Faktograficzne bazy danych dla biotechnologii*

Maciej Nawrot
Maciej Astriab
Tomasz Twardowski
Instytut Chemii Bioorganicznej
Polska Akademia Nauk
Poznań

1. Wstęp

Bioinformatyka jest jednym z głównych działów współczesnej biologii. Wielodyscyplinarny aspekt prac badawczych powoduje zróżnicowane wymagania w stosunku do bioinformatyki. Uogólniając, można stwierdzić, że jej celem i zadaniem winno być uporządkowanie i transfer informacji, tak aby naukowcy mogli interpretować dane, a nie — jak dotychczas — zajmować się przekazem i kolekcjonowaniem pojedynczych wyników. Warunkiem *sine qua non* takiego podziału pracy jest nie tylko powszechność technik komputerowych, ale także ich jednolitość i prostota obsługi.

W Europie podstawowe znaczenie dla informacji w biotechnologii ma European Molecular Biology Network (EMBNet) prowadzona przez European Molecular Biology Laboratory (EMBL) w Heidelbergu (RFN). Działalność finansowana jest w ramach European Molecular Biology Organization (EMBO), której członkami są: Belgia, Finlandia, Dania, Francja, Grecja, Hiszpania, Holandia, Irlandia, Izrael, Norwegia, Szwajcaria, Szwecja, Wielka Brytania, Włochy. Korzystanie z usług informacyjnych jest bezpłatne dla pracowników naukowych (także dla uczonych z innych państw). Oprócz ośrodka w Heidelbergu, który udostępnia szereg banków danych (tab. 1), istotne znaczenie odgrywają również:

— Human Genome Mapping Centre (HGMC, Wielka Brytania);

— International Centre for Genetic Engineering & Biotechnology (ICGEB, Włochy);

— European Center for Bioinformatics (ECBi, Wielka Brytania).

---

* Ekspertyza wykonana dla Komitetu CODATA pt. Stan i potrzeby w zakresie baz danych dla nauki i techniki, ich lokalizacja i rozpowszechnianie. Dział: Biotechnologia, listopad 1995 r. Przedruk za zgodą Krajowego Komitetu CODATA.

Biblioteka banków danych EMBL w Heidelbergu w ścisłej współpracy z GENBANK (Los Alamos, USA) i DDBJ (DNA Database, Japan) kolekcjonuje i udostępnia na cały świat dane dotyczące sekwencji kwasów nukleinowych i białek. EMBL oferuje wiele serwisów komputerowych do użytku dla środowisk naukowych i w ten sposób umożliwia natychmiastowy dostęp do tych tak bardzo ważnych danych.

Dział zbiorów danych EMBL wprowadził w pełni zautomatyzowany serwis dla użytku zewnętrznego. Jest on oparty na systemie komputerowym EMBL i umożliwia dogodne przeszukiwanie baz danych oraz uzyskanie odpowiednich informacji przy zastosowaniu poczty elektronicznej. Każdy użytkownik mający dostęp do międzynarodowej sieci komputerowej takiej jak Bitnet/EARN lub Internet może korzystać ze zbiorów, łącznie z aktualnymi bazami danych EMBL oraz GENBANK. Serwis ten stał się bardzo popularny w środowisku naukowym.

Biblioteka banków danych EMBL finansowana jest przez Unię Europejską (UE), pracuje w skali ogólnoświatowej i udziela miesięcznie około 1000 informacji (bezpłatnie). Oprócz baz danych EMBL, GENBANK, SwissProt i Brookhaven wprowadzone są do obsługi serwisowej inne bardzo ważne bazy danych: Prosite, ENZYME, enzymów restrykcyjnych REBASE, *E. coli* (ECD). Biblioteka banków danych EMBL prowadzi bezpłatną dystrybucję programów komputerowych dotyczących biologii molekularnej. Obecnie dostępne są programy komputerowe w systemach MS-DOS, Apple Macintosh, VAX/VMS i UNIX.

Przy badaniach naukowych i pracach aplikacyjnych w zakresie nowoczesnych biotechnologii, a w szczególności inżynierii genetycznej, najczęstszym wykorzystaniem baz danych dotyczących sekwencji jest porównywanie nowych struktur pierwszorzędowych oznaczanych genów lub białek z całą bazą w celu wykrycia podobieństwa lub różnicy w sekwencji. W związku z tym w bibliotece EMBL uruchomiono dwa dodatkowe serwisy, a mianowicie: Mail-Quicksearch oraz Mail-Fast A, które umożliwiają zdalne wyszukiwanie baz w systemach komputerowych EMBL.

Szczególnie cenne dla biotechnologii są informacje patentowe. Istnieje kilka banków danych o patentach biotechnologicznych. Polski Urząd Patentowy dysponuje taką bazą w systemie CD-ROM w wyniku kooperacji z EPO (European Patent Office, Monachium). Podobnie istotne dla ogólnego rozwoju są bazy danych określane potocznie jako „probiotechnologiczne", a dotyczące m.in. ochrony środowiska, regulacji prawnych, toksykologii, właściwości fizykochemicznych cząsteczek, katalogi mikroorganizmów. Bazy danych dostępne są w dwóch systemach: CD-ROM i on-line.

2. Bazy danych

TABELA 1
BANKI DANYCH SZCZEGÓLNIE WAŻNE DLA PROJEKTU „GENOM CZŁOWIEKA"

Nazwa i zakres tematyczny	Lokalizacja	Sponsor
Protein Data Bank PDB	Brookhaven National Laboratory, Upton, USA	Department of Energy, USA
Homology Database	Jackson Laboratory, Maine, USA	National Institutes of Health
Protein Identification	Georgetown University, Washington DC, USA	National Institutes of Health
GenBank	Los Alamos National Laboratory, New Mexico, USA	National Institutes of Health
Genome Database	John Hopkins University, Baltimore, USA	Department of Energy, USA
Nucleotide Sequence Data Library	EMBL, Heidelberg, Niemcy	UE
DNA Database of Japan	National Institute of Genetics, Mishima, Japonia	Agencja ds. Nauki i Technologii, Japonia

TABELA 2
ZAKRES TEMATYCZNY BANKU DANYCH EMBL W HEIDELBERGU

Bank	Zakres tematyczny
ECD	*E. coli*
Enzyme	enzymy
EPD	parametry eukariotyczne
DOC	dokumentacja ogólna
LIMB	lista banków danych
NUC	indeks
PROSITE	białka
PROTEIN	białka
PROTEINDATA	dane strukturalne
REBASE	enzymy restrykcyjne
REFLIST	referencje sekwencji
SOFTWARE	oprogramowanie

Bank danych EMBL w Heidelbergu dostepny jest w sieci Internet w serwisie WWW pod adresem: http://www.ebi.ac.uk/

3. Podstawowe bazy danych dostępne na nośniku CD-ROM

nazwa: INŻYNIERIA GENETYCZNA
słowa kluczowe: inżynieria genetyczna, patenty
język: angielski
nośnik\oprogramowanie: CD-ROM/CD Answer Retrieval Software
tematyka: Opisy patentowe USA dotyczące inżynierii genetycznej
aktualność: 1928 - 1992, nie aktualizuje się
właściciel: Urząd Patentowy USA
w posiadaniu na terenie Polski: Urząd Patentowy RP
cena: wg cennika Urzędu Patentowego

nazwa: BIOSIS GenRef on CD
słowa kluczowe: gen sequences, biotechnology, molecular biology
język: angielski
nośnik\oprogramowanie: CD-ROM\własne
tematyka: zawiera sekwencje genów różnych organizmów, informacje dla ośrodków władzy urzędów patentowych, zawiera informacje dotyczące genomów ludzi, roślin, zwierząt, mikroorganizmów
aktualność: 50 000 sekwencji do roku 1995
cena: 5500 USD do roku 1995, 2090 USD rok 1995

nazwa: AGRISEARCH
słowa kluczowe: agrobiotechnology food technology, biotechnology
język: angielski
nośnik\oprogramowanie: CD-ROM\Spirs
tematyka: baza przeznaczona dla przemysłu spożywczego i rolnego; szczególnie przydatna w biotechnologii żywności i agrobiotechnologii,
aktualność: od 1969 r. (dane z DR USA)
cena: I subskrypcja 795 USD, odnowienie subskrypcji 795 USD
komentarz: dystrybutor komercyjny na terenie Polski — STRATUS

nazwa: DNASIS/PROSIS
słowa kluczowe: DNA, białka, sekwencje
język: angielski
nośnik\oprogramowanie: CD-ROM\dołączone w pakiecie
tematyka: sekwencje genów i białek
dostępność: do 1994 r. w OIN PAN, obecnie nieznany dysponent
komentarz: dystrybutor komercyjny na terenie Polski — STRATUS

nazwa: AIDSLINE
słowa kluczowe: AIDS, Medical biotechnology, retroviruses
język: angielski
nośnik\oprogramowanie: CD-ROM/SPIRS
tematyka: dane dotyczące AIDS
aktualność: od 1980 r.
dystrybucja: w posiadaniu AM w Bydgoszczy
cena: pracownicy AM — bezpłatnie, inni — wg cennika AM, Bydgoszcz

nazwa: SEC: HEALTHCARE, PHARMACEUTICALS & BIOTECH ON SILVERPLATTER
słowa kluczowe: biotechnologia, przemysł biotechnologiczny
język: angielski
nośnik\oprogramowanie: CD-ROM/SPIRS
tematyka: informacje dotyczące przedsiębiorstw biotechnologicznych
cena: komercyjnie — I subskrypcja 1925 USD, odnowienie subskrypcji 1925 USD; indywidualnie — I subskrypcja 1320 USD, odnowienie 1320 USD
komentarz: dystrybutor komercyjny na terenie Polski — STRATUS

nazwa: NATURE OF GENES
słowa kluczowe: DNA, RNA, informacja genetyczna, kod genetyczny, translacja, białka, chromosomy
język: angielski
nośnik\oprogramowanie: CD-ROM\własne
tematyka: nie jest w pełni bazą danych, służy do multimedialnego przedstawienia cząsteczek kwasów nukleinowych, białek, procesów biochemicznych
cena: zakup jednorazowy: instytucje — 594 USD, indywidualnie — 210 USD
komentarz: wymagany komputer Macintosh dystrybutor komercyjny na terenie Polski — STRATUS

4. Podstawowe bazy danych dostępne on-line przez sieć Internet

nazwa: GEN BANK
adres internet: http://www.ncbi.nlm.nih.gov/Genbank/
słowa kluczowe: geny, genomy, projekty sekwencjonowania genomów
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: wszystkie do tej pory poznane sekwencje nukleotydowe kwasów nukleinowych oraz aminokwasowe białek
aktualność: aktualizowana na bieżąco, np. pocztą elektroniczną

nazwa: PROSITE
adres Internet: http://expasy.hcuge.ch/sprot/prosite.html
słowa kluczowe: sekwencje genów, porównania homologii cDNA
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: porównanie sekwencji genomowych bądź cDNA oraz nie scharakteryzowanych dotąd białek z większością znanych już sekwencji kwasów nukleinowych i białek
aktualność: aktualizowana na bieżąco, pocztą elektroniczną

nazwa: SWISS — PROT
adres internet: http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
słowa kluczowe: sekwencje białkowe
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: sekwencje białkowe
aktualność: luty 1995 r.

nazwa: European Molecular Biology Laboratory (EMBL)
adres internet: http://www.ebi.ac.uk/
słowa kluczowe: DNA, RNA, sekwencje, geny, genomy
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: sekwencje DNA i RNA
aktualność: aktualizowana na bieżąco

nazwa: DNA DATA BANK OF JAPAN (DDJB)
adres internet: http://www.nig.ac.jp/
słowa kluczowe: DNA, geny, sekwencje
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: sekwencje DNA
aktualność: codziennie, wymiana informacji z EMBL i GenBANK

nazwa: RNA WORLD IN JENA
adres internet: http://www.imb-jena.de/RNA.html
słowa kluczowe: RNA, sekwencje, struktury II- i III-rzędowe
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: szeroki zasób sekwencji RNA i struktur wyższego rzędu kwasów nukleinowych

nazwa: AMERICAN TYPE CULTURE CLLECTION (ATTC)
adres internet: gopher://culture.attc.org:70/11
słowa kluczowe: DNA, kultury mikroorganizmów, wirusy, linie komórkowe ludzi i zwierząt
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: przechowywanie i rozpowszechnianie kultur mikroorganizmów; wirusów; sondy DNA, ludzkie, zwierzęce i roślinne; kultury komórkowe
aktualność: aktualizowana na bieżąco

nazwa: PROTEIN DATA BANK (PDB)
adres internet: http://www.pdb.bnl.gov/

słowa kluczowe: białka, sekwencje, struktury II-rzędowe, dane z NMR i krystalografii
język: angielski
oprogramowanie: przeglądarki do serwisu WWW
tematyka: baza zawiera dane dotyczące sekwencji białek, DNA i RNA wirusów, zawiera też dane dotyczące struktur II-rzędowych białek, dane z krystalografii i NMR.
aktualność: co 3 tygodnie

Uwaga 1: Wykorzystując połączenia z podanymi bazami danych można uzyskać informacje i adresy dotyczące bardziej specjalistycznych baz.

Uwaga 2: Raz do roku ukazuje się specjalny zeszyt czasopisma „Nucleic Acid Research" poświęcony bazom danych kwasów nukleinowych. Umieszczone w nim artykuły mają na celu prezentację baz danych oraz metod ich analizy.

5. Wnioski

Wybór systemu informacyjnego (1,2) dla składowania, czy też przekazu informacji zdeterminowany jest przez takie czynniki jak:

1) warunki finansowe;

2) stan telekomunikacji i/lub wyposażenia lokalnego;

3) system musi spełniać warunki nowoczesności, także w przyszłości, czyli po 2000 r.; a w szczególności winien gwarantować: bezbłędność zapisu, przekazu i odczytu; użytkownicy prezentują bardzo zróżnicowany poziom przygotowania w zakresie znajomości technik informacji; prostota techniczna i programowa — to podstawowe zalety wybranego systemu.

Obecnie usługi w sieci Internet (poczta elektroniczna, serwisy WWW, ftp, gopher) są dla placówek akademickich pozornie bezpłatne, ponieważ wszelkie koszty pokrywane są przez odpowiednie resorty. Sytuacja ta z pewnością wkrótce ulegnie zmianie. Podobnie, dostęp do podstawowych zasobów informatycznych UE w Heidelbergu jest dla nas obecnie wolny od opłat. Jednakże należy sądzić, że w bliskiej przyszłości korzystanie z usług EMBL będzie związane albo z opłatami subskrypcyjnymi albo z opłaceniem składek członkowskich, np. European Molecular Biology Organization (EMBO). Cena użytkowania Internetu dodatkowo związana jest z kosztami eksploatacji łączy. Tylko w przypadku dysponowania przyłączem światłowodowym transmisja danych jest odpowiednio szybka i bezbłędna.

W przypadku baz danych na nośniku CD-ROM niezbędny jest zakup zarówno bazy jak i czytnika, oraz aktualizacja banku danych corocznie. Niewątpliwie zaletą jest pełna dyspozycyjność i nielimitowany czynnikami zewnętrznymi dostęp do bazy, co jest szczególnie cenne, np. przy szkoleniu i w dydaktyce.

Pierwszą instalację CD-ROM w Polsce uruchomiono w Ośrodku Informacji Naukowej PAN, w kooperacji z Instytutem Chemii Bioorganicznej PAN, w Po-

znaniu w roku 1988; był to bank danych DNASIS & PROSIS, zawierający sekwencje kwasów nukleinowych i białek; jest to bank danych niezmiernie ważny dla biotechnologii. Natomiast w 1992 r. w kraju dysponowaliśmy ponad stu stacjami w systemie CD-ROM z bazami danych dotyczącymi biotechnologii i dziedzin probiotechnologicznych (np. rolnictwo, chemia, farmacja), a w szczególności medycyny i ochrony środowiska (3) (nie przeprowadziliśmy sondażu dla aktualizacji danych zawartych w tym opracowaniu).

Przetłumaczenie dobrych koncepcji naukowych na efektywny (i efektowny) język technologii i przemysłu (z uwzględnieniem prawa) jest zadaniem trudnym, ale gwarantującym przyszłość biotechnologii. Z *Raportu o stanie polskiej biotechnologii* (4) wynika, że jest to technologia XXI w. Można również sądzić, że opóźnienie naszego kraju w tej dziedzinie jest mniejsze aniżeli w innych działach nauki i techniki. Natomiast, obecnie nie istnieje w Polsce nowoczesny przemysł biotechnologiczny. Aby mógł być możliwy rozwój przemysłu w przyszłości już dzisiaj konieczne jest zabezpieczenie infrastruktury. Narzędziem pracy w tym celu są bazy danych; niezbędne dla rozwoju i kontynuacji prac zostały wymienione w tym opracowaniu. Konieczne dla rozwoju polskiej biotechnologii banki danych nie są wzmiankowane w opracowaniu (5).

W sformułowaniu nazwy „biotechnologia" jest ukryte połączenie różnorodnych tematycznie zagadnień w całość, umożliwiającą przetworzenie badań podstawowych w technologie przemysłowe. Patrząc na osiągnięcia biotechnologii jako nowej dziedziny przemysłowej, konkurencyjnych do pewnego stopnia w stosunku do klasycznych dziedzin gospodarki, możemy dostrzec niewątpliwe sukcesy. Zmieniają one nie tyle w sposób ilościowy co jakościowy perspektywy nowych dziedzin naszego życia. Programowanie właściwości chemicznych molekuł, rolnictwo, diagnostyka medyczna i informacja — to dziedziny, w których możemy obserwować konkretne efekty. Niezbędnym elementem uzupełniającym są takie dyscypliny jak: prawo, etyka, czy też informacja (określana czasem terminem „bioinformatyka"). Właśnie informacja stanowi klucz do wielu osiągnięć nowoczesnych biotechnik. Konieczne są informacje zarówno prawne jak i techniczne oraz biologiczne dotyczące struktury i funkcji biomolekuł.

Literatura

1. Doughan L., Survey of CD-ROMs, (1995), Bio/Technology, 8-12.
2. Piątyszek M., Makałowski W., (1991), Biotechnologia, 13/14, 36-41.
3. Pianowska B., Twardowski T., (1992), Biotechnologia, 2, 89-112.
4. *Raport o stanie polskiej biotechnologii*, red. Zabża A., Kłaszewski S., (1995), Biotechnologia, 31, 13-59.
5. *Komputerowe bazy danych o nauce i technice*, (1995), Wyd.: Ośrodek Przetwarzania Informacji, Warszawa.

Databases for biotechnology

Summary

Bioinformatics is a very important part of biotechnology. The databases give us possibility to analyse the sequence and structure of biomolecules. It is very useful for research and industrial applications of biotechnology. In this paper we show the available database on CD-ROM and in on-line service in Poland useful for biotechnologists. This article based on the experts report written for CODATA Committee.

key words:
bioinformatics, database, CD-ROM, on-line database, biotechnology.

Adres do korespondencji:
Maciej Nawrot, Instytut Chemii Bioorganicznej PAN, Poznań, ul. Noskowskiego 12, 61-704 Poznań, fax: 52 05 32, e-mail: twardows@ibch.poznan.pl