Włodzimierz Trzebny[1]
Tomasz Twardowski[2]

[1] Ośrodek Informacji Naukowej PAN
[2] Instytut Chemii Bioorganicznej PAN
Poznań

Informacja w biotechnologii

Tendencje światowe w zakresie informacji naukowej (1) dobrze ilustruje analiza genomu człowieka. Ten problem badawczy – wyceniony wstępnie na 3 mld USD – będzie w swej finalnej formie polegał na archiwizacji 3 mld jednostek informacyjnych zawartych w określonej sekwencji nukleotydów.

1. Genom człowieka a informacja naukowa

Można przyjąć w pewnym uproszczeniu, że genom człowieka składa się z około 100 000 genów, które stanowią łańcuch uformowany z około $3x10^9$ nukleotydów. Zaawansowane są już badania nad analizą genomów relatywnie prostszych układów, np. *E. coli*, Arabidopsis i inne. Poznanie genomu niezależnie od materiału ująć można w trzech etapach:

1) utworzenie mapy genetycznej,
2) sformowanie mapy fizycznej,
3) określenie kompletnej sekwencji DNA.

Każdy z tych etapów wymaga precyzyjnego i łatwego w dostępie zapisu danych.

Mapa genetyczna polega na określeniu położenia poszczególnych genów względem siebie. Natomiast mapa fizyczna polega na określeniu wielkości (długości) i odległości pomiędzy poszczególnymi genami poprzez wprowadzenie specyficznych znaczników. Etap ten umożliwi bezwzględną lokalizację dowolnego fragmentu DNA w odniesieniu do innych genów. Ostatni etap polega na ustaleniu kolejności ułożenia nukleotydów w strukturze liniowej.

2. Problem zapisu danych

Kompletny opis genomu człowieka musi zawierać nie tylko sekwencje nukleotydowe, ale również ich interpretację i dane literaturowe. Zakłada się, że obecny zasób informacji zwiększy się 200–300 razy do roku 2000. Zapis posiadanych informacji prowadzony jest aktualnie zarówno na taśmach magnetycznych jak i w systemie CD–ROM. Pojedyncza płyta CD–ROM (o pojemności 500 milionów bajtów) mogłaby teoretycznie pomieścić całą sekwencję genomu człowieka. Jednakże uwzględniając opis uzupełniający, dane te zawarte zostaną na 6–7 płytach CD–ROM. Można jednak sądzić, że przy tak dużym postępie i rozwoju technik optycznego zapisu i odtwarzania danych w następnym stuleciu będzie można nabyć pojedynczą płytę CD–ROM z kompletnym zapisem indywidualnego genomu.

3. Światowe banki danych dla biotechnologii

Przy formułowaniu banków danych zawierających sekwencje DNA jedną z istotnych trudności – o kluczowym znaczeniu – jest bardzo szybki postęp i zwiększające się wymagania w stosunku do bioinformatyki.

Banki danych z zakresu biologii molekularnej – podstawowe dla problemu analizy genomu człowieka – zebrano w tab. 1 (wg 2).

Tabela 1

BANKI DANYCH

szczególnie ważne dla projektu „genom człowieka"*

Nazwa i zakres tematyczny	Lokalizacja	Sponsor
białka *Protein Data Bank*	Brookhaven National Laboratory, USA	Department of Energy, USA
homologie *Homology Database*	Jackson Laboratory Maine, USA	National Institute of Health USA
białka *Protein Identification*	Georgtown University Washington, D.C. USA	"
dane genetyczne *Online Mendelian Inheritance in Man*	Johns Hopkins University Baltimore, USA	"
sekwencje genowe *GenBank*	Los Alamos National Laboratory, New Mexico, USA	"
genom *Genome Database*	Johns Hopkins University, Baltimore, USA	Department of Energy, USA
Nematode worm	Laboratory of Molecular Biology, Cambridge, Anglia	Med. Research Council, Anglia
sekwencje kwasów nukleinowych i białek *Nucleotide Sequence Data Library*	EMBL, Heidelberg, RFN	EWG
sekwencje DNA *DNA Database of Japan*	National Institute of Genetics Mishima, Japonia	Agencja ds. Nauki i Technologii Japonia

* Niezależnie od już istniejących postuluje się utworzenie banku GENINFO, który stanowiłby „zapis baz danych", umożliwiając szybki dostęp do banków specjalistycznych. Problem lepiej ilustrują następujące liczby: w 1990 r. GenBank dysponował ok. 35 000 sekwencji kwasów nukleinowych; zakłada się, że „przyrost" będzie wynosił 60 000 – 100 000 sekwencji rocznie!

W zestawieniu (tab. 1) przedstawione są bazy danych o charakterze naukowo–badawczym. Brakuje natomiast baz o znaczeniu komercyjnym oraz formalnoprawnym, np. patentów czy też przepisów prawnych różnych państw z zakresu inżynierii genetycznej. Według amerykańskiej firmy Mitelligenetics patenty dotyczące sekwencji kwasów nukleinowych są praktycznie realizowane w 30 krajach, w myśl zasad legislacyjnych. Nowy bank danych GENESEQ opracowywany przez tę firmę w założeniu ma dotyczyć właśnie wszystkich kwestii patentowych. Przy bardzo powolnym trybie patentowania, co jednoznacznie rzutuje na swobodny przepływ informacji naukowej, ten bank danych ma stanowić wartościowe uzupełnienie standardowych informacji sekwencyjnych, jakimi dysponuje np. GenBank. GENESEQ ma zawierać dane dotyczące nie tylko sekwencji opatentowanych, ale także zgłoszonych do patentowania. Często różnica między nimi wynosi nawet 2 lata! W firmach zajmujących się tą problematyką uważa się, że proces patentowania sekwencji winien trwać tygodnie, a nie miesiące, czy lata. Będzie to jednakże możliwe tylko w przypadku szybkiej weryfikacji zgłoszonych sekwencji, np. poprzez jednoznaczne stwierdzenie czy jest to nowa sekwencja.

Z przedstawionego w tab. 1 zestawienia bardzo jednoznacznie wynika dominująca pozycja USA w zakresie gromadzenia informacji dla biotechnologii. Niewątpliwie jedną z przyczyn są znaczne nakłady finansowe konieczne dla realizowania szeroko rozumianej informacji naukowej. Firmy biotechnologiczne zlokalizowane w Europie Zachodniej podnoszą alarm dotyczący groźby braku dostępu do informacji naukowej w biotechnologii, gromadzonych w USA. Wiąże się to niewątpliwie z realiami ekonomicznymi, gdyż strona amerykańska zapowiedziała jednoznacznie, że dostęp do wyników związanych z badaniami nad genomem człowieka będzie pro-

porcjonalny do wkładu finansowego w te badania (3). Takie stanowisko reprezentuje m.in. James Watson, dyrektor Centrum Badania Genomu Ludzkiego przy National Institutes of Health (NIH, USA). Przy takim założeniu utajnienie badań w tym zakresie jest realne, gdyż dotychczas państwa Europy Zachodniej jedynie w niewielkim stopniu łożyły środki finansowe na te badania. Szereg państw ogłosiło – co prawda – rozpoczęcie prac badawczych w ramach tego problemu, np. w 1991 r. W. Brytania planuje przeprowadzenie badań o wartości 5 mln USD, a Francja zamierza przeznaczyć na ten cel 10 mln USD, lecz są to wszystko udziały skromne z chwilą, gdy zakłada się koszty rzędu 3 mld USD.

W Europie dominujące znaczenie ma biblioteka banków danych European Molecular Biology Laboratory, (EMBL) w Heidelbergu, która w ścisłej współpracy z GenBank i DDBJ (DNA – Database of Japan) kolekcjonuje i powszechnie udostępnia dane dotyczące sekwencji nukleotydów. EMBL oferuje także serwis komputerowy do użytku środowiska naukowego (3).

Od 1987 r. dział zbiorów danych EMBL wprowadził serwis dla użytku zewnętrznego (5). Jest on w pełni zautomatyzowany, oparty na systemie komputerowym EMBL. Umożliwa dogodne przeszukiwanie baz danych oraz uzyskanie odpowiednich informacji przy zastosowaniu elektronicznej poczty. 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.

Z chwilą jego wprowadzenia, serwis ten stał się bardzo popularny w środowisku naukowym. Biblioteka banków danych EMBL finansowana jest przez Europejską Wspólnotę Gospodarczą (EWG), pracuje w skali ogólnoświatowej i udziela bezpłatnie około 1000 informacji miesięcznie. Oprócz baz danych EMBL, GenBank, SwissProt i Brookhaven wprowadzane są obecnie do obsługi inne bardzo ważne zbiory danych, a mianowicie bazy danych PROSITE (6), ENZYME (5), bazę danych enzymów restrykcyjnych REBASE (7), bazę danych *E. coli* (ECD) (8). W tab. 2 zawarto zakres tematyczny banków danych EMBL*.

Tabela 2

Zakres tematyczny banków danych biblioteki EMBL w Heidelbergu

Baza	Zakres tematyczny
ECD	*E. coli*
ENZYME	enzymy
EPD	promotory układów eukariotycznych
DOC	dokumentacja ogólna
LIMB	lista baz danych
NUC EMBL, GenBank, DDBJ	dane sekwencyjne
PROSITE (*prosite pattern*)	białka
PROTEIN (SwissProt)	białka
PROTEINDATA (Brookhaven)	dane strukturalne
REBASE	enzymy restrykcyjne
REFLIST	referencje sekwencji
SOFTWARE	oprogramowanie

* Zasady korzystania i sposób dostępu do banków danych EMBL szczegółowo opisano w (4).

Pod koniec 1989 r. biblioteka banków danych EMBL rozpoczęła bezpłatną dystrybucję programów komputerowych *(software)* dotyczących biologii molekularnej. Obecnie stosowane są systemy komputerowe MS-DOS, Apple Macintosh i VAX/VMS, wkrótce dostępny będzie również UNIX.

Zachęca się autorów różnych programów do udostępnienia i włączenia tych programów do biblioteki EMBL w celu dalszego ich rozpowszechnienia. Tylko w ciągu 6 miesięcy 1990 r. udostępniono około 60 programów, które bezpłatnie zostały rozprowadzone do przeszło 2000 użytkowników.

Przy badaniach biologicznych najczęstszym wykorzystaniem baz danych dotyczących sekwencji jest porównywanie fragmentów sekwencji kwasów nukleinowych (np. genów) lub białek z całą bazą w celu wykrycia podobieństwa. W związku z tym w bibliotece EMBL uruchomiono dwa dodatkowe serwisy, a mianowicie: Mail–Quicksearch oraz Mail–Fast A, które umożliwiają wyszukiwanie danych w systemach komputerowych EMBL.

Obsługa serwisu zbiorów EMBL jest bardzo sprawna, jednakże jej funkcjonalność w powiązaniu z pocztą elektroniczną posiada pewne ograniczenia. Dlatego EMBL obecnie rozpatruje możliwość zastosowania innych metod dystrybucji oraz dostępu do baz danych. Bada się możliwość wykorzystania sieci TCP/IP (do której już jest podłączona EMBL), która zyskuje coraz większą popularność w Europie w celu zastosowania nowocześniejszego sposobu udostępniania zbiorów.

Aktualną sytuację Japońskiego Banku Danych DNA (DNA Data Bank of Japan, DDBJ) przedstawiono w (9). Obecnie DDBJ opracowuje tylko 3% sekwencji kwasów nukleinowych publikowanych na świecie. Oczekuje się, że w najbliższym roku liczba ta wzrośnie do 10%, w związku z planowanym rozwojem DDBJ oraz w wyniku zobowiązania się Japonii do opracowania wszystkich sekwencji ustalonych w ich kraju. Pozostałe 97% sekwencji kwasów nukleinowych (a w przyszłości 90%) są opracowywane w Stanach Zjednoczonych (GENBANK) oraz w Niemczech (EMBL). W analizach prezentowanych w literaturze międzynarodowej nie jest uwzględniany w ogóle radziecki system GenExpress. W marcu 1990 r. uruchomiono sieć łączności komuterowej pomiędzy ośrodkiem DDBJ zlokalizowanym w mieście Mishima via Tokyo University do Stanów Zjednoczonych (GenBank) i Europy (EMBL). Ilość informacji wymienianych pomiędzy tymi ośrodkami pod koniec 1989 r. wynosiła około 1 milion bitów dziennie, natomiast nowe łącze umożliwia przekaz 64 kilobitów na sekundę. Jednakże nadal nie zostaną rozwiązane zasadnicze trudności związane z kompatybilnością sprzętu. W Japonii powszechnie stosowane są komputery systemu NEC, rzadko używane w USA i Europie. Poza tym przekaz danych opiera się w Japonii na taśmach magnetycznych, podczas gdy GenBank i EMBL rozpoczęły stosowanie w dużej skali systemu CD–ROM.

4. Nowości z dziedziny CD–ROM a zapis sekwencji genomowych

Według ostatnich doniesień rozwój technologii systemu dysków optycznych umożliwia elektroniczny dostęp do dużej części biblioteki czasopism naukowych oraz banków danych, w tym także banków danych sekwencji. Zakłada się, że już w 1990 r. około 25% informacji naukowej było zapisywanych w systemie CD–ROM. W katalogu baz danych „The CD–ROM Directory 1989" (10) opisano 390 baz danych na płytach kompaktowych. Co więcej, obecnie zainspirowano projekt pod nazwą „Adonis", w myśl którego konsorcjum firm wydawniczych: Blackwells, Elsevier, Pergamon i Springer, planuje w ciągu 1991 r. wprowadzenie subskrypcji przeszło 400 tytułów czasopism naukowych na płytach kompaktowych (11). Płyty kompaktowe, zawierające pełną zawartość bieżących wydań czasopism będą przesyłane co tydzień subskrybentom. Obraz każdej strony czasopisma jest przedstawiony *in facsimile*. Jedna płyta zawiera ponad 5000 typowych stron z czasopisma naukowego.

Zakłada się, że pierwszymi klientami tej formy subskrypcji będą biblioteki firm farmaceutycznych. Wybór czasopism będzie dokonywany przez użytkownika. Należy podkreślić, że konsorcjum wydawnictw od dwóch lat prowadziło eksperymentalną dystrybucję ponad 200 tytułów

czasopism naukowych na płytach kompaktowych do kilkudziesięciu bibliotek na całym świecie w celu przetestowania projektu „Adonis".

Ostatnio coraz częściej rozważa się konieczność archiwizowania danych na płytach kompaktowych. Pracownicy National Aeronautics and Space Administration (NASA) twierdzą, że olbrzymia ilość danych magazynowana do tej pory na taśmach magnetycznych częstokroć uniemożliwia ich wyszukiwanie i opracowanie (12). Przykładowo w Jet Propultion Laboratory (JPL) w Pasadenie (Kalifornia), gdzie są gromadzone dane z badań planetarnych uważa się, że połowa z 135 000 taśm magnetycznych jest prawie bezużyteczna. Ta sytuacja musi być uwzględniona przy planowanym zapisie sekwencji genomu człowieka; jak już wspomniano EMBL i GenBank wprowadziły już zapis sekwencji genowych w systemie CD–ROM.

5. Trudności bioinformatyki

Zespoły badawcze obecnie trzech najważniejszych banków danych w zakresie sekwencji kwasów nukleinowych i białek: EMBL, GenBank i DNA Database of Japan, bardzo blisko współpracują ze sobą, ale wiele prac jest po prostu dublowanych, jak np. zapis, korekta i weryfikacja danych. Te operacje jako pierwsze muszą ulec modyfikacjom. Podstawowym problemem jest zapewnienie kompatybilności sprzętu i oprogramowania oraz zabezpieczenie przed niepożądanym dostępem do danych. Zapewne równoległym zadaniem – i to równie trudnym – będzie automatyzacja zapisu nowych danych. Aktualnie około 70% nowych sekwencji jest nadsyłanych do EMBL bezpośrednio przez badaczy w bardzo zróżnicowanej formie, bez zunifikowanego formatu zapisu. W konsekwencji automatyczne (np. laserowe) przeniesienie danych jest możliwe tylko w ograniczonym zakresie. Poza tym pozostałe 30% sekwencji jest wyszukiwana w pismach naukowych i wprowadzana ręcznie do banku danych. Niezależnie od tych technicznych trudności, zachodzi podstawowy problem merytoryczny: które z prezentowanych sekwencji są w pełni nowe, które nakładają się (*overlapping*) częściowo, a być może są wręcz błędne lub źle zlokalizowane. W opinii tak wybitnych ekspertów jak Walter Gilbert, David Botstein czy też Tom Caskey zagadnienie dopuszczalnego błędu w oznaczonej sekwencji nie jest jednoznaczne (13). Przypuszcza się, że aktualnie opracowane sekwencje zawierają kilka błędów na tysiąc oznaczonych nukleotydów. Wiadomo również, że realizacja sekwencjonowania przy założeniu „zero błędów" jest nierealna, a osiągnięcie dokładności 1 błędu na 100 000 byłoby bardzo kosztowne. Pragmatycy uważają, że poprawność rzędu 1 promila będzie satysfakcjonująca. Odrębny problem stanowi szacunek kosztów związanych z większą wiarygodnością oznaczenia; innymi słowy jaki będzie koszt sekwencjonowania w przeliczeniu na 1 zasadę przy rzetelności 95%, a jaki w przypadku 99,9%.

Wiadomo również, że pewne sekwencje powtarzają się w genomie dwukrotnie lub nawet wielokrotnie. Wówczas sekwencja, aczkolwiek już będąca w banku danych, jest jednakże jakościowo nową, bowiem istnieje w innym miejscu genomu i może nawet pełnić inną funkcję.

Grupa uczonych z Heidelbergu (gdzie jest zlokalizowany EMBL) dąży do zorganizowania europejskiego instytutu bioinformatyki (European Institute of Bioinformatics), podobnie jak już istniejące European Space Agency lub CERN. Instytut taki powinien być finansowany w ramach EWG, a mógłby zapewnić Europejczykom konkurencyjność z badaniami ośrodków amerykańskich.

Wymiana informacji gwarantuje jej rozwój w interesującej nas dziedzinie. Jest to jednak uzależnione od warunków technicznych i łożonych nakładów finansowych. W Polsce problem rozwoju informacji naukowo–technicznej jest wyjątkowo złożony. Skutki braku informacji wydają się oczywiste, natomiast jak temu zapobiec?

Etos uczonych w zakresie „informacji otwartej", dostępnej dla wszystkich zainteresowanych, napotyka na złożone i wielorakie trudności; ale w konsekwencji sprowadzają się one w dużym stopniu do problemów finansowych. Przykładowo, American Chemical Society (USA) publikuje przeglądy literatury w zakresie danych chemicznych *Chemical Abstracts,* zajmując w tej dziedzinie praktycznie pozycję monopolisty. Podobnie The National Library of Medicine (USA) opraco-

wująca najbardziej ogólną i uniwersalną bazę danych w zakresie medycyny Medline, może dysponować swym bankiem danych w sposób całkowicie samodzielny*.

Oczekuje się, że The National Library of Medicine zapewne przejmie kontrolę nad dwoma bankami danych podstawowych dla biotechnologii: GenBank (sekwencje kwasów nukleinowych) i PIR (sekwencje aminokwasowe białek). Dla rozwoju nauki i techniki dostęp do tych banków danych ma główne znaczenie. Konieczne jest przy tym, aby był to dostęp **legalny**, polegający na wykupieniu stosownych licencji. Bowiem korzystanie z banków danych poprzez nielegalne kopiowanie nie przynosi rzeczywistych korzyści naukowych, a perspektywicznie wręcz odwrotnie – straty. W najbardziej dogodnej pozycji, stwarzającej możliwości do korzystania z amerykańskich banków danych, są międzynarodowe konsorcja mające swe przedstawicielstwa w różnych krajach. Również nie przejawiają obaw o dostęp do banków danych ci, którzy mogą coś zaoferować w zamian, jak np. Excerpta Medica (wydawane przez Elsevier, Holandia) dokonująca przeglądu 4500 tytułów czasopism naukowych.

Techniki komputerowe i laserowe stanowią podstawę rozwoju nie tylko informacji naukowej, ale także innych dziedzin, jak np. bioinformatyki. Ścisły związek i wzajemna zależność nowoczesnej biotechnologii (rozpatrywanej tutaj na przykładzie analizy sekwencji genomu) z technikami informacyjnymi jest oczywisty.

---

* Sytuacja taka miała już miejsce w stosunku do ZSSR po inwazji na Afganistan; bank Medline został zablokowany dla wszystkich aktualnych użytkowników ze Związku Radzieckiego.

Literatura

1. Na temat informacji w biotechnologii publikowane były już artykuły w naszym piśmie, por: „Biotechnologia" 1–2/'88,
2. S. Watts, (1990), New Scientist, 4 VIII, 37–41.
3. Hodgson J., (1990), Biotechnology, 8, 15.
4. Fuchs R., et al., (1990), Nucl. Acids Res., 18, 4319–4323.
5. Stoehr P., Omond R., (1989), Nucl. Acids Res., 17, 6763–6764.
6. Bairoch A., (1990), University of Geneva, Geneva.
7. Roberts R.J., (1985), Nucl. Acids Res., 13, r165–r200.
8. Kroeger M., (1989), Nucl. Acids Res., 17, r283–r309.
9. Swinbanks D., (1990), Nature, 344, 92.
10. The CD-ROM Directory 1989, 3rd ed. edited by Evin Cormack, TFPL Publishing, London.
11. Maddox J., (1990), Nature, 344, 287.
12. Lindley D., (1990), Nature, 344, 182.
13. Roberts L., (1990), Science, 250, 1336–1338.

Information in biotechnology

Summary

Several aspects of information in Biotechnology are presented: availability of data from international data banks, CD-ROM system, limitations of the acces to the information centers. As the exemplification of the situation, the problems concerning bioinformatics in sequencing of human genom are shown.

Adres dla korespondencji:
Włodzimierz Trzebny, ul. Gorczyczewskiego 2/3, 60–554 Poznań.