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 был, с помощью PSI-BLAST были увеличены выборки, где это оказалось возможным, результат продемонстрирован на таблице 1.

Таблица 1: Результат увеличения выборки для таксономических групп Protostomia и Cnidaria с помощью PSI-BLAST.

Филогенетическая группа	Таксон высокого ранга	Видов до PSI-BLAST	Сколько видов добавил PSI-BLAST	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.

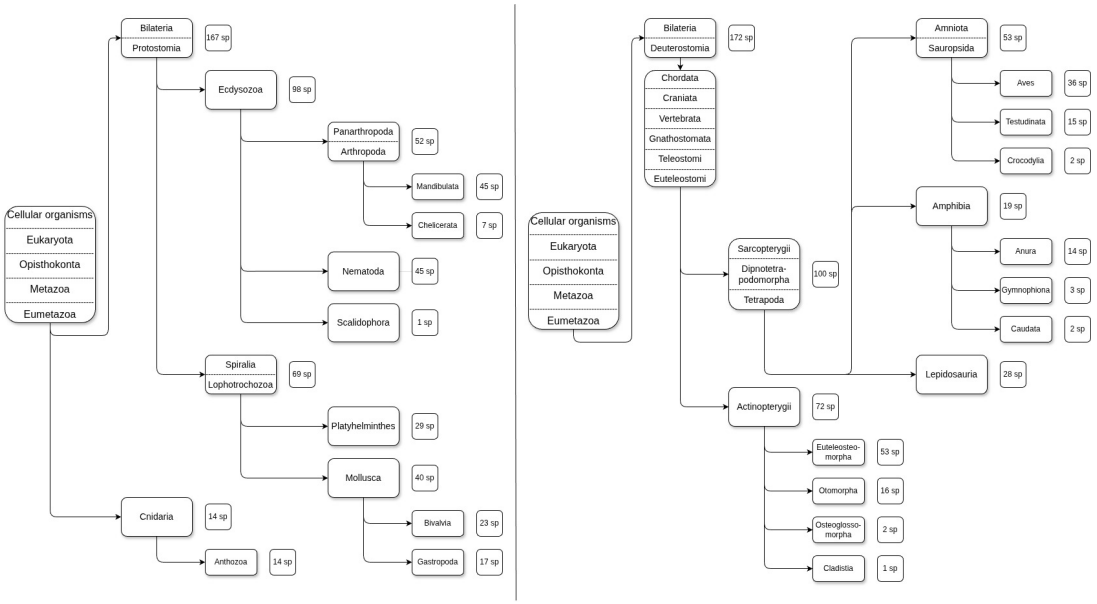

Рис. 1: Количество видов, взятых в анализ для Protostomia+Cnidaria и Deuterostomia.

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. Для анализа были взяты все найденные представители данной филогенетической группы - 72 вида.

На таблице 2 показана характеристика “консервативной кассеты“ для исследуемой группы.

Таблица 2: Сводная таблица с характеристикой кассетного интрона для таксономической группы Actinopterygii. 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**
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.

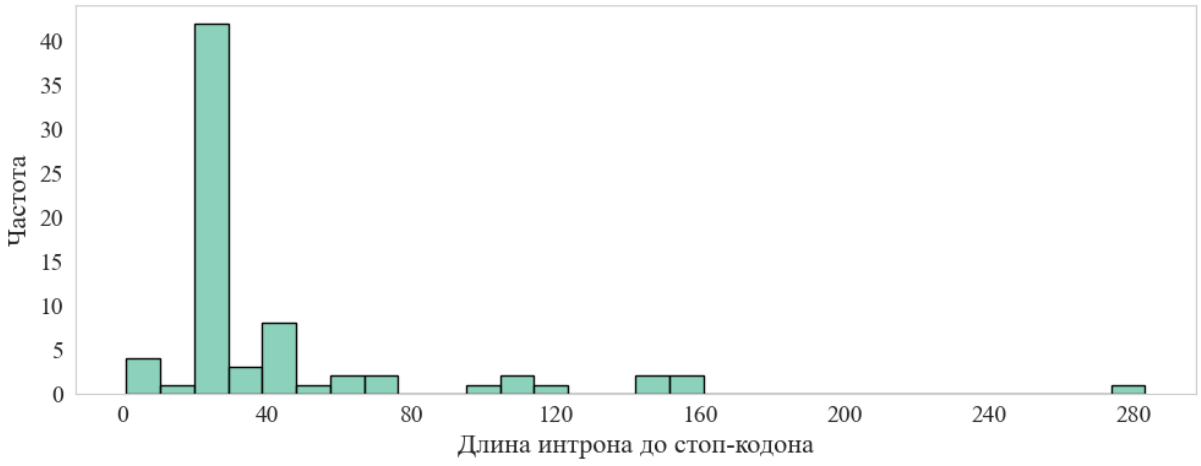

Рис. 2: Распределение длин части кассетного интрона до стоп-кодона у таксономической группы Actinopterygii

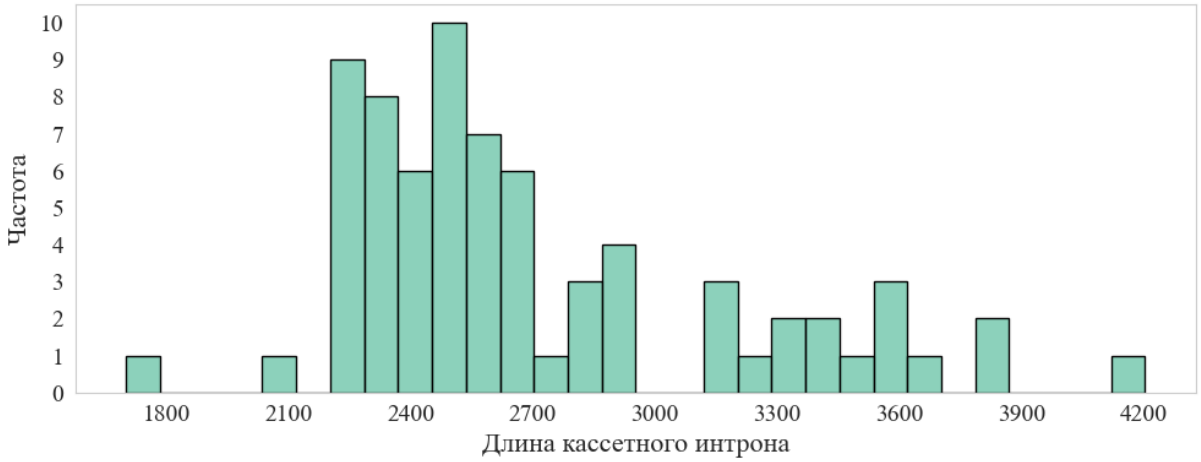

Рис. 3: Распределение длин кассетного интрона у таксономической группы Actinopterygii

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.

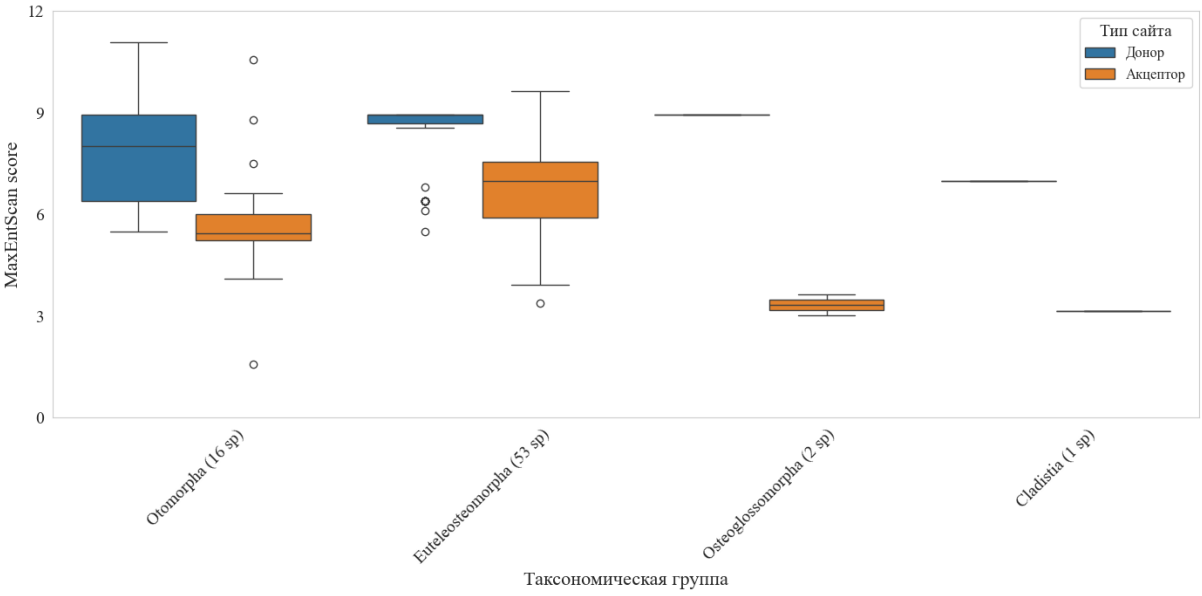

Рис. 4: Результаты проведения MaxEntScan для Actinopterygii.

Рисунок 5 демонстрирует результаты, полученные с помощью MEME Suite.

Найденные мотивы присутствуют не у всех 72 видов, 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 поиск найденных консервативных мотивов из “кассетного“ интроне в базе данных не дало статистически значимых ре-

зультатов.

Репрезентация вторичной структуры интрон-содержащего транскрипта с выделенным кассетным интроном и найденным мотивом показано 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.

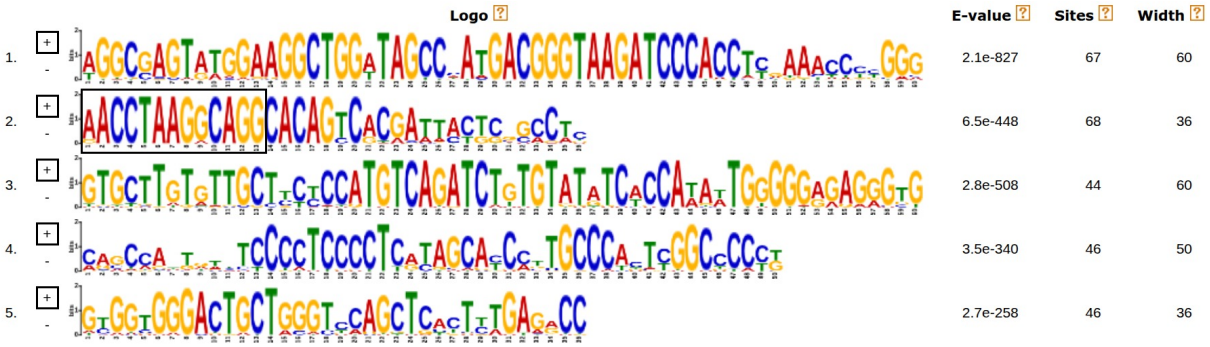

Рис. 5: Результат поиска мотивов внутри кассетного интрона с помощью MEME Suite для таксономической группы Actinopterygii. Черным прямоугольником выделен участок, похожий на консенсусную последовательность CTE 6 из статьи 2001 года.

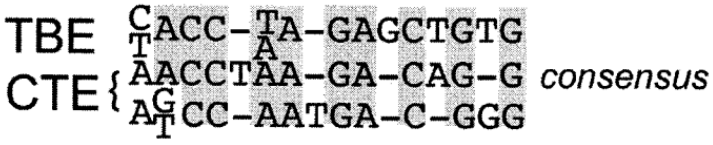

Рис. 6: Консенсусный CTE из статьи 2001 года

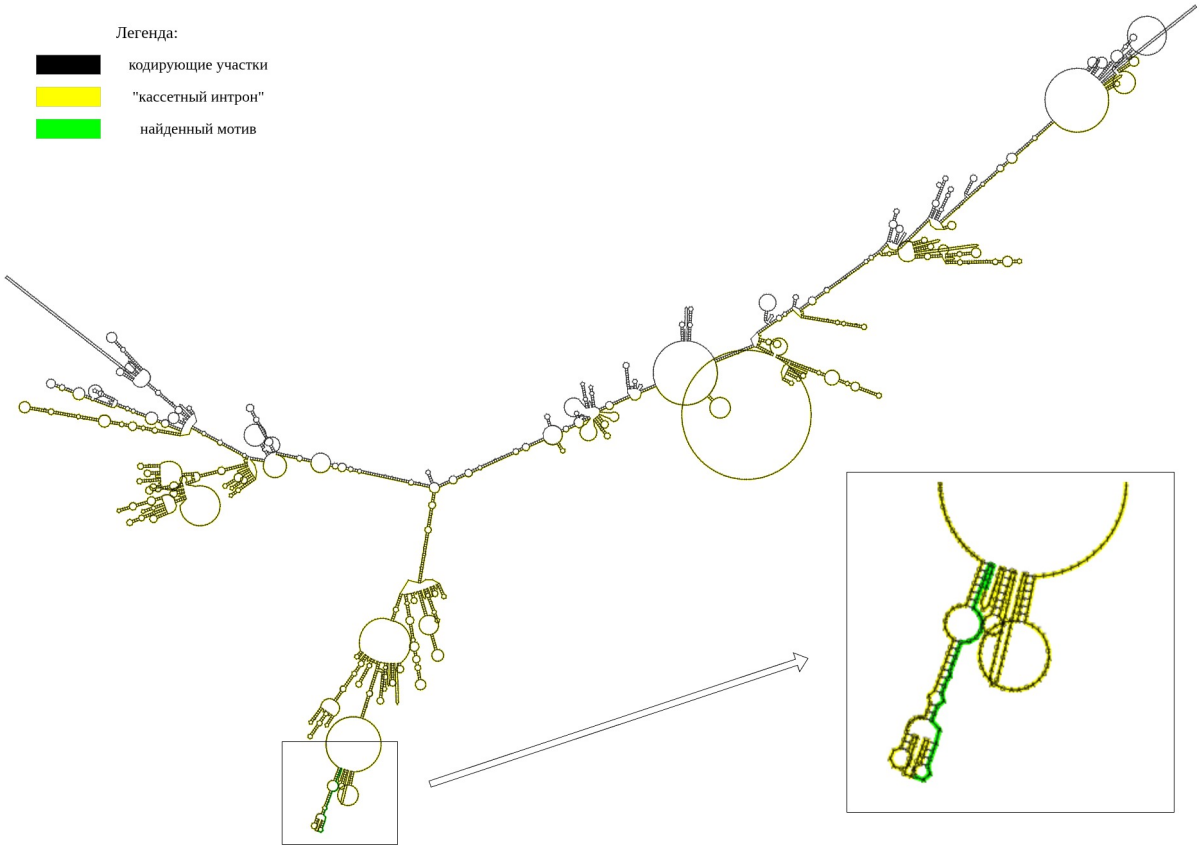

Рис. 7: Вторичная структура РНК-транскрипта для *Chanos chanos* из Otomorpha, содержащая кассетный интрон.

Рис. 8: Результаты множественного выравнивания для Actinopterygii.

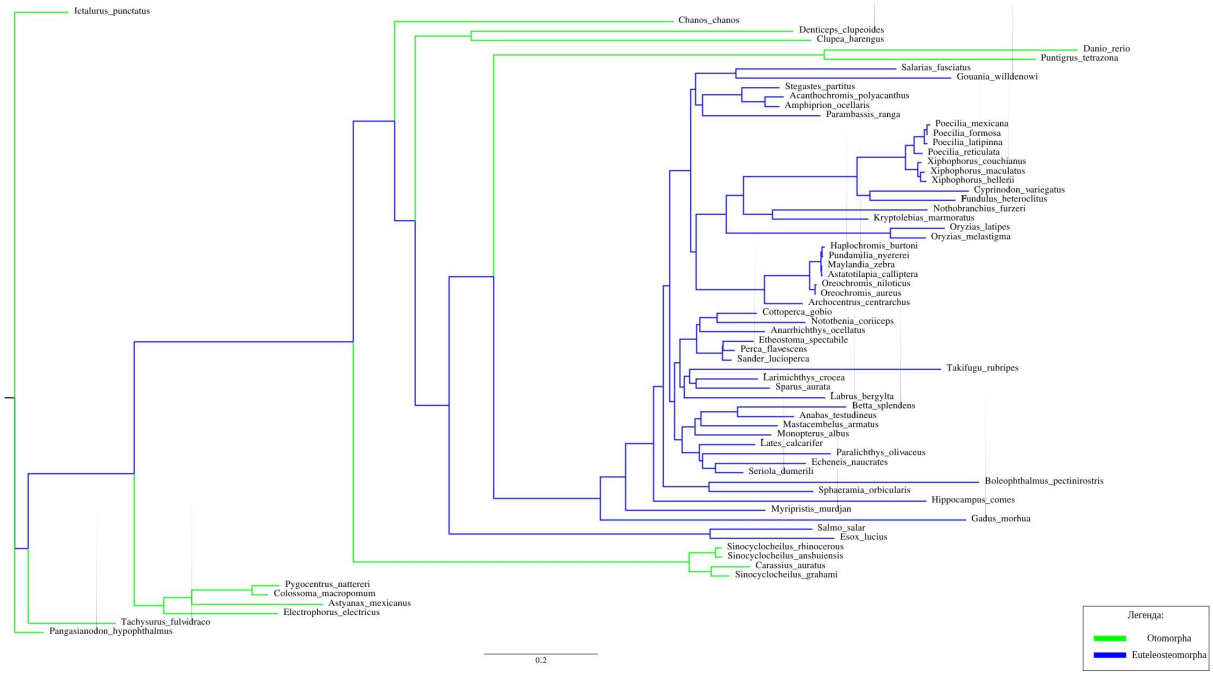

Рис. 9: Филогенетическое дерево для таксономической группы Actinopterygii

Обсуждение

Текст...

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.