Математическая биология и биоинформатика. 2014. Т. 9. № 2. С. 482–490.
URL: http://www.matbio.org/2014/Astakhova_9_482.pdf.

========== **БИОИНФОРМАТИКА** ==========

УДК: 577.21

Закономерности, связанные с распределением длин интронов

Астахова Т.В.[1], Ройтберг М.А.[*1,2,3], Цитович И.И.[2,3,4], Яковлев В.В.[1]

[1]*Институт математических проблем биологии, Пущино, Московская область, Россия*
[2]*НИУ Высшая школа экономики, Москва, Россия*
[3]*Московский физико-технический институт (ГУ), Долгопрудный, Московская область, Россия*
[4]*Институт проблем передачи информации им. А.А. Харкевича, Москва, Россия*

Аннотация. Изучены закономерности распределения длин интронов в геномах 17 организмов, принадлежащих к различным таксонам (насекомые, рыбы, земноводные, пресмыкающиеся, птицы, млекопитающие). Показано, что доля интронов, имеющих фазу 1, растет с ростом длины интрона. Кроме того, показано, что короткие и длинные интроны имеют тенденцию образовывать серии, например, доля коротких (длинных) интронов среди тех интронов, которые следуют за коротким (длинным) интроном, существенно выше, чем доля коротких (длинных) интронов в геноме. Эти закономерности показаны для всех рассмотренных геномов.

Ключевые слова: *интрон, экзон, фаза интрона, длина интрона.*

ВВЕДЕНИЕ

Проблема функционирования и эволюции генов эукариот, включая экзон-интронную структуру генов – одна из важнейших задач современной молекулярной биологии. С ростом объемов доступных данных возрастает роль биоинформатических подходов при изучении этой проблемы.

Изучение генов эукариот ведется, начиная с 80-х годов XX века. В это время большинство исследований были связаны с распознаванием белок-кодирующих участков ДНК. Обзор результатов первого периода исследований содержится в работах [1, 2]. К основным результатам этого периода можно отнести построение моделей сайтов сплайсинга в виде позиционных весовых матриц (ПВМ, англ. термин – PWSM), введение понятия кодирующего потенциала участка и постановку задачи распознавания белок-кодирующих участков как задачи выделения «оптимальных» путей в графах. В это же время был разработан ряд программных систем (GRAIL, GeneMark др.), которые обеспечивали точность распознавания на уровне около 70%. В последующие годы (конец девяностых – начало двухтысячных) в распознавании генов был достигнут существенный прогресс, и с практической точки зрения задачу распознавания белок-кодирующих областей можно считать решенной. Основными причинами такого прогресса являются использование сведений об экзон-интронной структуре уже известных генов, использование методов машинного обучения и построение скрытых марковских моделей для различных участков генов.

Изучение экзон-интронной структуры генов в качестве самостоятельного направления исследований, не связанного непосредственно с распознаванием генов,

[*]mroytberg@impb.psn.ru, mroytberg@lpm.org.ru

сформировалось во второй половине 90-х годов XX века. Работы, относящиеся к этому направлению, можно условно разделить на три класса: (1) статистические связи между свойствами элементов экзон-интронной структуры, см. например, [3, 4]; (2) закономерности, отражающие влияние свойств элементов экзон-интронной структуры генов на функциональные свойства генов, см., например, [5]; (3) сравнительный анализ экзон-интронных структур ортологичных генов разных видов и паралогичных генов в одном геноме, и построение моделей эволюции экзон-интронной структуры, см. обзор [6].

В качестве характеристик экзонов и интронов рассматриваются, как правило, фаза (остаток от деления суммы длин предшествующих транслируемых экзонов на 3), номер экзона/интрона в гене, длина экзона или интрона, нуклеотидная последовательность экзона/интрона. При этом закономерности, связанные с длинами интронов, изучены значительно хуже, чем, например, закономерности, связанные с их фазами.

Одним из наиболее известных фактов является повышенная длина 1-го интрона [7]. Также в работе [8] проведен систематический анализ длин интронов различных групп организмов. Изучалась связь между длиной интронов и протеканием различных внутриклеточных процессов в ходе эволюции. Работы [9, 10] посвящены времени протекания сплайсинга в зависимости от длины интрона. В работе [11] исследована корреляция между длиной интрона и степенью эволюционного отбора на аминокислотном уровне. А. Виноградов [12] показал, что интроны короче в конститутивных генах, чем в тканеспецифичных, или генах, отвечающих за развитие организма. Исследована зависимость между длиной интрона и уровнем экспрессии генов [13]. Показано, что соотношение между длиной интронов и GC-содержанием у различных видов может быть связано с изохорной структурой геномов [14]. В целом, GC-богатые изохоры позвоночных имеют короткие интроны, а GC-бедные изохоры – длинные интроны.

Целью нашей работы является изучение закономерностей строения экзон-интронной структуры, связанных с длиной и фазой интронов, которые в настоящее время изучены недостаточно. В частности, недостаточно изучена корреляция между длинами соседних интронов (в отличие от хорошо известной корреляции между фазами соседних интронов).

МАТЕРИАЛЫ И МЕТОДЫ

Данные

Анализировались аннотированные интроны 17 организмов (см. табл. 1), относящихся к 6 классам (насекомые, рыбы, земноводные, пресмыкающиеся, птицы, млекопитающие). Исходные данные были взяты с сайта ftp.ncbi.nih.gov/genomes, организованы в виде базы данных и доступны по адресу: http://server2.lpm.org.ru/~victor/introns_db/. Описание структуры базы и методики, использованной при ее построении, находятся по адресу http://server2.lpm.org.ru/~victor/introns_db/build_db/.

Ниже, говоря об интронах млекопитающих, птиц, насекомых и т. д., мы будем иметь в виду все интроны тех из указанных выше организмов, которые принадлежат к классу млекопитающих, птиц и т.д.

Таблица 1. Исследованные геномы (внутренние интроны)

№№	Организм	Класс	Кол-во интронов	Кол-во генов	Среднее кол-во интронов в гене
1	*Apis mellifera*	Insecta	29494	6015	4.9
2	*Drosophila melanogaster*	Insecta	19659	5385	3.65
3	*Nasonia vitripennis*	Insecta	31554	6816	4.63
4	*Tribolium castaneum*	Insecta	22703	5556	4.09
5	*Danio rerio*	Osteichthyes	114893	15525	7.4
6	*Xenopus tropicalis*	Amphibia	80533	10644	7.57
7	*Anolis carolinensis*	Reptilia	73632	9195	8
8	*Gallus gallus*	Aves	81916	9636	8.5
9	*Meleagris gallopavo*	Aves	51699	6263	8.25
10	*Taeniopygia guttata*	Aves	62216	7424	8.38
11	*Canis lupus familiaris*	Mammalia	93413	10962	8.52
12	*Mus musculus*	Mammalia	108206	13633	7.93
13	*Sus scrofa*	Mammalia	75050	10527	7.13
14	*Callithrix Jacchus*	Mammalia	78234	9907	7.9
15	*Macaca mulatta*	Mammalia	79200	10344	7.66
16	*Pan troglodytes*	Mammalia	90139	11552	7.8
17	*Homo sapiens*	Mammalia	94067	11578	8.12

Таблица 2. Пороги длин для различных долей длинных интронов и различных организмов

№	Организм	Пороги длин для различных долей длинных интронов					Самый длинный интрон
		50%	25%	10%	5%	1%	
1	*Apis mellifera*	110	325	1350	3530	22150	557152
2	*Drosophila melanogaster*	75	255	1340	3110	12500	132736
3	*Nasonia vitripennis*	85	210	880	2450	20110	264629
4	*Tribolium castaneum*	55	480	2770	4280	13880	163127
5	*Danio rerio*	870	2420	4750	8000	25320	383251
6	*Xenopus tropicalis*	940	2010	4710	8450	26910	374628
7	*Anolis carolinensis*	1410	2850	6840	13000	45440	420519
8	*Gallus gallus*	810	1780	4480	8600	30480	331673
9	*Meleagris gallopavo*	850	1850	4720	8970	29260	427402
10	*Taeniopygia guttata*	920	2040	5210	9940	34180	462377
11	*Canis lupus familiaris*	1290	3300	8350	15460	53180	700631
12	*Mus musculus*	1230	2900	7000	12860	49290	479338
13	*Sus scrofa*	1270	3170	7540	13890	44800	402757
14	*Callithrix jacchus*	1640	4140	10530	19340	65780	820705
15	*Macaca mulatta*	1530	3910	9920	18270	60030	963148
16	*Pan troglodytes*	1570	3980	9930	18190	63590	587523
17	*Homo sapiens*	1450	3650	9130	16820	58230	772625

Основные определения

С точки зрения положения в гене, можно выделить такие группы интронов: все интроны; первые интроны; последние интроны; внутренние интроны (все интроны, кроме первых и последних); интроны в двухэкзонных генах.

Данная работа посвящена изучению внутренних интронов, так как только для них возможно исследование связи интрона с его окружением.

Интрон называется *T-длинным*, если его длина не менее T нуклеотидов. Таблица 2 дает представление о распределении длин интронов для различных организмов.

РЕЗУЛЬТАТЫ

Длины смежных интронов

Анализ длин смежных интронов, т.е. интронов, окружающих один экзон, показал, что их длины зависят друг от друга. В частности, мы показали, что для всех рассматриваемых порогов длины интронов T и для всех исследованных таксонов вероятность того, что интрон, смежный с *T-длинным* (*T-коротким*) интроном, является также с *T-длинным* (*T-коротким*), значительно больше, чем вероятность того что он окажется *T-коротким* (*T-длинным*).

Данные для *H. sapience, G. gallus и D. melanogaster* представлены в таблицах 3а, 3б, 3в. В колонках этих таблиц обозначениями «*S→S*» и «*S←S*» (соответственно, «*SS→S*» и «*S←SS*») показаны доли коротких интронов среди всех таких интронов, для которых предыдущий (для столбца «*S→S*») или последующий (для столбца «*S←S*») также короткий. В колонках «*SS→S*» и «*S←SS*» показаны доли коротких интронов среди всех таких интронов, для которых два предыдущих (последующих) интрона – короткие. Столбцы «*L→L*», «*L←L*», «*LL→L*» и «*L←LL*» содержат аналогичные данные для длинных интронов. Например, для генома человека и порога $T = 1500$ п.н. геном содержит 51% коротких интронов, эмпирическая вероятность найти короткий интрон после другого короткого интрона составляет 65%, а вероятность найти короткий интрон после двух коротких интронов составляет 75%. Эмпирическая вероятность найти короткий интрон перед другим коротким интроном равна 62,4%, найти короткий интрон перед двумя короткими интронами равна 70,8%. Для длинных интронов соответствующие значения 49%, 62,5%, 68% и 49,0%, 65,2%, 72,0%. Наряду с приведенными наблюдениями следует отметить, что длинные интроны, как правило, имеют фазу 1, это находится в соответствии с эффектом цепей симметричных экзонов фазы 1, представленных в [15]. В таблицах 3б и 3в можно увидеть аналогичные таблицы для *Gallus gallus* и дрозофилы. Все данные приведены для внутренних интронов (см раздел «Материалы и методы»).

Аналогичные данные для других рассмотренных организмов доступны по адресу: http://server2.lpm.org.ru/static/introns_results/Appendix.htm.

В таблице 4 приведены Z-значения для увеличения количества соседних интронов сходной длины для порогов T, при которых *T-длинные* интроны составляют около 30% всех интронов. Для остальных порогов результаты, приведенные в таблицах 3а–3в, также являются статистически значимыми. Z-значения вычислялись по формуле

$$Z = \frac{N_{набл} - N \cdot p}{\sqrt{N \cdot p \cdot (1-p)}}.$$

Значения $N_{набл}$, N и p выбирались следующим образом (пояснения даются для *T-длинных* интронов, обозначения для *T-коротких* интронов аналогичны).

Таблица 3а. Доли *T-коротких* и *T-длинных* внутренних интронов в геноме *H. sapience* при различных порогах *T*

Порог	%коротких	S→S	SS→S	S←S	S←SS	%длинных	L→L	LL→L	L←L	L←LL
150	10.10%	26.20%	41.50%	25.00%	38.40%	89.90%	91.60%	92.70%	92.10%	93.50%
1000	40.20%	58.20%	71.50%	55.80%	67.50%	59.80%	71.10%	75.50%	73.10%	78.60%
1500	***51.00%***	***65.10%***	***75.00%***	***62.40%***	***70.80%***	***49.00%***	***62.50%***	***67.80%***	***65.20%***	***72.00%***
3000	70.20%	78.10%	83.20%	75.10%	79.20%	29.80%	46.40%	54.60%	50.60%	61.50%
5000	81.50%	86.70%	89.80%	83.80%	86.20%	18.50%	37.80%	48.00%	43.30%	57.70%
10000	91.00%	93.90%	95.50%	91.60%	92.60%	9.00%	31.80%	44.20%	39.60%	57.40%
20000	96.00%	97.40%	98.20%	95.90%	96.30%	4.00%	28.30%	41.10%	38.80%	54.40%
100000	99.60%	99.70%	99.80%	99.40%	99.40%	0.40%	14.90%	21.40%	25.40%	40.40%

Таблица 3б. Доли *T-коротких* и *T-длинных* внутренних интронов в геноме *G. gallus* при различных порогах *T*

Порог	%коротких	S→S	SS→S	S←S	S←SS	%длинных	L→L	LL→L	L←L	L←LL
150	11,05%	35,30%	56,30%	35,72%	56,23%	88,95%	92,01%	93,21%	91,87%	93,21%
1000	57,72%	71,61%	79,35%	69,66%	76,23%	42,28%	60,06%	66,84%	62,30%	70,45%
1500	**70,57%**	**80,92%**	**86,25%**	**78,55%**	**82,99%**	**29,43%**	**51,94%**	**60,37%**	**55,60%**	**65,76%**
3000	84,94%	90,38%	93,22%	88,27%	90,40%	15,06%	41,53%	51,48%	47,00%	59,69%
5000	91,04%	94,21%	95,96%	92,54%	93,73%	8,96%	35,77%	44,90%	42,18%	54,37%
10000	95,80%	97,19%	98,08%	96,11%	96,60%	4,20%	29,25%	39,12%	36,61%	49,50%
20000	98,27%	98,79%	99,22%	98,17%	98,36%	1,73%	23,22%	32,35%	31,44%	43,60%
100000	99,88%	99,90%	99,93%	99,84%	99,84%	0,12%	10,84%	26,67%	16,51%	36,36%

Таблица 3в. Доли *T-коротких* и *T-длинных* внутренних интронов в геноме *D. melanogaster* при различных порогах *T*

Порог	%коротких	S→S	SS→S	S←S	S←SS	%длинных	L→L	LL→L	L←L	L←LL
150	71,52%	80,17%	84,59%	73,36%	75,08%	28,48%	44,86%	53,71%	54,44%	66,30%
1000	89,53%	92,43%	93,64%	87,73%	87,24%	10,47%	26,88%	36,84%	38,55%	54,26%
1500	**92,12%**	**94,19%**	**95,15%**	**90,03%**	**89,46%**	**7,88%**	**23,17%**	**29,77%**	**35,12%**	**48,29%**
3000	95,54%	96,70%	97,31%	93,67%	93,13%	4,46%	18,82%	25,66%	31,41%	48,09%
5000	97,39%	98,01%	98,32%	95,67%	95,04%	2,61%	14,55%	19,55%	27,57%	43,21%
10000	98,90%	99,10%	99,26%	97,68%	97,19%	1,10%	9,16%	16,39%	20,93%	38,46%

Напомним, что мы рассматриваем только внутренние интроны. При оценке значимости появления двух длинных интронов подряд (столбцы *Z*2): $N_{набл}$ – количество пар двух длинных интронов подряд; N – количество длинных интронов, за которыми следует хотя бы один внутренний интрон; p – доля длинных интронов среди

всех внутренних интронов, перед которыми находится хотя бы один внутренний интрон. При оценке значимости появления трех длинных интронов подряд (столбцы *Z3a, Z3b*) мы использовали две статистические модели. В обоих случаях $N_{набл}$ – это количество троек длинных интронов, идущих подряд, N – количество пар длинных интронов, за которыми следует хотя бы один внутренний интрон. При модели, соответствующей столбцам *Z3a*, мы полагаем p равным доле длинных интронов среди всех внутренних интронов, перед которыми находятся хотя бы два внутренних интрона. При модели, соответствующей столбцам *Z3b*, мы полагаем p равным доле длинных интронов среди всех внутренних интронов, которые следуют за длинным интроном и перед которыми находятся хотя бы два внутренних интрона.

Таблица 4. Z-значения для данных из таблиц 3а–3в. Пояснения см. в тексте

Организм	Порог	%длинных интронов	*Z*-значения					
			Короткие интроны			Длинные интроны		
			Z2	***Z3a***	***Z3b***	***Z2***	***Z3a***	***Z3b***
H. sapiense	3000	29.80%	46.80	58.34	25.81	67.36	68.72	22.33
G. gallus	1500	29,43%	57.28	68.59	26.73	84.46	82.77	20.89
D. melanogaster	150	28,48%	25.61	28.08	11.50	35.24	33.35	8.39

Длины и фазы интронов

Известно, что во всех геномах существует избыток интронов в фазе 0, количества интронов в фазах 0,1,2 соотносятся примерно, как 5:3:2 [15]. Как показывают наши данные, это соотношение меняется, если рассматривать только относительно длинные интроны.

В таблице 5 показаны процентные соотношения для различных фаз при пороге *T*5 – таком пороге, что *T*5-*длинные* интроны составляют 5% всех интронов генома. В таблице 6 показаны соответствующие *Z*-значения; в таблице 7 – сведения, при каких порогах достигаются максимальные *Z*-значения для увеличения доли интронов в фазе 1. *Z*-значения вычислялись по формуле

$$Z = \frac{N_{Long}[f] - p_{All}[f] \cdot N}{\sqrt{N \cdot p_{All}[f] \cdot (1 - p_{All}[f])}}.$$

Здесь N – общее количество рассматриваемых интронов, $P_{All}[f]$ – доля интронов в фазе f среди всех интронов, N_{Long} – количество длинных интронов (при выбранном пороге) в фазе f.

Таблица 5. Доли интронов в разных фазах среди всех внутренних интронов и среди 5% наиболее длинных внутренних интронов. Порог *T*5 – это порог, при котором *T*5-*длинные* интроны составляют 5% всех интронов генома

№	Организм	К-во внутр. интронов (КВИ)	5% от КВИ	Порог T5	Процентное содержание фаз 0, 1 и 2					
					Все внутренние интроны			T5-длинные внутренние интроны		
					Фаза 0	Фаза 1	Фаза 2	Фаза 0	Фаза 1	Фаза 2
1	*Apis mellifera*	29494	1475	3530	44.49%	30.82%	24.69%	40.08%	38.56%	21.35%
2	*Drosophila melanogaster*	19659	983	3160	41.68%	31.13%	27.19%	37.93%	40.97%	21.10%
3	*Nasonia vitripennis*	31554	1578	2450	44.72%	30.85%	24.43%	41.02%	36.95%	22.03%
4	*Tribolium castaneum*	22703	1135	4270	43.77%	31.85%	24.38%	41.25%	36.94%	21.81%
5	*Danio rerio*	114893	5745	9160	45.48%	32.37%	22.16%	40.46%	37.53%	22.01%
6	*Xenopus tropicalis*	80533	4027	8440	46.54%	31.35%	22.11%	42.26%	36.13%	21.61%
7	*Anolis carolinensis*	73632	3682	12990	46.71%	30.08%	22.42%	41.93%	35.47%	20.89%
8	*Gallus gallus*	81916	4096	8590	46.22%	31.23%	22.55%	43.65%	35.52%	20.83%
9	*Meleagris gallopavo*	51699	2585	8970	46.89%	30.36%	22.75%	44.58%	34.09%	21.32%
10	*Taeniopygia guttata*	62216	3111	9940	46.36%	30.61%	23.03%	43.84%	34.32%	21.84%
11	*Canis lupus familiaris*	93413	4671	15450	46.19%	31.03%	22.78%	42.32%	34.92%	22.76%
12	*Mus musculus*	108206	5410	12850	45.96%	31.51%	22.53%	42.27%	35.98%	21.75%
13	*Sus scrofa*	75050	3753	13670	45.68%	31.49%	22.83%	42.70%	33.94%	23.36%
14	*Callithrix jacchus*	78234	3912	19080	45.84%	31.44%	22.72%	42.70%	36.09%	21.21%
15	*Macaca mulatta*	79200	3960	18260	45.97%	31.32%	22.71%	43.12%	35.02%	21.86%
16	*Pan troglodytes*	90139	4507	18170	45.92%	31.62%	22.46%	42.76%	34.93%	22.31%
17	*Homo sapiens*	94067	4703	16790	46.26%	31.22%	22.51%	43.19%	35.27%	21.53%

Таблица 6. *Z*-значения для данных таблицы

№	Организм	Порог T5	Z-значения		
			Фаза 0	Фаза 1	Фаза 2
1	*Apis mellifera*	3530	–3.41	6.36	–2.88
2	*Drosophila melanogaster*	3160	–2.43	6.71	–4.30
3	*Nasonia vitripennis*	2450	–2.95	5.22	–2.19
4	*Tribolium castaneum*	4270	–1.77	3.77	–2.05
5	*Danio rerio*	9160	–6.98	7.61	–0.21
6	*Xenopus tropicalis*	8440	–5.49	6.55	–0.72
7	*Anolis carolinensis*	12990	–5.79	7.11	–2.23
8	*Gallus gallus*	8590	–3.29	5.83	–2.55
9	*Meleagris gallopavo*	8970	–2.39	4.19	–1.75
10	*Taeniopygia guttata*	9940	–2.83	4.53	–1.61
11	*Canis lupus familiaris*	15450	–5.30	5.75	–0.04
12	*Mus musculus*	12850	–5.45	7.08	–1.37
13	*Sus scrofa*	13670	–3.67	3.21	0.80
14	*Callithrix jacchus*	19080	–3.89	6.28	–2.34
15	*Macaca mulatta*	18260	–3.59	5.03	–1.29
16	*Pan troglodytes*	18170	–4.26	4.79	–0.25
17	*Homo sapiens*	16790	–4.22	5.99	–1.61

Таблица 7. Значения порогов *TZmax*, при которых достигается максимальное *Z*-значение для увеличения доли интронов в фазе 1. Во всех случаях порог *T*Zmax меньше порога *T*5 (см. таб. 5, 6), т.е. *TZmax-длинные* интроны, составляют более 5% всех интронов.

№	Организм	Порог Tzmax	Z-значения		
			Фаза 0	Фаза 1	Фаза 2
1	*Apis mellifera*	1300	–5.35	8.77	–3.23
2	*Drosophila melanogaster*	360	–1.53	8.29	–6.94
3	*Nasonia vitripennis*	410	–4.30	7.07	–2.62
4	*Tribolium castaneum*	2290	–1.89	4.90	–3.13
5	*Danio rerio*	6850	–7.65	8.28	–0.16
6	*Xenopus tropicalis*	9960	–5.52	7.15	–1.36
7	*Anolis carolinensis*	6880	–6.42	7.85	–1.77
8	*Gallus gallus*	3050	–5.58	7.02	–1.13
9	*Meleagris gallopavo*	3500	–4.59	6.94	–2.14
10	*Taeniopygia guttata*	2500	–5.81	6.81	–0.57
11	*Canis lupus familiaris*	9990	–5.96	6.71	–0.31
12	*Mus musculus*	11190	–6.11	8.11	–1.73
13	*Sus scrofa*	11700	–3.42	3.83	–0.18
14	*Callithrix jacchus*	17100	–4.35	6.79	–2.36
15	*Macaca mulatta*	13330	–4.38	5.23	–0.58
16	*Pan troglodytes*	12580	–5.00	5.30	0.06
17	*Homo sapiens*	13800	–5.45	6.94	–1.20

ЗАКЛЮЧЕНИЕ

Показано, что у всех рассмотренных организмов процентное соотношение фаз среди *T*-длинных интронов меняется с увеличением порога *T*. Для всех видов организмов с ростом порога *T* доля интронов в фазе 1 растет, а доля интронов в фазе 0 убывает и при определенном значении порога доли интронов в фазах 1 и 0 сравниваются. Это значение порога различно для организмов различных таксонов. Для насекомых это значение равно примерно 20000 нуклеотидных пар (нп); для рыб и земноводных несколько ниже (соответственно 18000 нп и 19000 нп), у рептилий (~ 50000 нп), а также у птиц и млекопитающих (~ 110000 нп) – существенно выше. При этом указанный эффект изменения долей интронов в фазах 1 и 0 становится статистически значимым при существенно меньших значениях порога *T*. Для насекомых при $T = 350$ Z-значение для значимости увеличения количества интронов в фазе 1 составляет примерно 13 (общее количество интронов длины более 350 – более 20000). Для млекопитающих при $T = 14000$ *Z*-значение для значимости увеличения количества интронов в фазе 1 составляет примерно 15 (общее количество интронов длины более 14000 у млекопитающих более 35000).

Показано, что соседний интрон длинного (короткого) интрона склонен также быть длинным (коротким) интроном. Эффект был продемонстрирован для различных таксонов и порогов. Показано также, что эффект усиливается, если рассматривать не пары, а тройки интронов. При этом эффект для троек не сводится к наложению эффектов для пар. Следует отметить, что длинные интроны часто имеют фазу 1, что находится в соответствии с эффектом цепей симметричных экзонов фазы 1, представленных в [15].

Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 12-04-00944).

СПИСОК ЛИТЕРАТУРЫ

1. Fickett J.W. The gene identification problem: an overview for developers. *Computer & Chemistry*. 1996. V. 20. P. 103.
2. Burge C.B., Karlin S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol*. 1998. V. 8. P. 346–354.
3. Щепеткова И.Л., Гельфанд М.С. Некоторые статистические особенности сайтов сплайсинга позвоночных и беспозвоночных. *Биофизика*. 1997. Т. 42. № 1. С. 82–89.
4. Moss S.P., Joyce D.A., Humphries S., Tindall K.J., Lunt DH. Comparative analysis of teleost genome sequences reveals an ancient intron size expansion in the zebrafish lineage. *Genome Biol. Evol*. 2011. V. 3. P. 1187–1196. doi: 10.1093/gbe/evr090.
5. Chen D., Zhang J. Analysis of intron sequence features associated with transcriptional regulation in human genes. *PLoS ONE*. 2012. V. 7. № 10. P. e46784. doi: 10.1371/journal.pone0046784.
6. Rogozin I.B, Carmel L., Csuros M., Koonin E.V. Origin and evolution of spliceosomal introns. *Biol. Direct*. 2012. V. 7. P. 11. doi: 10.1186/1745-6150-7-11.
7. Bradnam K.R., Korf I. Longer first introns are a general property of eukaryotic gene structure. *PLoS ONE*. 2008. V. 3. № 8. Article No. e3093. doi:10.1371/journal.pone.0003093.
8. Shuang W., Zhang Z., Jun YU. Systematic analysis of intron size and abundance parameters indiverse lineages. *SCIENCE CHINA Life Sciences.* 2013. V. 56. № 10. P. 968–974.
9. Shepard S., McCreary M., Fedorov A. The peculiarities of large intron splicing in animals. *PLoS ONE*. 2009. V. 4. № 11. Article No. e7853. doi:10.1371/journal.pone.0007853.
10. Farlow A., Dolezal M., Hua L., Schlotterer C. The genomic signature of splicing-coupled selection differs between long and short introns. *Mol. Biol. Evol*. 2012. V. 29. № 1. P. 21–24.
11. Marais G., Nouvellet P., Keightley P.D., Charlesworth B. Intron size and exon evolution in drosophila. *Genetics*. 2005. V. 170. P. 481–485.
12. Vinogradov A. “Genome design” model: Evidence from conserved intronic sequence in human–mouse comparison. *Genome Res*. 2006. V. 16. № 3. P. 347–54.
13. Catania F., Lynch M. A simple model to explain evolutionary trends of eukaryotic gene architecture and expression: How competition between splicing and cleavage/polyadenylation factors may affect gene expression and splice-site recognition in eukaryotes. *Bioessays*. 2013. V. 35. № 6. P. 561-570. doi: 10.1002/bies.201200127.
14. Zhu L., Zhang Y., Zhang W., Yang S., Chen J.-Q., Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*. 2009. V. 10. P. 47–53.
15. Long M., Rosenberg C., Gilbert W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci*. 1995. V. 92. № 26. P. 12495–12499.

Материал поступил в редакцию 17.11.2014, опубликован 12.12.2014.