САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
КАФЕДРА ГЕНЕТИКИ И БИОТЕХНОЛОГИИ

Васильев Артем Викторович
Выпускная квалификационная работа

“Эволюционные особенности структуры гена *Nxf1* (nuclear export factor) у животных“

Научный руководитель:
к.б.н., доцент, кафедра генетики и биотехнологии,
Голубкова Елена Валерьевна

Рецензент:
заведующая лабораторией, ведущий научный сотрудник,
лаборатория эволюционной геномики и палеогеномики, ЗИН,
к.б.н., с.н.с.,
Абрамсон Наталья Иосифовна

Санкт-Петербург
2025

Оглавление

Материалы и методы

В качестве отправной точки был произведен поиск гена *Nxf1* внутри веб-сервиса NCBI [1]. Полученные данные были сохранены в текстовом формате и загружены в виде tsv-таблицы с помощью пакета pandas v2.2.3 [2] для языка программирования Python v3.12.6 [3]. Всего был найден 651 организм, содержащий анализируемый ген, большинство из которых относятся к Deuterostomia (Вторичноротые) - 436 видов. Таким образом, в качестве материалов выступали нуклеотидные и белковые последовательности гена *Nxf1* из открытых баз данных NCBI [1].

Большинство этапов последующего анализа реализовано в виде отдельных скриптов, разработанных в рамках данной работы, если не указано другое. Для логического разделения на блоки был использован Jupyter Notebook v1.1.1 [4].

По данным из полученной таблицы в разведывательных целях было построено филогенетическое дерево по найденным видам для оценки количества видов в таксонах более низкого ранга. Для глубокого анализа было принято решение сфокусироваться на организмах, относящихся к группе Protostomia (Первичноротые), Cnidaria (Стрекающие), а также на всех группах из Deuterostomia за исключением Mammalia (Млекопитающие).

Для найденных организмов с помощью пакета NCBI E-utilities из BioPython v1.85 [5] и NCBI Datasets Command-Line Interface (CLI) v18.0.2 [6] были загружены нуклеотидные последовательности гена, кодирующих участков и мРНК, а также аминокислотные последовательности белка в формате FASTA и аннотации для гена в GenBank-формате, необходимые для получения нуклеотидных последовательностей экзонов и поиска “консервативной кассеты“. Затем были получены и проанализированы интересующие нас участки экзон-интрон-экзонной структуры и созданы файлы со всеми экзонами и “кассетным“ интроном для всех организмов, у которых получилось найти “кассету“. Данные файлы будут необходимы для последующего анализа.

Учитывая очень маленькие выборки во многих анализируемых группах (например, Cnidaria - 4 вида, Spiralia - 9 видов), было принято решение по увеличению их количества. Для этой цели, учитывая разнообразия полученных генов даже внутри одной таксономической группы, самым эффективным вариантом оказалось использование PSI-BLAST [7]. В качестве запроса (Query), или референса, использовались белковые последовательности тех организмов, у которых была найдена “кассета“. Для проведения PSI-BLAST были выбраны настройки по-умолчанию за исключением параметра Organism: поиск проводился внутри таксономической группы, к которой принадлежал референс, также референс был исключен из поиска.

Парсинг результатов BLAST также осуществлялся с помощью пакета BioPython [5] и специально разработанных скриптов. Он включал в себя фильтрацию данных по параметрам процента покрытия (Query Coverage, QC), длине и сходству (Per. Ident) найденных последовательностей (Subject), а также загрузку нуклеотидных и белковых последовательностей, однако реализация отличалась из-за особенностей баз

данных NCBI [1]. Получение “кассеты“ было произведено по тому же принципу, но, опять же, с отличиями. Благодаря данному шагу удалось увеличить выборки суммарно на 117 видов. К сожалению, для некоторых таксономических групп увеличение выборки оказалось невозможным в связи с отсутствием у некоторых организмов интересующего нас участка.

Множественные выравнивания осуществлялись с помощью алгоритма MAFFT [8], 10 итераций, остальные настройки по-умолчанию, в программе Unipro UGENE v52.0 [9].

Анализ видов из Deuterostomia изначально шел более благоприятно за счет большого сходства последовательностей, в том числе интронных, и большего количества видов в группах. Для них также были загружены все необходимые файлы и произведен поиск и анализ “консервативной кассеты“. Мы решили сосредоточить свое внимание на организмах из Actinopterygii (Лучеперые рыбы), 72 вида, так как данных по ним ранее получено не было. Учитывая большую степень сходства интронных последовательностей, с помощью пакета инструментов MEME Suite v5.5.8 [10] локально был произведен поиск консервативных мотивов внутри “кассетного“ интрона. Найденные мотивы, у которых E-value < 0.05 также локально были проанализированы с помощью Tomtom [11] из того же пакета. Для описанного шага была взята база данных JASPAR2024 CORE (NON-REDUNDANT) DNA.

С помощью инструмента RNAfold v2.7.0 из пакета ViennaRNA [12] были построены вторичные структуры РНК для нуклеотидных последовательностей в двух вариантах (MFE и Centroid), содержащих экзоны и “кассетный“ интрон, т.к. мы предполагаем, что избегание интроном сплайсинга может быть опосредовано образованной им специфической вторичной структурой. Учитывая данное предположение, разумным шагом также являлся анализ “силы сайтов сплайсинга“, проведенный с помощью MaxEntScan [13]. Также с помощью скриптов цветом были выделены интронные последовательности внутри вторичной структуры и найденный мотив у Actinopterygii, который предположительно является CTE (Constitutive Transport Element).

Для Actinopterygii также был проведен филогенетический анализ, включающий построение и визуализацию деревьев. Для данной цели использовались самые популярные и проверенные временем инструменты. Построение деревьев осуществлялось с помощью IQ-TREE v2.4.0 [14], визуализация - с помощью Figtree v1.4.4 [15].

Работа проводилась в виртуальном окружении Mamba v1.5.5 [16], использованные пакеты и примеры анализа в Jupyter Notebooks можно найти в GitHub [17] репозитории автора: https://github.com/ArtemVaska/Diploma.

Для написания ВКР была использована система верстки LaTeX v4.76 [18], таблицы генерировались в веб-сервисе TablesGenerator [19]. Большинство картинок создано с помощью веб-сервиса draw.io [20]. Все шаги анализа проводились на базе операционной системы Linux Ubuntu 22.04 [21].

Результаты

Анализ всех найденных видов

Были проанализированы 413 нуклеотидных и белковых последовательностей гена *Nxf1* у представителей различных филогенетических групп из клад Cnidaria (Стрекающие) и Bilateria (Двусторонне-симметричные). Организмы, относящиеся к Mammalia, в анализ не были взяты в связи с уже имеющимися для них данными.

Для таксономических групп более низкого ранга с небольшим количеством видов в них с помощью PSI-BLAST были увеличены выборки, где это оказалось возможным, результат продемонстрирован на таблице 1.

Таблица 1: Результат увеличения выборки с помощью PSI-BLAST.

Филогенетическая группа	**Таксон высокого ранга**	**Видов до PSI-BLAST**	**Видов добавлено**	**Итого видов**
Bilateria→Protostomia	Ecdysozoa	56	42	98
	Spiralia	6	63	69
Cnidaria	Anthozoa	2	12	14

В итоге для 353 видов удалось найти “консервативную кассету“ и продолжить дальнейший анализ.

На рисунке 1 отображено распределение исследованных видов по таксонам высокого ранга.

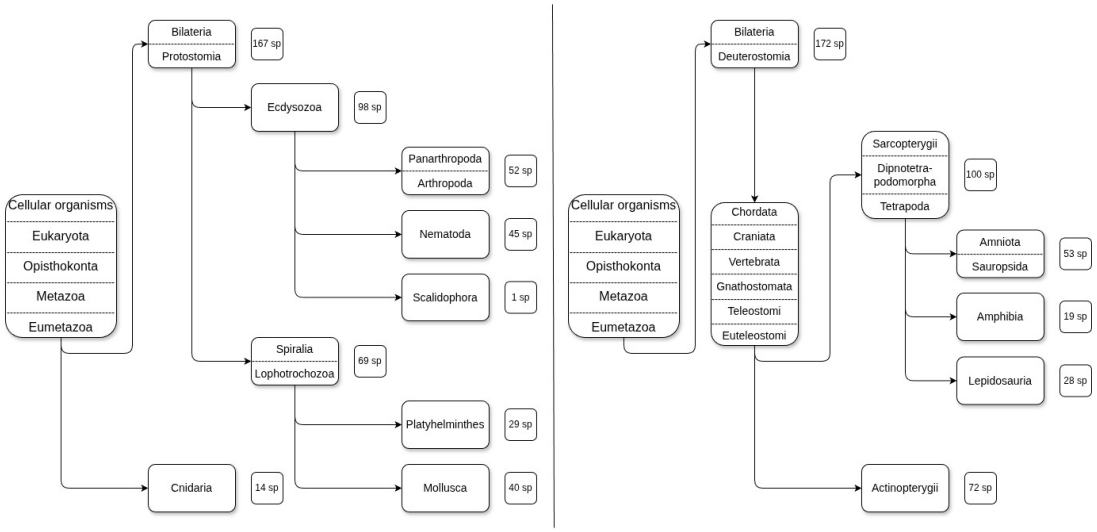

Рис. 1: Количество видов, взятых в анализ, для разных таксономических групп.

Для всех видов, имеющих “консервативную кассету“, были построены вторичные структуры для интрон-содержащего транскрипта с выделением цветом “кассетного“ интрона (предоставляется по запросу).

Подробный анализ Actinopterygii

Для таксономической группы Actinopterygii проводился более углубленный анализ, так как на текущий момент данных по гену *Nxf1* для них не было. Были взяты все найденные нуклеотидные последовательности гена у представители данной филогенетической группы - 72 вида.

На таблице 2 показана характеристика “консервативной кассеты“ исследуемой группы. Результаты по другим группам можно найти в приложении, таблицы 3–8.

Таблица 2: Сводная таблица с характеристикой кассетного интрона для таксономической группы Actinopterygii. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в “кассетном“ интроне.

Название организма	**Кол-во нуклеотидов до стоп-кодона в интроне**	**Длина 1-го экзона в кассете**	**Длина кассетного интрона**	**Длина 2-го экзона в кассете**
Chanos chanos	1	110	3568	37
Danio rerio	1	110	3580	37
Denticeps clupeoides	7	110	2629	37
Labrus bergylta	10	110	2684	37
Cottoperca gobio	16	110	2388	37
Xiphophorus couchianus	22	110	2227	37
Larimichthys crocea	22	110	2340	37
Lates calcarifer	22	110	2434	37
Notothenia coriiceps	22	110	2886	37
Betta splendens	22	110	2274	37
Poecilia reticulata	22	110	2262	37
Takifugu rubripes	22	110	2114	37
Salarias fasciatus	22	110	3855	37
Poecilia mexicana	22	110	2247	37
Stegastes partitus	22	110	2900	37
Clupea harengus	22	110	3219	37
Archocentrus centrarchus	22	110	2644	37
Esox lucius	22	110	2848	37
Monopterus albus	22	110	2353	37
Echeneis naucrates	22	110	2314	37
Paralichthys olivaceus	22	110	3148	37
Maylandia zebra	22	110	2565	37
Parambassis ranga	22	110	2484	37
Sander lucioperca	22	110	2494	37
Xiphophorus maculatus	22	110	2231	37
Nothobranchius furzeri	22	110	2290	37
Anabas testudineus	22	110	2352	37
Acanthochromis polyacanthus	22	110	2797	37
Anarrhichthys ocellatus	22	110	2355	37
Boleophthalmus pectinirostris	22	110	1702	37
Sparus aurata	22	110	2361	37
Oryzias melastigma	22	110	2212	37
Seriola dumerili	22	110	2494	37
Poecilia formosa	22	110	2259	37
Oreochromis niloticus	22	110	2580	37
Kryptolebias marmoratus	22	110	2556	37
Xiphophorus hellerii	22	110	2240	37
Poecilia latipinna	22	110	2261	37
Pundamilia nyererei	22	110	2527	37

Hippocampus comes	22	110	2622	37
Oreochromis aureus	22	110	2579	37
Amphiprion ocellaris	22	110	2752	37
Seriola lalandi dorsalis	22	110	2481	37
Austrofundulus limnaeus	22	110	2541	37
Puntigrus tetrazona	25	110	2440	37
Fundulus heteroclitus	25	110	2476	37
Cyprinodon variegatus	28	110	2533	37
Haplochromis burtoni	31	110	2535	37
Astatotilapia calliptera	31	110	2571	37
Gouania willdenowi	37	110	2616	37
Oryzias latipes	40	110	2331	37
Sphaeramia orbicularis	43	110	2376	37
Pygocentrus nattereri	46	110	2649	37
Astyanax mexicanus	46	110	2791	37
Colossoma macropomum	46	110	2644	37
Ictalurus punctatus	46	110	3166	37
Tachysurus fulvidraco	46	110	3493	37
Pangasianodon hypophthalmus	46	110	3348	37
Erpetoichthys calabaricus	55	110	3662	37
Perca flavescens	58	110	2378	37
Mastacembelus armatus	64	110	2371	37
Salmo salar	67	110	3553	37
Gadus morhua	67	110	3151	37
Etheostoma spectabile	97	110	2457	37
Scleropages formosus	112	110	3412	37
Myripristis murdjan	112	110	2492	37
Paramormyrops kingsleyae	121	110	2929	37
Carassius auratus	148	110	3854	37
Sinocyclocheilus grahami	148	110	3330	37
Sinocyclocheilus rhinocerous	154	110	3449	37
Sinocyclocheilus anshuiensis	154	110	4202	37
Electrophorus electricus	283	110	2874	37

На рисунках 2 и 3 показано распределение длин части “кассетного“ интрона до стоп-кодона и длин “кассетного“ интрона, соответственно.

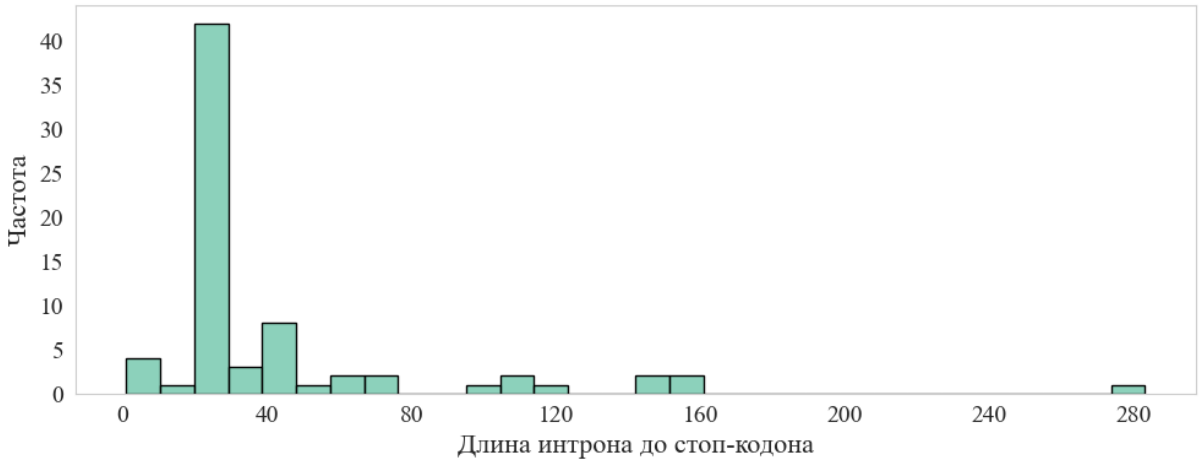

Рис. 2: Распределение длин части кассетного интрона до стоп-кодона у таксономической группы Actinopterygii

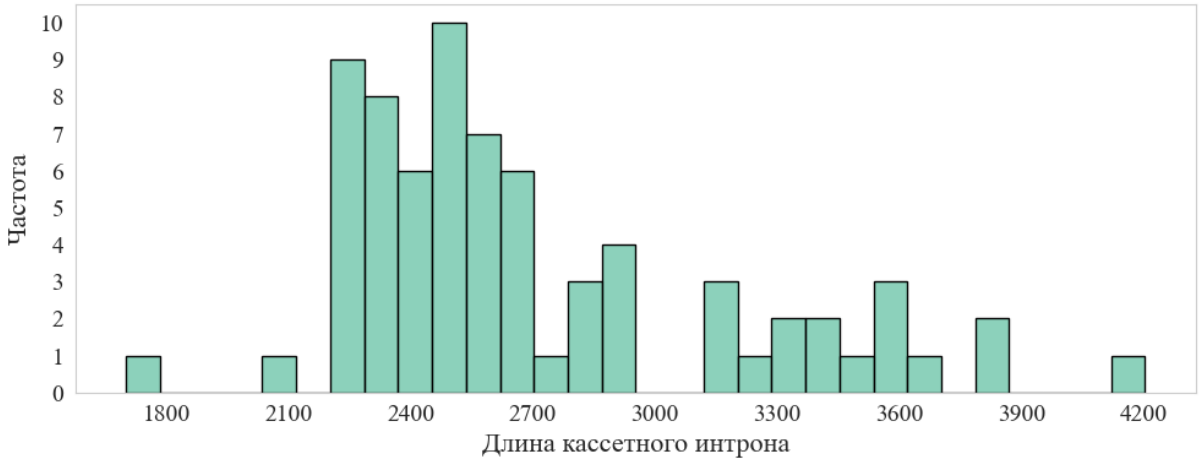

Рис. 3: Распределение длин кассетного интрона у таксономической группы Actinopterygii

На картинке 4 представлены результаты оценки “силы сайтов сплайсинга“ - “ящики с усами“, отображающие распределение MaxEntScan score для таксонов более низкого ранга внутри группы Actinopterygii. Разбиение на подгруппы основано на их удаленности друг от друга. Порядок групп на графике не несет смысловой нагрузки.

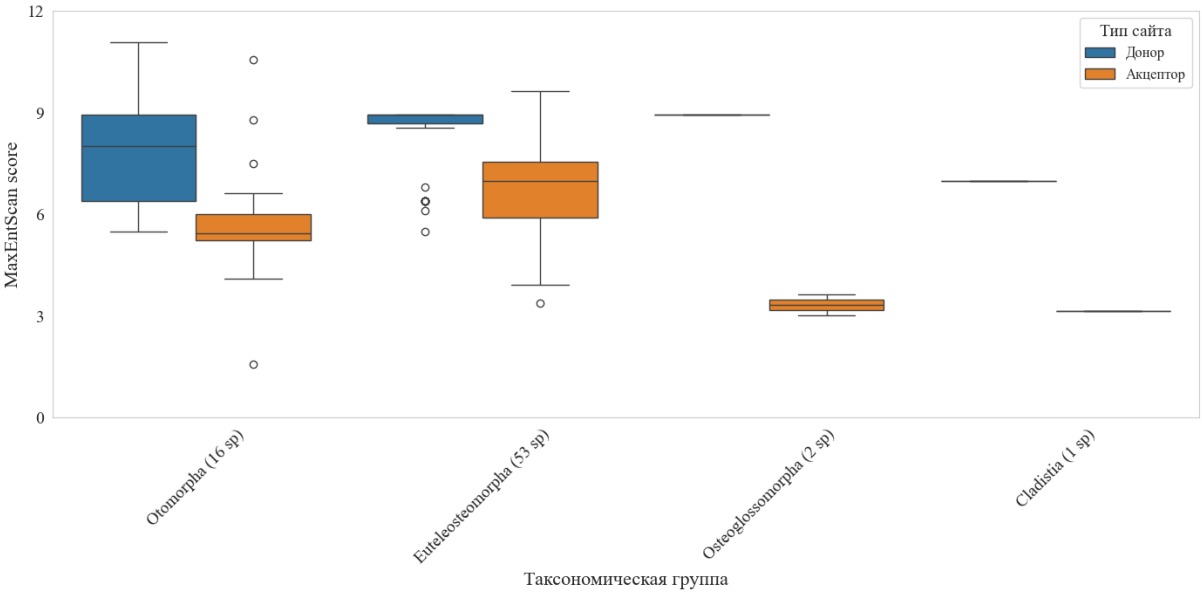

Рис. 4: Результаты проведения MaxEntScan для Actinopterygii.

Рисунок 5 демонстрирует результаты, полученные с помощью MEME Suite.

Найденные мотивы присутствуют не у всех видов, взятых в анализ изначально, их количество отображено в столбце Sites. Нас заинтересовал 2-й найденный мотив, так как его начало очень похоже на предложенную авторами статьи 2001 года [22] консенсусную последовательность для CTE из рисунка 6.

К сожалению, использование Tomtom для поиска найденных консервативных мотивов из “кассетного“ интрона в базе данных мотивов не дало статистически значи-

мых результатов.

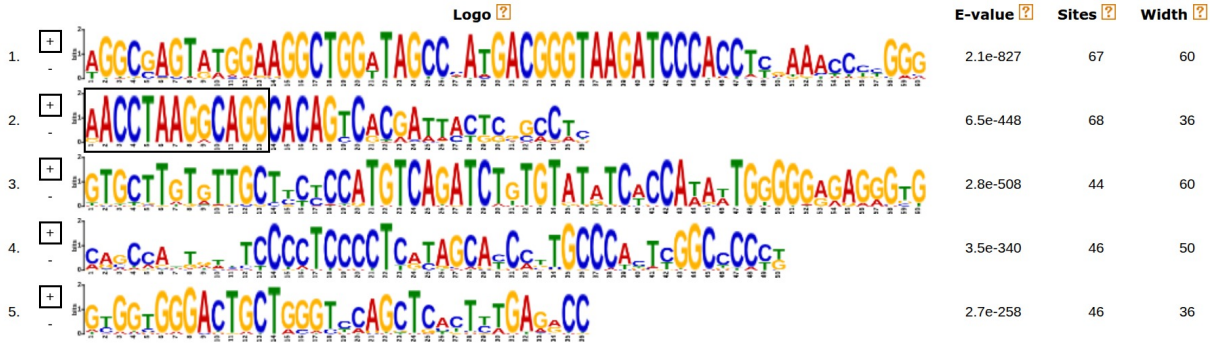

Рис. 5: Результат поиска мотивов внутри кассетного интрона с помощью MEME Suite для таксономической группы Actinopterygii.

Черным прямоугольником выделен участок, похожий на консенсусную последовательность CTE (рис. 6) из статьи 2001 года [22].

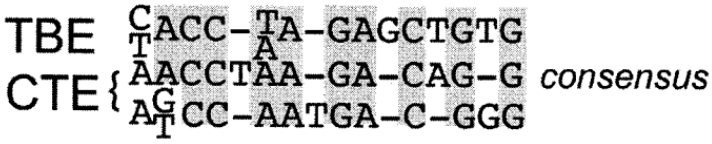

Рис. 6: Консенсусный конститутивный транспортный элемент [22].

Репрезентация вторичной структуры интрон-содержащего транскрипта с выделенным кассетным интроном и найденным мотивом показана на рисунке 7. Вид для демонстрации был выбран случайно.

Учитывая тот факт, что мотив с интересующим нас участком, был найден у 68 видов, именно для них был проведен последующий анализ.

Рисунок 8 отображает результаты множественного выравнивания, а на рисунке 9 представлено филогенетическое дерево, построенное по результатам этого выравнивания.

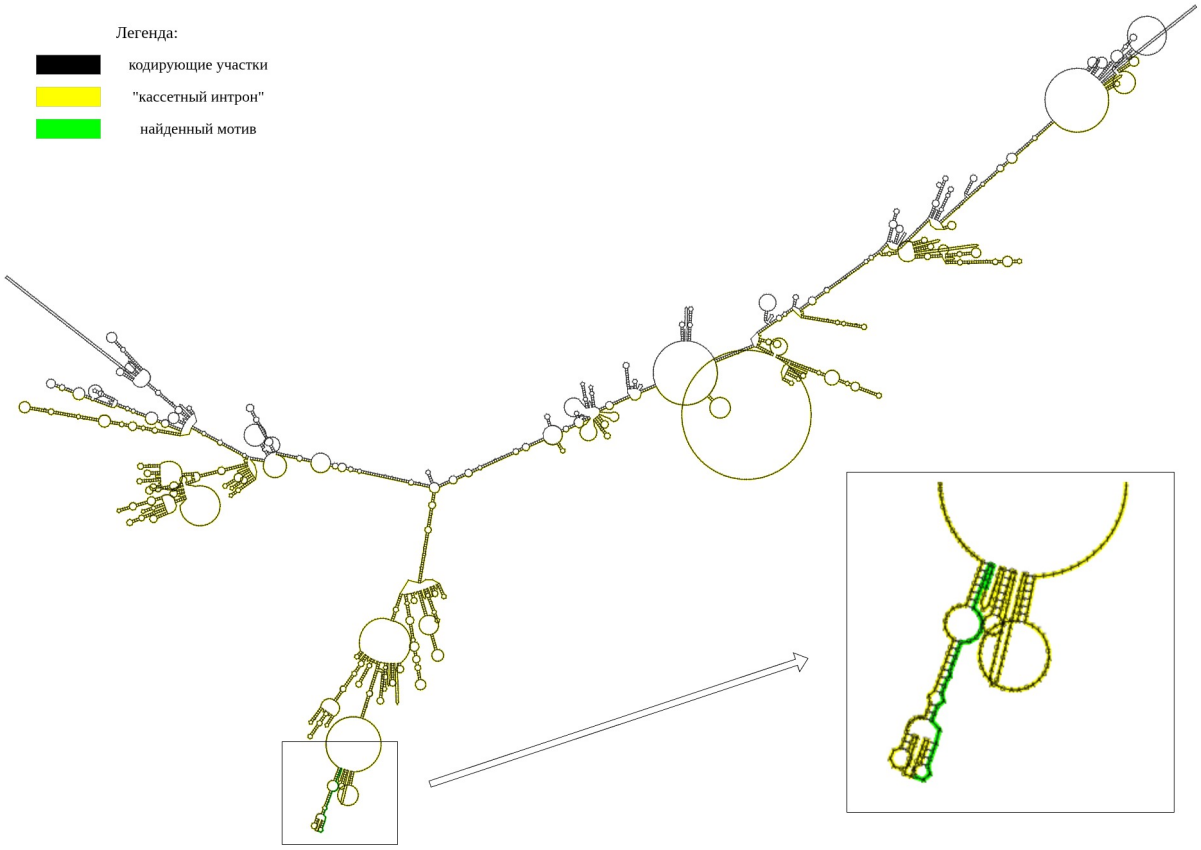

Рис. 7: Вторичная структура РНК-транскрипта для *Chanos chanos* из Otomorpha, содержащая кассетный интрон.

Рис. 8: Результаты множественного выравнивания для Actinopterygii.

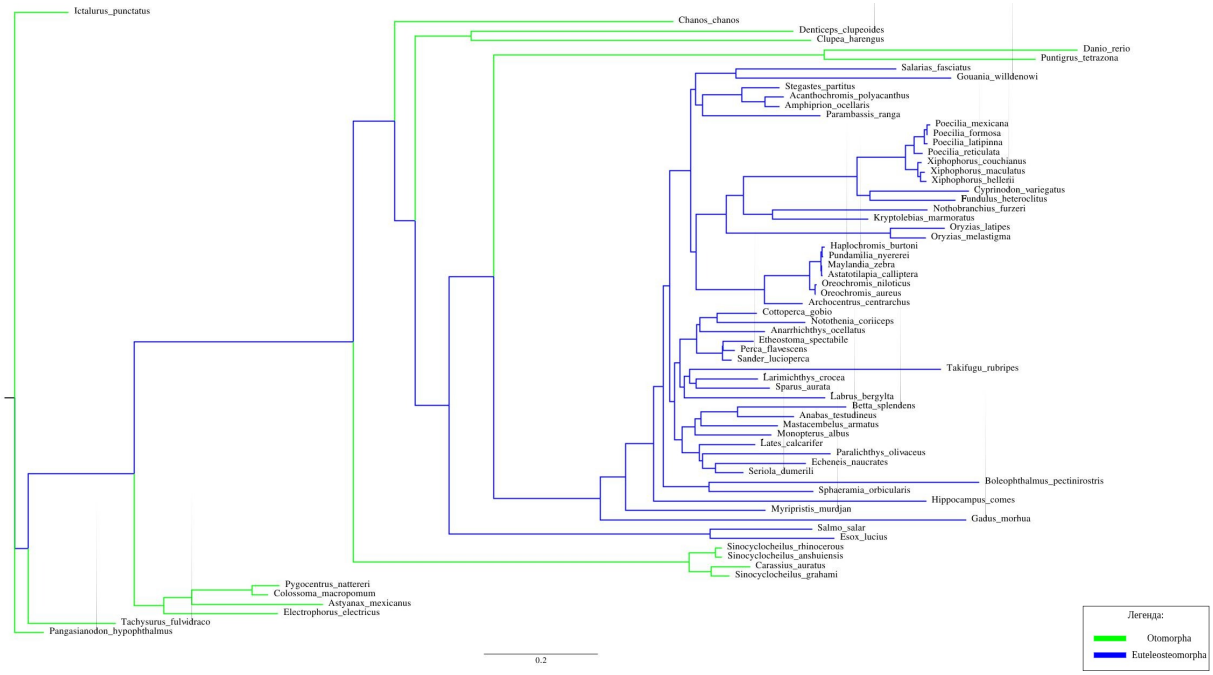

Рис. 9: Филогенетическое дерево для таксономической группы Actinopterygii.

Обсуждение

Анализ всех найденных видов

У всех проанализированных видов размер второго экзона из “консервативной кассеты“ равен 37 нуклеотидам, в то время как размер первого экзона варьирует в различных группах. На рисунке 10 показано распределение длины первого экзона из “кассеты“ для Protostomia.

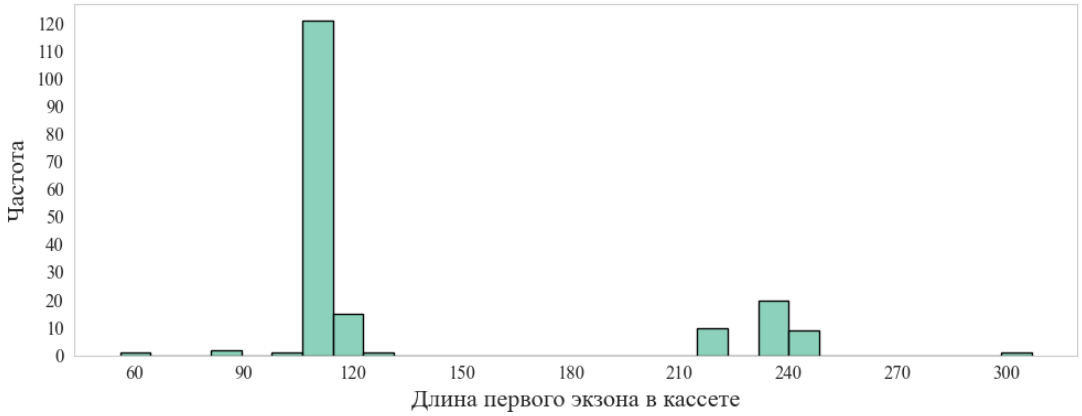

Рис. 10: Распределение длины первого экзона из “консервативной кассеты“ для Protostomia.

Для Ecdysozoa и Cnidaria первый экзон как правило размером 110 нуклеотидов, но встречаются и исключения. У Spiralia размер этого экзона гораздо больше и чаще всего составляет 239 нуклеотидов. По данному отличию и встречающимся уникальным вариантам размера экзона требуется углубленное исследование.

У Deuterostomia размер первого экзона в абсолютном большинстве случаев (171 из 172 исследованных видов) составляет 110 нуклеотидов, что также характерно и для млекопитающих.

Длина участка внутри интрона до стоп-кодона, как и длина самого интрона, варьирует в более широких пределах в разных группах. Тем не менее внутри отдельных групп, например Lepidosauria (таблица 8 в приложении), наблюдается высокая степень консервативности обоих параметров.

Также встречаются виды, у которых происходит частичная или даже полная трансляция “кассетного“ интрона, потому что в нем не встречается преждевременный стоп-кодон. Например, таким видом является давно известный *Caenorhabditis elegans*, у которого преждевременный стоп-кодон встречается в одном из экзонов после “кассетного“ интрона. В данном исследовании были найдены еще 2 вида, у которых интрон полностью считывается: Aves - *Vidua chalybeata*, Paraneoptera - *Rhopalosiphum maidis*. Упомянутые виды также требуют тщательного изучения.

Подробный анализ Actinopterygii

Данная группа организмов была исследована более подробно по перечисленным ранее причинам. Внутри группы размеры первого и второго экзона из “консервативной кассеты“ для всех исследованных видов составляют 110 и 37 нуклеотидов, соответственно. Длина участка внутри интрона до стоп-кодона у большинства видов составляет 22 нуклеотида (39 из 72 исследованных в работе). Размер “кассетного“ интрона варьирует от 1702 до 4202 нуклеотидов (в среднем 2705).

Анализ “сайтов силы сплайсинга“ 4 говорит о том, что практически у всех видов данный интрон успешно вырезается сплайсосомой. Учитывая большую выборку видов, взятую для анализа, было принято решение ориентироваться на эмпирическую интерпретацию результатов, которая выглядит следующим образом:

- 0–3: слабый сайт сплайсинга
- 3–6: умеренный сайт сплайсинга
- >6: сильный сайт сплайсинга

Так как у большинства видов значение MaxEntScan score больше или около 6, был сделан вывод, высказанный выше. Соответственно, невырезание сплайсосомой как минимум у данной группы не является причиной альтернативного сплайсинга с сохранением “кассетного“ интрона.

В связи с этим и было принято решение о поиске консервативных мотивов внутри “кассетного“ интрона. Несмотря на то, что на рисунке 5 представлено 5 найденных мотивов, их количество может быть больше, потому что данное значение мотивов было ограничением запуска MEME Suite. Учитывая высокую степень сходства начала 2-го найденного мотива с консенсусной последовательностью CTE из рисунка 6, можно предположить сохранение интрона благодаря этой и возможно другим структурам внутри интрон-содержащего транскрипта (рисунок 7).

Проведенное множественное выравнивание на рисунке 8 говорит о высокой степени консервативности как кодирующих участков (левый и правый крайние части диаграммы под выравниванием), так и некоторых участков внутри интрона (центр диаграммы под выравниванием). Филогенетическое древо (рисунок 9), построенное по результатам выравнивания, несмотря на наличие “кассетного“ интрона в последовательности, использованной для его построения, успешно разделяет виды на таксоны более высокого ранга - Otomorpha и Euteleosteomorpha.

Остальные группы, не включенные в подробный анализ, требуют его проведения.

Выводы

По результатам проведенной работы были сформулированы следующие выводы:

1. Внутри одной таксономической группы существуют преобладающие значения для характеристик “консервативной кассеты“:
    - длина первого и второго экзона;
    - длина “кассетного“ интрона;
    - длина участка внутри “кассетного“ интрона до стоп-кодона.
2. Внутри “кассетного“ интрона существуют участки, которые образуют особые структуры при формировании вторичной структуры интрон-содержащего транскрипта, и за счет их наличия возможно сохранение такого транскрипта и последующая трансляция укороченной формы белка.

Приложение

Таблица 3: Сводная таблица с характеристикой кассетного интрона для таксономической группы Ecdysozoa. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в кассетной интроне.

Название организма	Кол-во нуклеотидов до стоп-кодона в интроне	Длина 1-го экзона в кассете	Длина кассетного интрона	Длина 2-го экзона в кассете
Trichinella spiralis	1	83	417	37
Priapulus caudatus	1	110	2114	37
Galendromus occidentalis	1	110	1491	37
Ixodes scapularis	1	110	3567	37
Limulus polyphemus	1	110	915	37
Parasteatoda tepidariorum	1	110	1725	37
Cryptotermes secundus	1	110	4335	37
Maniola hyperantus	1	110	920	37
Cimex lectularius	1	110	4437	37
Vespa mandarinia	1	113	379	37
Zerene cesonia	1	110	1162	37
Pararge aegeria	1	110	2657	37
Myzus persicae	1	107	772	37
Halyomorpha halys	1	110	7270	37
Diuraphis noxia	1	107	742	37
Sipha flava	1	107	58	37
Manduca sexta	1	110	1796	37
Apis laboriosa	1	113	1254	37
Orussus abietinus	1	113	74	37
Danaus plexippus	1	110	1009	37
Colletes gigas	1	113	379	37
Ostrinia furnacalis	1	110	1946	37
Vespa crabro	1	113	381	37
Venturia canescens	1	113	621	37
Papilio polytes	1	110	1674	37
Vespa velutina	1	113	377	37
Cephus cinctus	1	113	75	37
Bombus pyrosoma	1	113	244	37
Papilio xuthus	1	110	999	37
Vanessa tameamea	1	110	2352	37
Megalopta genalis	1	113	373	37
Vespula pensylvanica	1	113	363	37
Leptopilina heterotoma	1	113	921	37
Acromyrmex echinatior	1	113	438	37
Aphidius gifuensis	1	113	240	37
Polistes fuscatus	1	113	400	37
Dirofilaria immitis	7	98	248	37
Odontomachus brunneus	10	113	498	37
Diploscapter pachys	10	110	662	37
Bactrocera dorsalis	13	110	1808	37
Drosophila melanogaster	13	110	1602	37
Ceratitis capitata	19	110	2023	37
Pediculus humanus corporis	19	110	631	37
Aphelenchoides avenae	19	110	441	37
Litomosoides sigmodontis	19	110	242	37
Acanthocheilonema viteae	19	110	225	37
Aethina tumida	19	110	1729	37

Lepeophtheirus salmonis	22	110	1555	37
Anoplophora glabripennis	22	110	3664	37
Varroa jacobsoni	22	110	3077	37
Varroa destructor	22	110	3077	37
Thelazia callipaeda	25	110	209	37
Bursaphelenchus xylophilus	25	110	638	37
Acyrthosiphon pisum	28	107	68	37
Anisakis simplex	30	219	665	37
Tetranychus urticae	31	122	648	37
Homarus americanus	31	110	9821	37
Bursaphelenchus okinawaensis	37	110	593	37
Globodera pallida	43	113	47	37
Amphibalanus amphitrite	73	110	369	37
Cotesia glomerata	73	116	236	37
Caenorhabditis angaria	79	110	96	37
Onchocerca ochengi	88	110	243	37
Brugia pahangi	91	110	232	37
Ditylenchus destructor	97	307	1167	37
Mesorhabditis belari	97	110	147	37
Melanaphis sacchari	97	107	71	37
Enterobius vermicularis	100	110	195	37
Pristionchus mayeri	103	110	131	37
Cercopithifilaria johnstoni	103	110	238	37
Steinernema carpocapsae	106	110	131	37
Wuchereria bancrofti	106	125	242	37
Parelaphostrongylus tenuis	112	110	228	37
Toxocara canis	115	110	1062	37
Necator americanus	136	110	243	37
Brugia malayi	139	110	243	37
Caenorhabditis auriculariae	145	110	156	37
Auanema sp. JU1783	145	110	80	37
Pristionchus entomophagus	151	110	154	37
Steinernema hermaphroditum	157	110	131	37
Caenorhabditis brenneri	175	110	130	37
Angiostrongylus cantonensis	181	110	213	37
Dictyocaulus viviparus	190	110	832	37
Caenorhabditis elegans	193	110	106	37
Cooperia oncophora	205	110	215	37
Caenorhabditis sp. 36 PRJEB53466	205	110	133	37
Caenorhabditis nigoni	214	110	142	37
Pristionchus pacificus	214	110	251	37
Trichostrongylus colubriformis	214	110	224	37
Caenorhabditis briggsae	217	110	145	37
Cylicocyclus nassatus	229	110	239	37
Haemonchus contortus	304	110	220	37
Caenorhabditis bovis	316	110	235	37
Nippostrongylus brasiliensis	316	110	235	37
Dracunculus medinensis	334	110	122	37
Mesorhabditis spiculigera	376	110	173	37
Pollicipes pollicipes	436	110	367	37
Rhopalosiphum maidis	1345	107	69	37

Таблица 4: Сводная таблица с характеристикой кассетного интрона для таксономической группы Spiralia. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в "кассетном" интроне.

Название организма	Кол-во нуклеотидов до стоп-кодона в интроне	Длина 1-го экзона в кассете	Длина кассетного интрона	Длина 2-го экзона в кассете
Schistosoma haematobium	1	239	652	37
Magallana gigas	1	110	1537	37
Mya arenaria	1	110	1727	37
Crassostrea virginica	1	110	1613	37
Aplysia californica	1	221	4146	37
Gigantopelta aegis	1	110	1869	37
Mercenaria mercenaria	1	110	1690	37
Dreissena polymorpha	1	110	2207	37
Ruditapes philippinarum	1	110	1646	37
Mactra antiquata	1	110	2319	37
Mytilus coruscus	1	110	1234	37
Potamilus streckersoni	1	110	4567	37
Saccostrea echinata	1	110	1556	37
Mytilus edulis	1	110	1360	37
Mytilus trossulus	1	110	1357	37
Pecten maximus	1	110	5000	37
Ostrea edulis	1	110	1643	37
Mizuhopecten yessoensis	1	110	4836	37
Saccostrea cucullata	1	110	1706	37
Ylistrum balloti	1	110	4649	37
Argopecten irradians	1	110	5057	37
Magallana angulata	1	110	1534	37
Mytilus californianus	1	110	1248	37
Pinctada imbricata	1	110	4144	37
Haliotis asinina	1	110	2375	37
Sinanodonta woodiana	1	110	4580	37
Haliotis cracherodii	1	110	2506	37
Haliotis rufescens	1	110	2505	37
Patella caerulea	1	110	1362	37
Patella vulgata	1	110	1384	37
Lymnaea stagnalis	1	221	2705	37
Batillaria attramentaria	1	110	8614	37
Schistosoma turkestanicum	1	239	905	37
Paragonimus westermani	1	239	13971	37
Pomacea canaliculata	1	56	255	37
Bradybaena similaris	1	221	3811	37
Elysia crispata	1	221	8063	37
Elysia chlorotica	1	221	7182	37
Bulinus truncatus	1	221	1873	37
Biomphalaria pfeifferi	1	221	1885	37
Biomphalaria glabrata	1	221	1889	37
Schistosoma guineensis	1	239	652	37
Schistosoma curassoni	1	239	652	37
Schistosoma bovis	1	239	652	37
Schistosoma margrebowiei	1	239	650	37
Schistosoma intercalatum	1	239	652	37
Schistosoma rodhaini	1	239	671	37
Schistosoma japonicum	1	239	847	37
Clonorchis sinensis	1	242	6006	37
Hydatigera taeniaeformis	1	242	375	37

Taenia crassiceps	1	242	278	37
Taenia asiatica	1	242	480	37
Heterobilharzia americana	1	239	2163	37
Trichobilharzia szidati	1	239	1336	37
Trichobilharzia regenti	1	239	996	37
Opisthorchis felineus	1	242	14603	37
Rodentolepis nana	1	242	222	37
Calicophoron daubneyi	1	239	4214	37
Taenia solium	1	242	480	37
Echinococcus granulosus	1	242	521	37
Fasciola hepatica	1	239	2631	37
Fasciola gigantica	1	239	2581	37
Schistosoma mattheei	1	239	649	37
Fasciolopsis buskii	1	239	1303	37
Dicrocoelium dendriticum	1	239	2612	37
Paragonimus heterotremus	1	239	18219	37
Hymenolepis diminuta	1	242	224	37
Solemya velum	4	110	2071	37
Littorina saxatilis	19	218	6746	37

Таблица 5: Сводная таблица с характеристикой кассетного интрона для таксономической группы Cnidaria. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в "кассетном" интроне.

Название организма	**Кол-во нуклеотидов до стоп-кодона в интроне**	**Длина 1-го экзона в кассете**	**Длина кассетного интрона**	**Длина 2-го экзона в кассете**
Actinia tenebrosa	10	116	173	37
Dendronephthya gigantea	10	116	328	37
Nematostella vectensis	25	116	991	37
Montipora foliosa	31	116	907	37
Pocillopora verrucosa	34	116	390	37
Acropora digitifera	40	116	670	37
Acropora millepora	40	116	682	37
Acropora muricata	40	116	679	37
Pocillopora damicornis	46	116	392	37
Pocillopora meandrina	46	116	392	37
Porites lutea	61	116	711	37
Porites evermanni	61	116	711	37
Exaiptasia diaphana	76	86	227	37
Xenia sp. Carnegie-2017	103	116	116	37

Таблица 6: Сводная таблица с характеристикой кассетного интрона для таксономической группы Sauropsida. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в "кассетном" интроне.

Название организма	**Кол-во нуклеотидов до стоп-кодона в интроне**	**Длина 1-го экзона в кассете**	**Длина кассетного интрона**	**Длина 2-го экзона в кассете**
Molothrus aeneus	1	110	745	37
Taeniopygia guttata	1	110	443	37
Lonchura striata	1	110	629	37
Gallus gallus	7	110	1616	37

Cygnus atratus	25	110	1257	37
Haliaeetus leucocephalus	25	110	1375	37
Phalacrocorax carbo	25	110	1345	37
Grus americana	25	110	1659	37
Haliaeetus albicilla	25	110	1378	37
Oxyura jamaicensis	25	110	1246	37
Anser cygnoides	25	110	1279	37
Ciconia boyciana	25	107	1459	37
Anas acuta	25	110	1346	37
Astur gentilis	25	110	1393	37
Aquila chrysaetos chrysaetos	25	110	1375	37
Aythya fuligula	25	110	1227	37
Struthio camelus	64	110	1405	37
Chelonia mydas	79	110	1674	37
Dermochelys coriacea	79	110	1661	37
Caretta caretta	79	110	1656	37
Ammospiza caudacuta	82	110	3942	37
Aphelocoma coerulescens	85	110	3626	37
Gopherus flavomarginatus	142	110	1655	37
Chelonoidis abingdonii	142	110	1645	37
Malaclemys terrapin pileata	142	110	1652	37
Mauremys mutica	142	110	1662	37
Mauremys reevesii	142	110	1661	37
Trachemys scripta elegans	142	110	1661	37
Chrysemys picta bellii	142	110	1662	37
Emys orbicularis	142	110	1650	37
Alligator sinensis	148	110	1497	37
Alligator mississippiensis	148	110	1618	37
Caloenas nicobarica	184	110	1245	37
Rissa tridactyla	205	110	1388	37
Terrapene triunguis	211	110	1662	37
Emydura macquarii macquarii	223	110	1647	37
Catharus ustulatus	241	110	3252	37
Gopherus evgoodei	301	110	1639	37
Strigops habroptila	457	110	1317	37
Neopsephotus bourkii	502	110	1245	37
Melopsittacus undulatus	517	110	1257	37
Apteryx rowi	541	110	1359	37
Apteryx mantelli	541	110	1359	37
Dromaius novaehollandiae	553	110	1365	37
Chroicocephalus ridibundus	562	110	1373	37
Pezoporus wallicus	568	110	1328	37
Pezoporus flaviventris	568	110	1328	37
Rhea pennata	568	110	1348	37
Pezoporus occidentalis	568	110	1319	37
Pelodiscus sinensis	640	110	1643	37
Phaenicophaeus curvirostris	892	110	2155	37
Camarhynchus parvulus	1360	110	2456	37
Vidua chalybeata	1519	110	678	37

Таблица 7: Сводная таблица с характеристикой кассетного интрона для таксономической группы Amphibia. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в "кассетном" интроне.

Название организма	**Кол-во нуклеотидов до стоп-кодона в интроне**	**Длина 1-го экзона в кассете**	**Длина кассетного интрона**	**Длина 2-го экзона в кассете**
Ambystoma mexicanum	1	110	10340	37
Pelobates fuscus	1	110	2424	37
Bufo bufo	7	110	3002	37
Bufo gargarizans	7	110	2879	37
Hyperolius riggenbachi	10	110	3902	37
Rana temporaria	10	110	3036	37
Pseudophryne corroboree	19	110	3561	37
Spea bombifrons	25	110	2840	37
Engystomops pustulosus	25	110	2004	37
Nanorana parkeri	25	110	3038	37
Hyla sarda	25	110	3029	37
Pyxicephalus adspersus	25	110	2917	37
Ranitomeya imitator	37	110	2650	37
Xenopus tropicalis	46	110	2596	37
Xenopus laevis	52	110	3791	37
Geotrypetes seraphini	55	110	3065	37
Rhinatrema bivittatum	103	110	4053	37
Pleurodeles waltl	151	110	3245	37
Microcaecilia unicolor	187	110	2784	37

Таблица 8: Сводная таблица с характеристикой кассетного интрона для таксономической группы Lepidosauria. Сортировка по возрастанию количества нуклеотидов до стоп-кодона в "кассетном" интроне.

Название организма	**Кол-во нуклеотидов до стоп-кодона в интроне**	**Длина 1-го экзона в кассете**	**Длина кассетного интрона**	**Длина 2-го экзона в кассете**
Python bivittatus	1	110	2374	37
Notechis scutatus	1	110	2507	37
Pseudonaja textilis	1	110	2519	37
Anolis sagrei	1	110	4667	37
Pituophis catenifer annectens	1	110	2420	37
Lacerta agilis	1	110	2499	37
Candoia aspera	1	110	2293	37
Sphaerodactylus townsendi	1	110	2825	37
Thamnophis elegans	1	110	2426	37
Ahaetulla prasina	1	110	2432	37
Gekko japonicus	1	110	2924	37
Crotalus tigris	1	110	3091	37
Pogona vitticeps	1	110	2746	37
Podarcis raffonei	1	110	2495	37
Protobothrops mucrosquamatus	1	110	3264	37
Varanus komodoensis	1	110	2658	37
Pantherophis guttatus	1	110	2411	37
Elgaria multicarinata webbii	1	110	2800	37
Rhineura floridana	1	110	2581	37
Podarcis muralis	1	110	2506	37
Heteronotia binoei	1	110	3002	37

Anolis carolinensis	1	110	4026	37
Erythrolamprus reginae	1	110	2638	37
Sceloporus undulatus	1	110	2380	37
Eublepharis macularius	1	110	2577	37
Euleptes europaea	1	110	2901	37
Hemicordylus capensis	1	110	2830	37
Zootoca vivipara	1	110	2516	37

Список литературы

1. Database resources of the National Center for Biotechnology Information / E. W. Sayers, E. E. Bolton, J. R. Brister, [et al.] // Nucleic Acids Research. — 2022. — Vol. 50, no. D1. — P. D20–D26. — DOI: 10.1093/nar/gkab1112. — URL: https://doi.org/10.1093/nar/gkab1112.
2. *McKinney W.* Data Structures for Statistical Computing in Python. — 2010.
3. *Python Software Foundation.* Python, Version 3.12. — 2023. — https://www.python.org/downloads/release/python-3120/.
4. Jupyter Notebooks – a publishing format for reproducible computational workflows / T. Kluyver [et al.]. — 2016. — DOI: 10.3233/978-1-61499-649-1-87. — URL: https://doi.org/10.3233/978-1-61499-649-1-87.
5. Biopython: Freely available Python tools for computational molecular biology and bioinformatics / P. J. A. Cock [et al.] // Bioinformatics. — 2009. — Vol. 25, no. 11. — P. 1422–1423. — DOI: 10.1093/bioinformatics/btp163. — URL: https://doi.org/10.1093/bioinformatics/btp163.
6. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets / N. A. O'Leary [et al.] // Scientific Data. — 2024. — Vol. 11, no. 1. — P. 732. — DOI: 10.1038/s41597-024-03571-y. — URL: https://doi.org/10.1038/s41597-024-03571-y.
7. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs / S. F. Altschul [et al.] // Nucleic Acids Research. — 1997. — Vol. 25, no. 17. — P. 3389–3402. — DOI: 10.1093/nar/25.17.3389. — URL: https://doi.org/10.1093/nar/25.17.3389.
8. *Katoh K., Standley D. M.* MAFFT multiple sequence alignment software version 7: improvements in performance and usability // Molecular Biology and Evolution. — 2013. — Vol. 30, no. 4. — P. 772–780. — DOI: 10.1093/molbev/mst010. — URL: https://doi.org/10.1093/molbev/mst010.
9. Unipro UGENE: a unified bioinformatics toolkit / K. Okonechnikov [et al.] // Bioinformatics. — 2012. — Vol. 28, no. 8. — P. 1166–1167. — DOI: 10.1093/bioinformatics/bts091. — URL: https://doi.org/10.1093/bioinformatics/bts091.
10. The MEME Suite / T. L. Bailey [et al.] // Nucleic Acids Research. — 2015. — Vol. 43, W1. — W39–W49. — DOI: 10.1093/nar/gkv416. — URL: https://doi.org/10.1093/nar/gkv416.
11. Quantifying similarity between motifs / S. Gupta [et al.] // Genome Biology. — 2007. — Vol. 8, no. 2. — R24. — DOI: 10.1186/gb-2007-8-2-r24. — URL: https://doi.org/10.1186/gb-2007-8-2-r24.

12. ViennaRNA Package 2.0 / R. Lorenz [et al.] // Algorithms for Molecular Biology. — 2011. — Vol. 6, no. 1. — P. 26. — DOI: 10.1186/1748-7188-6-26. — URL: https://doi.org/10.1186/1748-7188-6-26.

13. *Yeo G.*, *Burge C. B.* Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals // Bioinformatics. — 2004. — Vol. 20, no. 3. — P. 327–335. — DOI: 10.1093/bioinformatics/btg005. — URL: https://doi.org/10.1093/bioinformatics/btg005.

14. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era / B. Q. Minh [et al.] // Molecular Biology and Evolution. — 2020. — Vol. 37, no. 5. — P. 1530–1534. — DOI: 10.1093/molbev/msaa015. — URL: https://doi.org/10.1093/molbev/msaa015.

15. *Rambaut A.* FigTree v1.4.4. — 2018. — Institute of Evolutionary Biology, University of Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/.

16. *QuantStack*, *contributors mamba*. Mamba: The Fast Cross-Platform Package Manager. — 2024. — https://github.com/mamba-org/mamba.

17. *GitHub, Inc.* GitHub. — 2008. — URL: https://github.com.

18. *Lamport L.* LaTeX: A Document Preparation System. — 2nd ed. — Reading, Massachusetts : Addison-Wesley, 1994.

19. *TablesGenerator.com*. Tables Generator – LaTeX Tables Editor. — 2025. — URL: https://www.tablesgenerator.com.

20. *diagrams.net*. draw.io - Online Diagram Software. — 2025. — URL: https://www.diagrams.net/.

21. *Canonical Ltd.* Ubuntu 22.04 LTS (Jammy Jellyfish). — 2022. — https://releases.ubuntu.com/22.04/.

22. Replication of Human Herpesvirus 6A and 6B Is Associated with Distinct Nuclear Domains / F. Tajima [et al.] // Journal of Virology. — 2001. — Vol. 75, no. 12. — P. 5567–5575. — DOI: 10.1128/JVI.75.12.5567-5575.2001. — URL: https://doi.org/10.1128/JVI.75.12.5567-5575.2001.

Благодарности

Я хотел бы поблагодарить моего научного руководителя, к.б.н. Голубкову Елену Валерьевну, и моего куратора, Бондарука Дмитрия Денисовича, за постоянную поддержку и помощь в обсуждении результатов работы.

Отдельно я хотел бы поблагодарить Абрамсон Наталью Иосифовну за повторное рецензирование работы моего авторства.

Также хочу выразить благодарность преподавателям программы “Биоинформатика“ и кафедры генетики и биотехнологии СПбГУ, и коллективу преподавателей и ассистентов Института Биоинформатики за полученные знания в процессе обучения, с помощью которых стало возможным осуществление данной работы.