre J., Lipman D.: PubMed: bridging the information gap. CMAJ. 164(9), 2001, s. 1317÷1319.
7. Ostell J.: The Entrez Search and Retrievlal System. http://www.ncbi.nlm.nih.gov/ /bookshelf/ br.fcgi?book=handbook&part=ch15 [stan na 02.02.2011].

8. Sayer E., Wheeler D.: Building Customized Data Pipelines Using The Entrez Programming Utilities (eUtils). http://www.ncbi.nlm.nih.gov/bookshelf/ /br.fcgi?book=coursework&part= eutils [dostęp 02.02.2011].
9. Sayers E.: The E-Utilities In-Depth: Parameters, Syntax and More. http://www.ncbi-.nlm.nih.gov/bookshelf/br.fcgi?book=helpeutils&part=chapter4 [dostęp 02.02.2011].
10. SOAP Version 1.2 Part 1: Messaging Framework (Second Edition), http://www.w3-.org/TR/soap12-part1/ [dostęp 02.02.2011].

Recenzenci: Prof. dr hab. inż. Mieczysław Muraszkiewicz
Dr Ewa Romuk

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Since the rapid development of computer science, which we have seen over the past years, many other fields of science can now use the solutions and technologies, which this development is an undeniable cause.

Medicine and natural sciences began to be branches of science, whose research generates a huge amount of data. This unimaginably large amount of data had to be processed and stored in some way. The task involved specialists from the IT and bioinformatics who, having the appropriate database technologies and a good theoretical basis associated with the natural sciences, could solve the problem of storing and processing large amounts of biological information.

Due to the constantly growing collection of information, solutions and techniques that focus on storing specific medical data in databases have to be more fine-tuned to the purpose. One of the main reasons for establishing biological databases is a very large amount of genetic information, including nucleotide sequences, for which there is no better storage method than database systems. Practically since 1981, when the Sanger method of sequencing was invented, the problem of storing and processing genetic information is still up-to-date.

GenBank is one of the world's most famous database storing tens of millions of nucleotide sequences of DNA and RNA. In this article, we present a new system designed to explore genetic data in the GenBank database. The *search GenBank* system not only allows to search and browse biological data in the GenBank, but also combine the GenBank database

entries with items in other NCBI databases. Therefore, the *search GenBank* provides the cross-database exploration possibilities.

Adresy

Artur SIĄŻNIK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, artur.siaznik@gmail.com.
Bożena MAŁYSIAK-MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, bozena.malysiak@polsl.pl.
Dariusz MROZEK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, dariusz.mrozek@polsl.pl.