Katarzyna Tarant,
Marek Ziętarski

Instytut Chemii Bioorganicznej PAN
Poznań
Ośrodek Informacji Naukowej PAN
Odział w Poznaniu

Banki danych dla biotechnologii

Wstęp

Tempo z jakim docierają wciąż nowe doniesienia ze świata nauki zmusza nas do poszukiwania nowych bardziej efektywnych i wyrafinowanych technik porządkowania, przechowywania i wyszukiwania informacji. Zważmy, że już obecnie poznane sekwencje kwasów nukleinowych liczą ponad 31 milionów par zasad, a optymiści oczekują, iż w ciągu najbliższych 3 lat będzie znana pełna sekwencja genomu ludzkiego (około $3x10^9$ nukleotydów). Dogłębna analiza tych informacji, stanowiących tylko wąski wycinek wiedzy, jest niewyobrażalna bez wykorzystania systemów komputerowych, natomiast magazynowanie takiej ilości danych wymaga użycia najnowszych nośników o dużej pojemności. Właśnie ze względu na dużą pojemność i relatywnie szybki dostęp do informacji wiele firm oferuje bazy danych na płytach laserowych w systemie CD–ROM. Dzięki tej technice każdy pracownik nauki może na swoim komputerze wyszukiwać interesujące go informacje w bazie o pojemności około 0,5 GB zawartej na krążku o średnicy 4,72 cala. Sprzęt potrzebny do korzystania z banków na płytach laserowych to komputer klasy IBM XT, AT lub PS/2, drukarka oraz czytnik laserowy. Jest on łatwo dostępny i dość tani (7).

W naszym kraju banki danych na płytach laserowych dostępne są zaledwie w kilku instytucjach. Ośrodek Informacji Naukowej PAN Oddział w Poznaniu już od ponad 2 lat eksploatuje trzy takie banki dotyczące biotechnologii: HIBIO (DNASIS/PROSIS), ChemBank i LSC, które udostępniane są bezpłatnie zainteresowanym osobom. Wszystkie banki są uaktualniane przez wymianę płyty w drodze subskrybcji. Poniżej scharakteryzujemy pokrótce wszystkie bazy oraz dołączone do nich oprogramowanie.

HIBIO – Bank Inżynierii Genetycznej

Bank HIBIO (1) opracowywany przez japońską firmę Hitachi, a rozpowszechniany przez szwedzką firmę LKB obejmuje trzy banki sekwencji: GenBank, EMBL i PIR oraz dwa pakiety programów: DNASIS i PROSIS. GenBank i EMBL to kolekcje kwasów nukleinowych, natomiast PIR zawiera sekwencje białkowe. Zapis dotyczący jednej sekwencji obejmuje krótką charakterystykę cząsteczki, rodzaj i gatunek organizmu będącego jej źródłem, odsyłacze bibliograficzne, tablice opisujące regiony sekwencji mające znaczenie biologiczne i pełną sekwencję. Bazy te są powszechnie znane wśród biochemików dlatego nie będziemmy dokładniej ich omawiać (6).

Bardzo cenną zaletą banku jest to, że programy zebrane w obu pakietach umożliwiają wszechstronną analizę sekwencji zarówno pochodzących z baz jak i wprowadzonych przez badacza. Poniżej omawiamy krótko możliwości tych pakietów.

I. Pakiet DNASIS zawiera 7 grup programów:

Edycja sekwencji kwasów nukleinowych (DNAEDIT)

1. Edycja istniejącej sekwencji lub tworzenie nowej.
2. Porównywanie sekwencji DNA z jednoczesną możliwością przesuwania jednej z nich.
3. Operacje na plikach: mazanie, zmiana nazwy lub otrzymywanie sekwencji komplementarnej.
4. Druk sekwencji.

Analiza Struktury Pierwszorzędowej Kwasów Nukleinowych
(DNA *Primary Structure Analysis*)

1. Graficzna prezentacja alternatywnych sygnałów start i stop.
2. Translacja sekwencji DNA na kod aminokwasów (jedno- lub trzyliterowy).
3. Określenie w sposób graficzny lub za pomocą tabeli częstości występowania kodonów.
4. Odszukiwanie miejsc restrykcyjnych. Prezentacja graficzna lub tabelaryczna. Budowa własnych zbiorów restryktaz.
5. Graficzne porównywanie sekwencji metodą Harr, a następnie operacje na wydruku, np. powiększanie interesującego fragmentu wykresu.
6. Wyszukiwanie i graficzna prezentacja fragmentów cząsteczki DNA bogatych w pary G/C oraz porównanie ich z regionami zawierającymi A/T.
7. Określenie procentowej zawartości wszystkich zasad w strukturze analizowanego DNA w postaci histogramu.
8. Łączenie dwóch różnych sekwencji prowadzące do utworzenia nowej molekuły DNA z identyfikacją fragmentów określonych jako introny i exony.
9. Poszukiwanie homologii między dwoma cząsteczkami DNA.
10. Szukanie w danej sekwencji określonych trójek zasad, np. szukanie sekwencji kontrolnych i regulatorowych.

Analiza Struktury Drugorzędowej Kwasów Nukleinowych
(DNA *Secondary Structure Analysis*)

1. Przewidywanie struktury drugorzędowej poprzez określenie odwrotnych powtórzeń prowadzących do tworzenia pętli.
2. Wyszukiwanie miejsc powtarzalnych w genie tzw. transposonów.
3. Wyszukiwanie w DNA tzw. „lepkich końcy".
4. Określanie sekwencji palindromowych.

Łączenie Kwasów Nukleinowych (DNA *Connecting*)

1. Łączenie określonych sekwencji DNA z fragmentów uzyskanych w wyniku analizy restrykcyjnej.
2. Automatyczne łączenia fragmentów komplementarnych uzyskanych w wyniku analizy sekwencyjnej.

Budowa Map Restrykcyjnych *(Restriction Map Construction)*

1. Edycja danych niezbędnych do budowy mapy restrykcyjnej.
2. Konstrukcja kolistej mapy restrykcyjnej.
3. Konstrukcja linearnej mapy restrykcyjnej.
4. Anulowanie mapy lub map restrykcyjnych.
5. Wyświetlanie na monitorze wszystkich map restrykcyjnych i łączenie ich z nowymi danymi.
6. Wydruk liniowy struktur DNA z zaznaczonymi miejscami restrykcyjnymi.

Uaktywnienie Bazy GenBank *(Database Acces)*

1. Uaktywnienie baz GenBank i EMBL, która może być przeszukiwana wg numerów akcesyjnych, słów kluczowych, autorów, nazw wejścia lub fraz. Otrzymywane informacje mogą być

zapisywane na dysku, a następnie obrabiane za pomocą opisanych programów. Każdy z indeksów, wg których przeszukiwana jest baza może być wydrukowany.

2. Szukanie homologii cząsteczki DNA z jedną lub kilkoma cząsteczkami albo z całą bazą.

Programy Narzędziowe (DNASIS Utilities)

1. Przekształcanie formatu zapisu innych sekwencji DNA do formy umożliwiającej analizę w ramach programów banku DNASIS.

2. Formułowanie nowych tabeli kodów genetycznych, które umożliwiają translację nietypowych DNA, np. w przypadku układu mitochondrialnego.

II. Pakiet PROSIS zawiera 4 grupy programów. Trzy grupy, tj. programy edycyjne, narzędziowe i uaktywniające bazę mają swoje odpowiedniki w pakiecie DNASIS. Dlatego poniżej wyszczególniamy tylko programy czwartej grupy służące do analizy białek.

Analiza Białek *(Protein Analysis)*

1. Przekształcanie sekwencji DNA do białek i odwrotnie.

2. Określanie hydrofobowości i hydrofilowość białek na podstawie składu aminokwasowego metodą Hopp–Woods.

3. Określanie homologii białek na podstawie metody Harr.

4. Identyfikacja miejsc modyfikacji posttranslacyjnej.

5. Przewidywanie strukury drugorzędowej.

6. Kompozycja aminokwasów w odniesieniu do ciężaru cząsteczkowego.

7. Porównanie homologii pomiędzy dwoma cząsteczkami białka.

8. Szukanie homologii cząsteczki białka z jedną lub więcej cząsteczek albo z całą bazą zawartą na dysku kompaktowym.

CHEMBANK – bank faktograficzny z zakresu toksycznych substancji chemicznych

CHEMBANK rozpowszechniany przez brytyjską firmę SilverPlatter jest zestawem trzech baz: RTECS, OHMTADS, CHRIS – dotyczących problematyki bezpieczeństwa i higieny pracy w zakresie związków chemicznych. Bazy te wzajemnie się uzupełniają i charakteryzują chemikalia pod innym kątem. System zarządzania bazą zapewnia błyskawiczne wyszukiwanie informacji wg frazy lub słowa kluczowego w dowolnie wybranym polu rekordu z możliwością użycia operatorów logicznych, a także wydruk i zapis na dysku wyszukanych informacji. Ponadto istnieje możliwość przeglądania słownika terminów oraz korzystania z systemu podpowiedzi (3).

RTECS (*Registry of Toxic Effects of Chemical Substances*). Baza wydawana jest przez *The National Institute for Occupational Safety and Health* (NIOSH). Zawiera opis toksycznych oddziaływań 95 tys. substancji. W ich charakterystyce uwzględniono: nazwę związku (wraz z synonimami), jego definicję i wzór chemiczny, numer w wykazie *Chemical Abstract Service*, masę cząsteczkową, mutagenność, działanie rakotwórcze, wielkość dawki toksycznej odsyłacze do rejestru oddziaływań toksycznych i kancerogennych oraz normy zanieczyszczeń.

OHMTADS (*Oil and Hasardous Materials Technical Assistance Data System*) opracowywany jest przez *The Office of Water and Waste Management of the U.S. Environmental Protection Agency*. Obejmuje ponad tysiąc substancji opisując ich właściwości fizyczne, ze szczególnym uwzględnieniem negatywnego wpływu na jakość wody i sposobu postępowania w przypadku skażenia. Każdy rekord składa się ze 135 pól, które można podzielić na następujące grupy: pola identyfikujące substancję, informacje o warunkach transportu i przechowywania, dane dotyczą-

ce reaktywności chemicznej, metody wykrywania skażenia, podatność na ogień i wybuch, właściwości fizyczne i chemiczne, potencjalne zagrożenie dla przyrody, ogólne informacje o toksyczności, dane dotyczące szkodliwego oddziaływania na roślinność i na człowieka, wykaz czynności, które należy wykonać w przypadku zanieczyszczenia środowiska i inne dane z zakresu chemii środowiskowej.

CHRIS (*Chemical Hazard Response Information System*), który wydawany jest przez *The U.S. Departament of Transportation*, stanowi źródło informacji o zapobieganiu, środkach ostrożności i sposobach postępowania z 1077 niebezpiecznymi chemikaliami (styczeń, 1988). Podobnie jak w bazie OHMTADS i tutaj substancja charakteryzowana jest bardzo wszechstronnie, gdyż każdy rekord podzielony jest na 103 pola, w których można znleźć takie informacje jak: ogólna charakterystyka substacji; dane o postępowniu w nagłych wypadkach; oznaczenia chemiczne; właściwości fizykochemiczne; dane o zagrożeniu dla zdrowia, o niebezpieczeństwie wystąpienia ognia, reaktywności chemicznej, dotyczące skażenia wody.

LSC – bank bibliograficzny dotyczący nauk biologicznych

LSC (*Live Sciences Collection*) jest wydawany przez *Cambridge Scientific Abstracts* (CSA), natomiast wersję na dysku kompaktowym rozpowszechnia firma Microinfo Ltd. z Wielkiej Brytanii. Obejmuje ona informacje z ponad 5 tysięcy profesjonalnych czasopism, książek, monografii, doniesień konferencyjnych i opracowań patentowych ukazujących się na świecie. Nasz Ośrodek dysponuje wydaniami począwszy od roku 1982. Baza obejmuje następujące dziedziny nauki: etologię, biochemię aminokwasów, peptydów, białek i błon biologicznych, ekologię, entomologię, genetykę, immunologię, mikrobiologię przemysłową, algologię, mykologię, protozoologię, neurologię, toksykologię i wirusologię. Osoba poszukująca publikacji z tych dziedzin, posługując się bankiem LSC, uzyskuje następujące informacje: nazwisko autora, adres jego miejsca pracy, tytuł pracy w języku angielskim, miejsce publikacji, język oryginału, słowa kluczowe, streszczenie, nazwę, datę i miejsce konferencji, tytuł oryginalny publikacji, międzynarodowy znormalizowany kod wydawnictwa ciągłego (ISSN), międzynarodowy znormalizowany numer książki (ISBN), numer patentu oraz inne numery identyfikacyjne (4,5).

Literatura

1. HIBIO DNASIS, (1987), DNA sequence input and analysis software system. Reference manual. Hitachi America Ltd.
2. HIBIO PROSIS, (1987), Protein input and analysis software system reference manual, Hitachi America Ltd.
3. Getting started, (1988), SilverPlatter Information System.
4. Life Sciences Collection. User's manual, (1987), Microinfo Ltd.
5. On line Data bases in the Medical and Life Sciences, (1987), New York: Caudra, Elsevier.
6. Ciesiołka J., Popenda M., Krzyżosiak W., (1988), Komputerowe banki sekwencji kwasów nukleinowych i białek, Biotechnologia 1/2.
7. Twardowski T., (1988), Charakterystyka systemu CD-ROM i ON-LINE, Biotechnologia 1/2.

Data banks for biotechnology

Summary

The paper presents databases available and used by the Center of Scientific Information in Poznań. These databases have been bought to meet the demands of biotechnological research. These are: GenBank – bank of gene sequences, NBRFPIR – bank of protein sequences (both of them recorded on the CD-ROM disc), RTECS – Registry of Toxic Effects of Chemical Substances, OHMTADS – Oil and Hazard-

ous Materials Technical Assistance Data System, CHRIS – Chemical Hazard Response Information System (all three on the same CD–ROM disc) and LSC – bank of sciences of life which at present consists of 4 volumes.

Adres dla korespondencji:
Katarzyna Tarant, Instytut Chemii Bioorganicznej PAN, ul. Noskowskiego 12, 61-704 Poznań.

NOWOŚCI!

Inżynieria genetyczna w uzyskiwaniu czynnika VIII krwi ludzkiej

Pierwszy zapis świadczący o istnieniu choroby nazwanej później hemofilią A znajdujemy w księgach Talmudu. W początkach XIX w. stwierdzono, że jest ona sprzężona z płcią, w 1840 r. wykazano, że białko obecne w krwi, nazwane czynnikiem VIII, bierze udział w procesach krzepnięcia krwi.

Stosowana obecnie terapia hemofilii A polega na częstym podawaniu drogą pozajelitową większej ilości preparatu czynnika VIII uzyskanego z ludzkiej plazmy. Ze względu na obawy o jednoczesne przekazanie przy transfuzji patogennych wirusów (HIV, w zapaleniu wątroby) wprowadzono dodatkową procedurę przy produkcji tych preparatów, co czyni je bardzo kosztownymi, a zabiegi terapeutyczne stosuje się tylko w sytuacjach krytycznych. Preparaty takiej jakości wytwarzają w Stanach Zjednoczonych firmy Armour Pharmaceutical Co. (preparat Monoclate) i Hyland Div., Baxter Health–care Corp. (Hemofil M).

W 1984 r. izolowano ludzki gen czynnika VIII, a w latach 1984–1988 syntetyzowano cDNA kodujący ten gen, wprowadzono go do linii komórek chomika (CHO) i uzyskano jego ekspresję. Pierwotnym produktem ekspresji w organizmie ludzkim jest pojedynczy polipeptyd, który ulega dojrzewaniu po wydzieleniu z komórek, w plazmie krwi – udział w tym procesie biorą trombina i/lub czynnik białkowy X. W prekursorowym peptydzie istnieją trzy powtórzenia domeny A (330 aminokwasów), domena B (980 aminokwasów) i domena C (150 aminokwasów).

Ekspresja genu w komórkach chomika jest o 2–3 rzędów niższa niż innych obcych genów tak samo wprowadzonych do tych komórek. Wynika to z: a) niskiego poziomu transkrypcji, b) braku białek, które towarzyszą ekspresji czynnika VIII *in vivo*, chroniących go przed proteolizą, c) oraz zatrzymanie większości produktu ekspresji w *reticulum* endoplazmatycznym. Zatem: a) do transfekcji użyto wektor indukowalny transkrypcyjnie, b) kotransformowano komórki genem czynnika Willebranda, spełniającego m.in. rolę ochronną i c) nowa konstrukcja komórek ekspresyjnych wzbogacona została antysensowym RNA obniżającym poziom ekspresji białek wiążących czynnik VIII w endoplazmatycznym *reticulum*. W wyniku tych modyfikacji układu poziom ekspresji genu czynnika VIII podniósł się kilkadziesiąt razy.

W dwu firmach farmaceutycznych zbadano cechy produktu inżynierii genetycznej – czynnika VIII. Uzyskano preparat homogenny, dobrze tolerowany przez pacjentów i o analogicznym, do naturalnego produktu, działaniu farmakologicznym. Stwierdzono również, że czynnik VIII pozbawiony domeny B nie różni się w aktywności biologicznej od pełnego białka, za to ekspresja takiego genu *in vitro* jest bardziej wydajna. Tak modyfikowany gen wprowadzono również do wektora pochodzącego z retrowirusa i uzyskano dobre wyniki transfekcji komórek ludzkich. To ostatnie doświadczenie otwiera być może perspektywy terapii genowej ludzi chorych na hemofilię A.

M.F.

Opracowano na podstawie: (1989), Nature, 342, 6246, 207–208.