КОМПЬЮТЕРНАЯ МАТЕМАТИКА

Математические модели в биологии и медицине

На основе аппарата цепей Маркова и байесовских процедур проведено распознавание известных экзонно-интронных фрагментов генов организма Caenorhabditis elegans.

УДК 519.681

И.И. АНДРЕЙЧУК

РАСПОЗНАВАНИЕ ФРАГМЕТОВ ГЕНОВ В ДНК

Введение. Симметрия в записи оснований, подсчитанных по нитям в хромосомах ДНК, исследовалась в работах [1, 2]. Соотношения симметрии приведены в виде коротких формул, что значительно упрощает восприятие этих результатов и является основой построения математического аппарата для получения новых результатов. Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания свойств участков оснований (генов), в том числе генетических заболеваний.

В данной работе показано, что в геноме червяка Caenorhabditis elegans наблюдается явная ассимметрия в записи оснований интронно-экзонных фрагментов генов. На осно-ве этих свойств построены эффективные процедуры распознавания известных фрагментов данного генома.

ДНК имеет форму двойной спирали, информация записана в четырехбуквенном алфавите оснований: аденин (A), цитозин (C), гуанин (G), тимин (T). Известно, что C – G, A – T – комплементарные пары оснований, связывающие две цепи. Запись и считывание оснований по первой комплементарной нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$, по второй – справа налево в направлении $5' \to 3'$ (рис. 1).

Для оснований, записанных по одной нити ДНК хромосомы, выполняются приближенные соотношения

$$n(\mathrm{A}) = n(\mathrm{T}), \quad n(\mathrm{C}) = n(\mathrm{G}), \qquad (1)$$

где $n(i)$ – количество оснований i, $i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, вычисленных по одной нити.

Таким образом, имеет место симметрия относительно записи оснований по каждой нити ДНК.

$$5' \_\_\_\_\_ AC \_\_\_\_\_ GT \_\_\_ \rightarrow \_\_ 3' \quad \text{1-я нить}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \updownarrow\updownarrow \ldots\ldots\ldots\ldots\ldots\ldots$$
$$3' \_\_\_\_\_ TG \_\_\_\_\_ CA \_\_\_\_ \leftarrow \_\_ 5' \quad \text{2-я нить}$$

РИС. 1. Условная запись двух нитей хромосомы (модель Уотсона – Крика)

Расчеты показали, что для пар оснований выполняются соотношения

$$n(ij) = n(\bar{j}\bar{i}), \qquad (2)$$

где $i, j \in \{A, C, G, T\}$, $\overline{A} = T$, $\overline{C} = G$, $\overline{T} = A$, $\overline{G} = C$.

Из соотношения (2) вытекает симметрия относительно записи 16 пар оснований по каждой нити ДНК: $n(ij,1) = n(ij,2)$, где $i, j \in \{A, C, G, T\}$ [1, 2].

Материалы и методы. Общая выборка фрагментов генома червяка Caenorhabditis elegans сформирована на основе версии W190 из сайта NCBI [3]. Структура гена и процесс формирования зрелой мРНК(CDS) из первичной мРНК представлен на диаграмме (рис. 2). Ген всегда начинается с экзона (UTR или CDS фрагмент) при этом последовательности экзонов чередуются с интронами [4].

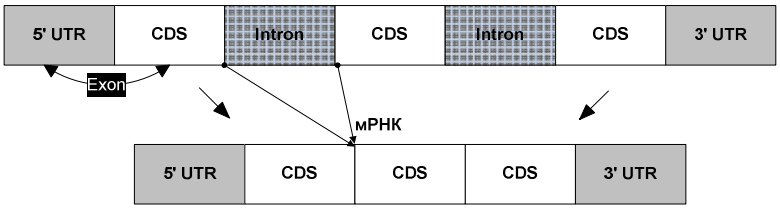

РИС. 2. Структура гена и мРНК

Классы распознавания. Рассмотрим три класса интронно-экзонных фрагментов генома: UTR – транскрибирующиеся, но не транслируемые области генома, CDS – транслируемые в белок области генома и INTRON – не транслируемые области генома, которые удаляются из первичной мРНК в процессе сплайсинга. Таким образом, класс EXON состоит из двух подклассов отличающихся по своим функциям в формировании мРНК и синтезе белков: UTR и CDS.

Данные о начальной и конечной позиции интервалов для фрагментов UTR и CDS, ориентации нити и их принадлежность к определенному гену взяты из файла seq_gene.md [3]. Последовательности фрагментов формируются из файла хромосомы в формате FASTA, используя позиции начала и конца фрагмента (chr_start, chr_stop).

Для хромосомы I средняя длина фрагмента CDS по числу оснований – 213, INTRON – 402, UTR – 126 при средней длине гена – 3446 оснований. Средние значения числа фрагментов по геному равны: 6.11 INTRON-ов на ген, 6.85 CDS-ов на ген, 1.84 UTR-ов на ген соответственно.

Хромосома 5 содержит наибольшие количества фрагментов в гене: 66 CDS и 65 INTRON, соответственно хромосома X содержит в гене 25 фрагментов UTR. Фрагменты максимальной длины в основаниях находятся: CDS – 14975 в хромосоме 4, UTR – 4530 в хромосоме 2, INTRON – 100913 в хромосоме X.

Процедура распознавания фрагмента y последовательности оснований $x = (x_1, x_2, ..., x_n)$, $x_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ строится на основе формулы Байеса

$$P(y|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|y)P(y)}{P(x_1, ..., x_n)}, \tag{3}$$

где $y \in \{\text{intron, cds, utr}\}$.

В модели однородной цепи Маркова

$$p(x_1, x_2, ..., x_n) = p(x_1)\, p(x_1, x_2) ... p(x_{n-1}, x_n), \tag{4}$$

где $p(x_1)$ – начальное распределение, $p(i, j) = p(x_k = j | x_{k-1} = i)$ – переходные вероятности. В численных расчетах используем оценки переходных вероятностей, построенных в виде частот:

$$\hat{p}(ij) = \frac{n(ij)}{\sum_j n(ij)}, \tag{5}$$

где $n(ij)$ – число пар (ij) оснований в суммарной последовательности класса $y \in \{\text{intron, cds, utr}\}$.

В табл. 1 приведены подсчеты числа оснований и их частот по фрагментам генов в хромосоме 1. Для хромосом 2–5 частоты оснований аналогичны частотам для хромосомы I. Знак «+» или «–» обозначают, что частоты подсчитаны по 1-й и 2-й нитям ДНК соответственно. В столбцах без знака, частоты подсчитывались по фрагментам, расположенным на обеих нитях.

Из табл. 1 видно, что для фрагментов $y \in \{\text{intron, cds, utr}\}$ симметрия (1) не выполняется, при этом для всей нити ДНК она имеет место. В табл. 2 записано количество пар оснований и оценки переходных вероятностей (о.п.в.), подсчитанные по фрагментам генов и ДНК по одной нити. Аналогично табл. 1 симметрия (2) для пар оснований отсутствует, а для всей нити ДНК она выполняется.

Результаты распознавания. Результаты распознавания для хромосомы 3 приведены в табл. 3. В первом столбце указаны классы распознавания. В вычислениях используются логарифмы от значений переходных вероятностей в (4), чтобы избежать проблем в вычислениях с бесконечно малыми значениями; строка lh обозначает, что априорная вероятность класса $P(y)$ в (3) не учитывается; а в строке pclass_lh она вычисляется.

Интересно, что фрагменты utr, не принимающие участия в синтезе белка, в ситуации CDS/INTRON распознаются в 70 % случаях как INTRON, что говорит о схожести utr и intron фрагментов.

ТАБЛИЦА 1

Осно-вание	**intron число**	**intron частота**	**cds число**	**cds частота**	**utr число**	**utr частота**	**cds+ число**
A	2119211	0.334400	1261728	0.307918	158469	0.283964	635295
C	1027014	0.162057	853033	0.208178	110473	0.197959	427076
G	1022900	0.161408	916644	0.223702	80198	0.143708	460074
T	2168235	0.342135	1066206	0.260202	208921	0.374369	539866
	utr- число	**utr- частота**	**utr+ число**	**utr+ частота**	**cds- число**	**cds- частота**	**cds+ частота**
A	79302	0.284672	79167	0.283257	626433	0.307784	0.30805
C	55051	0.197618	55422	0.198298	425957	0.209285	0.207086
G	40148	0.14412	40050	0.143298	456570	0.224326	0.223087
T	104072	0.37359	104849	0.375147	526340	0.258606	0.261777
	intron- число	**intron- частота**	**intron+ число**	**intron+ частота**	**DNA+ число**	**DNA+ частота**	
A	1054142	0.335282	1065069	0.333531	4835939	0.320847	
C	507744	0.161494	519270	0.162612	2695879	0.178862	
G	507601	0.161448	515299	0.161368	2692150	0.178614	
T	1074557	0.341775	1093678	0.34249	4848453	0.321677	

ТАБЛИЦА 2

Пары букв	**intron число**	**intron о.п.в.**	**cds число**	**cds о.п.в.**	**utr число**	**utr о.п.в.**	**DNA+ число**	**DNA+ о.п.в.**
AA	988702	0.466547	437705	0.348420	59322	0.378235	2049736	0.423855
AC	267960	0.126444	224876	0.179005	25033	0.159610	704042	0.145585
AG	294225	0.138838	254398	0.202505	19767	0.126034	758857	0.156920
AT	568304	0.268170	339277	0.270070	52717	0.336122	1323304	0.273640
CA	348548	0.339384	293857	0.344903	33902	0.308716	909654	0.337424
CC	201872	0.196565	157821	0.185236	24770	0.225559	521368	0.193395
CG	184086	0.179246	183496	0.215371	16597	0.151135	503521	0.186774
CT	292495	0.284805	216824	0.254489	34547	0.31459	761335	0.282407
GA	343641	0.341505	355057	0.391833	26493	0.334170	953424	0.354150
GC	190521	0.189337	179035	0.197579	15002	0.189228	515116	0.191340
GG	182455	0.181321	178787	0.197306	11521	0.145320	518674	0.192662
GT	289638	0.287838	193264	0.213282	26264	0.331282	704936	0.261849
TA	438301	0.202148	164805	0.154938	36783	0.177024	923125	0.190396
TC	366647	0.169100	288942	0.271643	45172	0.217398	955353	0.197043
TG	345492	0.159344	295194	0.277520	31534	0.151763	911097	0.187915
TT	1017780	0.469408	314743	0.295899	94296	0.453815	2058878	0.424646

ТАБЛИЦА 3

Классы распознавания	Распознано число	Не распознано число	Число	Распознано %	Не распознано %
cds/utr/intron					
lh	36848	7794	44642	0.8254	0.1746
pclass_lh	34173	10469	44642	0.7655	0.2345
cds/intron					
lh	36765	3414	40179	0.9150	0.0850
pclass_lh	36969	3210	40179	0.9201	0.0799
utr/intron					
lh	18950	4301	23251	0.8150	0.1850
pclass_lh	16747	6504	23251	0.7203	0.2797
cds/utr					
lh	22354	3500	25854	0.8646	0.1354
pclass_lh	21996	3858	25854	0.8508	0.1492
utr как cds/intron	**cds**	**intron**		**Распознано как cds**	**Распознано как intron**
lh	1286	3177	4463	0.2881	0.7119
pclass_lh	1542	2921	4463	0.3455	0.6545

Заключение. На основе байесовского подхода и модели однородных цепей Маркова построены простые в вычислительном плане процедуры распознавания известных фрагментов генома организма Caenorhabditis elegans с процентом распознавания порядка 80–90 %. В настоящее время аналогичные задачи решаются на геноме человека. Полученные результаты открывают широкие возможности применения байесовских процедур на моделях цепей Маркова для распознавания свойств участков оснований (генов), в том числе генетических заболеваний.

І.І. Андрійчук

РОЗПІЗНАВАННЯ ФРАГМЕНТІВ ГЕНІВ У ДНК

За допомогою апарата ланцюгів Маркова та байєсівських процедур проведено розпізнання відомих інтронно-екзонних фрагментів генів організму Caenorhabditis Elegans.

I.I. Andriychuk

RECOGNITION OF GENE FRAGMENTS IN DNA

On the basis of instrument of Bayesian approach and Markov chain, recognition procedure for gene fragments is obtained. The procedure has high percentage of recognition (80–90 %) and is simple and effective in computation. Its applicability is described by using Markov chains of 1-th order for the entire array of intron-exon sequences of the genome of a Caenorhabditis Elegans organism.

1. *Гупал А.М., Сергиенко И.В.* Оптимальные процедуры распознавания. – Киев: Наук. думка, 2008. – 232 с.
2. *Гупал А.М., Гупал Н.А., Островский А.В.* Симметрия и свойства записи генетической информации в ДНК // Проблемы управления и информатики. – 2011. – № 3. – С. 120–127.
3. NCBI ftp resource for Caenorhabditis Elegans gene regions (CDS, UTR) data
4. ftp://ftp.ncbi.nih.gov/genomes/MapView/Caenorhabditis_elegans/sequence/current/initial_release/seq_gene.md.gz
5. *Batzoglou S., Alexandersson M., Pachter L., Saxonov S.* / Lecture – Gene Recognition, http://ai.stanford.edu/~serafim/CS262_2006/Slides/CS262_2006_Lecture16.ppt

Получено 22.04.2011

Об авторе:

Андрейчук Иван Иванович,
младший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.
vanya@ukr.net.