1-2'88

# Przegląd pakietów programów komputerowych stosowanych w analizie kwasów nukleinowych i białek

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

## WPROWADZENIE

W ostatnich kilku latach, wraz z gwałtownym rozwojem techniki komputerowej, nastąpił intensywny wzrost zastosowań komputerów w naukach biologicznych. Wyrazem tych tendencji jest przeznaczenie trzech zeszytów czasopisma Nucleic Acids Research (1-3) na prezentację programów komputerowych napisanych dla potrzeb biologii molekularnej oraz powstanie w 1985 r. czasopisma CABIOS (Computer Applications in the Biosciences), którego zadaniem jest upowszechnianie nowych programów dotyczących problemów szeroko rozumianych nauk biologicznych (4).

Jednocześnie na rynku pojawiło się wiele pakietów programów analitycznych w wersjach przeznaczonych głównie do współpracy z mikrokomputerami. Programy te stosować można do szerokiej gamy teoretycznych i praktycznych problemów biologii molekularnej. Są one zazwyczaj proste w obsłudze, gdyż przeznaczone są dla użytkowników nie posiadających jakichkolwiek wcześniejszych doświadczeń w pracy z komputerami.

W artykule tym przedstawiony zostanie przegląd wybranych, podstawowych funkcji pełnionych przez większość pakietów i omówione zostaną skrótowo niektóre ogólne kryteria jakości oprogramowań. Ponadto bliżej przedstawione będą trzy popularne systemy programów analitycznych: IntelliGenetics, DNASTAR i IBI/PUSTELL, a także załączona zostanie lista aktualnie dostępnych pakietów programów, zawierająca ich krótką charakterystykę funkcjonalną.

## 1. Niektóre podstawowe funkcje programów analitycznych

### 1.1. Edycja sekwencji

Stosując komputer w pracach z zakresu genetyki molekularnej często zachodzi potrzeba wprowadzania nowej sekwencji do pliku lub edycji plików utworzonych wcześniej, np. wybranych sekwencji z banków danych. Funkcje te pełni edytor sekwencji czyli program, który odczytuje sekwencje, modyfikuje je zgodnie z instrukcjami użytkownika oraz wpisuje zmodyfikowane sekwencje. Istniejące edytory różnią się znacznie sposobem operowania i łatwością użytkowania. Ważną funkcją jaką może pełnić edytor sekwencji jest symulacja niektórych eksperymentów z zakresu genetyki molekularnej, np. konstruowanie nowych plazmidów poprzez dzielenie jednej lub kilku sekwencji wyjściowych na części i odpowiednie składanie fragmentów zgodnie z życzeniami użytkownika.

### 1.2. Manipulowanie sekwencjami

Jednym z przykładów jest tworzenie sekwencji komplementarnych. Mimo że cząsteczki DNA składają się z dwóch komplementarnych nici o przeciwnej polarności, w bankach sekwencji przyjęło się zachowywać tylko sekwencję jednej nici (nici + ) w kierunku od 5' do 3', który jest zgodny z kierunkiem transkrypcji i translacji. Tworzenie nici komplementarnej jest funkcją, w jaką zaopatrzone są dobre edytory sekwencji, której nić musi być brana pod uwagę w takich funkcjach analitycznych jak poszukiwanie sekwencji homologicznych lub wyszukiwanie otwartych ramek odczytu.

Bardzo typową funkcją jest translacja. Polega ona na tworzeniu sekwencji białka poprzez zamianę każdego trypletu kodonowego na odpowiedni aminokwas. W programach translacji istnieje zazwyczaj szereg opcji, gdyż każda nić sekwencji DNA może być tłumaczona w trzech fazach i wyrażana jedno- lub trójliterowym kodem aminokwasów. Większość sekwencji DNA wymaga użycia standardowego kodu genetycznego. Wyjątek stanowią jednak sekwencje mitochondrialne oraz mutanty supresorowe. Niektóre programy translacji dostosowane są do używania alternatywnych kodów genetycznych, inne takich możliwości nie posiadają.

### 1.3. Określanie składu sekwencji

Funkcja ta jest dość prosta i polega na obliczaniu częstości występowania poszczególnych zasad w sekwencji. Na tej podstawie możliwe jest wyznaczanie składu nukleotydowego

również dla nici komplementarnej, ciężaru cząsteczkowego każdej i obydwu nici. Znajomość składu sekwencji pozwala na określenie temperatury topnienia DNA w oparciu o stosunek par G-C i A-T. Bardziej zaawansowanym przykładem zastosowania analizy składu nukleotydowego jest badanie asymetrii sekwencji, tj. regionów bogatych w pary G-C lub A-T i przedstawianie rezultatów w formie graficznej.

Ponadto stosowane jest obliczanie częstości występowania każdego z 16 dwunukleotydów oraz każdego z 64 kodonów w sekwencji. Analiza kodonów posiada głębszy sens, gdy dotyczy regionów kodujących białka i wtedy oznacza ona częstotliwość występowania poszczególnych aminokwasów.

1.4. Mapowanie sekwencji

1.4.1. Analiza restrykcyjna

Jest to jedna z podstawowych funkcji analitycznych, znajdowana w ogromnej większości pakietów. Programy analizy restrykcyjnej zaopatrzone są zazwyczaj w bazy enzymów restrykcyjnych obejmujące przynajmniej ich nazwę, sekwencję rozpoznawaną i miejsce hydrolizy. Niektóre programy pracują w oparciu o bazy zawierające wszystkie znane enzymy restrykcyjne inne natomiast zawierają wyłącznie enzymy dostępne w handlu. W tym ostatnim przypadku ważną sprawą jest uaktualnianie ich bazy przez dystrybutora programu lub użytkownika. Obecnie znanych jest ponad 700 enzymów restrykcyjnych (5), podczas gdy około 25% tej liczby stanowią enzymy dostępne w handlu (ustalenia autorów artykułu). Analizę restrykcyjną można prowadzić z użyciem jednego enzymu, spośród wszystkich znajdujących się w bazie lub też dowolnie wybranej grupy enzymów. Rezultaty analizy mogą być przedstawione bądź w trybie tekstowym (lista), bądź graficznym (mapa). Niektóre pakiety posiadają też programy zdolne do tworzenia kolistych map restrykcyjnych genomów na podstawie rezultatów trawień.

1.4.2. Mapowanie innych cech sekwencji

Praktycznie każda cecha sekwencji, która zostanie odpowiednio ściśle zdefiniowana, może być mapowana w podobny sposób jak miejsca restrykcyjne czy podsekwencje. Niestety, nie zawsze sprawą prostą jest sprecyzowanie reguł opisujących właściwości takich regionów sekwencji, które pomimo pełnienia tej samej funkcji wykazują dość dużą zmienność. Dotyczy to w szczególności sekwencji promotorowych, regionów oddzielających części kodujące genów od intronów oraz miejsc wiązania mRNA do prokariotycznych i eukariotycznych rybosomów.

1.5. Porównywanie sekwencji

Poznanie nowej sekwencji DNA rodzi zazwyczaj wiele pytań, a jednym z nich jest to, czy sekwencja ta jest identyczna lub podobna do już znanych. Przedmiotem porównywania mogą być również sekwencje RNA i białek. Ponadto istotną sprawą jest czy sekwencja zawiera regiony wewnętrznie parujące się, regiony palindromowe lub wewnętrzne powtórzenia. Programy porównywania sekwencji opierają się na dwóch rodzajach algorytmów. Pierwszy, opracowany przez Wilburga i Lipmana (6) wykorzystywany jest do szybkiego poszukiwania zadanych sekwencji w dużych bazach danych sekwencyjnych np. GenBank, EMBL czy PIR. Wyszukiwane sekwencje nie muszą być identyczne z zadaną, a niewielkie różnice mogą dotyczyć różnego, lecz zdefiniowanego poziomu punktowych modyfikacji, insercji czy delecji. Poziom tolerancji określany jest przez użytkownika, a efektem działania programu jest wydruk wszystkich pozycji z przeszukiwanej bazy danych, które mieszczą się w założonym zakresie podobieństwa.

Programy opierające się na algorytmie Needlemana i Wuncha (7) pracują wolniej, jednakże dostarczają precyzyjniejszych informacji. Dotychczasowa praktyka wykazuje, że 2/3 czasu pracy dużych komputerów pracujących dla potrzeb genetyki molekularnej w USA wykorzystywane jest na porównywanie nowych sekwencji z tymi, które zawarte są już w bazach danych.

1.6. Konwersja sekwencji

W ramach każdego pakietu programów stosowany jest ten sam format zapisu sekwencji tak, że przechodząc z jednego programu w drugi zbędna jest konieczność zamiany formatów. Występuje to wówczas, gdy zamiennie stosowane są programy pochodzące z różnych pakietów. Problem ten pojawia się także, gdy analizie ma być poddana sekwencja wypisana z innego banku sekwencji niż ten z którym program współpracuje. W niektórych pakietach powyższe zagadnienia nie stanowią kwestii, gdyż zaopatrzone są one w specjalne programy konwersji formatów. W innych funkcje te może pełnić edytor sekwencji.

1.7. Przegląd funkcji realizowanych przez znane pakiety programów

Manual najnowszego wydania GenBank (Release 48.0) zawiera

dołączoną krótką charakterystykę, techniczną i funkcjonalną, ponad 50 znanych aktualnie pakietów oprogramowań. W tabeli 1 przytoczona jest część tej informacji, która dotyczy funkcji pełnionych przez poszczególne pakiety.

TABELA 1
Funkcje pełnione przez programy

1 - liczba porządkowa
2 - nazwa pakietu
3 - edit
4 - compare
5 - search
6 - translate
7 - measurments
8 - molecular structure
9 - plot display print out
10 - seq. det. toots
11 - exp. design toots

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DNA SEQUENCE ANALYSIS/Mount | X | X | X | X | X | | X | | |
| 2 | Cornell DNA Sequence Analusis Package | X | X | X | X | X | | X | | |
| 3 | Cold Spring arbor/ Roberts | | X | X | X | | | X | X | X |
| 4 | STADEN | X | X | X | X | X | X | X | X | |
| 5 | IBI/Pustell Sequence Analysis Programs | X | X | X | X | X | | X | X | X |
| 6 | MicroGenie Sequence Software | X | X | X | X | | | X | X | |
| 7 | CompuGene DNALYSIS | | X | X | X | X | | X | X | |
| 8 | MOLGEN Sequence Software Handling Program | X | X | X | X | | | | | |
| 9 | Sequence Analysis Program/Larson and Messing | X | X | X | X | X | | X | X | |
| 10 | IDEAS | | X | X | X | | X | | | |
| 11 | DNASTAR | X | X | X | X | X | X | X | X | X |
| 12 | SPAC | X | X | X | X | X | | X | | |
| 13 | S.A.S.I.P.(Sequence Analysis System of Institut Pasteur) | X | X | X | X | X | X | X | X | |
| 14 | SEQ | X | X | X | X | | X | X | X | |
| 15 | PROPIIET | X | X | X | X | X | X | X | | |
| 16 | DNATOOLS | X | X | X | X | X | | X | | |
| 17 | The DNA Inspector II | X | X | X | X | X | | X | X | |
| 18 | EMBLSCAN | | X | | | | | | | |
| 19 | GENAS | | X | X | X | X | X | X | X | |
| 20 | Protein Comparison (FASTP) | | X | X | | | | | | |
| 21 | IBM/CMS Sequence Analysis Program | | X | X | X | | X | X | | |
| 22 | MATILDA | X | | X | | | | X | | |
| 23 | ALIGN | | X | X | | | | X | | |
| 24 | DNACE-The Lilly DNA Computing Environment | X | X | X | X | X | X | X | X | |
| 25 | UWGCG - Sequence Analysis Software Package | X | X | X | X | X | X | X | X | |
| 26 | Protein Identification Resource(PIR) | X | X | X | X | X | | X | | |
| 27 | Recombinant Tootkit | X | | X | X | X | | X | | |
| 28 | Apple Sequence Analysis Program (ASAP) | | | X | | | | | | |
| 29 | BIONET: National Computer Resource for Molecular Biology | X | X | X | | X | X | X | X | X |
| 30 | GENEPRO | X | X | X | X | X | | X | | |
| 31 | DNASIS | X | X | X | X | X | | X | | |
| 32 | Nucleotide and Protein Sequence Analysis Software (NPSAS) | X | X | X | X | X | X | X | X | |
| 33 | IntelliGeneties Engineering Software Package | X | X | X | X | X | X | X | X | X |
| 34 | MTX | | X | X | X | | | X | | |
| 35 | MBIS | X | X | X | X | X | X | X | X | |
| 36 | Analyseq | X | X | X | X | | | X | | |
| 37 | SEQAID II (TM) | X | X | X | X | X | X | X | X | |
| 38 | Sequence Analysis Programs for the IBM-PC | X | X | X | X | X | | X | | |
| 39 | PCS | X | X | X | X | X | | X | | |
| 40 | Restriction Fragment Sizing and Sequence Entry Programs | | | | | | X | | X | |
| 41 | Restriction Mapping and Analysis Utilities | | | | | | | X | | |
| 42 | Evolutionary Trees | | X | | | | | | | |
| 43 | ZPEP and ZSEQ | | X | X | X | X | | | | |
| 44 | LDNA | X | | X | X | X | | X | | X |
| 45 | SSEDIT | X | | X | X | | | | X | |
| 46 | VTUTIN/NBATIN | X | | X | | | | | X | |
| 47 | Computer Pragrams by Fitch | | X | | | | | | | |
| 48 | ACNUC | X | | | | | | | | |
| 49 | SEQALIGN | | X | | | | | | X | |
| 50 | Lang and Burger^s Sequence Software | | X | X | X | | X | X | X | X |
| 51 | PROBIND PAS | | | | X | | | X | | X |
| 52 | GENEUS | X | X | X | X | | X | X | X | X |
| 53 | GENEXPERT | | X | X | X | | X | X | X | |
| 54 | SEQEXT/PEPTTR | | | | X | | X | | | |
| 55 | SEQANAL and SEQTALK | | X | X | | | X | X | | |
| 56 | General DNA analysis program for IIP Model 86/87 | X | X | X | X | X | | X | | X |
| 57 | COMPSEQ | X | X | X | X | X | X | X | X | |
| 58 | MacGene | X | X | X | X | X | X | X | X | X |

2. Niektóre kryteria jakości oprogramowań

Istnieje cały szereg kryteriów jakości oprogramowań. W różnym stopniu spełniają je poszczególne pakiety programów. Jednym z podstawowych mierników służących za podstawę oceny jest łatwość w opanowaniu użytkowania programu, czyli w jak "przyjazny dla użytkownika" (user friendly) sposób opracowany został program. Jest rzeczą naturalną, że początkujący użytkownik ma prawo popełniać błędy pracując z programem i komputerem. Istotną sprawą jest wówczas to, czy program posiada system informowania o błędach (error message), a także w jakim stopniu pomaga użytkownikowi w wyborze właściwej odpowiedzi poprzez podawanie przykładów.

Innym kryterium jest niezawodność programu; oznacza to częstotliwość zablokowania działań w konsekwencji natrafienia na zbiór nietypowych danych. Przykładem może być napotkanie dłuższej sekwencji niż możliwa do analizy przez dany program. Istotne jest wtedy, czy program pracuje dostarczając błędnych danych, czy też ostrzega i informuje użytkownika o znalezieniu zbyt długiej sekwencji, która nie może być w tej formie analizowana. Niezawodność programu zależna jest w dużym stopniu od tego czy był on testowany na obszernej grupie różnych sekwencji. Jeżeli ten warunek został spełniony istnieje duże prawdopodobieństwo, że program nie zawiedzie przy analizie kolejnych sekwencji. Wiele oprogramowań posiada górną granicę długości sekwencji, które mogą analizować. Granica ta może zmieniać się w szerokim zakresie i zależy przede wszystkim od wielkości pamięci mikrokomputera. Ograniczenie długości analizowanej sekwencji nie stanowi zbyt wielkiego problemu, gdyż większość pakietów posiada możliwość analizowania fragmentów bardzo długich sekwencji.

Pisemna dokumentacja (opis możliwości i zasad funkcjonowania programów) powinna towarzyszyć każdemu pakietowi bez względu na to, w jak "przyjazny dla użytkownika" sposób napisany został program. Dokumentacja taka stanowić powinna nie tylko przewodnik dla użytkownika, lecz także powinna tłumaczyć niektóre trudniejsze aspekty algorytmów stosowanych przez różne programy oraz wyjaśniać mniej oczywiste komendy programów i dostarczać pełnej specyfikacji zawartości pakietu. Opis taki nie powinien być jednak zbyt obszerny, gdyż wiele informacji program może posiadać "on line" w formie informacji HELP wywoływanej bezpośrednio na ekran monitora.

3. Charakterystyka wybranych pakietów programów analitycznych

W części tej bliżej przedstawione zostaną trzy bardzo popularne na rynku amerykańskim systemy analizy sekwencji DNA i białek: IntelliGenetics, DNASTAR i IBI/PUSTELL.

3.1. IntelliGenetics (8)

IntelliGenetics jest amerykańską firmą, której oprogramowanie z zakresu biologii molekularnej i biotechnologii było pierwszym, jakie ukazało się na rynku. Pakiet programów IntelliGenetics uważany jest za najpełniejszy i najwszechstronniejszy ze wszystkich dostępnych. Składa się on z ponad pięćdziesięciu indywidualnych programów, zgrupowanych w dziewięciu blokach wg podobieństwa pełnionych funkcji.

Obecnie oprogramowanie IntelliGenetics służy ponad 2,5 tys. naukowcom i ponad 350 instytucjom naukowym na całym świecie. Żaden z konkurencyjnych systemów nie posiada większego grona użytkowników.

W 1985 r. NIH (National Institutes of Health, USA) przyznał firmie fundusze na prowadzenie programu BIONET - National Computer Resource for Molecular Biology (9). Aktualnie oprogramowanie IntelliGenetics stanowi trzon biblioteki prowadzonej przez BIONET. Wraz z oprogramowaniem dostarczane są trzy główne bazy sekwencji kwasów nukleinowych i białek: GenBank, EMBL i NBRF-PIR, bazy wektorów oraz enzymów restrykcyjnych. Oprogramowanie i bazy uaktualniane są cztery razy w roku.

Programy IntelliGenetics mogą być używane na pięciu rodzajach komputerów: VMS VAX, MicroVax II, DEC 2060 i BION-Sun Microsystems.

Środowiska akademickie USA i reszty świata mogą używać oprogramowanie przyłączając się do programu BIONET. Od strony technicznej wymagane jest wtedy jedynie posiadanie terminalu w postaci mikrokomputera zaopatrzonego w przystawkę (modem) i telefon. Łączność z komputerami DEC 2060 zlokalizowanymi w Mountain View w Kalifornii odbywa się za pomocą lokalnej lub międzynarodowej sieci telekomunikacyjnej.

3.2. DNASTAR (10)

DNASTAR jest niewielką amerykańską firmą, której zadaniem jest dostarczanie kompletnego i satysfakcjonującego zaplecza mikrokomputerowego laboratoriom pracującym w dziedzinie biologii i genetyki molekularnej. Ponad 40 programów

wchodzących w skład pakietu jest rezultatem kilkuletniej pracy zespołu sześciu programistów i genetyków.

Kompletny system mikrokomputerowy oferowany przez firmę obejmuje:

- mikrokomputer IBM AT z pamięcią 640 Kb,
- monitor monochromatyczny,
- stację dysków elastycznych 360 Kb,
- stację dysków elastycznych 1.2 Mb,
- wysokiej jakości drukarkę "dot-matrix",
- SEQ - EASY digityzer z syntezatorem głosu, podświetlaczem i klawiaturą,
- twardy dysk 60 Mb,
- stację taśm typu Tallgrass 60 Mb,
- kompletne oprogramowanie składające się z 43 indywidualnych programów.

Pakiet DNASTAR zawiera dwa kompletne banki sekwencji GenBank i PIR, które podobnie jak programy są okresowo uaktualniane i za niewielką stosunkowo opłatą przesyłane na taśmie użytkownikom.

Poważną zaletą programów jest brak ograniczeń co do długości analizowanych sekwencji. Mogą one pracować z najdłuższą opublikowaną sekwencją wirusa EBV zawierającą 172.282 nukleotydy, bez konieczności dzielenia jej na fragmenty.

Istotną "wadą" kompletnego systemu jest dla potencjalnego nabywcy jego wysoka cena, około 30 tys USD, z czego połowa przypada na samo oprogramowanie.

3.3. IBI/PUSTELL (11)

IBI, International Biotechnologies Inc., jest młodą i prężną amerykańską firmą biotechnologiczną, w której ofercie handlowej - obok biochemikaliów i drobnego sprzętu labolatoryjnego - znajduje się pakiet analizy sekwencji kwasów nukleinowych i białek IBI/PUSTELL. Cały pakiet składa się z ponad trzydziestu programów. Do oprogramowania dołączona jest baza sekwencji białek PIR(NBRF), a wszystkie programy mogą współpracować z bazą GenBank w wersji taśmowej lub dyskietkowej. Długość analizowanych sekwencji ograniczona jest jedynie pamięcią stosowanego komputera. Oprogramowanie jest przewidziane do współpracy z mikrokomputerami IBM PC i innymi w pełni zgodnymi z pracującymi w systemie operacyjnym MS-DOS 2.0 lub wyższym, posiadającymi pamięć 256 Kb lub więcej. W handlu znajduje się również wersja oprogramowania dla komputerów Apple. Wszystkie programy analityczne mieszczą się na pięciu dyskach elastycznych i ich cena jest konkurencyjna, gdyż wynosi 800 USD.

---

Literatura

1. Nucleic Acids Research (1982) 10, 1-456.
2. Nucleic Acids Research (1984) 12, 1-428.
3. Nucleic Acids Research (1986) 14, 1-619.
4. CABIOS, published by IRL Press Ltd., P.O.Box 1, Eynsham, Oxford OX 8 1 JJ, England.
5. Kessler,C. and Hotke, H.J. "Specificity of Restriction Endonucleases and Methylases" Gene (1986) 47, 1-153.
6. Wilbur,W.J. and Lipman,D.J. (1983) Proc.Nat.Acad.Sci.USA 80, 726.
7. Needleman,S. and Wunsch,C. (1970) J.Mol.Biol. 48, 443.
8. IntelliGenetics,Inc 1975 El Camino Real West Mountain View, California 94040-2216, USA.
9. Smith,D.H., Brutlag,D., Friedland,P. and Kedes,L.H. "BIONET : National Computer Resource for Molecular Biology" Nucleic Acids Research (1986) 14, 17-20.
10. DNASTAR, Inc. 1801 University Ave., Madison, WI 53705, USA.
11. International Biotechnologies, Inc. 275 Winchester Avenue P.O. Box 9558, New Haven, CT 06535, USA.