МЕТОДИ БІОФІЗИЧНИХ ДОСЛІДЖЕНЬ

УДК577.3

КАК ПОЛУЧИТЬ НУКЛЕОТИДНЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ ДНК ИЗ БАЗЫ ДАННЫХ ГЕНЕТИЧЕСКОГО БАНКА

Д.Р. Дуплий [1], В.В. Калашников [2], Н.А Чащин[1]

[1] *Институт молекулярной биологии и генетики НАНУ, ул. Заболотного 150, Киев, duplijd@gmail.com*

[2] *Харьковский национальный экономический университет, пр. Ленина 9А, Харьков 61001, hw@ksue.edu.ua*

Поступила в редакцию 1 апреля 2010г.

Принята 4 июня 2010 г.

Предложен метод извлечения нуклеотидных последовательностей мРНК генов человека из данных внутреннего формата генетического банка Национального центра биотехнологической информации США (NCBI [1]). Метод основан на применении регулярных выражений языка PERL на основе расширенных нормальных форм Бэкуса-Науэра. Предлагаемый в работе набор регулярных выражений, может быть использован для анализа последовательностей других геномов, представленных в формате “gbk” (Gene Bank flat format).

КЛЮЧЕВЫЕ СЛОВА: NCBI, локус, мРНК, нуклеотидные последовательности, интроны, экзоны.

ЯК ОТРИМАТИ НУКЛЕОТИДНІ ПОСЛІДОВНОСТІ ДНК З БАЗИ ДАНИХ ГЕНЕТИЧНОГО БАНКУ

Д.Р. Дуплій [1], В.В. Калашніков [2], Н.А Чащин[1]

[1] *Інститут молекулярної біології і генетики НАНУ, вул. Заболотного 150, Київ, duplijd@gmail.com*

[2] *Харківський національний економічний університет, ін. Ленина 9А, Харків 61001, hw@ksue.edu.ua*

Запропонований метод вилучення нуклеотидних послідовностей мРНК генів людини з даних внутрішнього формату генетичного банку Національного центру біотехнологічної інформації США (NCBI [1]). Метод заснований на застосуванні регулярних виразів мови PERL на основі розширених нормальних форм Бэкуса-Науэра. Пропонований в роботі набір регулярних виразів, може бути використаний для аналізу послідовностей інших геномів, представлених у форматі "gbk" (Gene Bank flat format).

КЛЮЧОВІ СЛОВА: NCBI, локус, мРНК, нуклеотидні послідовності, інтрони, екзони

HOW TO OBTAIN THE NUCLEOTIDE SEQUENCES FROM GENE BANK DATA BASE?

D.R. Duplij [1]., V.V. Kalashnikov [2]., N.A. Chashyn [1]

[1]*Institute of Molecular Biology and Genetics, 150 Ac. Zabolotny Str., 03143 Kiev, Ukraine, duplijd@gmail.com*

[2] *Kharkiv National Economic University, 9A Lenina Street, 61077 Kharkiv, Ukraine, hw@ksue.edu.ua*

A method of extraction nucleotide sequences of human mRNA from Gene Bank flat format of NCBI was proposed. The method is based on using regular expressions PERL language on the basis of Extended Backus–Naur Forms (EBNF). Suggested for work set of regular expressions could be used for analysis other genomes' sequences that are presented in Gene Bank flat format (gbk).

KEY WORDS: NCBI, contig, mRNA, CDS, nucleotide sequences, introns, exons

На сегодняшний день, практически ни одно исследование в области молекулярной биологии, не обходится без компьютерного анализа последовательностей ДНК. Информация о геномных ДНК, представленных в виде электронных баз данных [2, 3] позволяет изучать структуру и функцию первоосновы наследственности.

Для работы с нуклеотидными последовательностями имеется ряд программных средств выполняющих, например, поиск сходства (BLAST_N), выявление семейств повторов (RepeatMasker), анализ частот встречаемости кодонов (CodonW), а также

инструменты на базе языков C++ [4], MATLAB [5, 6] и пр. [7]. В то же время, каждая программа предназначена для решения той или иной узкой задачи, и наверно, каждый исследователь, сталкивался с проблемой извлечения фрагментов ДНК из геномного локуса. Предлагаемые инструменты, громоздки и имеют некоторые ограничения. Целью данной работы явилась разработка программного подхода для извлечения заданных нуклеотидных последовательностей из файлов геномной ДНК генбанка. Для решения выбран язык PERL (Practical Extraction and Reporting Language), имеющий синтаксис похожий на C++, но также объединяющий лучшие приемы некоторых других языков [8], что делает PERL исключительно эффективным при работе с символьными последовательностями. Кроме того PERL предоставляет пользователю возможность определять собственные подпрограммы, которые можно загружать с помощью системных функций [8]. Это дает пользователю достаточную свободу в дизайне исследования.

МАТЕРИАЛЫ И МЕТОДЫ

Генетический банк Национального центра биотехнологической информации США (NCBI National Center of the Biotechnological Information) хранит данные о первичных последовательностях ДНК геномов тысяч таксонов. Web-ресурс NCBI [1] снабжен удобной системой навигации ENTREZ. В то же время исчерпывающая информация о геноме человека представлена на сервере NCBI [9]. Наиболее подходящим форматом для выделения нуклеотидных последовательностей мРНК, является внутренний формат генбанка – “gbk”(Gene Bank flat format) [10], поскольку содержит аннотированную информацию о структуре и протяженности генов [11, 12].

Геном человека представлен набором локусов, полученных путем соединения коротких фрагментов в более протяженные при условии гомологии концов их последовательностей (перекрывания по типу черепичной кровли – «tiling path») [12]. Каждый локус имеет уникальный номер доступа (Accession number), благодаря которому всегда возможно выяснить кем, где и когда данная последовательность была секвенирована. Пример физического соответствия набора секвенированных локусов с цитогенетической идеограммой приведен на рис.1. Благодаря хромосомным координатам можно заметить, что покрытие локусов не является сплошным. Например, между центромерным локусом NT_024862 и дистальным локусом короткого плеча NT_010718 имеется расстояние в 20 100 000 пар нуклеотидов (п.н.), а между NT_024862 и локусом длинного плеча NT_010799 - 100 000 п.н.. Это объясняется тем, что целью международного консорциума по секвенированию геномов IHGSC (International Human Genome Sequencing Consortium) [10] было секвенирование эухроматической части генома, поэтому некоторые локусы гетерохроматиновых блоков, например, вблизи центромер, остались несеквенированными.

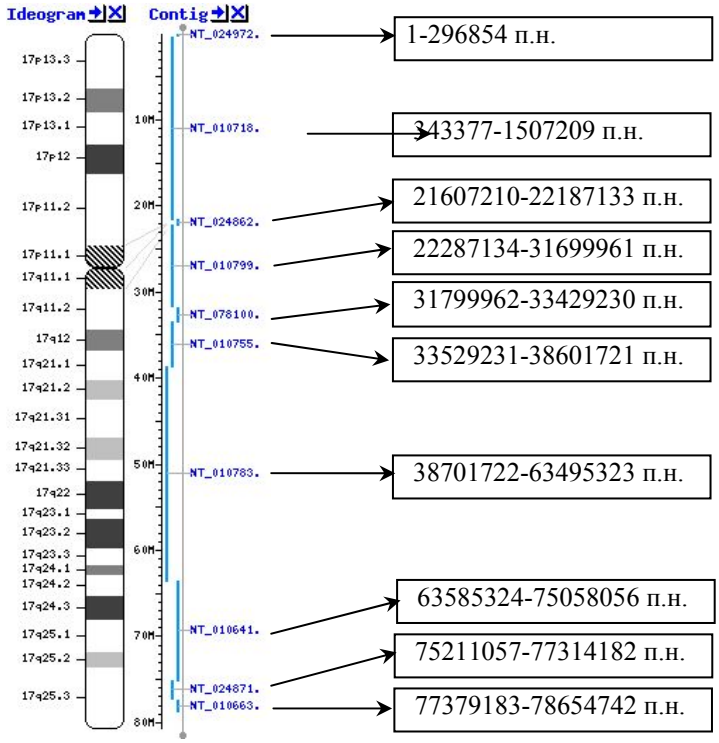

Рис. 1. Пример физического расположения локусов, представленных в электронной базе данных. Слева изображена идеограмма хромосомы (17), с обозначениями цитогенетических бэндов. Посередине показаны длины и номера доступа, соответствующих локусов. Справа указаны хромосомные координаты (п.н.) локусов "от" и "до": начиная сверху, от апикальной части теломеры короткого плеча 17p13.3, вниз до теломеры длинного плеча 17q25.3. Все локусы имеют 5′→3′ направление.

В архиве генбанка представлены два основных формата файлов – “gbk” и “fasta”[10]. В документе формата “gbk” можно условно выделить три части.

В первой части документа (рис.2 “DEFINITION”) содержатся выходные данные последовательности: название, номер доступа, длина (п.н.), тип (геномная ДНК, кДНК, мРНК и пр.), дата последней модификации, таксономическое положение организма, из которого данная последовательность получена, ключевые слова и ссылки на публикации. Во второй описательной части (рис.2 “FEATURES”) даны источники получения последовательности (с указанием лаборатории, названий и точных длин клонов, их координат в локусе), границы функциональных фрагментов и комментарии к ним. Третья часть содержит собственно нуклеотидную последовательность (рис.2 “ORIGIN”). Последовательность ДНК представлена в виде блока строчных букв c, g, t, a, с длиной строки 60 букв, причем каждые 10 разделены пробелом. Слева, в начале каждой строки указана нумерация первых символов строки (1, 61, 121, 181 и т.д.), от начала данной последовательности локуса (рис. 2).

```
LOCUS       NT_029490       490233 bp    DNA     linear   CON 20-AUG-2004
DEFINITION  Homo sapiens chromosome 21 genomic contig.
ACCESSION   NT_029490
VERSION     NT_029490.4  GI:51475310
SOURCE      Homo sapiens
ORGANISM  Homo sapiens; Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;  Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 490233)
AUTHORS   International Human Genome Sequencing Consortium.
TITLE     The DNA sequence of Homo sapiens
JOURNAL   Unpublished (2003)
COMMENT     GENOME ANNOTATION REFSEQ:  Features on this sequence have
been  produced for build 35 version 1 of the NCBI's genome annotation
[see documentation]. On Aug 20, 2004 this sequence version replaced
gi:22067441. The DNA sequence is part of the third release of the
finished human  reference genome. It was assembled from individual clone
sequences  by the Human Genome Sequencing Consortium in consultation with
NCBI staff…. …
FEATURES:  Location/Qualifiers
             … …source          1..208420
                      /organism="Homo sapiens"
                      /mol_type="genomic DNA"
                      /db_xref="taxon:9606"
                      /clone="CTD-2503J9"
                      /note="Accession AF254982 sequenced by Institute
                       for  Molecular Biotechnology, Jena Germany" … …
  gene             164726..165171 … …
  mRNA             join (164726..164771,164855..165171)
                      /gene="LOC390530"
                      /product="similar to immunoglobulin heavy-chain-2
                         light-chain-2 VH segment"
                      /note="Derived by automated computational analysis
                   using  gene prediction method: GNOMON."
                      /transcript_id="XM_372543.2"… …
  CDS              join(164726..164771,164855..165171)
                      /gene="LOC390530"
                      /codon_start=1
                      /protein_id="XP_372543.2"… …
ORIGIN:   1 aattctgaga aacttctttg tgagggttgg attcatttca cacatttgaa catttctttg
         61 attgaagatt tggaaacagt ctttttgtaa aatctataaa gggataattg tgaacccttt
        121 gaggcctagg gtgaagtagg aaatatcttc acataaaaac tacacagaaa ttttctgaga
… …
     490141 gtctgcatgg cagcagttgg acctcacaat gtggattgtg ccttcaccgt ggaatgttta
     490201 taccctatcg ccatggtgat gggattaggg atc
//
```

Рис. 2. Пример файла формата “gbk”. Символом «... ...» обозначены пропуски повторяющихся данных, несущественных для рассмотрения структуры формата.

Согласно [11] в геномной ДНК различают следующие определения функциональных фрагметов: гены (gene), мРНК (mRNA), транслируемые последовательности (CDS), и пр. Описание структуры гена включает информацию об экзон-интронной организации каждого транскриптного варианта мРНК, а также транслируемых экзонов, если такие имеются.

Документ формата “fasta” [11] состоит из двух частей: первой строки, обязательно начинающейся символом “>” за которым следует номер доступа и короткое название последовательности, а также блока нуклеотидной последовательности всей хромосомы. Блок содержит прописные буквы A, T, C, G, с длиной строки 70 символов без пробелов и нумерации. Такой формат совместим с известными программами анализа: BLAST_N, MegaBLAST, RepeatMasker, CodonW и пр. [3]. Несеквенированные участки (см. выше) обозначены в файлах “fasta” символами “N”, для сохранения общей протяженности локуса. Файлы формата “gbk” завершенных последовательностей не содержат несеквенированных областей и, следовательно, символов “N”.

Таким образом, формат “gbk” позволяет извлекать функциональные фрагменты ДНК согласно их координатам в локусе.

РЕЗУЛЬТАТЫ

Структурный разбор файлов локусов осуществляли при помощи регулярных выражений языка PERL [8]. Регулярные выражения строили по описаниям формата файлов банка NCBI [11] на основе EBNF грамматик (расширенных нормальных форм Бэкуса-Науэра) [13]. В таблице 1 приведены шаблоны для получения:1 – десктрипторов генов (G); 2 – локусов (L); 3 – номеров транскриптных вариантов мРНК; 4 – последовательностей генов (GENE); 5, 6 – экзонных и интронных последовательностей первичной мРНК (mRNA) и зрелой (CDS).

Таблица 1.

Образцы шаблонов

№	Содержание шаблона
1	G ::= gene <gene name>
2	L ::=LOCUS NT_<xxxxxx>
3	номер транскрипта ::= transcript_ID=NM_<xxxxxxx.x> |transcript_ID=XM_<xxxxxxx.x>
4	GENE ::= gene <gene_begin> .. <gene_end> |gene complement <gene_begin>..<gene_end>
5	mRNA ::= mRNA join (<mrna_begin_1> .. <mrna_end_1> [, <mrna_begin_N> .. <mrna_end_N>] {repeat N} | mRNA complement join(<mrna_begin_1> .. <mrna_end_1> [, <mrna_begin_N> .. <mrna_end_N>] {repeat N}
6	CDS ::= CDS join (<cds_begin_1> .. <cds_end_1> [, <cds_begin_N> .. <cds_end_N>] {repeat N}) | CDS complement join (<cds_begin_1> .. <cds_end_1> [, <cds_begin_N> .. <cds_end_N>] {repeat N})

Перевод шаблонов в форму регулярных выражений является достаточно прямолинейным. В таблице 2 приведены примеры трансляции шаблонов для выделения имени гена (G) и экзон-интронных последовательностей CDS в код на языке PERL.

Таблица 2.

Примеры перевода шаблонов на язык PERL

шаблон	текст программы на языке PERL
G	if (m{gene\s+(\S+) }xi) {$gene_name=$1;...}

CDS	if (m{CDS \s+join \s*\((.*?\)}xi
	or m{CDS \s+ complement \s+ join \s*\((.*?) \)}xi
	) {
	@CDS_begin_end=map {(m{(\d+) \s\.\.\s*9\d+)}xi)}
	...
	}

На начальном этапе файлы “gbk” полных хромосом разделяли по шаблону "LOCUS NT_<XXXXXX>" на составляющие его локусы в отдельные файлы, где XXXXXX- шестизначный номер доступа локуса. В дальнейшем работали с файлами локусов. Разработанные шаблоны позволяют формировать на основе аннотаций локусов списки генов по таким параметрам: имя гена, номер доступа транскриптного варианта мРНК, количество вариантов мРНК, номер локуса и номер хромосомы, содержащий данный ген.

Согласно требованиям номенклатурного комитета генов человека IHGSC ген определяется как «фрагмент ДНК, который имеет вклад в фенотип или функцию. Ген с неустановленной функцией характеризуется последовательностью, транскрипцией, гомологией. Имя (дескриптор) гена описывается заглавными латинскими буквами или их сочетанием с арабскими цифрами, причем длина имени гена, должна быть не более шести символов. Имя гена должно начинаться с буквы, не содержать знаков препинания, буквы «G» для обозначения гена и символов видовой принадлежности (например, «H/h» для человека)» [13]. Несмотря на эти требования, в 35-й версии генома встречаются несоответствующие им имена генов: e(y)2, 13CDNA73, 182-FIP, 3'HEXO, OK/SW-cl.56, FCA/MR, DKFZP434D177-like, Beta4GalNAc-T4 и пр. Таким образом, была выявлена неоднозначность между описанием формата [11] и рабочими материалами. Это позволило учесть в программе дополнительные критерии разбора и осуществить корректное и полное извлечение нуклеотидных последовательностей генов.

Извлечение последовательностей мРНК проводили согласно описанным координатам, используя шаблоны 5, 6 (Табл.1). Блок-схема программы получения последовательностей мРНК изображена на рис. 3.

Поскольку в документах “gbk” даны границы экзонов [10], то экзонные части и кодирующие последовательности вырезали, с включением границ, а интронные части и межгенные промежутки вырезали с исключением границ. Если указывалось, что ген транскрибируется с противоположной цепи (имелся комментарий «complement»), то символы в последовательности заменяли на комплементарные, а направление последовательности меняли на обратное. Блок схема подпрограммы, выделяющей последовательности экзонов и интронов изображена на рис.4.

Если для одного гена было описано несколько мРНК, то выбирали самый протяженный вариант. Кроме внутригенных фрагментов выделяли нетранскрибируемые, прилегающие к гену области (по 1000 п.н.). Область на 5'-конце вырезали влево относительно сайта старта транскрипции, а на 3'-конце - вправо от последнего нуклеотида последнего экзона мРНК.

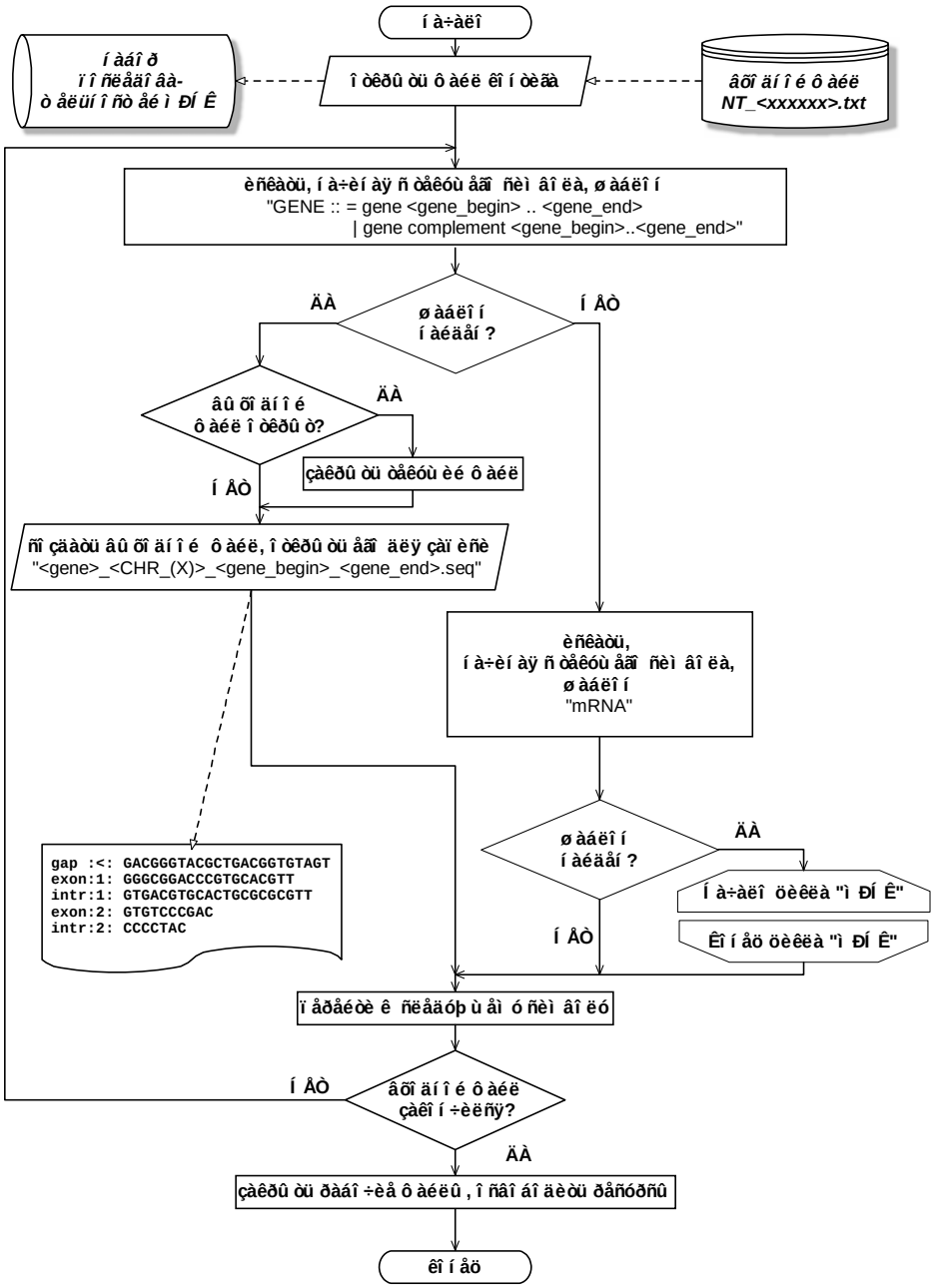

Рис. 3. Блок-схема программы получения последовательностей мРНК.

Извлеченную последовательность мРНК сохраняли в отдельный файл, которому присваивали имя, включающее дескриптор гена, его координаты в локусе и номер хромосомы. В описание файла сохраняли название соответствующего гену белка.

Наличие ключевой информации в имени файлов позволило в дальнейшем формировать различные выборки генов. Тело файла представляло блок символов, строки которого начинались маркером экзонного (exon:<x>) или интронного (intr:<x>) фрагмента и его номером (x). Прилегающие к гену области 5’-конца обозначали «gap:<», а 3’-конца «gap:>». В таком виде были получены 22777 последовательностей мРНК для всех генов Build 35 , систематизированные по локусам и хромосомам (рис.5).

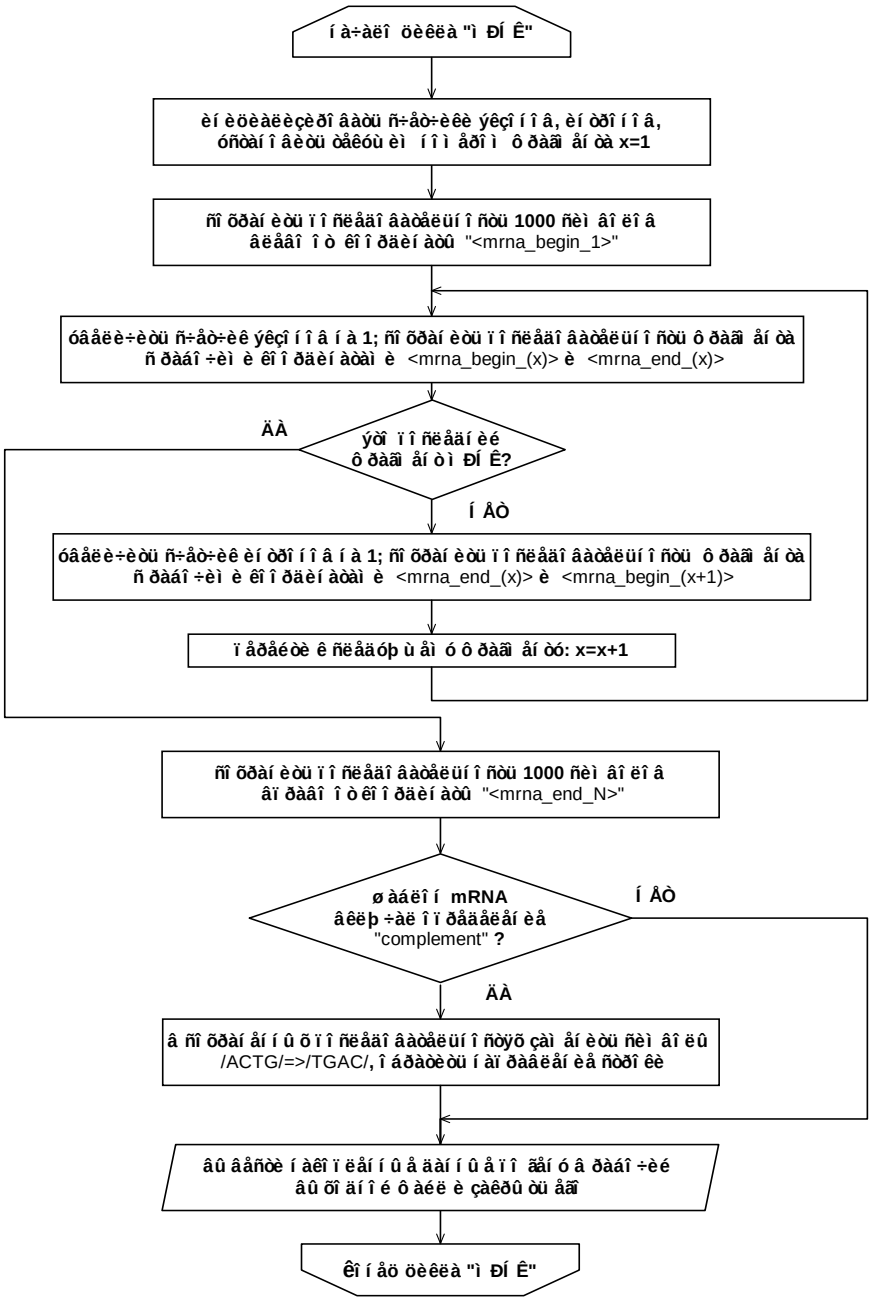

Рис. 4. Блок-схема подпрограммы, выделяющей последовательности экзонов и интронов.

Отдельным важным результатом также было извлечение последовательностей нетранслируемых концов. Известно, что первичная мРНК, кроме интронов, содержит лидерную и трейлерную последовательности, не участвующие в кодировании белка. В документах “gbk” границы нетранслируемых областей не аннотированы. С помощью разработанного подхода можно определить 5’-нетранслируемую область (5’-UTR) как разницу координат начала гена (шаблон GENE) и начала кодирующей последовательности (шаблон CDS). Соответственно 3’-нетранслируемый конец (3’-UTR) определяли разницей кординат 3’- конца CDS и 3’- концом гена. Если для гена описывалось несколько CDS, то выбирали вариант наиболее близко расположенный к 5’-концу гена.

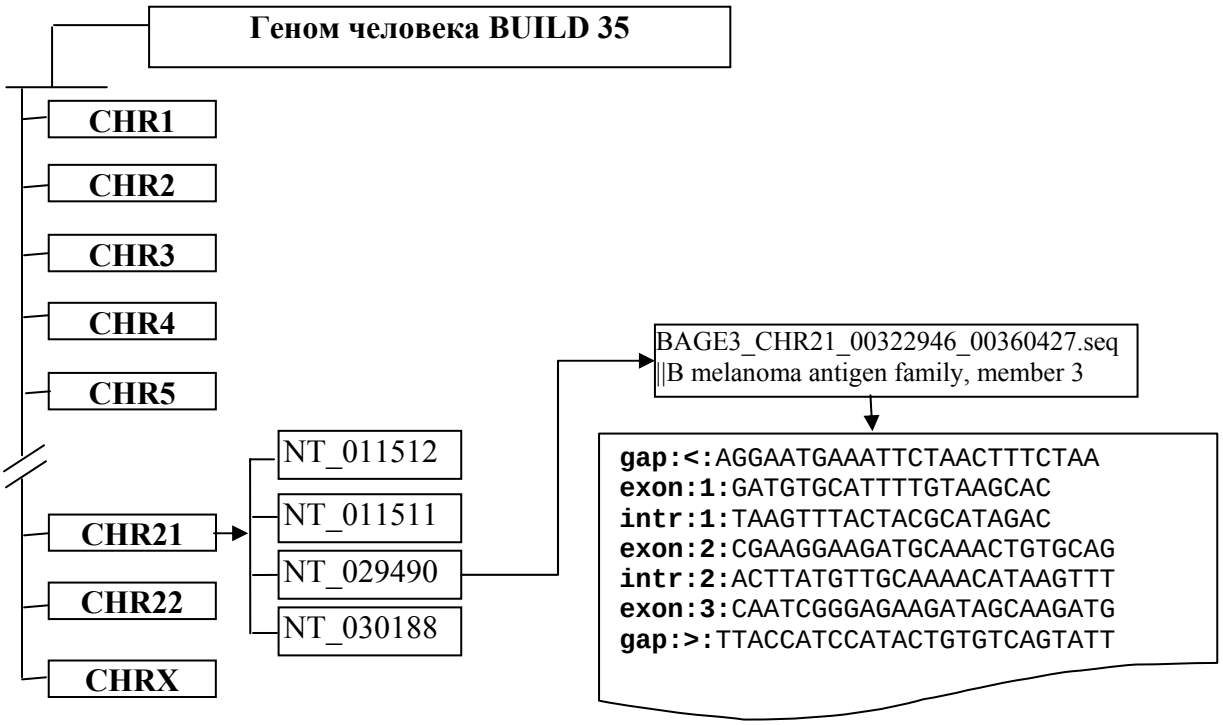

Рис. 5. Структура результирующей базы последовательностей мРНК.

ВЫВОДЫ

Разработан подход для извлечения функциональных фрагментов геномной ДНК из данных внутреннего формата генетического банка НЦБИ США с помощью регулярных выражений языка PERL, разработанных на основе расширенных нормальных форм Бэкуса-Науэра.

Предложен набор шаблонов, позволяющих извлекать из файлов геномной ДНК последовательности генов, первичной и зрелой мРНК, нетранслируемых 5’ и 3’-концов генов, межгенных промежутков. Извлеченные последовательности мРНК имеют формат, удобный для анализа внешними программами.

Предлагаемый набор регулярных выражений может быть применен в удобной для исследователя программной среде и использован для извлечения и анализа нуклеотидных последовательностей других геномов, представленных во внутреннем формате генбанка.

СПИСОК ЛИТЕРАТУРЫ

1. *Web-resource: http://www.ncbi.nih.gov/*
2. Cochrane G. R., Galperin M. Y. The 2010 Nucleic Acids Research Database Issue and online database collection: a community of data resources // Nucl. Acids Res. – 2010.– V.38.– D1-D4; doi:10.1093/nar/gkp1077.
3. Sayers E. W., Barrett T., Benson D. A. [et al.]. Database resources of the National Center for Biotechnology Information // Nucl. Acids Res. – 2010.– V. 38.– D5-D16; doi:10.1093/nar/gkp967.
4. The NCBI C++ Toolkit [Internet] / edit. Vakatov D., Siyan, K., Ostell J. – Bethesda (MD): National Library of Medicine (US), NCBI; 2004.
5. Gilat A. MATLAB: An Introduction with Applications 2nd Edition. – John Wiley & Sons. ISBN 978-0-471-69420-5.– 2004.

6. MATLAB technical documentation [Access] http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/ref/seqtool.html
7. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/INDEX.HTML
8. Водолазкий В,Семериков В. Энциклопедия PERL. СПб.: Питер. - 2002.-576С.
9. ftp://ftp.ncbi.nih.gov/genbank/genomes/H_sapiens
10. International Human Genome Sequencing Consortium. The NCBI Handbook [Web resource]: Bethesda National Library of Medicine (US), NCBI 2002-2005 / edit. J. McEntyre, J. Ostell . – [Access]: http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook
11. The DDBJ/EMBL/GenBank Feature Table: Definition [Web resource]: Bethesda.– 2007 / International Sequence Databank Collaboration.– [Access]: http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html#7.4
12. NCBI Help Manual.– Bethesda (MD): National Library of Medicine (US), NCBI; 2005-2009.
13. Компиляторы: принципы, технологии и инструментарий / А. В. Ахо, М. С. Лам, Р. Сети, Д. Д. Ульман. – М.: Вильямс. –2008. –2-е издание. –1184 с.
14. Wain H. M., Bruford E. A., Lovering R. C., Lush M. J., Wright M. W. and Povey S. Guidelines for Human Gene Nomenclature//Genomics.-2002.-№79(4). - P. 464-470.
15. Пат. 43786 Україна, МПК (2007) G 06 F, 17/21. Спосіб перетворення структури послідовностей генів бази даних / Дуплій Д.Р., Калашніков В.В., Чащин Н.А., заявник та патентовласник Київ. Ін-т Молекулярної Біології и Генетики; Держ.реєстр. від 25.08.09.