Komputerowe banki sekwencji kwasów nukleinowych i białek

Jerzy CIESIOŁKA, Mariusz POPENDA, Włodzimierz KRZYŻOSIAK
Instytut Chemii Bioorganicznej PAN
Poznań

I. WPROWADZENIE

Sekwencjonowanie kwasów nukleinowych i białek stanowi jedną z najważniejszych metod badawczych stosowanych w biologii molekularnej. Poznane sekwencje organizowane są w zbiory zwane bankami sekwencji, które obok pełnionej funkcji magazynowania danych sekwencyjnych stanowią wyjściowe źródło informacji dla tworzenia i testowania hipotez o organizacji, funkcji i ewolucji cząsteczek biologicznych.

Kolekcje sekwencji białek rozpoczęto tworzyć już w latach sześćdziesiątych. Opracowanie szybkich technik sekwencjonowania DNA (1,2) w końcu lat siedemdziesiątych, doprowadziło do gwałtownego wzrostu liczby znanych sekwencji kwasów nukleinowych. Obecnie przybywa ponad 1,5 mln zasad w ciągu roku i rozważane są nawet możliwości oznaczenia sekwencji całego genomu ludzkiego, liczącego około $3x10^9$ nukleotydów (3).

Nagromadzona ilość danych sekwencyjnych stworzyła konieczność zastosowania techniki komputerowej do ich przechowywania i analizy. W różnych ośrodkach badawczych zaczęły powstawać komputerowe banki wyselekcjonowanych grup kwasów nukleinowych i białek. Znana jest m.in. kompilacja struktur tRNA i genów tRNA (4,5), bank sekwencji organizmów prokariotycznych (6), kompilacja map genetycznych (7), bank enzymów restrykcyjnych (8). Wkrótce podjęto również próby stworzenia ogólnodostępnych banków wszystkich dotychczas poznanych sekwencji kwasów nukleinowych i białek. Wysiłki te doprowadziły do powstania na początku lat osiemdziesiątych dwóch banków sekwencji nukleotydowych: w Heidelbergu, w European Molecular Biology Labolatory (bank EMBL) i w Los Alamos National Labolatory (GenBank) oraz banku sekwencji białek: Protein Identification Resource przy National Biological Research Foundation (bank NBRF-PIR).

Wymienione banki sekwencji stanowią obecnie najważniejsze źródło informacji dotyczących sekwencji kwasów nukleinowych i białek, docierają również do najszerszego kręgu odbiorców. Z tego względu przedstawiony w artykule przegląd komputerowych banków sekwencji zawężony został do przedstawienia organizacji i formatu zapisu informacji stosowanego w banku EMBL, GenBank oraz NBRF-PIR.

Zaprezentowane zostaną również założenia systemu CODATA będącego wynikiem tendencji zmierzających do ujednolicenia zróżnicowanych formatów zapisu informacji.

Artykuł uzupełniony został także o krótki opis podstawowych sposobów korzystania z informacji zawartych w bankach sekwencji nukleotydowych i białek.

II. SYMBOLE ZASAD KWASÓW NUKLEINOWYCH I AMINOKWASÓW

Symbole zasad kwasów nukleinowych, jakie stosowane są w bankach sekwencji, przedstawione zostały na rys. 1a. Są one zgodne z zaleceniami Komisji IUPAC-IUB. Dozwolone jest stosowanie zarówno dużych jak i małych liter alfabetu. Modyfikowane zasady, występujące głównie w cząsteczkach tRNA, oznaczane są za pomocą symboli analogicznych do używanych w atlasie Sprinzla i Gaussa (4).

a)

```
G = Guanine
A = Adenine
T = Thymine
C = Cytosine
R = Purine              (A or G)
Y = Pyrimidine          (C or T or U)
M = Amino               (A or C)
K = Ketone              (G or T)
S = Strong interaction  (C or G)
W = Weak interaction    (A or T)
H = Not-G               (A or C or T) H follows G in the alphabet
B = Not-A               (C or G or T) B follows A
V = Not-T (not-U)       (A or C or G) V follows U
D = Not-C               (A or G or T) D follows C
N = Any                 (A or C or G or T)
U = Uracil (in RNA)
```

b)

```
A = Ala          = Alanine
B = Asx          = Aspartic Acid or Asparagine
C = Cys          = Cysteine
D = Asp          = Aspartic Acid
E = Glu          = Glutamic Acid
F = Phe          = Phenylalanine
G = Gly          = Glycine
H = His          = Histidine
I = Ile          = Isoleucine
K = Lys          = Lysine
L = Leu          = Leucine
M = Met          = Methionine
N = Asn          = Asparagine
P = Pro          = Proline
Q = Gln          = Glutamine
R = Arg          = Arginine
S = Ser          = Serine
T = Thr          = Threonine
V = Val          = Valine
W = Trp          = Tryptophane
X = X            = any amino acid
Y = Tyr          = Tyrosine
Z = Glx          = Glutamine or Glutamic Acid
```

Rys.1. Symbole stosowane w bankach sekwencji kwasów nukleinowych i białek: a) zasad kwasów nukleinowych, b) aminokwasów.

Symbole aminokwasów zestawione na rys.1b odpowiadają regułom pisowni przyjętej przez Komisję Nazewnictwa Biochemicznego IUPAC-IUB. Stosowane są oznaczenia jedno- bądź trójliterowe; w banku sekwencji białek NBRF-PIR są to symbole jednoliterowe. Dla oznaczenia kodonu terminującego przyjęte zostało oznaczenie " " lub "." .

III. BANKI SEKWENCJI NUKLEOTYDOWYCH

1. Bank EMBL

W programie powstałej w 1980 r. biblioteki EMBL znalazły się trzy podstawowe zadania:

- stworzenie powszechnie dostępnego zbioru wszystkich, znanych sekwencji kwasów nukleinowych;
- popieranie zamierzeń prowadzących do standaryzacji i szerokiej wymiany informacji z zakresu biologii molekularnej;
- pełnienie roli europejskiego centrum informacji komputerowej służącego potrzebom biologii molekularnej (9).

Opracowane w ramach przyjętego programu pierwsze wydanie banku sekwencji nukleotydowych ukazało się w kwietniu 1982 r. Do grudnia 1986 r. ukazało się dziesięć kolejnych wydań banku EMBL.

Tabela 1 przedstawia liczbę sekwencji nukleotydowych zebranych w kolejnych wydaniach banku, od siódmego do dziesiątego, z podziałem sekwencji na kilka grup strukturalno-funkcjonalnych.

Wydanie dziesiąte banku zawiera zbiór 8817 sekwencji, złożonych z blisko 10 mln zasad, pochodzących z 6192 publikacji źródłowych.

Bank EMBL rozpowszechniany jest na taśmie magnetycznej. Sekwencje nukleotydowe uporządkowane są w ciąg jednostkowych pozycji zapisu informacji (w terminologii angielskiej określanych mianem "entry"). Każda z nich zawiera oprócz pojedynczej sekwencji nukleotydowej podstawową jej charakterystykę oraz spis publikacji źródłowych na podstawie których została wprowadzona do banku EMBL (rys.2).

```
ID  HSIL03     standard; DNA; 462 BP.
XX
AC  X00201;
XX
DT  01-FEB-1984  (first entry)
XX
DE  Human interleukin 2 (IL2) gene fragment
XX
KW  interleukin.
XX
OS  Homo sapiens (man, homme, Mensch)
OC  Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
OC  Eutheria; Primates.
XX
RN  <1>   (bases 1-462; enum. 1 to 462)
RA  Degrave W., Tavernier J., Duerinck F., Plaetinck G., Devos R.,
RA  Fiers W.;
RT  "Cloning and structure of the human interleukin 2 chromosomal
RT  gene";
RL  EMBO J. 2:2349-2353(1983).
XX
CC  Data kindly reviewed (01-JUN-1984) by W. Fiers
XX
FH  Key         From      To        Description
FH
FT  IVS           <1      259       intron II
FT  CDS          260      403       fragment of IL2 gene
FT  IVS          404     >462       intron III
XX
SQ  Sequence    462 BP;  173 A;  82 C;  79 G;  128 T;
TGCAGAAAGT CTAACATTTT GCAAAGCCAA ATTAAGCTAA AACCAGTGAG TCAACTATCA
CTTAACGCTA GTCATAGGTA CTTGAGCCCT AGTTTTTCCA GTTTTATAAT GTAAACTCTA
CTGGTCCATC TTTACAGTGA CATTGAGAAC AGAGAGAATG GTAAAAACTA CATACTGCTA
CTCCAAATAA AATAAATTGG AAATTAATTT CTGATTCTGA CCTCTATGTA AACTGAGCTG
ATGATAATTA TTATTCTAGG CCACAGAACT GAAACATCTT CAGTGTCTAG AAGAAGAACT
CAAACCTCTG GAGGAAGTGC TAAATTTAGC TCAAAGCAAA AACTTTCACT TAAGACCCAG
GGACTTAATC AGCAATATCA ACGTAATAGT TCTGGAACTA AAGGTAAGGC ATTACTTTAT
TTGCTCTCCT GGAAATAAAA AAAAAAAAGT AGGGGGAAAA GT
```

Rys.2. Jednostkowa pozycja zapisu informacji ("entry") pochodząca z banku EMBL.

Tabela 1

Liczba sekwencji nukleotydowych oraz sumaryczna liczba zasad zawartych w różnych podgrupach kwasów nukleinowych (w czterech kolejnych wydaniach banku EMBL) (10)

Data Library Growth

	Release 7		Release 8		Release 9		Release 10	
References cited	4007		4410		5319		6192	
Sequences	Entries	Bases	Entries	Bases	Entries	Bases	Entries	Bases
Artificial	153	46073	163	52672	182	68540	192	71237
Chloroplast	118	119852	131	128219	149	153786	167	465378
Genetic elements	41	37804	51	42990	54	43857	54	43857
Mitochondrial	275	303549	296	316905	307	346721	322	376392
Prokaryotic	750	763188	857	879633	1065	1130637	1175	1305116
Viral/Phage	887	1344597	946	1446696	1195	1689681	1335	1975030
Eukaryotic	3557	2997022	3944	3475686	4668	4364797	5540	5465260
Unclassified	8	10553	7	10239	10	15195	32	64678
Total	5789	5622638	6395	6353040	7630	7813214	8817	9766948

Stosowany jest znormalizowany zapis informacji; każda pozycja zapisu złożona jest z linii tekstu, rozpoczynających się dwuliterowym wyróżnikiem, wskazującym na charakter informacji zawartej w danej linii.

Rysunek 3 przedstawia uszeregowanie stosowanych w banku wyróżników linii oraz hasłowe objaśnienie charakteru informacji w nich zawartych.

ID - identification	(begins each entry)
AC - accession number(s)	(use first one in citations)
DT - date	
DE - description	
KW - keywords	
OS - organism species	
OC - organism classification	
RN - reference number	
RA - reference author	
RT - reference title	
RL - reference location	
CC - comments or notes	
FH - feature table header	(for readability)
FT - feature table data	
SQ - sequence header	
- (blanks) sequence data	
XX - spacer line	(for readability)
// - termination line	(ends each entry)

Rys.3. Wyróżniki linii stosowane w banku EMBL oraz hasłowe objaśnienia ich znaczenia (10).

Krótki opis poszczególnych linii podany jest poniżej:

ID: pierwsza linia każdej pozycji w banku. Złożona jest z:

- nazwy, jednoznacznie przyporządkowanej danej sekwencji w określonym wydaniu banku,
- określenia czy informacje zawarte w danej pozycji zapisu zostaną poddane weryfikacji w kolejnych wydaniach banku (ang. standard, unreviewed, preliminary),
- określenia typu cząsteczki (RNA lub DNA),
- liczby zasad w cząsteczce;

DT: data wprowadzenia lub ostatniego uaktualnienia informacji;

DE: ogólny opis sekwencji zawierający nazwę kodowanego genu, regionu genomu, z którego pochodzi oraz inne informacje ułatwiające jej identifikację;

KW: zbiór słów kluczowych służących do tworzenia katalogów grupujących sekwencje zawarte w banku wg wspólnych, ważnych cech funkcjonalnych bądź strukturalnych;

OS, OC: nazwa i klasyfikacja taksonomiczna organizmu - źródła pochodzenia sekwencji;

RN, RA, RT, RL: spis publikacji stanowiących podstawę wprowadzenia sekwencji do banku;

CC: krótki komentarz opisujący sekwencję i jej najważniejsze cechy charakterystyczne;

FH, FT: ujęty w formie tabeli opis regionów sekwencji bądź pojedynczych jej nukleotydów charakteryzujących się szczególnie istotnym znaczeniem biologicznym;

SQ: pierwsza linia bloku informacyjnego zawierającego sekwencję nukleotydową; podana jest w niej długość sekwencji (liczba zasad) oraz wyszczególniony skład nukleotydowy;

XX: linia nie zawiera żadnej informacji, służy do rozdzielenia poszczególnych bloków informacyjnych;

//: linia oznaczająca zakończenie każdej jednostkowej pozycji zapisu w banku sekwencji.

2. GenBank

Bank sekwencji genowych - GenBank utworzony został w 1982 r., a jego głównym sponsorem jest Narodowy Instytut Zdrowia (National Institutes of Health) w Stanach Zjednoczonych. Pierwsze wydanie banku ukazało się w lipcu 1982 r. Zbieraniem informacji i opracowywaniem kolejnych wydań banku zajmuje się Los Alamos National Laboratory, natomiast jego dystrybucją BBN Laboratories Inc.

Sekwencje nukleotydowe zebrane w GenBank podzielone zostały według kryterium ich pochodzenia na 13 grup. W tabeli 2 przedstawiona jest liczba sekwencji nukleotydowych i sumaryczna liczba zasad zawarta w poszczególnych grupach sekwencji, w dwóch kolejnych wydaniach GenBank 44.0 i 48.0.

Tabela 2

Liczba sekwencji nukleotydowych oraz sumaryczna liczba zasad, zawarta w poszczególnych grupach sekwencji (w dwóch kolejnych wydaniach GenBank: 44.0 z sierpnia 1986 r. oraz 48.0 z lutego 1987 r.) (11)

	Release	44.0		Release 48.0	
Division No.	Code Description	Number of Entries	Bases	Number of Entries	Bases
1	PRI Primate Seq.	1028	1240779	1337	1602436
2	ROD Rodent Seq.	1272	1111622	1612	1460441
3	MAM Other Mammalian Seq.	245	244554	310	325323
4	VRT Other Vertebrate Seq.	474	400509	513	440085
5	INV Invertebrate Seq.	605	435280	685	542525
6	PLN Plant Seq.	594	643365	689	819740
7	ORG Organelle Seq.	368	485666	423	840194
8	BCT Bacterial Seq.	749	1031546	916	1268642
9	RNA Structural RNA Seq.	637	69232	916	72838
10	VRL Viral Seq.	1076	1517025	1160	1684316
11	PHG Bacterial Seq.	160	271817	169	283756
12	SYN Synthetic Seq.	224	72029	251	78943
13	UNA Unannotated Seq.	1374	918933	2189	1542126
	Overall Summary:	8823	8442357	10913	10961365

GenBank udostępniany jest użytkownikom w różnych formach: zapisany na taśmie magnetycznej, na dyskietkach lub poprzez system "on line" (system konwersacyjny) w uaktualnianej co 6 tygodni wersji. Ponadto corocznie, we współpracy z bankiem EMBL, publikowane jest kompendium zbiorcze.

Najszerzej rozpowszechnioną formą dystrybucji GenBank jest wersja zapisana na dyskietkach formatu 5 1/4 cala, współpracująca z mikrokomputerami systemu IBM PC. W wersji tej sekwencje nukleotydowe oraz wyselekcjonowane informacje opisowe zapisane zostały w systemie zagęszczonym, odczytywanym za pomocą dołączonego do banku pakietu programów użytkowych. Podobnie jak w banku EMBL, zbiór sekwencji nukleotydowych oraz towarzyszących im informacji opisowych podzielony został na jednostkowe pozycje zapisu informacji. Złożone są one w tym przypadku nie z pojedynczych linii, lecz z bloków informacyjnych rozpoczynających się słowami - wyróżnikami np. locus, definition, keyword itp. Zapis pojedynczej pozycji pochodzący z wydania 44.0 GenBank przedstawia rys. 4.

W tabeli 3 zestawione zostały odpowiadające sobie wzajemnie określenia wyróżników poszczególnych bloków informacji, stosowane w bankach EMBL, GenBank oraz w systemie CODATA (przedstawiony w rozdz. II.3). Znaczenie słów - wyróżników jest podobne do odpowiadających im wyróżników stosowanych w banku EMBL (por. rozdz. II.1).

W skład GenBank wchodzi również kilka programów wyszukujących, uruchamianych poprzez GenBank Menu (zob. rys.5). Programy służą do selektywnego przeglądania informacji zawartych w banku np. według wybranego słowa kluczowego bądź nazwiska autora publikacji, w której opisana jest dana sekwencja. Ponadto dołączony jest program umożliwiający translację dowolnie wybranej z banku sekwencji nukleotydowej na kod aminokwasowy (wg reguł tzw. ogólnego kodu genetycznego) oraz program obliczający częstotliwość występowania poszczególnych kodonów w obrębie danej sekwencji.

```
**********************************************************
*                    GenBank Menu                        *
*   1.  Display entry information on screen              *
*   2.  Create file in GenBank tape format               *
*   3.  Keyword phrase search                            *
*   4.  Author name search                               *
*   5.  Accession number search                          *
*   6.  Translate a sequence                             *
*   7.  Codon usage calculation of a sequence            *
*   8.  Exit from menu                                   *
**********************************************************

Your choice:
```

Rys.5. "GenBank Menu" - wyszczególnienie opcji oprogramowania wchodzącego w skład banku sekwencji GenBank (zapisanego na dyskietkach formatu 5 1/4 cala) (11).

3. System zapisu informacji CODATA

Sposoby zapisu informacji stosowane w bankach sekwencji EMBL i GenBank chociaż podobne, są różne. Stanowi to istotną niedogodność dla użytkowników korzystających z oprogramowania komputerowego, wykorzystującego banki sekwencji jako podstawową bazę danych, bowiem różniące się formatem zapisu banki nie mogą być stosowane zamiennie.

```
LOCUS       HUMA1AT4     292 bp    DNA             updated   03/12/84
DEFINITION  Human alpha 1-antitrypsin gene: 3' terminus.
ACCESSION   J00067
REFERENCE   1
  AUTHOR    Kurachi,K., Chandra,T., Friezner Degen,S.J., White,T.T.,
            Marchioro,T.L., Woo,S.L.C., Davie,E.W.
  JOURNAL   Proc Nat Acad Sci USA 78, 6826-6830 (1981)
REFERENCE   2
  AUTHOR    Leicht,M., Long,G.L., Chandra,T., Kurachi,K., Kidd,V.J.,
            Mace,M.Jr., Davie,E.W., Woo,S.L.C.
  JOURNAL   Nature 297, 655-659 (1982)
REFERENCE   3
  AUTHOR    Rogers,J., Kalsheker,N., Wallis,S., Speer,A., Coutelle,CH.,
            Woods,D., Humphries,S.E.
  JOURNAL   Biochem Biophys Res Commun 116, 375-382 (1983)
FEATURES        from  to/span     description
    pept        <     1     211   alpha 1-antitrypsin (exon 4, partial)
BASE COUNT        72 a     94 c     59 g     67 t
ORIGIN
   1 acccctgaag ctctccaagg ccgtgcataa ggctgtgctg accatcgacg agaaagggac
  61 tgaagctgct ggggccatgt ttttagaggc catacccatg tctatccccc ccgaggtcaa
 121 gttcaacaaa ccctttgtct tcttaatgat tgaacaaaat accaagtctc ccctcttcat
 181 gggaaaagtg gtgaatccca cccaaaaata actgcctctc gctcctcaac ccctcccctc
 241 catccctggc cccctccctg gatgacatta aagaagggtt gagctggtcc ct
//
```

Rys.4. Przykład jednostkowej pozycji zapisu informacji ("entry") pochodzący z GenBank (wydania 44.0 na dyskietkach 5 1/4 cala).

Nie wszystkie pakiety programów analizujących zawierają podprogramy umożliwiające korzystanie z różniących się formatem zapisu danych sekwencyjnych. Korzystne byłoby wobec tego przyjęcie w różnych bankach sekwencji ujednoliconej formy zapisu informacji.

Znormalizowany format zapisu informacji dotyczących sekwencji kwasów nukleinowych i białek, określany jako system CODATA, zaproponowała w 1984 r. grupa naukowców powołana do tego celu przez International Congress of Scientific Unions (9). Propozycja ta ściśle przypomina formaty zapisu stosowane przez banki EMBL i GenBank. Porównanie formatów zapisu stosowanych w bankach EMBL, GenBank oraz w systemie CODATA (zob. tab. 3). Natomiast przykład zapisu informacji w systemie CODATA pochodzący z banku sekwencji białek NBRF-PIR (zob. rys.6).

Tabela 3

Porównanie formatów zapisu informacji ("struktury linii") jednostkowych pozycji w bankach sekwencji EMBL i GenBank oraz formatu stosowanego w systemie CODATA

EMBL	*GenBank*	*CODATA*
ID	LOCUS	ENTRY
		#Length
		#Checksum
AC	ACCESION	ACCESION
DE	DEFINITION	NAME
		ALTERNATE-NAME
		INCLUDES
		GENE-NAME
		MAP-POSITION
	SEGMENT	
DT		DATE
OS	SOURCE	SPECIES
	ORGANISM	
		#Strain
		#Plasmid
		#Clone
		#Tissue
		#Life-cycle
		HOST
		TAXONOMY
		SUPERFAMILY
	REFERENCE	REFERENCE
RN		#Number
RA	AUTHORS	#Authors
RL	JOURNAL	#Citation
RT	TITLE	#Title
CC	COMMENT	COMMENT
KW	KEYWORDS	KEYWORDS
FH	FEATURE	FEATURE
	SITES	
FT		
		INTRONS
		START-CODON
	BASE COUNT	SUMMARY
		#Molecular-weigth
		#Length
		#Checksum
XX		
SQ		SEQUENCE
	ORIGIN	
//	//	//

IV. BANK SEKWENCJI BIAŁEK NBRF-PIR

Bank sekwencji białek znany pod nazwą banku NBRF-PIR lub PIR (The Protein Identification Resource at National Biological Research Foundation) działa od roku 1984, stanowiąc kontynuację działalności prowadzonej przez Margaret Dayhoff (13,14). Działalność banku finansowana jest przez National Institutes of Health USA.

Przykład zapisu w systemie CODATA jednostkowej pozycji z banku NBRF-PIR (zob. rys.6).

```
ENTRY          OKBOG          Protein   #Length 670 #Checksum 5530
NAME           cGMP-dependent protein kinase (EC 2.7.1.37) - Bovine
DATE           17-May-1985 #Sequence 17-May-1985  #Text 27-Nov-1985
SPECIES        Bos taurus #Common-name ox
REFERENCE      Sequences of residues       1-17, 89-374, and 407-670
  #Authors     Takio, K., Wade R.D., Smith S.B., Krebs E.B., Walsh, K.A.
                 Titani, K.
  #Journal     Biochemistry (1984) 23: 4207-4218
REFERENCE      Sequence of residues 13-104
  #Authors     Takio,K., Smith, S.B., Walsh,K.A., Krebs, E.G., Titani, K.
  #Journal     J. Biol. Chem. (1983) 258:5531-5536
REFERENCE      Sequence of residues 373-409
  #Authors     Hashimoto, E., Takio, K., Krebs, E.G.
  #Journal     J. Biol. Chem. (1982) 257: 727-733
COMMENT        The protein,isolated from lung,is a dimer of identical
    chains.
SUPERFAMILY
  #Name        cAMP-dependent protein kinase regulatory chain
                 #Residues 102-340
  #Name        kinase-related transforming protein
                 #Residues 475-599
KEYWORDS       acetylation\ phosphoprotein\ cGMP\
                 serine-specific protein kinase
FEATURE
1                  #Modified-site acetylated amino end\
42                 #Disulfide-bonds interchain\
58                 #Binding-site phosphate\
1-101              #Domain dimerization <DIM>\
102-219            #Domain cGMP-binding 1 <GB1>\
320-340            #Domain cGMP-binding 2 <GB2>\
341-474            #Domain ATP-binding <APB>\
475-599            #Domain catalytic <CAT>
COMMENT        These boundaries are approximate.
SUMMARY        #Molecular-weight 76287 #Length 670 #Checksum 5530
SEQUENCE

                5         10        15        20        25        30
  1 S E L E E D F A K I L M L K E E R I K E L E K R L S E K E E
 31 E I Q E L K R K L H K C Q S V L P V P S T H I G P R T T R A
 61 Q G I S A E P Q T Y R S F H D L R Q A F R K F T K S E R S K
 91 D L I K E A I L D N D F M K N L E L S Q I Q E I V D C M Y P
121 V E Y G K D S C I I K E G D V G S L V Y V M E D G K V E V T
151 K E G V K L C T M G P G K V F G E L A I L Y N C T R T A T V
181 K T L V N V K L W A I D R Q C F Q T I M M R T G L I K H T E
211 Y M E F L K S V P T F Q S L P E E I L S K L A D V L E E T H
241 Y E N G E Y I I R Q G A R G D T F F I I S K G K V N V T R E
271 D S P N E D P V F L R T L G K G D W F G E K A L Q G E D V R
301 T A N V I A A E A V T C L V I D R D S F K H L I G G L D D V
331 S N K A Y E D A E A K A K Y E A E A A F F A N L K L S D F N
361 I I D T L G V G G F G R V E L V Q L K S E E S K T F A M K I
391 L K K R H I V D T R Q Q E H I R S E K Q I M Q G A H S D F I
421 V R L Y R T F K D S K Y L Y M L M E A C L G G E L W T I L R
451 D R G S F E D S T T R F Y T A C V V E A F A Y L H S K G I I
481 Y R D L K P E N L I L D H R G Y A K L V D F G F A K K I G F
511 G K K T W T F C G T P E Y V A P E I I L N K G H D I S A D Y
541 W S L G I L M Y E L L T G S P P F S G P D P M K T Y N I I L
571 R G I D M I E F P K K I A K N A A N L I K K L C R D N P S E
601 R L G N L K N G V K D I Q K H K W F E G F N W E G L R K G T
631 L T P P I I P S V A S P T D T S N F D S F P E D N D E P P P
661 D D N S G W D I D F
//
```

Rys.6. Jednostkowa pozycja zapisu informacji pochodząca z banku sekwencji białek NBRF-PIR (przedstawiona w systemie CODATA) (12).

V. KORZYSTANIE Z BANKÓW SEKWENCJI

Różnorodne sposoby wykorzystania informacji zawartych w bankach sekwencji podzielić można na trzy zasadnicze grupy:

1. Przeszukiwanie banku sekwencji. Jest to najpowszechniej stosowana forma korzystania z banków sekwencji. Polega na szukaniu podobieństw między sekwencją badaną a innymi sekwencjami znajdującymi się w banku. Do przeszukiwania wykorzystywane są nazwy sekwencji nadane im w banku ("entry names") oraz zbiory słów kluczowych ("keyword index"). Ponadto w szukaniu podobieństw strukturalnych bardzo użyteczne są tablice cech charakterystycznych sekwencji ("feature tables"). Do przeszukiwania służą programy dostarczane wraz z bankiem (np. GenBank) lub programy opracowane przez użytkowników, zgodne z ich potrzebami.

2. Korzystanie ze specjalnych programów analizujących sekwencje kwasów nukleinowych i białek. Banki sekwencji stanowią podstawową bazę informacyjną szeregu, dostępnych handlowo, pakietów programów analizujących z zakresu biologii molekularnej (np. IntelliGenetics, DNASTAR, IBI/PUSTELL ; bliższe informacje o tych programach zawarte są w artykule: M.Popenda, J.Ciesiołka, W.J.Krzyżosiak, *Przegląd* ...).

3. Wykorzystanie banków jako zbiorów informacji bibliograficznych. Banki sekwencji oprócz roli podstawowej jaką pełnią, służą również jako uporządkowane, specjalistyczne zbiory bibliograficzne, ułatwiając dostęp do publikacji źródłowych.

VI. UWAGI KOŃCOWE

Wykorzystanie w bankach sekwencji współczesnej techniki komputerowej sprawiło, że wielokrotnie wzrosła szybkość i niezawodność przetwarzania informacji dotyczących sekwencji kwasów nukleinowych i białek. Niekiedy, a w szczególności podczas posługiwania się sekwencjami o bardzo dużej długości, banki sekwencji są podstawowym, niemalże niezastąpionym źródłem danych sekwencyjnych. Również w szeregu komputerowych programów planowania i analizy eksperymentów korzysta się z banków sekwencji jako wyjściowej bazy danych.

Banki sekwencji stanowią ponadto odzwierciedlenie tempa i kierunków rozwoju badań w biologii molekularnej. Tak np. w tabeli 4 przedstawiono podział sekwencji nukleotydowych zebranych w wydaniu GenBank z września 1985 r. według kryterium pełnionych przez te sekwencje funkcji biologicznych (15). Podobne zestawienie dokonane na podstawie kolejnych wydań banku może stanowić miarę wzrostu lub spadku zainteresowań badaczy poszczególnymi funkcjami kwasów nukleinowych.

Tabela 4

Podgrupy funkcjonalne kwasów nukleinowych, których sekwencje zawarte są w GenBank

Funkcja	Udział procentowy[a]
Sekwencje kodujące:	
białka	46.82
rybosomalne RNA	2.62
transferowe RNA	1.67
małe jądrowe RNA	0.06
inne RNA	0.10
Introny	12.24
Sekwencje niekodujące[b]	36.49

a) obliczono na podstawie GenBank (wrzesień 1985); udział procentowy określono biorąc pod uwagę sumaryczną liczbę zasad zawartych w każdej grupie sekwencji; b) grupa sekwencji zawiera również sekwencje, których funkcja nie jest dotychczas znana; wśród nich mogą znajdować się regiony DNA o potencjalnych właściwościach kodujących.

W tymże wydaniu GenBank zawarte są kompletne sekwencje ponad 80 genomów, przede wszystkim wirusów, plazmidów, bakteriofagów. Spośród organizmów szczególnie interesujących biologię molekularną, poznano dotychczas 9% genomu E.coli, 1% genomu Saccharomyces cerevisiae i mniej niż 0.1% genomu Drosophila melanogaster, myszy i człowieka. Odzwierciedla to skalę trudność związanych z planami oznaczenia sekwencji genomu człowieka, jak i wskazuje przekonująco, że skoordynowany kilkuletni wysiłek mógłby doprowadzić do poznania kompletnej sekwencji genomu E.coli.

Paroletni okres działalności banków sekwencji ujawnił nie tylko ich możliwości, ale również niedogodności i ograniczenia. Jednym z nich jest opóźnienie, z jakim opublikowana w pracy źródłowej sekwencja dociera do szerokiego kręgu odbiorców za pośrednictwem banku sekwencji. Czas ten uległby skróceniu po przyjęciu odpowiednich porozumień między czasopismami naukowymi a bankami sekwencji. Na przykład na mocy takiego porozumienia, uzgodnionego

pomiędzy GenBank i Nucleic Acids Research, autorzy publikacji proszeni są o dostarczenie informacji o sekwencji, którą zamierzają opublikować również w znormalizowanym formacie zapisu ułatwiającym wprowadzenie jej do banku.

Natomiast trudności związane z odmiennym, w różnych bankach, sposobem zapisu informacji zostaną usunięte po powszechnym zaakceptowaniu systemu CODATA.

Jak można przewidywać, prowadzenie banków sekwencji w obecnym kształcie może zostać istotnie ograniczone z uwagi na wzrastającą w szybkim tempie ilość informacji. Tak np. ilość informacji zawartych w GenBank ulega podwojeniu co dwa lata, a wydanie 48.0 z lutego 1987 r., w zredukowanej formie, o zagęszczonym zapisie, zajmuje 33 dyskietki formatu 5 1/4 cala. Korzystnym rozwiązaniem byłoby zatem posługiwanie się bankami zawężonymi do określonej grupy kwasów nukleinowych. W GenBank (ale nie w banku EMBL) podział sekwencji nukleotydowych na grupy już istnieje.

Rozważając możliwości oznaczenia całkowitej sekwencji genomu ludzkiego należy uwzględnić fakt, że jego długość jest blisko 300-krotnie większa od sumarycznej długości wszystkich dotąd poznanych sekwencji kwasów nukleinowych. Przedsięwzięcie to wymaga zatem zarówno usprawnienia metod oznaczania sekwencji DNA, jak i znacznego udoskonalenia komputerowych technik gromadzenia i przetwarzania informacji.

LITERATURA

1. Maxam A.M. and Gilbert W. (1977) Proc.Natl.Acad.Sci.USA, 74, 560.
2. Sanger F.,Nicklen S. and Coulson,A.R. (1977) Proc.Natl.Acad. Sci.USA, 74, 5463.
3. Lewin R. (1986) Science, 232, 1598.
4. Sprinzl M. and Gauss D.H. (1984) Nucleic Acids Res., 12, r1.
5. Sprinzl M. and Gauss D.H. (1984) Nucleic Acids Res., 12,r59.
6. Schneider T.D., Stormo G.D., Haemer J.S. and Gold L. (1982) Nucleic Acids Res., 10, 3013.
7. O'Brien S.J.(ed.) (1984) Genetic Maps: A Compilation of Linkage and Restriction Maps of Genetically Studied Organisms, Vol.3, published by Cold Spring Harbor Laboratory Press, NY.
8. Roberts R.J. (1985) Nucleic Acids Res., 13, suppl.
9. Lesk A.M., (1985) Nature, 314, 318.
10. EMBL Nucleotide Sequence Data Library, Release 10 (December 1986) - Manual.
11. GenBank Floppy Diskette User's Guides, Release 44.0 (August 1986) and Release 48.0 (February 1987).
12. Bishop M.J., Ginsburg M., Rawlings C.J. and Wakeford R. (1987) in "Nucleic Acid and protein sequence analysis", Bishop M.J., Rawlings C.J. Eds., p.83, IRL Press,Oxford,Washington DC.
13. Orcutt B.C., George D,G. and Dayhoff M.O. (1983) Annu. Rev. Biophys.Bioeng., 12, 419.
14. George D.G., Barker W.C. and Hunt L. (1986) Nucleic Acids Res., 14, 11.
15. Burks C., Fickett J.W., Goad W.B., Kanehisa M., Lewitter F.I., Rindone W.P., Swindell C.D., Tung C.-S. and Bilofsky H.S. (1985) CABIOS, 1, 225.