Mieczysław Piątyszek
Wojciech Makałowski

Zakład Biochemii Biopolimerów
Uniwersytet im. Adama Mickiewicza
Poznań

Bazy danych sekwencji kwasów nukleinowych i białek: stan obecny i perspektywy rozwoju

1. Wprowadzenie

W ostatnich kilku latach nastąpił gwałtowny wzrost efektywności badań naukowych z zakresu biologii molekularnej i biotechnologii. Sytuacja ta jest rezultatem nie tyle wzrostu współzawodnictwa czy pojawienia się nowych narzędzi badawczych, ale uzyskania przez te dyscypliny poziomu rozwoju umożliwiającego praktyczne zastosowania ich osiągnięć. Pociągnęło to za sobą konkretne zamówienia z przemysłu, rolnictwa i medycyny, które z jednej strony wymuszają ukierunkowanie badań, a z drugiej, poprzez nie spotykane jak dotąd rozliczanie z rezultatów, jeszcze bardziej zwiększają ich efektywność.

Zjawisku temu towarzyszy dokonująca się na naszych oczach rewolucja w przepływie informacji naukowej i formach współpracy. Manifestacją dokonujących się zmian są powstałe ogromne bazy danych sekwencji kwasów nukleinowych i białek, bazy danych obejmujące szczepy mikrobiologiczne, enzymy restrykcyjne, wektory służące do klonowania materiału genetycznego, struktury makrocząsteczek biologicznych, mapy genomów. Powstały także międzynarodowe biuletyny informacyjne i krajowe sieci danych, których powiązanie umożliwia uzyskiwanie czasopism naukowych metodą *on-line* (np. *Biochemistry)*.

2. Rozwój genetycznych baz danych na przykładzie bazy GenBank

Na początku lat osiemdziesiątych w Stanach Zjednoczonych, Republice Federalnej Niemiec, Francji, Japonii i Związku Radzieckim rozpoczęto tworzenie baz danych sekwencji kwasów nukleinowych i białek (tzw. genetycznych baz danych). Ze względu na zróżnicowaną dynamikę rozwoju poszczególnych baz danych oraz trudności w dostępie do niektórych z nich obecnie znaczenie międzynarodowe uzyskały jedynie cztery największe – zob. tab. 1 (1–4).

Najczęściej wykorzystywaną bazą danych jest GenBank, baza sekwencji nukleotydowych Krajowego Instytutu Zdrowia (NIH) USA. Baza ta powstała w 1980 r. Popularność GenBank jest wynikiem wielkości bazy danych* oraz wszechstronnych możliwości komunikacji z bazą. Rozwój bazy GenBank mierzony liczbą zgromadzonych sekwencji w ciągu ostatnich dziewięciu lat przedstawia rys. 1.

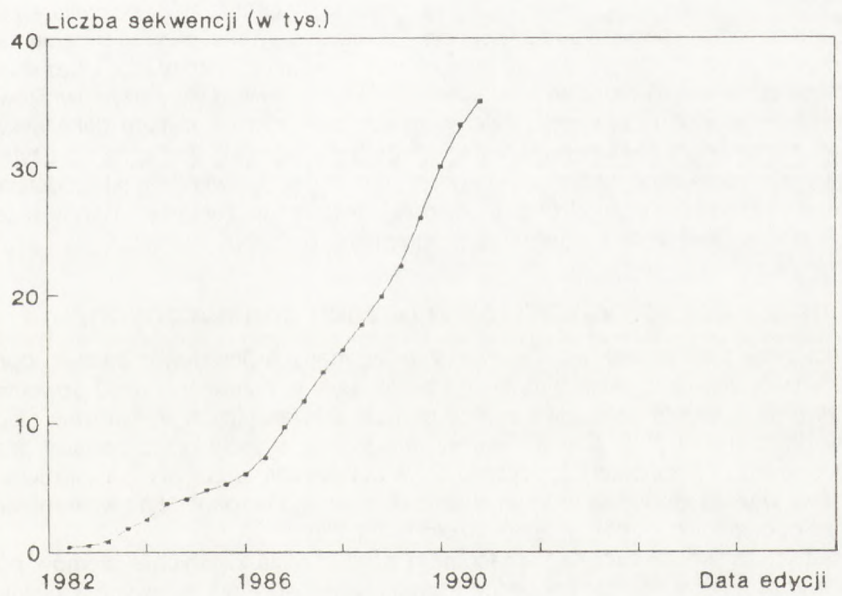
GenBank obecnie – jak widać na rys. 1 – znajduje się w wykładniczej fazie rozwoju. Przypuszcza się, że w 1995 r. baza ta będzie zawierała ok. 30 razy więcej sekwencji nukleotydowych aniżeli w 1990 r., a liczba nowo wprowadzanych danych rocznie wzrośnie kilkadziesiąt razy. Aktualnie GenBank zawiera kompletne sekwencje 150 genomów wirusowych, mitochondrialnych i chloroplastowych. W GenBank zdeponowano także sekwencje różnych regionów w genomie *Escherichia coli*, które pod względem długości odpowiadają już 28% całego genomu tej bakterii. Nukleotydowe sekwencje DNA *Homo sapiens* zdeponowane w GenBank stanowią natomiast 0,2% wielkości genomu ludzkiego (5). Należy spodziewać się, że w latach dziewięćdziesiątych, w związku z realizacją programu Human Genom Project, liczba wprowadzanych sekwencji z genomu ludzkiego będzie gwałtownie wzrastać.

* W marcu 1991 r. obejmowała 55,2 mln par zasad.

Tabela 1

Międzynarodowe genetyczne bazy danych

Nazwa bazy danych	Kraj	Adres przesyłania danych do bazy
GenBank	USA	GenBank Submissions Mail Stop K710 Los Alamos National Laboratory Los Alamos, NM 87545 USA E–mail: gb–sub%life@lanl.gov Tel: (505) 665–2177
Protein Identifcation Resource amino acid data bank (PIR)	USA	PIR Data Submission National Biomedical Research Foundation, Georgetown Univ. Medical Center, 3900 Reservoir Road N.W., Washington, D.C. 20007, USA E–mail: pirsub@gunbrf.bitnet
EMBL Data Library	RFN	EMBL Data Library Submissions Postfach 10.2209 D–6900 Heidelberg tel. (6221) 387258 E–mail: datasubs@embl.bitnet
DNA Data Bank of Japan (DDBJ)	Japonia	DNA Data Bank of Japan Laboratory of Genetic Information Analysis Center for Genetic Information Research Natl. Inst. Gentics Mishima, Shizuoka 411 Japan E–mail:ddbjsub@ddbj.nig.ac.jp



Rys. 1. Wzrost liczby sekwencji w kolejnych edycjach genetycznej bazy danych GenBank. Przyrost liczby nukleotydów w bazie danych jest jeszcze większy aniżeli przyrost liczby sekwencji, ponieważ przeciętna długość wprowadzanych sekwencji systematycznie rośnie (5).

3. Komunikowanie się z bazami

Dane z GenBank można uzyskać poprzez zakup kolejnych edycji banku na taśmach, dyskietkach lub płytach kompaktowych (CD). Istnieje też możliwość dostępu do GenBank za pomocą elektronicznych sieci Telenet i BIONET. Ta opcja komunikowania określana jest jako GenBank *on–line* Service (GOS). Po dokonaniu kwartalnie wnoszonej opłaty, wynoszącej 500 USD, użytkownik uzyskuje 1 Mb pamięci w bazie, dostęp do codziennie uaktualnianych baz GenBank, GenPept i EMBL oraz podstawowego oprogramowania umożliwiającego poruszanie się w obrębie baz i analizę danych. Opłata ta umożliwia realizację połączeń z GenBank przez 15 godz i 20 min. Opłata za każdą dodatkową godzinę połączeń wynosi 15 USD. Ten sposób komunikowania się z bazą gwałtownie zyskuje na popularności. W listopadzie 1989 r. zanotowano 1200 sesji połączeniowych z GenBank, a w marcu 1990 r. było ich już 6150 (6). Użytkownicy wykorzystują te połączenia do pobierania danych. Niestety, bezpośredni dostęp do ciągle uaktualnianej bazy GenBank nie jest możliwy dla osób pracujących w polskich pracowniach. Aby skontaktować się z GenBank, osoby spoza USA muszą najpierw połączyć się z krajową siecią komputerową, a następnie poprzez nią znaleźć się w amerykańskiej sieci danych „Telenet". Zagraniczni użytkownicy muszą dodatkowo płacić za użytkowanie sieci lokalnej. W Polsce brakuje krajowej sieci komputerowej, a także biorąc pod uwagę wysokie opłaty telekomunikacyjne, nawet istnienie sieci lokalnej nie pozwalałoby na korzystanie z usługi GOS.

GenBank, w porównaniu z pozostałymi bazami danych, od grudnia 1989 r. oferuje szczególnie dogodny i efektywny sposób przesyłania danych do bazy. Należy w tym miejscu przypomnieć, że czasopisma naukowe warunkują publikowanie prac zawierających sekwencje białek lub kwasów nukleinowych od wcześniejszego zdeponowania tych sekwencji w jednej z wymienionych w tab. 1 baz danych i podania tzw. *accession number*. Dotychczas, sekwencjom przesyłanym do baz danych na dyskietkach towarzyszyć musiał kilkustronicowy opis obejmujący wszystkie dane dotyczące wysyłanej sekwencji. Obecnie każdy zainteresowany może uzyskać z GenBank bez jakiejkolwiek opłaty program AUTHORIN[1], który pozwala – przesyłaną sekwencję oraz wszelkie dodatkowe informacje – zamieścić we właściwym formacie na dyskietce lub przesłać pocztą elektroniczną do jakiejkolwiek z wymienionych w tab. 1 baz danych. Należy podkreślić, że jest to program typu *user friendly*, podpowiadający użytkownikowi kolejne kroki kompletnego opisu sekwencji. Gwarantuje to bowiem, że wysyłane dane sekwencyjne będą natychmiast zasymilowane przez bank i po 24 godz staną się dostępne dla użytkowników GOS. Dla porównania dane nadsyłane w sposób tradycyjny pojawiają się jako dostępne w bazach danych najwcześniej po 30 dniach. Autorzy opracowania zachęcają osoby przesyłające sekwencje do baz danych do korzystania z programu AUTHORIN.

4. Analiza danych zawartych w bazach genetycznych

Równolegle z pojawieniem się genetycznych baz danych powstawać zaczęło oprogramowanie do analizy sekwencji nukleotydowych i białek. Były to przeważnie dość powolne programy umożliwiające jedynie częściową analizę danych sekwencyjnych wymieniane nieodpłatnie pomiędzy pracowniami (7,8). Szybko zostały one jednak wyparte przez zestawy programów opracowywanych w zespołach specjalistów i sprzedawane przez wyspecjalizowane firmy. Szczegółowy wykaz tego typu oprogramowania dostępnego w roku 1988 zamieszczony został w jednym z poprzednich numerów „Biotechnologii–P.I." (9).

W ostatnich latach zaznaczyła się tendencja do tworzenia zwartych systemów pozwalających na wyczerpującą analizę sekwencji i struktury kwasów nukleinowych i białek. Cechą

[1] Program AUTHORIN rozpowszechnia bezpłatnie: GenBank, c/o IntelliGenetics, Inc., 700 East El Camino Real, Mountain View, CA 94040, USA.

charakterystyczną jest bardzo rozbudowana prezentacja graficzna wyników oraz obecność kilkukanałowego systemu komunikacji programu z użytkownikiem (np. wprowadzanie danych za pomocą głosu, wybieranie opcji za pomocą tzw. myszy). Dodatkową zaletą tych systemów jest to, że zawierają równocześnie bazy danych, tak przetworzone, iż mogą być one bezpośrednio wykorzystane przez oferowane oprogramowanie. W tab. 2 przedstawiono najpowszechniej stosowane zestawy oprogramowania wraz z orientacyjnymi cenami (z sierpnia 1990 r.).

Tabela 2

Zestawy oprogramowania do analizy genetycznych baz danych

Nazwa zestawu	Dystrybutor	Orientacyjna cena w USD
IBI Pustell Sequence Analysis Software + Gel Reader System*	International Biotechnologies Inc.	6000
Staden Plus™	Amershram	5800
DNASTAR Software*	DNASTAR, Inc.	8500
DNASIS + PROSIS*	Stratagene, Gmbh	6000

* Zestaw zawiera genetyczne bazy danych.

5. Uwagi końcowe

Rozwój technik sekwencjonowania materiału genetycznego, a ostatnio ich automatyzacja (10) przyczyniają się do lawinowego wzrostu danych sekwencyjnych. Zjawisku temu towarzyszy wykładniczy rozwój genetycznych baz danych (zob. rys. 1). Udział w tych bazach sekwencji poznanych w polskich pracowniach jest znikomy. Przewiduje się, że w ciągu najbliższych pięciu lat rozmiary baz danych wzrosną kilkudziesięciokrotnie. Stosowane obecnie dyski kompaktowe przestaną być adekwatnym nośnikiem informacji. Oczekiwać zatem należy nowych rozwiązań opierających się także na optycznym odczycie danych. Konieczność dostępu do najnowszych danych, czego nie gwarantują mechaniczne nośniki informacji, powoduje, że rośnie zainteresowanie korzystaniem z danych deponowanych w bankach na zasadzie *on–line*. Pobierane wówczas są tylko wybrane bloki danych, a ich analiza przebiega w pracowni. Kraje, w których nie ma genetycznych baz danych organizują krajowe centra wchodzące w skład europejskiej sieci danych biologii molekularnej (EMBnet). W centrach krajowych dane są uaktualniane codziennie. Taka struktura dostępu do danych źródłowych usprawnia komunikację lokalnych użytkowników oraz znacznie obniża koszty dostępu do danych. W ten sposób rozwiązany został problem dostępu do baz danych między innymi w Grecji, Danii, Hiszpanii i Norwegii. Takie rozwiązanie w Polsce jest jednak w najbliższym czasie nierealne, między innymi ze względu na fatalny stan ogólnodostępnej sieci telekomunikacyjnej. W tej sytuacji duże znaczenie dla polskiego środowiska naukowego będą miały lokalne ośrodki informacji naukowej dysponujące zestawami oprogramowania wraz z genetycznymi bazami danych w systemie CD ROM. Jeden z takich ośrodków jest zlokalizowany w Poznaniu[2].

[2] Ośrodek Informacji Naukowej PAN, Oddział w Poznaniu, Stary Rynek 77, 61–722 Poznań, tel. 525–954, telex 413618 oin pl. Będący w dyspozycji OIN PAN system był opisany w „Biotechnologii–P.I." (1988), 1; por. również „Biotechnologia–P.I." (1991), 2.

Literatura

1. Burks C., et al., (1985), Comput. Appl. Biosci., 1, 22–27.
2. Cameron G., (1988), Nucleic Acids Res., 16, 1865–1868.
3. Sidman K. E., et al., (1988), Nucleic Acids Res., 16, 1869–1873.
4. Miyazawa S., (1989), in: Computers and DNA (G. I. Bell and T. Marr, eds.) 47–51, Addison–Wesley, New York.
5. Burks C., et al., (1990), in: Molecular Evolution: Computer Analysis of Protein and Nucleic Acids Sequences (R. F. Doolittle, ed.) 3–22, Academic Press, Inc., San Diego.
6. Youdin K., (1990), News From GenBank, 3, 1.
7. (1982), Nucleic Acids Res., 10, 1–456.
8. (1984), Nucleic Acids Res., 12, 1–428.
9. Popenda M., et al., (1988), „Biotechnologia–P.I." 1, 19–23.
10. Cathart R., (1990), Nature, 347, 310.

Databases of Nucleic Acids and Proteins Sequences: Current Status and Perspectives

Summary

In this review we have described the genetic databases with emphasis on GenBank database. Ways of communication with databases are described in details. Review contains information about the most popular software packages commercially available.

Adres dla korespondencji:
Wojciech Makałowski, Zakład Biochemii Biopolimerów, Uniwersytet im. Adama Mickiewicza, ul. Fredry 10, 61–701 Poznań, E-mail: WM2BB@PLP UAM11BITNET.

Nota dodana w czasie korekty

Rozwój bazy GenBank spowodował podjęcie decyzji o zaprzestaniu rozprowadzania Banku Genów na dyskach elastycznych. Ostatnia „dyskietkowa" edycja GenBank zostanie udostępniona w kwietniu 1992r. Począwszy od jesieni przyszłego roku GenBank będzie rozprowadzany wyłącznie na płytach kompaktowych.

W ostatnim czasie naukowcy polscy uzyskali od dawna oczekiwany dostęp do światowej łączności komputerowej. Polska została bowiem włączona do sieci EARN (European Academic and Research Net), która jest integralną częścią amerykańskiej sieci BITNET. Umożliwia to, poprzez tzw. pocztę elektroniczną, komunikowanie się praktycznie z całym światem, gdyż sieć BITNET poprzez system tzw. bramek ma połączenie rónież z innymi sieciami komputerowymi. W Polsce główny węzeł sieci EARN znajduje się na Uniwersytecie im. Adama Mickiewicza w Poznaniu. Sieć EARN umożliwia bezpośredni dostęp do komputera EMBL (European Molecular Biology Laboratory) w Heidelbergu. Autorzy opracowania z powodzeniem wykorzystywali już ten bardzo wygodny sposób dostępu do danych. Z bazą EMBL można się kontaktować poprzez adres NETSERV@EMBL.

Więcej informacji na temat sieci EARN w Polsce można usyskać w Centrum Informatyki UW (mgr inż. A. Smreczyński, E-mail: OEKO5@PLEARN), a także w ośrodkach lokalnych (w Poznaniu, np. mgr inż. A. Stolarski, Ośrodek Informatyki UAM).