

Mariusz Popenda, Jerzy Ciesiołka, Włodzimierz J. Krzyżosiak
Instytut Chemii Bioorganicznej PAN
Poznań

PROGRAM ANALIZY SEKWENCJI KWASÓW NUKLEINOWYCH NASEQ

WPROWADZENIE

Jakkolwiek celowość dysponowania nowoczesnym zapleczem komputerowym w pracach badawczych z zakresu biologii i genetyki molekularnej nie jest obecnie przez nikogo kwestionowana i w krajowych laboratoriach pojawiło się ostatnio wiele wysokiej klasy mikrokomputerów, kontrowersyjna pozostaje sprawa, czy korzystniejsze jest prowadzenie badań w oparciu o istniejące już gotowe pakiety programów, czy też pisanie oprogramowania we własnym zakresie.

Gwałtowny rozwój badań w dziedzinie genetyki molekularnej, ciągłe pojawianie się nowych problemów, podejść i technik badawczych powoduje szybkie przekraczanie ram elastyczności wielu gotowych programów i chociaż ambicja dystrybutorów tych opracowań jest ciągłe ich uaktualnienie lub stworzenie takich możliwości użytkownikowi, korzystanie z gotowych programów nie wydaje się rozwiązaniem optymalnym. Przemawia za tym również mnogość istniejących obecnie programów opracowanych przez wyspecjalizowane firmy komputerowe i wiele ośrodków uniwersyteckich krajów wysoko rozwiniętych. Obserwuje się silną tendencję do posiadania własnych niezależnych opracowań programowych, gdyż stopień zaangażowania technik komputerowych w proces badawczy

decyduje w znacznej mierze o efektywności prowadzonych badań, a nierzadko o ich powodzeniu. Wyniki analiz komputerowych są często stymulatorami nowych interesujących kierunków badań, a symulacja komputerowa różnych wariantów eksperymentu pozwala na wybór optymalnej drogi jego przeprowadzenia. Można więc stwierdzić, że stopień komputeryzacji badań naukowych we wspomnianych dziedzinach w poszczególnych ośrodkach naukowych decyduje w znacznej mierze o konkurencyjności tych badań.

Powyższe fakty przemawiają za celowością pisania własnych programów, elastycznych i ciągle uaktualnianych, dostosowanych do potrzeb badawczych placówki naukowej czy środowiska, szybko reagujących na nowe pomysły i rozwiązania.

Prezentowany w tym opracowaniu pakiet programów analizy sekwencji kwasów nukleinowych NASEQ stanowi pierwszy krok na drodze do stworzenia własnego, komputerowego zaplecza analitycznego dla potrzeb biologii i genetyki molekularnej. Powstaje on w Zakładzie Chemii Bioorganicznej PAN na zaproszenie pracujących tam zespołów badawczych. Założeniem autorów tego przedsięwzięcia jest udostępnienie aktualnego niewielkiego pakietu, a w przyszłości jego rozszerzonych wersji wszystkim zainteresowanym placówkom badawczym.

W artykule tym przedstawiony zostanie aktualny kształt programu NASEQ, jego wymagania dotyczące sprzętu komputerowego i zasada funkcjonowania.

1. Opis programu NASEQ

1.1. Struktura plików

Program NASEQ w jego aktualnie prezentowanej wersji składa się z 13 programów użytkowych, programu systemu menu i pliku README.NAS z krótkim opisem. Pliki wynikowe programów użytkowych, które mogą być dalej przetwarzane mają zarezerwowane nazwy rozszerzeń. Podane są one w tabeli.

Tabela

Typ pliku	Objaśnienie
NAS	Pliki wynikowe programów: CHANGE1, CHANGE2, CHANGE3 i MANIPUL. Zawierają sekwencje kwasów nukleinowych przedstawione w odpowiednim formacie
RLS	Pliki wynikowe programu RESSEAR, które podają nazwy enzymów, sekwencje rozpoznawane oraz miejsca restrykcyjne w analizowanej sekwencji
TRN	Pliki wynikowe programu TRANSL zawierające wynik translacji sekwencji zasad na łańcuch aminokwasów zapisany w kodzie jednoliterowym

Cały program funkcjonalny pakietu NASEQ podzielić można na dwie części. W pierwszej zawarte są programy umożliwiające tworzenie własnego zbioru plików typu ".NAS" z zapisanymi sekwencjami kwasów nukleinowych. Druga część stanowią programy analityczne, które opisane zostaną w dalszej części tego artykułu.

1.2. Programy do tworzenia własnej bazy analizowanych sekwencji DNA

Sekwencje zasad analizowane przez program NASEQ muszą być odpowiednio zakodowane i zapisane w pamięci zewnętrznej komputera w postaci pliku typu ".NAS".

Do zapisu sekwencji przyjęto następujące symbole:

- "A" - dla adenozyiny,
- "C" - dla cytydyny,
- "G" - dla guanozyiny,
- "T" - dla tymidyny lub urydyny,
- "1" - dla oznaczenia końca sekwencji liniowej,
- "2" - dla oznaczenia końca zapisu sekwencji cyklicznej.

Inne znaki, które bardzo rzadko występują w sekwencjach pochodzących z banków danych traktowane są przez programy analityczne pakietu NASEQ jako niezidentyfikowane nukleotydy. Pierwszy element pliku typu ".NAS" składający się z osiemdziesięciu znaków może zawierać dowolny tekst wprowadzony przez użytkownika (np. dodatkowe informacje dotyczące badanej sekwencji). Do tworzenia plików typu ".NAS" służą cztery programy użytkowe o nazwach: CHANGE1, CHANGE2, CHANGE3, MANIPUL.

Programy CHANGE1 i CHANGE2 umożliwiają bezpośrednie korzystanie ze zbiorów sekwencji DNA zawartych w bazach GenBank i EMBL. Programy te przetwarzają automatycznie sekwencje zawarte w tych bazach danych na kod, który jest akceptowany przez pozostałe programy analityczne pakietu NASEQ. Program CHANGE1 wykonuje automatycznie następujące operacje:

- wyszukuje sekwencje z GenBank po uprzednim wprowadzeniu nazwy grupy oraz podaniu nazwy sekwencji, pod która skatalogowana jest w bazie danych,
- przetwarza wyszukana sekwencje na własny kod,
- dodaje znak określający koniec sekwencji,
- umożliwia wprowadzenie dodatkowego tekstu z klawiatury (do

osiemdziesięciu znaków),

- całość zapisuje do pliku wynikowego typu ".NAS".

Podobne operacje wykonuje program CHANGE2 z tą tylko różnicą, że pliki źródłowe pochodzą z bazy EMBL.

Użytkownik może również utworzyć plik na dowolnym edytorze tekstu (np. WORDSTAR, PCWRITE, EDLIN, CHIWRITER), a następnie przetworzyć go przez program CHANGE3, który wykonuje następujące operacje:

- zamienia wszystkie małe litery na duże,
- kasuje wszystkie znaki interpunkcyjne, sterujące, symbole specjalne i cyfry,
- w zależności od sekwencji wprowadza na końcu łańcucha znak "1" lub "2",
- umożliwia wprowadzenie z klawiatury krótkiego opisu do osiemdziesięciu znaków i zapisuje na dysku plik z rozszerzeniem ".NAS".

Do wprowadzania własnej sekwencji do bazy danych służy program MANIPUL. Jest on procesorem tekstu ukierunkowanym na edycję kwasów nukleinowych. Może pracować w trybie INSERT lub OVERWRITE.

W trybie INSERT znaki znajdujące się od pozycji kursora są przesuwane w lewo, a nowe znaki wprowadzane z klawiatury lub z dysku (opcja MERGE) są umieszczane w aktualnym położeniu kursora. W trybie OVERWRITE nowe znaki są stawiane na istniejące znaki, co powoduje zastąpienie starego łańcucha znaków na nowy.

Program MANIPUL posiada następujące opcje:

- wprowadzanie zawartości pliku lub jego fragmentu do pamięci operacyjnej komputera z jednoczesnym usunięciem starej sekwencji (LOAD),

- zachowanie całej sekwencji lub wybranego fragmentu na dysku w postaci pliku typu ".NAS" (SAVE),
- kasowanie pliku z dysku (KILL),
- wymazywanie fragmentu lub całej sekwencji z pamięci operacyjnej komputera (DELETE),
- dołączanie do sekwencji znajdującej się w pamięci operacyjnej nowej sekwencji zapisanej uprzednio na dysku (MERGE).
- wyszukiwanie w danej sekwencji dowolnego łańcucha zasad o długości do pięćdziesięciu znaków (FIND),
- wyświetlanie zawartości dysku (FILES),
- przeglądanie zawartości pliku ".NAS"(DISPLAY),
- drukowanie całej sekwencji DNA lub jej fragmentu w trzech różnych formatach (PRINT)(rys.2)
- wyprowadzenie programu do menu (END).

Sequence pBR322 (fragment)

1. TTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATGCGGTAGTTTATCACAGTTAAATTGCTAAC
 71. AGTCAGGCACCGTGTATGAAATCTAACAATGCGCTCATCGTCATCCTCGGCACCGTCACCCTGGATGC
 141. TAGGCATAGGCTTGGTTATGCCGGTACTGCCGGGCTTTCGCGGGATATCGTCCATTCCGACAGCATC
 211. CAGTCACTATGGCGTGTCTAGCGCTATATGCGTTGATGCAATTTCTATGCGCACCCGTTCTCGGAG
 281. CTGTCCGACCGCTTTGGCCGCGCCAGTCCTGCTCGCTTTCGCTACTTGGAGCCACTATCGACTACGC
 351. TCATGGCGACCACACCCGTCCTGTGGATCCTCTACGCCGGACGCATCGTGGCCGGCATCACCGGCGCC
 421. AGGTGCGGTTGCTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGC
 491. ATGAGCGCTTGTTCGGCGTGGGTATGGTGGCAGGCCCGTGGCCGGGGGACTGTTGGGCGCCATCTC
 561. TGCATGCACCATTCTTTCGGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATG
 631. GGAGTCGCATAAGGGAGAGCGTCGACCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGT
 701. GCGCGGGGCATGACTATCGTCGCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACAGGT
 771. CGGCAGCGCTCTGGGTCAATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGATGATCGGCCTGTCTG
 841. TGCGGTATTTCGGAATCTTGC

Sequence pBR322 (fragment)

```

*           *           *           *           *           *
1. TTCTCATGTTTGACAGCTTATCATCGATAAGCTTTAATGCGGTAGTTTATCACAGTTAAATTGCTAAC
71. AGTCAGGCACCGTGTATGAAATCTAACAATGCGCTCATCGTCATCCTCGGCACCGTCACCCTGGATGC
141. TAGGCATAGGCTTGGTTATGCCGGTACTGCCGGGCTCTTGGCGGGATATCGTCCATTCCGACAGCATC
211. CAGTCACTATGGCGTGCTGCTAGCGCTATATGCGTTGATGCAATTTCTATGCGCACCCGTTCTCGGAG
281. CTGTCCGACCGCTTTGGCCGCGCCAGTCCTGCTCGCTTCTGCTACTTGGAGCCACTATCGACTACGC
351. TCATGGCGAACCACACCCGTCCTGTGGATCCTCTACGCCGAGCATCGTGGCCGGCATCACCCGGCC
421. AGGTGCGGTTGCTGGCGCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGC
491. ATGAGCGCTTGTTCGGCGTGGGTATGGTGGCAGGCCCGTGGCCGGGGGACTGTTGGGGCGCCATCTC
561. TGCATGCACCATTCCTTGGCGCGGCGGTGCTCAACGGCCTCAACCTACTACTGGGCTGCTTCCTAATG
631. GGAGTCGCATAAGGGAGAGCGTCGACCCGATGCCCTTGAGAGCCTTCAACCCAGTCAGCTCCTTCCGGT
*           *           *           *           *           *
701. GCGCGGGGCATGACTATCGTCGCCGCACTTATGACTGTCTTCTTTATCATGCAACTCGTAGGACAGGT
771. CGGCAGCGCTCTGGGTCAATTTTCGGCGAGGACCGCTTTCGCTGGAGCGCGACGATGATCGGCCTGTGC
841. TCGGGTATTCGGAATCTTGC

```

Sequence pBR322 (fragment)

```

1   TTCTCATGTT TGACAGCTTA TCATCGATAA GCTTTAATGC GGTAGTTTAT CACAGTTAAA
61  TTGCTAACGC  AGTCAGGCAC CGTGTATGAA ATCTAACAAT GCGCTCATCG TCATCCTCGG
121 CACCGTCACC  CTGGATGCTG TAGGCATAGG CTTGGTTATG CCGGTAAGTC CGGGCCTCTT
181 GCGGGATATC  GTCCATTCCG ACAGCATCGC CAGTCACTAT GCGGTGCTGC TAGCGCTATA
241 TGCGTTGATG  CAATTTCTAT GCGCACCCGT TCTCGGAGCA CTGTCCGACC GCTTTGGCCG
301 CCGCCAGTC   CTGCTCGCTT CGCTACTTGG AGCCACTATC GACTACGCGA TCATGGCGAC
361 CACACCGGTC  CTGTGGATCC TCTACGCCGG ACGCATCGTG GCCGGCATCA CCGGCGCCAC
421 AGGTGCGGTT  GCTGGCGCCT ATATCGCCGA CATCACCGAT GGGGAAGATC GGGCTCGCCA
481 CTTCGGGCTC  ATGAGCGCTT GTTTCGGCGT GGGTATGGTG GCAGGCCCGG TGGCCGGGGG
541 ACTGTTGGGC  GCCATCTCCT TGCATGCACC ATTCCTTGCG GCGGCGGTGC TCAACGGCCT
601 CAACCTACTA  CTGGGCTGCT TCCTAATGCA GGAGTCGCAT AAGGGAGAGC GTCGACCGAT
661 GCCCTTGAGA  GCCTTCAACC CAGTCAGCTC CTTCGGGTGG GCGCGGGGCA TGACTATCGT
721 CGCCGCACTT  ATGACTGTCT TCTTTATCAT GCAACTCGTA GGACAGGTGC CGGCAGCGCT
781 CTGGGTCAAT  TTCGGCGAGG ACCGCTTTCG CTGGAGCGCG ACGATGATCG GCCTGTGCTC
841 TCGGGTATTC  GGAATCTTGC

```

Rys. 2. Wydruk z programu MANIPUL fragmentu sekwencji pBR322 w trzech różnych formatach.

1.3. Programy analityczne

W zależności od funkcji programy analityczne znajdujące się w oprogramowaniu NASEQ można podzielić na trzy grupy:

- analizy restrykcyjnej,
- translacji,
- obliczania częstości występowania poszczególnych zasad w sekwencji.

Programy do analizy sekwencji DNA o restrykcyjnej

Analizę restrykcyjną można przeprowadzić za pomocą programu SITELOOK lub programu RESSEAR. Program SITELOOK jest szczególnie przydatny, gdy przeprowadzamy analizę krótkiego fragmentu DNA. Program wyprowadza na monitor lub drukarkę analizowaną sekwencję z numeracją oraz zaznacza na niej występowanie miejsc restrykcyjnych (rys.3).

```

Sequence name: HUMUR1A

10      20      30      40      50      60      70
*      *      *      *      *      *      *
GATACTTACCTGGCAGGGGAGATACCATGATCACGAAGGTGGTTTTCCAGGGCGAGGCTTATCCATTGC
^  ^      ^  ^      ^  ^      ^  ^      ^  ^      ^  ^
ApyI      BclI      ApyI
BstNI      DpnI      BstNI
EcoRII      MboI      EcoRII
MvaI      NdeII      MnlI
ScrFI      NlaIII      MvaI
          Sau3AI      ScrFI

80      90      100      110      120      130      140
*      *      *      *      *      *      *
ACTCCGGATGTGCTGACCCCTGCGATTTCCCCAAATGTGGGAAACTCGACTGCATAATTTGTGGTAGTGG
^  ^      ^  ^      ^  ^      ^  ^      ^  ^
AccIII      FokI      TagI
BspMII      TthHB8I
HapII
HpaII
MspI

```

Rys.3. Przykładowy wydruk analizy restrykcyjnej z programu SITELOOK dla fragmentu sekwencji HUMUR1A.

Drugi program do analizy restrykcyjnej o nazwie RESSEAR zapisuje wszystkie wyszukane miejsca restrykcyjne dla poszczególnych enzymów łącznie z nazwami i sekwencjami rozpoznawanymi do pliku typu ".RSL". Pliki te mogą być dalej przetwarzane przez programy: SITELIST, SITEMAP i FRAGSIZE.

SITELIST odczytuje z pliku ".RSL" i wyprowadza na ekran

lub drukarkę informację zawierającą nazwy enzymów, które były używane w analizie, ich sekwencje rozpoznawane i miejsca restrykcyjne w danej sekwencji DNA (rys.4).

AccII	CG/CG	1	3	154	
AccIII	T/CCGGA	1	3	74	
ApyI	CC/WGG	2	3	11	50
BclI	T/GATCA	1	3	29	
BspMII	T/CCGGA	1	3	74	
BstNI	CC/WGG	2	3	11	50
CfoI	GCG/C	1	3	156	
DpnI	GA/TC	1	3	31	
EcoRII	/CCWGG	2	3	9	48
FnuDII	CG/CG	1	3	154	
FokI	GGATG(9/13)	1	3	89	
HapII	C/CGG	1	3	75	
HhaI	GCG/C	1	3	156	
HinP1I	G/CGC	1	3	154	
HpaII	C/CGG	1	3	75	
MboI	/GATC	1	3	29	
MnlI	CCTC(7/7)	1	3	51	
MspI	C/CGG	1	3	75	
MvaI	CC/WGG	2	3	11	50
NdeII	/GATC	1	3	29	
NlaIII	CATG/	1	3	30	
Sau3AI	/GATC	1	3	29	
ScrFI	CC/NGG	2	3	11	50
TagI	T/CGA	1	3	117	
ThaI	CG/CG	1	3	154	
TthHB8I	T/CGA	1	3	117	

Rys. 4. Przykładowy wydruk analizy restrykcyjnej z programu SITEMAP dla całej sekwencji HUMURA.

Program SITELIST może wprowadzać wyniki analizy restrykcyjnej dla całej sekwencji lub dla wybranego przez użytkownika fragmentu sekwencji DNA w postaci tabelarycznej. Program SITELIST umożliwia wprowadzenie wyników przeprowadzonej analizy restrykcyjnej dla wszystkich enzymów restrykcyjnych, enzymów pozostawiających lepkie końce po stronie 3', enzymów pozostawiających lepkie końce po stronie 5',

Non-cutting enzymes

AatI	AvrII	BstI	EcoO109	HinfII	NlaIV	PvuII	SspI
AatII	BalI	BstXI	EcoRI	HpaI	NotI	RsaI	SstI
AccI	BamHI	CauII	EcoRV	HphI	NruI	RsrII	SstII
AcyI	BanI	Cfr10I	EspI	KpnI	NsiI	SacI	StuI
AflIII	BanII	Cfr13I	Fnu4HI	MaeI	NspBII	SacII	StyI
AhaII	BanIII	CfrI	FspI	MaeII	NspI	SalI	Tth111I
AhaIII	BbeI	ClaI	GdiII	MaeIII	NspII	Sau96I	XbaI
AluI	BbiII	CvnI	HaeII	MboII	NspIII	SauI	XcyI
AosI	BbvI	DdeI	HaeIII	MluI	NspIV	ScaI	XhoI
ApaI	BcnI	DraI	HgaI	MstI	NspV	SfaNI	XhoII
ApaLI	BglI	DraII	HgiAI	MstII	PaeR7	SfiI	XmaI
Asp700	BglIII	DraIII	HgiCI	NaeI	PalI	SinI	XmaIII
Asp718	BsmI	EaeI	HgiDI	NarI	PfMI	SmaI	XmnI
AsuI	Bsp128	EagI	HincII	NciI	PpuMI	SnaBI	XorII
AsuII	BspMI	Eco47I	HindII	NcoI	PssI	SpeI	
AvaI	BssHII	Eco47IIII	HindIII	NdeI	PstI	SphI	
AvaII	BstEII	Eco81I	HinfI	NheI	PvuI	SplI	

Rys. 4. Przykładowy wydruk analizy restrykcyjnej z programu SITEMAP dla całej sekwencji HUMUR1A.

Program SITELIST może wyprowadzać wyniki analizy restrykcyjnej dla całej sekwencji lub dla wybranego przez użytkownika fragmentu sekwencji DNA w postaci tabelarycznej.

Program SITELIST umożliwia wyprowadzenie wyników przeprowadzonej analizy restrykcyjnej dla:

- wszystkich enzymów restrykcyjnych,
- enzymów pozostawiających lepkie końce po stronie 3',
- enzymów pozostawiających lepkie końce po stronie 5',

- enzymów pozostawiających tępe końce,
- enzymów rozpoznających sekwencje niepalindromowe,
- dowolnie wybranych enzymów,
- enzymów przecinających daną sekwencję w n miejscach (gdzie n=0,1,2...).

Program SITEMAP przedstawia wyniki w postaci graficznej (rys.5). Posiada podobne opcje jak program SITELIST.

Sequence name: HUMUR1A

Analysed range:	from base number	1	to base number	165	step:3	
Enzymes	30	60	90	120	150	180
AccII	-----]	-----]	-----]	-----]	-----]	*-----]
AccIII	-----]	-----]	*-----]	-----]	-----]	-----]
ApyI	---*-----]	-----*--]	-----]	-----]	-----]	-----]
BclI	-----*-----]	-----]	-----]	-----]	-----]	-----]
BspMII	-----]	-----]	*-----]	-----]	-----]	-----]
BstNI	---*-----]	-----*--]	-----]	-----]	-----]	-----]
CfoI	-----]	-----]	-----]	-----]	-----]	*-----]
DpnI	-----]	*-----]	-----]	-----]	-----]	-----]
EcoRII	--*-----]	-----*--]	-----]	-----]	-----]	-----]
FnuDII	-----]	-----]	-----]	-----]	-----]	*-----]
FokI	-----]	-----]	-----*--]	-----]	-----]	-----]
HapII	-----]	-----]	*-----]	-----]	-----]	-----]
HhaI	-----]	-----]	-----]	-----]	-----]	*-----]
HinP1I	-----]	-----]	-----]	-----]	-----]	*-----]
HpaII	-----]	-----]	*-----]	-----]	-----]	-----]
MboI	-----*-----]	-----]	-----]	-----]	-----]	-----]
MnlI	-----]	-----*--]	-----]	-----]	-----]	-----]
MspI	-----]	-----]	*-----]	-----]	-----]	-----]
MvaI	---*-----]	-----*--]	-----]	-----]	-----]	-----]
NdeII	-----*-----]	-----]	-----]	-----]	-----]	-----]
NlaIII	-----*-----]	-----]	-----]	-----]	-----]	-----]
Sau3AI	-----*-----]	-----]	-----]	-----]	-----]	-----]
ScrFI	---*-----]	-----*--]	-----]	-----]	-----]	-----]
TagI	-----]	-----]	-----]	-----*--]	-----]	-----]
ThaI	-----]	-----]	-----]	-----]	-----]	*-----]
TthHB8I	-----]	-----]	-----]	-----*--]	-----]	-----]

Rys.5. Przykładowy wydruk analizy restrykcyjnej z programu SITEMAP dla całej sekwencji HUMUR1A.

Program FRAGSIZE na podstawie danych zawartych w plikach ".RLS" wyznacza długości fragmentów powstałych w wyniku działania poszczególnych enzymów restrykcyjnych lub grupy enzymów na cząsteczkę DNA (rys.6). Program posiada podobne

opcje do programów SITELIST i SITEMAP.

13	AsuI		G/GNCC				
1	1461	1951 ...	3411	8	191	4344 ...	172
2	616	3730 ...	4345	9	189	1762 ...	1950
3	352	147 ...	525	10	179	1262 ...	1440
4	279	1483 ...	1761	11	124	1138 ...	1261
5	275	526 ...	800	12	88	801 ...	888
6	249	889 ...	1137	13	79	3412 ...	3490
7	222	3508 ...	3729	14	42	1441 ...	1482
				15	17	3491 ...	3507
51	FnuDII		CG/CG				
1	581	2525 ...	3104	12	122	1418 ...	1539
2	493	3435 ...	3927	13	115	705 ...	819
3	452	4259 ...	347	14	103	2080 ...	2182
4	372	1637 ...	2008	15	97	1540 ...	1636
5	356	349 ...	704	16	69	2009 ...	2077
6	341	2183 ...	2523	17	66	1042 ...	1107
7	332	3928 ...	4259	18	61	981 ...	1041
8	330	3105 ...	3434	19	27	949 ...	975
9	145	1247 ...	1391	20	26	1392 ...	1417
10	129	820 ...	948	21	10	1237 ...	1246
11	129	1108 ...	1236	22	5	976 ...	980
				23	2	2078 ...	2079

Rys.6. Przykładowy wydruk analizy restrykcyjnej dla całej sekwencji pBR322 z programu FRAGSIZE dla dwóch wybranych enzymów.

Programy RESSEAR i SITELOOK wykorzystują w trakcie wykonywania analizy restrykcyjnej plik o nazwie "ENZYMES.ALL", który jest zestawieniem wszystkich dostępnych handlowo enzymów restrykcyjnych i zawiera nazwy tych enzymów, ich sekwencje rozpoznawane z zaznaczonymi miejscami hydrolizy. Do oznaczenia tych sekwencji zastosowano następującą symbolikę przyjętą przez IUB:

- "A" dla adenozyiny,
- "C" dla cytydyny,
- "G" dla guanozyiny,
- "T" dla tymidyny (urydyny),
- "M" dla adenozyiny lub cytydyny,
- "K" dla guanozyiny lub tymidyny (urydyny),
- "Y" dla cytydyny lub tymidyny (urydyny),
- "R" dla adenozyiny lub guanozyiny,
- "W" dla adenozyiny lub tymidyny (urydyny),
- "S" dla cytydyny lub guanozyiny,
- "D" dla adenozyiny lub guanozyiny lub tymidyny (urydyny),
- "H" dla adenozyiny lub cytydyny lub tymidyny (urydyny),
- "N" dla niezidentyfikowanej zasady,
- "/" dla oznaczenia miejsca cięcia,

Dla sekwencji niepalindromowych wprowadzono dodatkowe oznaczenia. Na przykład sekwencja rozpoznawana przez enzym BbvI ma postać GCAGC(8/12). Taki zapis jest równoznaczny sekwencjom GCAGC>NNNNNNNN/ i /NNNNNNNNNNNGCTGC, które są wyszukiwane w analizowanej cząsteczce DNA. Użytkownik za pomocą programu EDITENZ może uaktualniać listę enzymów restrykcyjnych, która zapisana jest w pliku ENZYMES.ALL. Program EDITENZ ma następujące opcje:

- wyprowadzenie na ekran lub drukarkę aktualnej listy enzymów ułożonej alfabetycznie według nazw lub sekwencji rozpoznawanych,
- wprowadzenie nowego enzymu,
- skasowanie enzymu z listy,
- wyszukiwanie enzymu z listy po nazwie lub sekwencji,
- wydruk enzymów pozostawiających lepkie końce po stronie 3',
- wydruk enzymów pozostawiających lepkie końce po stronie 5',
- wydruk enzymów o tępych końcach,
- wydruk enzymów, które rozpoznają sekwencje niepalindromowe.

Programy do translacji (w trakcie realizacji)

Do translacji przewidywane są dwa programy o nazwach: TRANS i BACKTRAN. Pierwszy z nich przeznaczony jest do przeprowadzenia translacji dla całej sekwencji lub jej fragmentu na łańcuch aminokwasów. Wynik translacji zapisany w kodzie jednoliterowym może być zachowany w pliku o nazwie z rozszerzeniem ".TRN". Program przewiduje również możliwość wyprowadzenia wyniku translacji w kodzie trójliterowym dla oznaczenia poszczególnych aminokwasów na ekran lub drukarkę. Translacja przeprowadzana jest według ogólnego kodu genetycznego w trzech ramkach odczytu. Drugi program BACKTRAN przeprowadza translację odwrotną.

Program FREQBASE

Program ten służy do analizy statystycznej składników sekwencji. Program FREQBASE wylicza liczbę poszczególnych zasad, par zasad i kodonów dla całej sekwencji DNA lub wybranego fragmentu. Wyniki są wyprowadzane na ekran lub drukarkę w liczbach bezwzględnych lub w procentach.

1.4. Program systemu menu

Dostęp do poszczególnych programów użytkowych odbywa się poprzez dwustopniowy system menu. Po inicjacji z systemu operacyjnego MS-DOS programu NASEQ na ekranie pojawia się MAIN MENU. Użytkownik po naciśnięciu odpowiedniego klawisza przechodzi do menu podrzędnego z którego może wywołać poszczególne programy użytkowe. Każdy program użytkowy ma wyjście do tego menu z którego został wywołany. Program

wychodzi również do systemu menu, gdy w trakcie jego wykonywania pojawił się błąd typu I/O. Program systemu menu posiada dodatkowo 10 aktywnych klawiszy funkcyjnych, które ułatwiają obsługę całego systemu. Schemat systemu menu przedstawiony jest na rys.7.

MAIN MENU

(Nucleic acid sequence analysis software)

- 1 - SEQUENCE LIBRARY FORMATION
- 2 - RESTRICTION ANALYSIS
- 3 - TRANSLATION MENU
- 4 - NUKLEOTIDE FREQUENCIES

SEQUENCE LIBRARY FORMATION

(specify source of sequence)

- 1 - GenBank CHANGE1
- 2 - EMBL CHANGE2
- 3 - File of text CHANGE3
- 4 - EDYTOR MANIPUL

RESTRICTION ANALYSIS

(Nucleic acid sequence analysis software)

- 1 - SEARCH RESTRICTION SITES AND CREATE *.LSR FILE RES_SEAR
- 2 - PRINT RESTRICTION SITES FROM *.LSR FILE SITELIST
- 3 - DISPLAY RESTRICTION SITES GRAPHICALLY FROM *.LSR FILE SITEMAP
- 4 - DISPLAY RESTRICTION SITES FROM *.LSR FILE ON SEQUENCE SITELOOK
- 5 - DISPLAY FRAGMENTS SIZE FROM *.LSR FILE
- 6 - EDIT RESTRICTION ENZYME DATA

TRANSLATION MENU

(Nucleic acid sequence analysis software)

- 1 - CREATE *.TRN FILE OF TRANSLATION PRODUKT TRANSL
- 2 - DISPLAY TRANSLATION FROM *.TRN AND SEQUENCE TRANSL
- 3 - BACK TRANSLATION *.TRN FROM *.TRN FILE BACKTRAN

NUCLEOTIDE FREQUENCIES

(Nucleic acid sequence analysis software)

- 1 - BASE COMPOSITION
- 2 - DINUCLEOTIDE FREQUENCIES
- 3 - CODON USAGE
- 4 - SHOW AT/GC RICH REGIONS

Rys.7. Schemat systemu menu.

1.5. Oprzyrządowanie

Program NASEQ napisany został w języku GWBASIC w systemie operacyjnym MS-DOS (wersja 2.11) na mikrokomputerze Olivetti M24 z pamięcią RAM 256 KB. i wymaga następującego oprzyrządowania:

- mikrokomputer IBM PC XT/AT lub kompatybilny z kolorową kartą graficzną CGA lub EGA,
- kolorowy monitor z wysoką rozdzielczością,
- dwie stacje dysków lub jedna z dodatkowym sztywnym dyskiem,
- drukarkę typu Epson.

Wspomniany program NASEQ jest przeznaczony do zastosowania w laboratoriach. Pięć najbardziej popularnych programów umożliwiających zakładanie osobistych baz danych zostało szczegółowo omówione w Science (1).

Celem tego artykułu jest przedstawienie komputerowego systemu informacji naukowej (INFOSYSTEM) opracowanego i funkcjonującego od 1988 r. w Zakładzie Chemii Biorганиcznej w Poznaniu.

Jedyną, jak dotychczas, poważnie dostępną w Polsce i stosunkowo pełnym źródłem informacji bibliograficznych jest seria Current Contents (CC) - ISI Philadelphia USA. Wydawnictwo to jest dobrze znane i, jak się wydaje, nie wymaga omówienia.

Opracowany przez nas program komputerowy umożliwia założenie bazy danych bibliograficznych, na którą składają się informacje wyspecjalizowane z CC przez pracowników naukowych. Następnie zostają one wprowadzone do pamięci komputera w odpowiednio przygotowanej postaci oraz uporzędkowane poprzez nadanie numerów katalogowych, a także