Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи

Зверков Олег Анатольевич

Функции и эволюция РНК-полимераз в митохондриях и пластидах

03.01.09 – Математическая биология, биоинформатика

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д.ф.-м.н. профессор В. А. Любецкий

Москва – 2014

СОДЕРЖАНИЕ

ВВЕДЕНИЕ **4**

1. Общая характеристика работы 4

2. Основные результаты и выводы 8

Публикации автора по теме диссертации *9*

3. Используемые сведения о митохондриях и пластидах 12

3.1. Митохондрии у хордовых: лягушки, человека и крысы *12*

3.2. Структура и взаимное расположение промоторов *14*

3.3. Влияние белковых факторов на уровни транскрипции *15*

3.4. mTERF-зависимая терминация транскрипции *16*

3.5. Белок-независимый терминатор транскрипции *16*

3.6. MELAS болезни *17*

3.7. Время полураспада РНК *17*

3.8. Пластиды растений и водорослей *19*

3.9. Конкуренция РНК-полимераз *19*

3.10. Нокауты генов σ-субъединиц РНК-полимераз *20*

3.11. Тепловой шок изолированных хлоропластов *21*

3.12. Анализ других экспериментальных данных *21*

3.13. Заключение *22*

ГЛАВА 1. ВЗАИМОДЕЙСТВИЕ РНК-ПОЛИМЕРАЗ В МИТОХОНДРИЯХ И ПЛАСТИДАХ **23**

1. Примеры локусов в митохондриях и пластидах 23

2. Модель взаимодействия РНК-полимераз 24

3. Параметры модели 28

3.1. Параметры РНК-полимеразы бактериального типа (PEP) *28*

3.2. Параметры PEP-промоторов и число абортивных попыток *28*

3.3. Параметры РНК-полимеразы фагового типа (NEP) *30*

4. Экспериментальные данные об уровнях транскрипции генов и временах полураспада 31

4.1. Данные о митохондриях *31*

4.2. Данные о пластидах *37*

5. Оценка согласия с опытом 39

6. Методика моделирования ... 40
6.1. Обоснование модели ... *40*
6.2. Случай митохондрий ... *41*
7. Компьютерная реализация модели ... 44
8. Результаты о митохондриях ... 46
9. Результаты о пластидах ... 48
10. Обсуждение результатов о митохондриях ... 50
11. Обсуждение результатов о пластидах ... 55
12. Заключение ... 57

ГЛАВА 2. СЕМЕЙСТВА БЕЛКОВ, КОДИРУЕМЫХ В ПЛАСТИДАХ ... 58
1. Введение и постановка задачи ... 58
1.1. Пластиды родофитной ветви ... *60*
1.2. Пластиды хлорофитной ветви ... *63*
1.3. Пластиды цветковых растений ... *64*
2. Результаты ... 64
2.1. Алгоритм кластеризации ... *64*
Пример работы алгоритма ... 71
2.2. Кластеризация белков родофитной ветви пластид ... *74*
2.2.1. Характеристика кластеров пластомных белков родофитной ветви ... 75
2.2.2. Поиск РНК-полимераз в ядерных геномах споровиков ... 77
2.2.3. Обсуждение результатов кластеризации для родофитной ветви ... 80
2.3. Кластеризация белков хлорофитной ветви пластид ... *81*
2.3.1. Характеристика кластеров пластомных белков хлорофитной ветви ... 81
2.3.2. Обсуждение результатов кластеризации для хлорофитной ветви ... 83
2.3.3. Дополнительное исследование кластеров CysA и CysT ... 85
2.4. Кластеризация пластомных белков однодольных растений ... *87*
2.5. Кластеризация пластомных белков цветковых растений ... *88*

ГЛАВА 3. СОПРЯЖЕНИЕ ТРАНСЛЯЦИИ И ПРОЦЕССИНГА мРНК В ПЛАСТИДАХ ... 91
1. Введение и постановка задачи ... 91
2. Материалы и методы ... 97
3. Результаты ... 98
4. Обсуждение ... 100

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ... 103

ВВЕДЕНИЕ

1. Общая характеристика работы

Актуальность темы

В биоинформатике велико значение быстрых и эффективных алгоритмов, поскольку зачастую возникают входные данные весьма большого объёма. Известные и новые методы вычислений требуют адаптации к работе на многопроцессорных вычислительных комплексах (суперкомпьютерах), которые стали в последнее время значительно доступнее.

К настоящему времени известны сотни полностью секвенированных геномов пластид, тысячи геномов митохондрий, скорость пополнения баз данных геномной информации растёт экспоненциальными темпами. Возникает такой объём информации, что доля геномов, доступных биохимическому исследованию, становится всё меньше. Поэтому возникает потребность в эффективных и быстрых алгоритмах компьютерного анализа данных, а также в создании специализированных баз данных. Существенно, чтобы алгоритмы опирались на «точные модели», т.е. было доказано, что они приводят к глобальным экстремумам соответствующих функционалов, имели низкую вычислительную сложность (полином 2–3 степени) и допускали эффективное распараллеливание.

Моделирование клеточных процессов требует нетривиальных алгоритмов и является важным инструментом биоинформатического исследования. Оно позволяет предсказать значения параметров биохимических процессов (например, инициации, элонгации и терминации транскрипции), которые трудно измерить непосредственно, а также – решить нетривиальную обратную задачу: выбрать значения параметров, которые соответствуют экспериментальным зависимостям.

Экспериментальные исследования, в том числе проведённые в Институте физиологии растений им. К. А. Тимирязева РАН (Зубо и др.), позволили предположить важную роль взаимодействия РНК-полимераз в процессе транскрипции пластомов растений и в ответе пластид на тепловой шок. Для проверки этого предположения и предсказания параметров, не определяемых в экспериментах, была поставлена задача моделирования процесса транскрипции в пластидах с одновременным участием многих РНК-полимераз, факторов и вторичных структур, взаимодействующих друг с другом. Затем задача была расширена на моделирование транскрипции в митохондриях.

Использование кластера MVS-100K в Межведомственном суперкомпьютерном центре РАН позволило впервые провести моделирование транскрипции для всей кольцевой ДНК митохондрий человека, крысы и лягушки, а также для существенных локусов пластид.

Построение близких по последовательности и минимальных по содержанию паралогов белковых семейств (кластеризация белков) позволяет уточнять аннотации белков, судить о работоспособности белковых комплексов, например РНК-полимераз бактериального типа. (В случае отсутствия последних транскрипция выполняется РНК-полимеразами фагового типа, что придаёт этому процессу другие черты.) Известно несколько баз данных семейств ортологичных белков [1]. Однако большинство из них содержат небольшое число видов с пластидами или вовсе не содержат их. Например, (по состоянию на 1 июля 2013) OrthoDB [2] не содержит растений и простейших, OrthoMCL [3] включает только 11 водорослей и 14 споровиков; GeneDB [4] – только 7 споровиков; в RoundUp [5] и InParanoid [6] таких видов ещё меньше; OMA [7] и EggNOG [8] почти не содержат видов с пластидами; в COG и KOG [9] представлено два растения и ни одного споровика. Поэтому была поставлена задача: предложить эффективный алгоритм кластеризации белков и получить базы данных пластомных белков.

Изучение пластид споровиков (апикопластов) значимо, поскольку споровики вызывают опасные заболевания человека и животных, в том числе токсоплазмоз и малярию. Исследование регуляции экспрессии генов, кодируемых в апикопластах, важно для понимания роли апикопластов в передаче инфекции, а также в механизмах действия лекарственных средств на апикопласты, которые являются главной мишенью антибиотиков, не оказывающих прямого воздействия на экспрессию ядерных и митохондриальных генов хозяина. В частности, *Theileria* и *Babesia* переносятся иксодовыми клещами и вызывают заболевания крупного рогатого скота: *B. bigemina* и *B. bovis* – бабезиоз крупного рогатого скота, *Th. annulata* – тейлериоз крупного рогатого скота, *Th. parva* – лихорадку Восточного Берега; *Eimeria tenella* вызывает эймериоз кур; *Toxoplasma gondii* – токсоплазмоз, в том числе у человека; различные виды рода *Plasmodium* вызывают малярию у людей (*P. falciparum*, *P. vivax*) и других животных. Некоторые споровики, например *Cryptosporidium parvum*, не имеют пластид.

Исследование митохондрий человека, крысы и лягушки значимо для понимания молекулярных механизмов MELAS болезней человека (митохондриальная энцефаломиопатия, лактатацидоз, инсультоподобные эпизоды), болезней, связанных с недостаточностью гормона щитовидной железы, и т.д.

Цели работы

1. Разработать модель взаимодействия и конкуренции РНК-полимераз в митохондриях и пластидах, которая должна предсказывать уровни транскрипции всех генов. На её основе объяснить изменения уровней транскрипции генов: в митохондриях человека с MELAS-мутацией; в митохондриях крысы с эпигенетическими нарушениями, вызванными недостатком тиреоидного гормона; в пластидах растений после нокаутов минорных σ-субъединиц или теплового шока.

2. Разработать алгоритм построения сходных по последовательности и минимальных по содержанию паралогов семейств белков (кластеризации данного множества белков). Применить алгоритм к множествам белков, кодируемых в пластидах родофитной и хлорофитной ветвей и цветковых растений. На основе полученных семейств: рассмотреть вопрос о присутствии полноценной РНК-полимеразы бактериального типа у споровиков; указать белки, характерные для узких таксономических групп («филогенетические подписи»).

3. Предсказать белковые сайты и вторичные структуры мРНК, ответственные за задержку инициации трансляции до завершения процессинга мРНК в пластидах.

Методы исследования

В работе использованы методы теорий алгоритмов и массового обслуживания, методы моделирования и организации вычислительных экспериментов с использованием известных и оригинальных программ, в том числе для параллельных вычислений на суперкомпьютерах, методы математической биологии и биоинформатики.

Научная новизна

Моделирование взаимодействия РНК-полимераз, по крайней мере на длинных локусах ДНК, ранее не выполнялось. Моделирование основано на новом математическом и алгоритмическом подходе к изучению большой системы одновременно взаимодействующих объектов. Кластеризация получена на основе оригинального алгоритма в теории графов. Все полученные алгоритмы имеют низкую оценку вычислительной сложности, а биоинформатические результаты являются новыми.

Практическая значимость работы

Работа носит теоретический характер. В то же время, исследование может иметь прикладное значение.

Предложенные алгоритмы и их программные реализации могут применяться для исследования широкого класса задач. А именно, в медицинских исследованиях могут

быть полезны разработанные методы количественной оценки влияния мутаций и эпигенетических нарушений на уровни транскрипции генов в митохондриях, предложенные нами объяснения механизма MELAS-синдрома у человека и нарушения метилирования мтДНК у крысы с недостатком гормона щитовидной железы.

Для создания новых видов растений, в том числе с ксенопластидами, могут быть полезны предложенные механизмы отклика на тепловой шок изолированных пластид и на нокауты транскрипционных факторов в пластидах.

Апробация работы

Компьютерные программы тестировались на биологических данных с экспериментально известными ответами, а также в процессе решения биологических задач. Результаты работы опубликованы и докладывались на следующих конференциях:

- Международная конференция “Moscow Conference on Computational Molecular Biology”: MCCMB'07 (Москва, 27–31 июля 2007), MCCMB'13 (Москва, 25–28 июля, 2013);
- 32-я, 33-я, 35-я, 37-я конференция «Информационные технологии и системы»: ИТиС'09 (Бекасово, 15–18 декабря 2009), ИТиС'10, (Геленджик, 20–24 сентября 2010), ИТиС'12 (Петрозаводск, 19–25 августа 2012), ИТиС'13 (Калининград, 1–6 сентября 2013);
- 7-я международная конференция “Bioinformatics of Genome Regulation and Structure\Systems Biology” BGRS\SB'10 (Новосибирск, 20–27 июня 2010);
- 51-я, 53-я, 54-я научная конференция МФТИ (Москва, 28–30 ноября 2008, 24–29 ноября 2010, 25–26 ноября 2011);
- 3-я Московская международная конференция “Molecular Phylogenetics” (Москва, 31 июля – 4 августа 2012).
- 8-я Международная конференция «Современные информационные технологии и ИТ-образование» (Москва, МГУ им. М. В. Ломоносова, 8–10 ноября 2013).

Работа также докладывалась на научных семинарах механико-математического факультета Московского государственного университета им. М. В. Ломоносова и на семинаре по Математической биологии и биоинформатике Института проблем передачи информации им. А. А. Харкевича РАН.

Публикации

По теме диссертации опубликовано 9 статей и 13 тезисов докладов на конференциях (см. список в конце пункта 2). Все результаты, включённые в диссертацию, получены лично автором.

Структура и объём работы

Работа состоит из введения, трёх глав и списка литературы. Список литературы содержит 127 наименований. Объём работы составляет 112 страниц, включая 21 таблицу и 29 рисунков.

2. Основные результаты и выводы

Разработана математическая и компьютерная модель взаимодействия РНК-полимераз между собой, с вторичными структурами и белковыми факторами в процессах инициации и элонгации транскрипции. Модель применена к локусам пластид и митохондрий, и находится в согласии практически со всеми опытными данными, относящимися к пластидам растений и митохондриям, включая данные об изменениях уровней транскрипции генов после нокаутов σ-субъединиц РНК-полимераз и после теплового шока изолированных пластид, данные об относительных количествах РНК и временах их полураспада в митохондриях лягушек, человека здорового и с MELAS-мутацией, крысы здоровой и с пониженным уровнем тиреоидного гормона.

На основе модели предсказаны характеристики транскрипции в митохондриях хордовых животных: доли РНК-полимераз, завершающих транскрипцию на mTERF-зависимом терминаторе в одном и другом направлениях (поляризация); интенсивность связывания регуляторного белка mTERF с сайтом терминации на ДНК; интенсивности инициации транскрипции на промоторах в пластидах растений и митохондриях лягушки, человека, включая случай MELAS-мутации, крысы, включая гипотиреоида. На основе модели предсказаны значения уровней транскрипции всех генов, в то время как в опытах известны лишь их относительные значения и только для некоторых генов.

На основе модели предположен механизм влияния на фенотип MELAS-мутации: снижение концентраций как фенилаланиновой и валиновой тРНК, так и рРНК, а главное – резкое изменение времени полураспада определённых мРНК.

На основе модели показана корреляция между изменениями метилирования сайта связывания mTERF и промоторов с интенсивностями связывания с ними mTERF и РНК-полимераз.

Разработан алгоритм кластеризации множества белковых последовательностей. На его основе получены семейства сходных по последовательности и минимальных по содержанию паралогов белков, кодируемых в пластомах багрянок и видов с пластидами, родственными пластидам багрянок (родофитная ветвь); белков, кодируемых в пластомах рано отделившихся ветвей зелёных водорослей и видов с родственными им пластидами: Viridiplantae, эвгленовые, *Bigelowiella natans* (хлорофитная ветвь); белков, ко-

дируемых в пластомах цветковых и отдельно однодольных растений. На этой основе найдены белки, специфичные для пластомов небольших таксономических групп водорослей и простейших.

Полученная кластеризация позволила заключить, что у споровиков *Toxoplasma gondii* и *Plasmodium falciparum* присутствует полноценная РНК-полимераза бактериального типа. У *Neospora caninum* и *Plasmodium* spp. найдены α- и σ-субъединицы, кодируемые в ядре. Напротив, у споровиков таксономической группы Piroplasmida α- и σ-субъединицы РНК-полимеразы бактериального типа не найдены, а её субъединицы, обычно кодируемые в пластидах, значительно изменены или фрагментированы. Это позволяет предположить глубокое различие видов Piroplasmida с другими содержащими пластиды споровиками в части транскрипции в пластидах.

На основе оригинальной компьютерной программы (поиска мотива путём определения клики в многодольном графе с учётом GC-состава) предположен механизм задержки инициации трансляции до завершения редактирования транскриптов генов *accD* и *atpH* в пластидах растений видов *Adiantum capillus-veneris* и *Anthoceros formosae*. Механизм вовлекает длинные шпильки в 5'-лидерной области около сайта связывания рибосомы. Найдены консервативные сайты перед шестью генами *atpF*, *clpP*, *petB*, *psaA*, *psbA*, *psbB* у трёх видов *Chara vulgaris*, *Zygnema circumcarinatum*, *Physcomitrella patens*, которые в части случаев также участвуют в задержке инициации трансляции до завершения сплайсинга или редактирования.

Публикации автора по теме диссертации

Статьи:

1. Lyubetsky V.A., Zverkov O.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase interaction in mitochondria of chordates // *Biology Direct*. 2012. 7:26.
2. Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level // *Biology Direct*. 2011. 6:3.
3. Любецкий В.А., Селиверстов А.В., Зверков О.А. Построение разделяющих паралоги семейств гомологичных белков, кодируемых в пластидах цветковых растений // *Математическая биология и биоинформатика*. 2013. Т. 8, № 1. С. 225–233.
4. Зверков О.А., Селиверстов А.В., Любецкий В.А. Белковые семейства, специфичные для пластомов небольших таксономических групп водорослей и простейших // *Молекулярная биология*. 2012. Т. 46, № 5. С. 799–809.

5. Lyubetsky V.A., Seliverstov A.V., Zverkov O.A. Transcription regulation of plastid genes involved in sulfate transport in Viridiplantae // *BioMed Research International*. 2013. Vol. 2013. Article ID 413450. 6 pages.
6. Зверков О.А., Русин Л.Ю., Селиверстов А.В., Любецкий В.А. Изучение вставок прямых повторов в микроэволюции митохондрий и пластид растений на основе кластеризации белков // *Вестник Московского университета*. Серия 16: Биология. 2013. № 1. С. 8–13.
7. Зверков О.А., Селиверстов А.В., Любецкий В.А. Усредненная энтропия как характеристика консервативности участков генома // *Вестник Тамбовского университета. Серия: Естественные и технические науки*. 2013. Т. 18, Вып. 5. С. 2529–2531.
8. Lyubetsky V.A., Korolev S.A., Seliverstov A.V., Zverkov O.A., Rubanov L.I. Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism // *Computational Biology and Chemistry*. 2014. Vol. 49. P. 7–13.
9. Seliverstov A.V., Zverkov O.A., Lyubetsky V.A. Translation of some chloroplast genes is checked to allow for splicing and editing // *Biophysics*. 2006. Vol. 51, S. 1. P. 18–22.

Тезисы докладов:

1. Lyubetsky V.A., Seliverstov A.V., Zverkov O.A. RNA Structures upstream *leuA* Genes in α-proteobacteria // *Proceedings of the International Moscow Conference on Computational Molecular Biology: MCCMB'07*. July 27–31 2007. P. 191–192.
2. Зверков О.А. Программный комплекс для согласования набора эволюционных деревьев и выявления эволюционных событий // *Труды 51-й научной конференции МФТИ*. Москва, 2008. С. 133–136.
3. Лопатовская К.В., Зверков О.А., Селиверстов А.В., Любецкий В.А. Транскрипция генов синтеза пролина у бактерий родов *Marinobacter*, *Pseudomonas* и *Shewanella* регулируется белком семейства tetR // *Труды 32-й конференции «Информационные технологии и системы»*. 15–18 декабря 2009. С. 278–281.
4. Зверков О.А., Селиверстов А.В., Рубанов Л.И., Любецкий В.А. Моделирование конкуренции РНК-полимераз: влияние нокаута сигма субъединицы и температуры на экспрессию генов // *Труды 32-й конференции «Информационные технологии и системы»*. Бекасово, 15–18 декабря 2009. С. 328–331.
5. Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Interaction between nucleome and plastome: heat shock response regulation in plastids of plants // *Pro-*

ceedings of the Seventh International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology. Novosibirsk, June 20–27 2010. P. 161.

6. Зверков О.А., Селиверстов А.В., Любецкий В.А. Позиционная связь генов пластомов растений и водорослей // *Труды 33-й конференции «Информационные технологии и системы»*. г. Геленджик, 20–24 сентября 2010. С. 326–330.
7. Зверков О.А., Селиверстов А.В., Любецкий В.А. Об одном алгоритме кластеризации белков // *Труды 53-й научной конференции МФТИ*, Часть I. Радиотехника и кибернетика, Т. 1, М.: МФТИ, 2010. С. 118–119.
8. Зверков О.А., Горбунов К.Ю., Селиверстов А.В., Любецкий В.А. Кластеризация белков с учётом их доменной структуры // *Труды 54-й научной конференции МФТИ*. Т. 2. М.: МФТИ, 2011. С. 88–89.
9. Зверков О.А., Селиверстов А.В., Любецкий В.А. Семейства белков, кодируемых в пластомах Chlorophyta, Euglenozoa и Rhizaria // *Труды 35-й конференции «Информационные технологии и системы»*, 19–25 августа 2012. С. 298–302.
10. Zverkov O.A., Korolev S.A., Seliverstov A.V., Lyubetsky V.A. Transcription regulation of plastid genes *cysT* and *cysA* in Viridiplantae // *Contributions to the 3rd Moscow International Conference* "*Molecular Phylogenetics*". July 31 – August 4, 2012. P. 85.
11. Зверков О.А. Использование быстрых алгоритмов в задаче кластеризации последовательностей // *Сборник избранных трудов VIII Международной научно-практической конференции* «*Современные информационные технологии и ИТ-образование*». Москва, МГУ им. М.В.Ломоносова, 8–10 ноября 2013. С. 757–763.
12. Зверков О.А., Селиверстов А.В., Любецкий В.А. Построение разделяющих паралоги семейств гомологичных белков, кодируемых в пластидах цветковых растений // *Труды 37-й конференции* «*Информационные технологии и системы*». Калининград, 1–6 сентября 2013. С. 172–177.
13. Kobets N.V., Goncharov D.B., Seliverstov A.V., Zverkov O.A., Lyubetsky V.A. Comparative analysis of apicoplast-targeted proteins in *Toxoplasma gondii* and other Apicomplexa species // *Proceedings of the International Moscow Conference on Computational Molecular Biology: MCCMB'13*, July 25–28, 2013.

3. Используемые сведения о митохондриях и пластидах

3.1. Митохондрии у хордовых: лягушки, человека и крысы

Многие эукариотические клетки содержат митохондрии – полуавтономные органеллы с сильно редуцированным геномом. В митохондриях хордовых животных в кольцевой хромосоме длиной 15–18 т.п.н. закодированы 22 тРНК, 2 рРНК и 13 белков. Транскрипция осуществляется РНК-полимеразами фагового типа, гомологичными РНК-полимеразам бактериофагов T7 и T3. Инициация транскрипции требует участия вспомогательных белковых факторов и происходит на нескольких (до пяти) различных промоторах. Транскрипты могут превосходить по длине хромосому.

К числу транскрипционных факторов относятся белки mtTFA и mtTFB, связывающие полимеразу на каждом промоторе [10, 11]. Эти факторы отделяются от полимеразы после транскрипции первых тринадцати нуклеотидов. Первый фактор имеет несколько изоформ, связанных с альтернативным сплайсингом [12]. У человека второй фактор имеет два варианта: mtTFB1 и mtTFB2 – оба могут участвовать в инициации транскрипции.

В инициации транскрипции участвуют и другие белки, например важный белок mTERF [13], который одновременно осуществляет терминацию транскрипции посредством кооперативного связывания. Этот фактор играет важную роль в предложенной нами модели. Свойства РНК-полимераз фагового типа исследовались в работах [14–17]. В частности, при лобовом столкновении две РНК-полимеразы, движущиеся навстречу друг другу по комплементарным цепям ДНК, могут миновать друг друга с образованием дуплекса [14].

Мы сосредоточимся на митохондриях человека *Homo sapiens* (GenBank: NC_012920.1), крысы *Rattus norvegicus* (GenBank: NC_001665.2) и шпорцевой лягушки *Xenopus laevis* (GenBank: NC_001573.1), а также используем сведения о митохондриях мыши *Mus musculus* (GenBank: NC_005089.1), имеющих тот же порядок генов. Эти модельные организмы были выбраны из-за доступности довольно полных наборов опытных данных о концентрациях РНК и временах их полураспада РНК, которые можно использовать для определения уровней транскрипции генов (т.е. частот их транскрипции). Митохондриальные геномы лягушки, человека и крысы приведены на рисунках 0.1–0.3 соответственно.

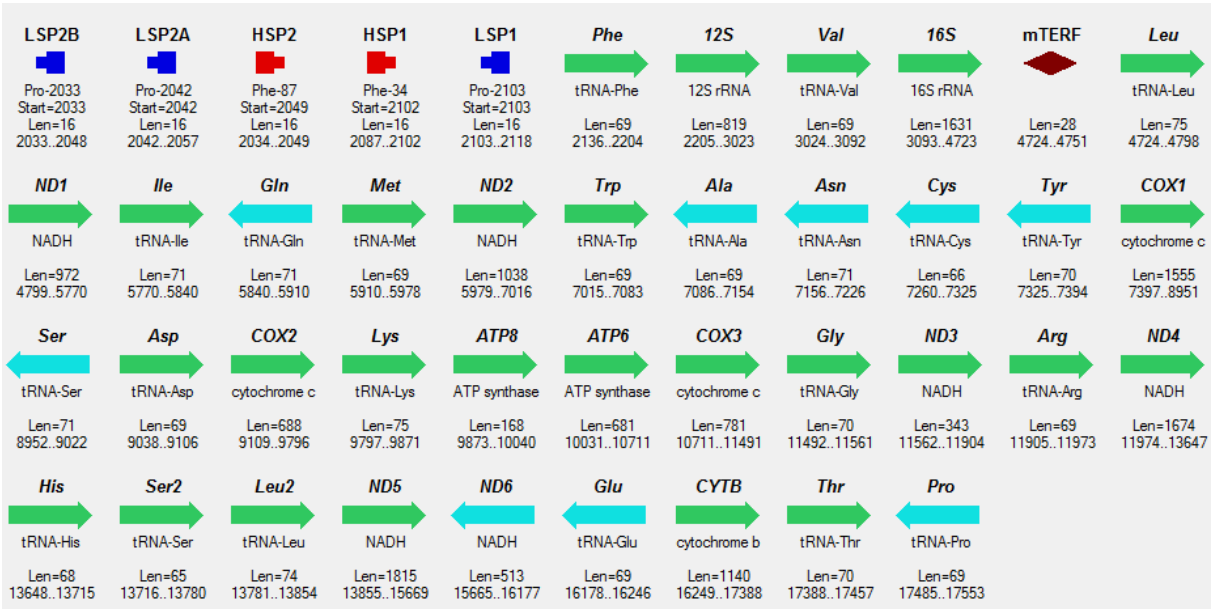


Рисунок 0.1. Митохондриальный геном *Xenopus laevis*
Полная кольцевая ДНК представлена последовательно в четырёх строках. Гены на H-цепи обозначены стрелками, направленными вправо; гены на L-цепи – стрелками, направленными влево. Гены показаны на смысловой цепи. Обозначения: HSP1 и HSP2 – два промотора на H-цепи. LSP1, LSP2A и LSP2B – три промотора на L-цепи. mTERF – сайт связывания белкового фактора mTERF служащего терминатором транскрипции. Координаты указаны по H-цепи.

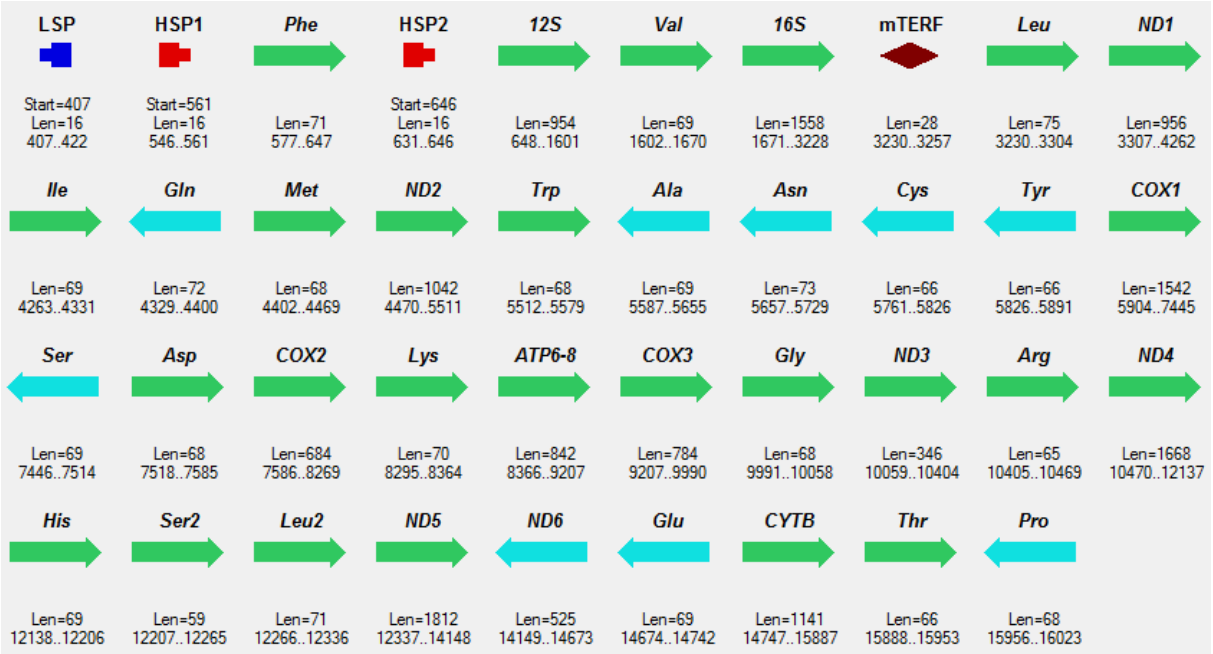


Рисунок 0.2. Митохондриальный геном *Homo sapiens*
Обозначения те же, что на рисунке 0.1.

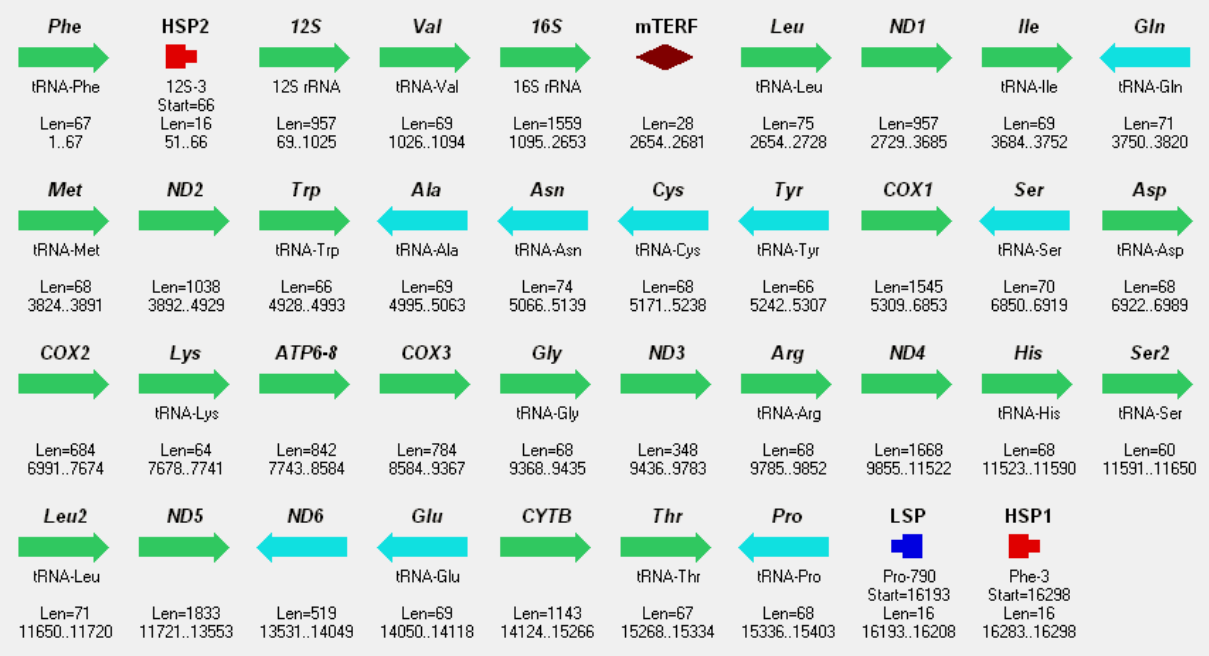


Рисунок 0.3. Митохондриальный геном *Rattus norvegicus*
Обозначения те же, что на рисунке 0.1.

3.2. Структура и взаимное расположение промоторов

Положения митохондриальных промоторов заметно различаются у разных видов. Экспериментальные сведения о расположении промоторов у человека, крысы и лягушки собраны в таблице 0.1. В митохондриях человека известны три промотора: HSP1, HSP2 и LSP. Промоторы HSP1 и LSP имеют консервативный бокс `5′-CANACC(G)CC(A)AAAGAPyA-3′`, [18]. Сайт инициации транскрипции располагается внутри этого бокса за 6–8 нуклеотидов до 3′-края. Сайты инициации транскрипции располагаются: HSP1 – в позиции 561 (перед геном tRNA-Phe), HSP2 – в позиции 646 (перед геном 12S rRNA), [19] и LSP – в позиции 407, [20]. Существенное влияние на качество промоторов оказывают участки: –16..+7 для HSP1 и –28..+16 для LSP, [21].

В митохондриях крысы также имеется три промотора [22]. Сайты инициации транскрипции: HSP1 – в позиции 16298 (15 п.н. перед геном tRNA-Phe), HSP2 – в позиции 66 (перед геном 12S rRNA) и LSP – в позиции 16193 (перед геном tRNA-Pro).

В митохондриях лягушки описаны пять промоторов: HSP1, LSP1, HSP2, LSP2A и LSP2B, все они расположены перед геном фенилаланиновой тРНК, [23, 24]. Транскрипция инициируется внутри консервативной нуклеотидной последовательности `ACRTTATA`. Для дикого типа лягушки и некоторых её мутантных вариантов определены относительные интенсивности инициации транскрипции [25], которые для удобства читателя воспроизведены в таблице 0.2.

Таблица 0.1. Сайты инициации транскрипции у митохондрий
В скобках указаны позиции начал сайтов, соответствующих промоторам на L-цепи.

Вид	Последовательность	Сайт	Позиция
Homo sapiens	Genbank:NC_012920.1	HSP1	561
		HSP2	646
		LSP	(407)
Rattus norvegicus	Genbank:NC_001665.2	HSP1	16298
		HSP2	66
		LSP	(16193)
Xenopus laevis	Genbank:NC_001573.1	HSP1	2102
		HSP2	2049
		LSP1	(2103)
		LSP2A	(2042)
		LSP2B	(2033)

Таблица 0.2. Интенсивности инициации транскрипции в митохондрии лягушки относительно интенсивности инициации на промоторе LSP1

Промотор	Интенсивность
HSP1	13.6 %
HSP2	60.0 %
LSP1	100.0 %
LSP2A	16.6 %
LSP2B	38.2 %

3.3. Влияние белковых факторов на уровни транскрипции

В ходе раннего эмбрионального развития лягушки наблюдается продолжительное увеличение концентрации транскрипционного фактора mtTFA, [26] и согласованное с ним увеличение уровней экспрессии генов [27]. В начале этого периода в митохондриях лягушки репликация и транскрипция почти не происходят и исходный большой запас митохондрий распределяется между делящимися клетками.

Уровень гормонов, характер метилирования определённых областей мтДНК, [22, 28] и мутации мтДНК существенно влияют на уровни транскрипции генов.

3.4. mTERF-зависимая терминация транскрипции

В митохондриях человека имеется два терминатора с различными механизмами действия. В первом механизме белок mTERF связывается с сайтом на ДНК длиной 28 п.н., расположенным непосредственно после гена 16S rRNA и внутри гена tRNA-Leu. Этот терминатор поляризован и вызывает почти 100% терминацию транскрипции по лёгкой цепи, но пропускает часть РНК-полимераз по тяжёлой цепи [27]. Второй механизм описан в следующем пункте 3.5.

Существуют две гипотезы о механизме регуляции транскрипции на тяжёлой цепи у млекопитающих [13, 22]. По первой – транскрипция, инициированная на HSP1, прерывается после транскрипции гена 16S рРНК, а более длинные транскрипты инициируются только на HSP2. По другой – длинные транскрипты могут начинаться с любого промотора и некоторая доля РНК-полимераз прерывает транскрипцию на mTERF независимо от промотора.

Подчеркнём, что у млекопитающих белок mTERF связывается кооперативно с сайтом терминатора и с сайтом активатора, расположенным вблизи промотора HSP1, выступая таким образом одновременно в роли терминатора и активатора [19].

3.5. Белок-независимый терминатор транскрипции

В митохондриях человека mTERF-независимый терминатор расположен в позициях 282..300 на лёгкой цепи, вызывая терминацию около 65% транскриптов, начинающихся с LSP, [29]. Этот терминатор является строго поляризованным, поскольку терминация обусловлена формированием гуанилового (или: G-) квадруплекса (тетрамера) на РНК, за которым следует полиурациловый участок. В митохондриях человека такая последовательность содержит 12 остатков «G» с одним «A» в середине. Терминация происходит, когда формируется гуаниловый квадруплекс на РНК вблизи РНК-полимеразы.

Белок-независимые терминаторы универсальны для всех РНК-полимераз фагового типа. Предполагаемые области терминатора у трёх модельных видов показаны в таблице 0.3. У крысы и лягушки они предсказаны нами биоинформатически. Вероятно, терминация происходит примерно на 10–15 нуклеотидов ниже этого участка, как это наблюдается у человека. Известно, что вблизи этого G-богатого участка происходит разрезание (процессинг) длинной мРНК у лягушки [30].

Таблица 0.3. Белок-независимый терминатор транскрипции (G-квадруплекс): G-богатые участки в митохондриях. Позиции в скобках относятся к кодирующему участку, расположенному на L-цепи.

Вид	Последовательность	Положение	Состав
Homo sapiens	Genbank:NC_012920.1	(16086..16098)	GGGGGAGGGGGGG
Rattus norvegicus	Genbank:NC_001665.2	(303..315)	GGGGGTGGGGGGG
Xenopus laevis	Genbank:NC_001573.1	(1808..1819)	GGGGGGTAGGGGG

3.6. MELAS болезни

Синдром MELAS – (митохондриальная энцефаломиопатия, лактатацидоз, инсультоподобные эпизоды) наиболее распространенная наследуемая по материнской линии митохондриальная болезнь. В более 80% случаев MELAS вызывается транзицией A→G в позиции 3243 в середине сайта связывания белка-терминатора mTERF, что существенно снижает связь mTERF с последовательностью ДНК. У человека эта мутация вызывает: (i) незначительное снижение уровня транскрипции рРНК (12S и 16S), (ii) не более чем 20% снижение концентрации tRNA-Leu, (iii) не более чем 50% снижение tRNA-Lys, (iv) небольшое снижение общего числа мРНК и (v) заметное изменение объёма белковых продуктов [31].

Подчеркнём, что у млекопитающих белок mTERF связывается кооперативно с сайтом терминатора и с сайтом активатора, расположенным вблизи промотора HSP1, выступая таким образом одновременно в роли терминатора и активатора [19].

Сайт mTERF-зависимого терминатора консервативен и расположен ниже гена 16S рРНК в митохондриях многих видов животных [32]. Известно, что в ядерных геномах многих животных кодируются белки, гомологичные mTERF.

3.7. Время полураспада РНК

В работах [33, 34] исследована стабильность митохондриальных РНК человека. Времена полураспада (в минутах) мРНК, кодируемых на тяжёлой цепи в митохондриях здорового человека (значение ± стандартное отклонение) таковы: ND1 – 219 ± 22, ND2 – 142 ± 3, COX1 – 204 ± 91, COX2 – 297 ± 97, ATP6/8 – 424 ± 104, ND3 – 59 ± 1, ND5 – 120 ± 27, CYTB – 132 ± 24. Времена полураспада рРНК составляют несколько часов, таблица 0.4.

В изолированных митохондриях крысы времена полураспада РНК измерены как у крысы с нормальным уровнем гормона щитовидной железы – эутиреоид, так и при недостатке этого гормона – гипотиреоид [22]. В нормальных условиях времена полураспада составили («значение ± стандартное отклонение» в минутах): 44.48 ± 6.34 у 16S

rRNA, 46.00 ± 10.41 у ND5, 84.41 ± 27.49 у ND4/4L и COX1, 63.70 ± 7.82 у CYTB, 78.14 ± 21.05 у ATP6/8 и COX3. Это существенно ниже, чем у человека, таблица 0.5. При недостатке гормона эти времена увеличивались в среднем в 2.13 раза.

Для лягушки времена полураспада неизвестны, но это не мешает сравнивать результаты моделирования с экспериментальными данными в части относительных уровней экспрессии генов, не зависящих от скорости распада РНК.

Таблица 0.4. Экспериментальные данные по митохондриальным транскриптам здорового человека. Уровни в стационарном состоянии представлены как процент от уровней ND1: значение ± доверительный уровень. Периоды полураспада представлены как значение ± стандартное отклонение. Данные взяты из [33, 34].

Ген	Необработанные клетки		Клетки, обработанные тиамфениколом		Длина гена
	Уровень в стационарном состоянии	Время полураспада (мин.)	Время полураспада (мин.)	Относительное изменение	
16S		180 ± 30			1558
ND1	100 ± 4	219 ± 22	273 ± 21	1.25	956
ND2	91 ± 11	142 ± 3	296 ± 22	2.09	1042
COX1	97 ± 19	204 ± 91	236 ± 65	1.15	1542
COX2	234 ± 19	297 ± 97	277 ± 78	0.94	684
ATP6/8	177 ± 69	424 ± 104	506 ± 51	1.19	842
ND3	28 ± 1	59 ± 1	132 ± 16	2.23	346
ND5	102 ± 17	120 ± 27			1812
CYTB	139 ± 16	132 ± 24	406 ± 27	3.06	1141

Таблица 0.5. Экспериментальные данные по митохондриальным транскриптам крысы. Данные взяты из [22] и представлены в виде: значение ± стандартное отклонение. По каждому гену отношения мРНК/рРНК были нормализованы, принимая за 100% значение эутиреоида.

Ген	Эутиреоид		Гипотиреоид	
	Отношение мРНК/рРНК	Период полураспада (мин.)	Отношение мРНК/рРНК	Период полураспада (мин.)
16S		44.48 ± 6.34		87.50 ± 27.52
COX1	100 ± 16	84.41 ± 27.49	86 ± 13	235.12 ± 48.68
ATP6/8	100 ± 19	78.14 ± 21.05	59 ± 9	277.52 ± 31.58
COX3	100 ± 19	78.14 ± 21.05	59 ± 9	277.52 ± 31.58
ND4/4L	100 ± 16	84.41 ± 27.49	86 ± 13	235.12 ± 48.68
ND5	100 ± 25	46.00 ± 10.41	52 ± 11	60.52 ± 5.92
CYTB	100 ± 27	63.70 ± 7.82	57 ± 7	204.30 ± 28.64

3.8. Пластиды растений и водорослей

Пластиды – полуавтономные органеллы растений, которые обладают, в том числе, собственной транскрипционной системой. В пластидах растений и водорослей транскрипцию осуществляют РНК-полимеразы разных типов: одна–две – фагового типа (NEP) и одна – бактериального типа (PEP). NEP – моносубъединичные полимеразы ядерного кодирования, которые связываются с соответствующими NEP-промоторами, а PEP – многосубъединичные РНК-полимеразы пластидного кодирования, которые связываются с PEP-промоторами. В случае PEP в инициации транскрипции участвует одна из нескольких σ-субъединиц, кодируемых и регулируемых в ядре. Интенсивность связывания холофермента РНК-полимеразы с PEP-промотором и процесс инициации транскрипции, вообще говоря, зависит от типа σ-субъединицы [35]. Под интенсивностью понимается частота связывания полимеразы со свободным промотором, не занятым другой полимеразой или фактором транскрипции. Эта ситуация даёт пример регуляторной системы, основанной на взаимодействии ядерного и пластидного геномов. Недавно описаны последовательности ДНК, кодирующие σ-субъединицы у растений; в частности, *Arabidopsis thaliana* обладает шестью σ-субъединицами: Sig1–Sig6. Одни σ-субъединицы достаточно универсальные, например Sig1, другие – специфичные, например Sig5 для светозависимого промотора гена *psbD*, [36]. В целом NEP-промоторы разных типов более изучены, чем PEP-промоторы, особенно в случае минорных σ-субъединиц. Во многих случаях положения NEP- и PEP-промоторов не были заранее известны и определялись нами по множественному выравниванию соответствующих лидерных областей аналогично тому, как это описано в [37].

3.9. Конкуренция РНК-полимераз

Конкуренция РНК-полимераз, в основном, происходит либо при столкновении встречных полимераз, вызывающем прекращение транскрипции, либо при блокировке промотора ранее связавшейся с ним полимеразой или фактором. Итак, связывание полимеразы с промотором возможно, лишь в случае, если в момент попытки связывания промотор не занят другой полимеразой или фактором транскрипции. Если промоторы расположены столь близко, что связывание с ними стереохимически взаимно исключается, то также возникает конкуренция. Принципиальное значение имеют инициация транскрипции (особенно для PEP) и взаимодействие полимеразы со вторичными структурами нуклеиновых кислот и белковыми факторами. Одновременно происходящее множество связываний и движений PEP и NEP позволяет объяснить опубликованные численные результаты экспериментов.

Важность математических и соответственно компьютерных моделей фундаментальных процессов в клетке отмечается во многих работах. Однако, насколько автор может судить, известно немного таких не узко специализированных моделей. Среди них отметим модель кинетики вторичной структуры РНК, [38, 39] и модель аттенюаторной регуляции [40].

Из работ, более близких к главе 1, отметим, например, [41–43]. В этих работах моделируется формирование замкнутого, открытого и элонгационного комплекса РНК-полимеразы, взаимодействие РНК-полимераз у *E.coli* и в паузе в ходе транскрипции и регуляция этих процессов белками, связывающими ДНК. Показано, что элонгация РНК-полимеразы может ингибировать связывание других полимераз с промоторами, а также активаторов – с сайтами на ДНК, лежащими перед ней (downstream). В этих работах показано, что, вопреки нашему исследованию, элонгация РНК-полимераз не приводит к заметному взаимодействию между противоположно направленными промоторами в бактериофаге λ. У РНК-полимеразы в момент транскрипции промотора наступает пауза, что показано *in vivo* и подтверждено в указанных работах моделированием. Регуляция генов посредством удлинения паузы при элонгации носит общий характер и может быть широко распространенной. В этих работах высказано предположение, что даже редкая транскрипция РНК-полимеразами как при встречном, так и при сонаправленном движении может приводить к значительному подавлению транскрипции.

Отметим ещё одну работу [44]: у фага Ф29 сенной палочки лобовое столкновение РНК-полимеразы и осуществляющей репликацию ДНК-полимеразы не приводит к терминации ни того, ни другого процесса. Это позволяет думать о существовании механизма, разрешающего такой конфликт. Однако у этого фага, по-видимому, нет аналогичного механизма разрешения конфликта при сонаправленном столкновении РНК- и ДНК-полимераз.

Автору неизвестны работы, в которых рассматривается одновременная инициация и элонгация РНК-полимераз на многих промоторах вместе с их взаимодействием с разнообразными факторами произвольного локуса, что является предметом главы 1.

3.10. Нокауты генов σ-субъединиц РНК-полимераз

У *Arabidopsis thaliana* и других растений сравнивались уровни транскрипции многих генов в диком типе и в мутантах по *sig3* или *sig4*. Точнее, в случае нокаута *sig4*, [45] и *sig3*, [46] в экспериментальных испытаниях оценивались усреднённые (по массе пластид) отношения MT/WT уровня транскрипции ряда генов у мутантного типа (MT) к таковому у дикого типа (WT) и их дисперсии.

3.11. Тепловой шок изолированных хлоропластов

В опытах с тепловым шоком оценки усреднённого отношения HT/WT и его дисперсии (HT – уровень транскрипции после теплового шока, WT – в диком типе) формально схожи с исследованием нокаута σ-субъединиц. Ответ на тепловой шок существенно различается у хлоропластов в составе эукариотических клеток и в изолированном состоянии, как экспериментально показано в [47]. Из этой работы известны отношения уровней транскрипции ряда генов после теплового шока к уровням их транскрипции в контрольном материале (без теплового шока) в изолированных хлоропластах. В последних уровни транскрипции генов зависят в основном от скорости элонгации полимераз и интенсивностей связывания промоторов, что снижает влияние ядра на изменения концентраций σ-субъединиц.

3.12. Анализ других экспериментальных данных

Помимо опытов с нокаутом σ-субъединицы и тепловым шоком, модель позволяет объяснить и данные хроматограмм [48], которые, однако, менее надёжны в количественном отношении. Хроматограммы могут использоваться для сравнения уровней транскрипции генов с разных промоторов или перед и после нокаута РНК-полимеразы фагового типа. Меньшая надёжность связана с невысокой точностью блот-метода, малым числом повторений опыта (не более двух в [48]) и неоднозначностью численной интерпретации хроматограмм. Например, наше измерение хроматограммы, приведённой в [48] обнаруживает различие в уровнях транскрипции гена *ycf1* с разных промоторов: RpoTp-зависимый промотор *ycf1*-39 более эффективен, чем RpoTmp-зависимый промотор *ycf1*-104 и вдвое более эффективен, чем PEP-зависимый *ycf1*-34/33. Эти данные хорошо согласуются с предсказаниями модели. При нокауте RpoTp (когда не происходит связывания с промотором *ycf1*-39) уровень транскрипции с *ycf1*-104 остаётся прежним, а с *ycf1*-34/33 даже увеличивается. В данной работе нокаут RpoTp не обсуждается из-за недостаточности экспериментальных данных.

Для численной оценки параметров модели использовались данные из независимых исследований: влияние мутаций PEP-промоторов на интенсивность связывания субъединиц Sig1–3, [49], влияние мутаций RpoTp-промотора фагового типа на интенсивность связывания NEP, [15] и другие исследования PEP- и NEP-промоторов пластид [48, 50–52].

3.13. Заключение

В главе 1 сделан шаг к моделированию механизма конкуренции РНК-полимераз. Предсказаны значения интенсивностей попыток связывания полимераз с промоторами, при которых имеется хорошее согласие с опытными данными по изменению уровней транскрипции генов в митохондриях хордовых и в пластидах растений; предположены механизмы ряда физиологических явлений и болезней человека. Модель также может также служить для предсказания из опытных данных трудно измеримых в непосредственных опытах характеристик РНК-полимераз и процесса транскрипции: интенсивность связывания холофермента с промотором в зависимости от его нуклеотидного состава и типа σ-субъединицы, среднее число абортивных попыток инициации транскрипции и т.д.

ГЛАВА 1. ВЗАИМОДЕЙСТВИЕ РНК-ПОЛИМЕРАЗ В МИТОХОНДРИЯХ И ПЛАСТИДАХ

1. Примеры локусов в митохондриях и пластидах

В качестве примеров рассмотрены три локуса пластид растений.

Первый локус из *Arabidopsis thaliana* (рисунок 1.1*a*): N1–N2–P1–*ycf1*–(*ndhF*–P2)–*rpl32*, где используются следующие обозначения промоторов: P1 = *ycf1*–33/34, P2 = *ndhF*–320, N1 = *ycf1*–104, N2 = *ycf1*–39. В скобках указываются объекты, расположенные на комплементарной цепи. Здесь и далее PEP-промоторы обозначаются буквой P, NEP-промоторы – буквой N. Отметим, что в пластоме содержится две копии участка N1–N2–P1–*ycf1*, в одной из которых короткий ген *ycf1* повторяет начало длинного гена *ycf1*; эти копии находятся в существенно разных окружениях. Уровень транскрипции *ycf1* является суммой уровней транскрипции двух копий.

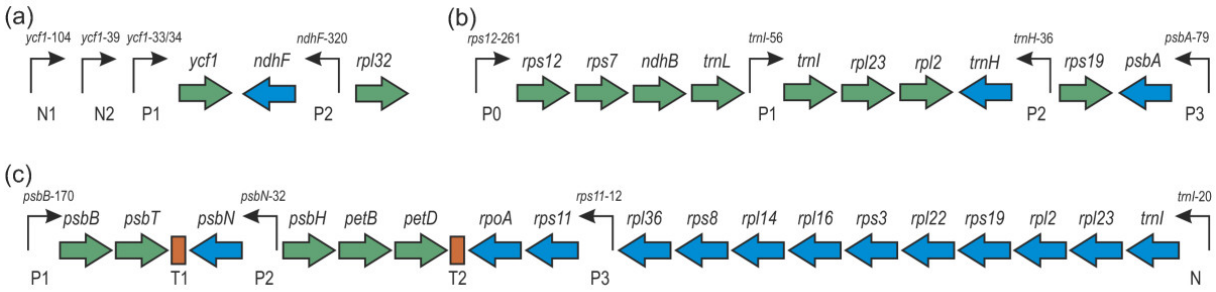


Рисунок 1.1. Расположение промоторов и генов для локусов 1–3
Локусы 1 и 3 принадлежат *Arabidopsis thaliana*, локус 2 – *Hordeum vulgare*. P# – PEP-промоторы, N# – NEP-промоторы, T# – найденные нами терминаторы. Указаны координаты сайта инициации транскрипции генов относительно их инициирующего кодона: (a) – локус 1, (b) – локус 2, (c) – локус 3.

В первой копии промоторам N1 и N2 предшествуют интенсивно транскрибируемые гены на комплементарной цепи, что практически блокирует доступ полимераз к этому участку. Во второй копии перед N1 также расположены интенсивно транскрибируемые гены на комплементарной цепи, а за *ycf1* следует длинный оперон на той же цепи, что делает эту копию участка практически независимой от окружающих промоторов. Этот локус исследовался в экспериментах с нокаутом гена *sig4* при температуре +23°C.

Второй локус из *Hordeum vulgare* содержит два участка. Первый участок (рисунок 1.1*b*): P0–*rps12*–*rps7*–*ndhB*–*trnL*$_{CAA}$–P1–*trnI*$_{CAU}$–*rpl23*–*rpl2*–(*trnH*–P2)–*rps19*–(*psbA*–P3), и второй участок: P0–*rps12*–*rps7*–*ndhB*–*trnL*$_{CAA}$–P1–*trnI*$_{CAU}$–*rpl23*–*rpl2*–(*trnH*–P2)–*rps19*–*rpl22*–*rps3*–*rps16*, где P0 = *rps12*–261, P1 = *trnI*–56, P2 = *trnH*–36, P3 = *psbA*–79 –

PEP-промоторы. В первом участке полимераза начинает транскрипцию с P0 и P1, и с P2 и P3 – на комплементарной цепи; во втором участке отсутствует ген *psbA* и его промотор P3. В обоих копиях промотору P0 предшествует активно транскрибируемые гены тРНК на комплементарной цепи, что практически изолирует P0 от апстрима. В первом участке перед P3 расположены гены, транскрибируемые в том же направлении, поэтому рассматривается совокупная транскрипция с P3, т.е. полимеразами, начавшими транскрипцию с этого промотора и с вышележащих промоторов на комплементарной цепи. Второй участок примыкает к 5'-концу большого оперона, расположенного на той же цепи, что блокирует инициацию транскрипции *trnH* из вышележащей относительно P2 области. Этот локус изучался в опытах с тепловым шоком: растения выращивали в течение 6–7 дней при температуре 21°C и затем подвергали воздействию температуры 40°C в течение 1.5 часа. Контрольные растения не подвергали нагреванию. В течение следующих 0.25 часа при температуре 25°C оценивался объём полных транскриптов относительно контрольных растений. Поскольку уровень транскрипции генов *rpl23* и *rpl2* измерялся совокупно, то же было сделано и в модели.

Третий локус из *Arabidopsis thaliana* (рисунок 1.1*c*): P1–*psbB*–*psbT*–T1–(*psbN*–P2)–*psbH*–*petB*–*petD*–T2–(*rpoA*–*rps11*–P3–*rpl36*–*rps8*–*rpl14*–*rpl16*–*rps3*–*rpl22*–*rps19*–*rpl2*–*rpl23*–*trnI*–N), где P1 = *psbB*–170, P2 = *psbN*–32, P3 = *rps11*–12 – PEP-промоторы; N = *trnI*–20 – NEP-промотор; T1 и T2 – терминаторы (вероятно, крест-шпильки на ДНК), предсказанные моделью на участках: T1 – *psbT*+22...*psbN*–1, T2 – *petD*+47...*rpoA*–139. Интенсивно транскрибируемый ген *clpP* расположен выше P1 на комплементарной цепи, а активный ген *ycf2* расположен ниже N на основной цепи, из-за чего локус практически не транскрибируется извне. Локус изучался в опытах по нокауту генов *sig3* и *sig4* при температуре +23°C. Нокаут *sig3* и *sig4* моделировался при тех же значениях интенсивностей связывания с промоторами РНК-полимераз посредством остальных σ-субъединиц, как и в диком типе.

При изучении митохондрий рассматривались полные митохондриальные геномы лягушки, человека и крысы, полученные из базы данных GenBank NCBI, [53], см. рисунки 0.1–0.3 и другие сведения во введении.

2. Модель взаимодействия РНК-полимераз

Транскрипция генов фиксированного локуса ДНК может выполняться одновременно многими РНК-полимеразами, которые связываются со своими промоторами, а затем движутся каждая вдоль своей цепи, возможно, навстречу друг другу. В нашей модели для каждого промотора задаётся *интенсивность* попыток связывания его какой-то

РНК-полимеразой. Значения интенсивностей обычно не известны из экспериментов и вычисляются в модели, как обратная задача: по совокупности опытных данных (в основном, об изменениях уровней транскрипции генов) найти неизвестные интенсивности и, возможно, другие параметры модели. Интервалы времени между такими попытками описываются пуассоновским процессом, каждая попытка считается успешной, если в момент, когда она произошла, промотор не занят другой РНК-полимеразой или любым другим фактором: регуляторным белком, вторичной структурой и т.д. Итак, каждому NEP-промотору и каждому PEP-промотору (причём последний берётся в паре с фиксированной группой σ-субъединиц) сопоставляется свой пуассоновский процесс с параметром λ. Ниже используются следующие группы: все σ-субъединицы и все σ-субъединицы кроме одной, нокаутируемой. В опыте с локусом 1 (рисунок 1.1*a*) в качестве нокаутируемой σ-субъединицы бралась Sig4, а в опыте с локусом 3 (рисунок 1.1*c*) – Sig3 или Sig4; локус 2 (рисунок 1.1*b*) не связан с опытами по нокауту σ-субъединицы, поэтому здесь для всех PEP-промоторов рассматривается одна группа, состоящая из всех σ-субъединиц.

Таким образом, каждому NEP-промотору соответствует свой стохастический процесс, который определяет промежутки времени между попытками связывания с NEP. Это время равно $-(\ln\xi)/\lambda_N$, где ξ – равномерно распределённая случайная величина, заданная на интервале от 0 до 1. Параметр λ_N – искомое значение для этого промотора. Аналогично определяются стохастические процессы для каждого PEP-промотора. Промежутки времени также вычисляются как $-(\ln\xi)/\lambda$, где $\lambda=\lambda_P$ для PEP в паре с группой всех σ-субъединиц и $\lambda=\lambda_4$ для PEP в паре с группой всех σ-субъединиц кроме нокаутируемой Sig4. Здесь Sig4 появляется в связи с локусом 1, а для локуса 3 фигурируют Sig3 или Sig4, в соответствии с нокаутами в экспериментах. Итак, используются пары параметров, соответствующие каждому в отдельности PEP-промотору локуса: λ_P и λ_4 (локус 1), λ_P и либо λ_3, либо λ_4 (локус 3). Для краткости все эти параметры λ, свои для каждого промотора, называются *интенсивностями связывания* промотора. Здесь важно: определив интенсивности связывания в диком типе, мы используем их без изменения при описании нокаутов по разным σ-субъединицам и при описании теплового шока в том же или даже в близком виде. Интенсивности измеряются в с^{-1} (обратных секундах).

Каждому белковому фактору транскрипции F соответствует аналогичный стохастический процесс с параметром λ_F, который определяет промежутки времени между попытками связывания фактора со своим сайтом на ДНК. Такая попытка считается

успешной, если в момент её совершения сайт связывания свободен от всех РНК-полимераз и любых факторов. Наконец, каждому терминатору транскрипции (крест-шпильке на ДНК) соответствует бернуллиевская случайная величина с параметром p, описывающая терминацию транскрипции на каком-либо нуклеотиде плеча шпильки.

Для моделирования процесса элонгации нужно задать значения параметров v_N и v_P – скорости элонгации NEP и PEP соответственно. Эти скорости зависят от температуры, нуклеотидного состава ДНК и вторичных структур, образующихся на РНК в процессе транскрипции [54, 40]. Результаты работы получены в предположении постоянной скорости РНК-полимеразы (при фиксированной температуре) и без учёта вторичной структуры РНК, так что элонгация моделируется как детерминированный процесс.

Если PEP связала PEP-промотор, то сначала моделируется *абортивный* процесс, а затем процесс *элонгации* полимеразы. Для абортивного процесса нужно определить число абортивных попыток и длину каждой из абортивных РНК, которые в модели находятся следующим образом. Длительность t всего абортивного процесса задаётся как $t = -(\ln \xi) \cdot t_0$, где t_0 – среднее время абортивного процесса (например, $t_0 = 0.4$ с). Число абортивных попыток k определяется как наибольшее число слагаемых в левой части неравенства $-(\ln \xi_1 + ... + \ln \xi_i + ... + \ln \xi_k) \le t \cdot v_P / r_0$, при котором оно остаётся верным. Параметр r_0 – средняя длина одной абортивной РНК (например, $r_0 = 4$). При каждой i-й абортивной попытке появляется РНК, длина которой равна целому числу, ближайшему к числу $-r_0 \cdot (\ln \xi_i)$. Таким образом, величина $-(\ln \xi_i)$ имеет смысл случайной поправки к среднему времени r_0 / v_{PEP}, уходящему на одну абортивную попытку, где v_P – скорость PEP.

Для моделирования опытов по изменению уровня транскрипции после теплового шока (локус 2, рисунок 1.1*b*) в модель введены следующие известные из опыта параметры: в течение времени t_1 растение находится при температуре T_1; затем в течение времени t_2 у одной массы изолированных хлоропластов температура повышается до T_2, а у другой такой же массы она остаётся равной T_1; затем в течение времени t_3 у обоих масс температура меняется на новое значение T_3, и в этом последнем промежутке времени измеряется отношение числа завершённых транскрипций некоторых генов в материале после шока к таковому в контрольном материале [47]. В опыте эти параметры имели следующие значения: t_1 = 6–7 суток (при моделировании можно брать t_1 равным 3 часам, так как за это время модель выходит на стационарный режим, и дальней-

шее увеличение t_1 не меняет результата), $T_1 = 21°C$, $t_2 = 1.5$ часа, $T_2 = 40°C$, $t_3 = 15$ минут, $T_3 = 25°C$.

Модель допускает самые разные **дисциплины взаимодействия**, но приводимые результаты были получены при следующих условиях: если передние края двух полимераз (транскрибирующих комплементарные цепи) занимают одну и ту же позицию, то в модели принимается, что элонгация обеих прекращается. Если на одной цепи ДНК полимераза *X* передним краем вплотную примыкает к полимеразе *Y*, то *X* не может обогнать *Y*. То же самое относится к холоферменту и абортивному процессу. Взаимодействие РНК-полимеразы с терминаторами транскрипции описаны отдельно, ниже.

Кажется, что принятая дисциплина взаимодействия, по существу, содержит мало произвола; мы варьировали её в биологически разумных пределах и получали практически те же результаты. Например, РНК-полимеразы одного типа имеют в модели одинаковую скорость элонгации, и, если движутся по одной цепи ДНК, то практически не сталкиваются с впереди идущей полимеразой. Особый случай, когда фаговая полимера движется вслед за бактериальной полимеразой. Однако и в этом случае можно думать, что лёгкая полимераза не сталкивает тяжелую и не сама не диссоциирует с ДНК. Нетривиальный экспериментальный результат [14] об РНК-полимеразах фагового типа, движущихся навстречу друг другу, также фактически не противоречит нашей модели: хотя движущиеся навстречу полимеразы могут миновать друг друга, при этом образуется дуплекс, который не позволяет увеличиться числу транскриптов и, можно думать, приводит к диссоциации разминувшихся полимераз. Детали описания взаимодействия РНК-полимераз с терминаторами разной природы (см. ниже), также оказывают небольшое влияние. Например, изменение параметра *p* взаимодействия полимеразы со шпилькой приводит, в основном, к изменению места терминации транскрипции на плече шпильки на несколько нуклеотидов.

В целом мы исходили из того, что согласие модели с обширным корпусом разнообразных экспериментов является достаточным на этой стадии исследования.

Крест-шпильки на ДНК, характерные для пластид [55] и бактерий [56] отсутствуют в рассмотренных митохондриях. В случае митохондрий факторами являются многофункциональный регуляторный белок mTERF и G-квадруплекс на РНК. В модели учитывается терминация транскрипции при столкновении РНК-полимеразы с белковым фактором mTERF. Если белок mTERF пытается связаться со своим сайтом, попытка считается успешной, если сайт свободен от полимераз и ранее связавшихся копий этого белка. Если mTERF связался с сайтом и к нему приходит РНК-полимераза, то либо она проходит дальше, а комплекс mTERF·ДНК диссоциирует («протекание терминатора»),

либо она терминирует, а комплекс сохраняется («непротекание терминатора»). Частота протекания в одну и другую стороны не предполагаются равными. Протекание G-квадруплекса описывается известным из опыта понижающим коэффициентом для числа полимераз, проходящих по одной из цепей; в рассматриваемом случае – это L-цепь. В остальном дисциплина взаимодействия объектов остаётся прежней, как выше, в пластидах.

3. Параметры модели

3.1. Параметры РНК-полимеразы бактериального типа (PEP)

Скорости элонгации PEP при разных температурах соответствуют скоростям РНК-полимеразы *E. coli*, так как соответствующие субъединицы этих полимераз – близкие гомологи [36]. Для *E. coli* известны две линейные зависимости скорости элонгации от температуры [57, 58] и одно прямое наблюдение 42.5 нт/с при 37°C, [59]. Зависимость с бóльшими значениями из [58] рассматривается как содержащая систематическую ошибку [57, 59]. Поэтому мы использовали зависимость из [57]. Из неё следует, что при 21°C (нормальная температура выращивания *Hordeum vulgare* в опыте) и 23° (аналогичная температура для *Arabidopsis thaliana*) значения скоростей соответственно равны 9.2 и 12.1 нт/с. Во время теплового шока (локус 2) температура поднималась до 40°C, а затем падала до 25°C, что соответствует скоростям 36.8 и 15 нт/с.

Размеры PEP брались такие же, как у бактериальной РНК-полимеразы, а размеры NEP – такие же, как у РНК-полимеразы бактериофага T7 (NEP рассматриваются в п. 3.3). У *E. coli*, *Thermus thermophilus* и в хлоропластах *Sinapis alba* классические опыты с футпринтингом [60], рентгеноструктурным анализом [61] и мутациями в области промотора [49, 62] дают размеры: 35 нт (от –15 до +20 относительно положения сайта инициации транскрипции) для кор-фермента, 29 нт (от –44 до –16) для области ДНК, покрываемой холоферментом без учёта кор-фермента; что даёт оценку от –44 до +20 для холофермента, 64 нт. Размер промотора можно также оценить: от небольшого участка ниже –10-боксом PEP-промотора до небольшого участка выше –35-бокса промотора, вплоть до позиции –44. Если учитывать связывание с ДНК α-субъединиц [62], то левую границу холофермента можно ещё отодвинуть до –60, но мы принимали предыдущие значения для PEP.

3.2. Параметры PEP-промоторов и число абортивных попыток

Интенсивности связывания в модели можно получить из данных об изменениях уровней транскрипции генов и других, решая обратную задачу, см. ниже. Но также зна-

чения интенсивностей связывания для некоторых генов и видов можно оценить сверху, что даёт дополнительную информацию при решении обратной задачи. Например, для оценки интенсивности связывания холофермента PEP с промотором можно использовать данные об интенсивности связывания с оптимальным промотором гена *rrn* у *E. coli* с последующим переносом полученного значения интенсивности на оптимальный и единственный промотор гена *psbA* у *Arabidopsis thaliana*, а затем с переносом этого значения на интересующие нас PEP-промоторы у ортологичных генов и близких видов. У *E. coli* эта оценка получается из опытных данных о числе рибосом, времени между делениями и числе копий генов рибосомной РНК: у *E. coli* в условиях аэрирования на среде с глюкозой при 37° со временем генерации 40 минут получены следующие количества рРНК в клетке [63]: 23S рРНК, 16S рРНК, 5S рРНК – по 18700 каждой. Отсюда простой подсчёт даёт приблизительно 0.9 секунды между последовательными инициациями. Поэтому интенсивность связывания может быть оценена сверху числом 1.12 $с^{-1}$. Соответственно, в модели интенсивность λ выбиралась до этой границы.

Чтобы оценить интенсивность связывания более точно, учтём, что эти 0.9 сек состоят из времени на собственно связывание и времени на абортивный процесс; например, 0.9=0.5+0.4. В модели варьировались разные варианты разбиения числа 0.9 на два слагаемых с точностью до 0.1 секунды (например, выше упоминалось, что среднее время абортивного процесса 0.4 с). Для рассмотренного выше примера интенсивность связывания для оптимального PEP-промотора перед геном *psbA* в хлоропластах *Arabidopsis thaliana* и *Hordeum vulgare* получается равной 0.5 и теперь (например, для генов 1-го локуса, рисунок 1.1*а*) нужно перейти к промоторам *ycf1*–33/34 и *ndhF*–320. Для этого интенсивность 0.5 у *psbA* умножается на понижающий коэффициент, который отражает более низкое качество последовательностей промоторов *ycf1*–33/34 и *ndhF*–320 по сравнению с последовательностью промотора *psbA*–77. Это возможно на основе экспериментальных оценок влияния на интенсивность связывания нуклеотидных замен в составе промотора *psbA*–77 в хлоропластах горчицы [49]. Заметим, что у всех фотосинтезирующих цветковых растений промотор перед геном *psbA* высококонсервативен [37, 64]. В результате получаем, что интенсивность связывания с *ycf1*–33/34 равна 0.09 $с^{-1}$, а интенсивность связывания с *ndhF*–320 равна 0.15 $с^{-1}$.

Для оценки числа абортивных попыток нужно знать кроме среднего времени t_0 на весь абортивный процесс – среднюю длину r_0 абортивной РНК. РНК·ДНКовый гибрид имеет длину около 9 нт, а возможно, несколько меньше за счёт закрытия канала

σ-субъединицей [61, 62]. Поэтому длина одной абортивной РНК находится в пределах от 1 до 8–9 нт и в модели перебирались эти значения, например r_0=4.

3.3. Параметры РНК-полимеразы фагового типа (NEP)

Автору не известны экспериментальные данные о скорости элонгации NEP. Этот параметр является важнейшим в нашей модели. Скорость репликации у *E. coli* равна 1500 нт/с; это значение было принято за максимальную скорость NEP. Нижняя оценка скорости элонгации NEP может быть косвенно получена из соотношения длины *E* первого экзона и длины *I* первого интрона в генах, кодирующих белки в пластидах растений и водорослей из таксономической группы Streptophyta. Поскольку в пластидах транскрипция и трансляция сопряжены, транскрипция первого интрона должна завершиться до начала трансляции первого экзона.

Таким образом, если нет специальной регуляции инициации трансляции (подобной той, что рассматривается в главе 3), отношение скоростей элонгации РНК-полимеразы и рибосомы больше, чем (*E*+*I*)/*E*. Для генов с очень короткими первыми экзонами должна иметь место регуляция, ведущая к задержке инициации трансляции. Поэтому для получения нижней оценки скорости элонгации NEP нужно использовать гены, для которых нет оснований предполагать задержку (регуляцию или транссплайсинг). У *Arabidopsis thaliana* такими генами, по-видимому, являются *rpoC1*, *infA*, *ndhA* и *ndhB*. Первый из них преимущественно транскрибируется NEP, [50], а перед генами *infA*, *ndhA* и *ndhB* не найдено хороших кандидатов на PEP-промотор, что позволяет предполагать, что они также преимущественно транскрибируются NEP. Максимальные (по разным видам) отношения (*E*+*I*)/*E* для этих четырёх генов равны: 1.08 (для *infA* у *Cucumis melo*), 3.71 (для *ndhA* у *Chara vulgaris*), 3.75 (для *rpoC1* у *Zygnema circumcarinatum*), 3.93 (для *ndhB* у *Olea europaea*). Наибольшее значение 3.93 соответствует нижней границе скорости NEP, таким образом, равной 177 нт/с. Ещё большие отношения получаются при рассмотрении генов водорослей, для которых, однако, характер транскрипции менее ясен: 7.86 (для *rpl2* у *Chara vulgaris*), 7.94 (для *ycf3* у *Zygnema circumcarinatum*) и 10.27 (для *ycf66* у *Zygnema circumcarinatum*). Если использовать наибольшее из этих отношений, то скорость элонгации NEP превышает 462 нт/с. Знание скорости элонгации NEP может улучшить точность модели.

У фага T7 для ортолога NEP из мутаций в области промотора гена *rpoB* в хлоропластах табака имеем координаты промотора от –14 до +1 относительно сайта инициации транскрипции [15]. Позиция –15 также оказывает, хотя и малое, влияние на качество промотора [15]. Из опытов по определению участка ДНК, защищенного NEP (фут-

принтинга), известно число 15 защищённых нуклеотидов ДНК; из других опытов известно число 11 неспаренных нуклеотидов ДНК, [16]. Значение 15 нуклеотидов получается из анализа кристаллической структуры РНК-полимеразы фага T7, [17]. Итак, в модели размер NEP принимался равным от –15 до +1.

4. Экспериментальные данные об уровнях транскрипции генов и временах полураспада

Решением называется набор неизвестных параметров в модели. Для ряда генов из опытов известны *относительные* (к нулевому моменту времени того же гена или к «эталонному» гену) количества РНК в стационарном состоянии, и в некоторых случаях известны времена полураспада этих РНК.

4.1. Данные о митохондриях

Для митохондрий *решением* является набор параметров модели, состоящий из интенсивностей попыток связывания с каждым из имеющихся промоторов, условных вероятностей p и q протекания в обе стороны mTERF-зависимого терминатора и интенсивности λ попыток связывания фактора mTERF. Здесь λ включает и процесс спонтанной диссоциации комплекса mTERF·ДНК. Характеристики квадруплекса берутся из опыта; роль mTERF как активатора не учитывается.

Для лягушек из опыта известны такие количества u_{ij} для j-го гена в i-й момент времени, отнесённые к количеству РНК того же гена в нулевой момент времени, т.е. $u_{ij} = \frac{2z_{ij} \cdot t_j}{2z_{0j} \cdot t_j} = \frac{z_{ij}}{z_{0j}}$, где z_{ij} – уровень транскрипции j-го гена в i-й момент времени, t_j – время полураспада j-го гена. Сами времена t_j здесь не известны. В приведены опытные отношения u_{ij} (их опытная погрешность не определялась); они сравниваются с отношениями $\frac{z_{ij}}{z_{0j}}$ средних значений z_{ij} и z_{0j}, *вычисленных в модели*, таблица 1.1 и рисунок 1.2. Последние практически не имеют погрешности. Здесь используется моделируемое время для связи времени в модели и в эксперименте.

Уровень транскрипции гена в модели определяется как число транскриптов, деленное на время. Число транскриптов определялось за 9 часов модельного времени после стабилизации модели, которая обычно происходила через 1 час после начала моделирования. Числа транскриптов для 48 часовых эмбрионов лягушек, а также для человека и крысы приведены в таблице 1.2.

Для человека известны *относительные* (к эталонному гену ND1) количества u_j РНК в стационарном состоянии и времена полураспада этих РНК, т.е. $u_j = \frac{2z_j \cdot t_j}{2z_0 \cdot t_0}$, где z_j – уровень транскрипции *j*-го гена, а t_j – время полураспада *j*-го гена. Из модели известно отношение $\frac{z_j}{z_0}$, которое сравнивается с опытным значением

$$u_j \cdot \frac{t_0}{t_j} \qquad (1)$$

(таблица 1.3, верхняя половина; особая ситуация с геном COX1 обсуждается ниже).

Таблица 1.1. Результаты для трёх лягушек, полученные в модели и в опыте.
Два параметра решения – интенсивность связывания mTERF с сайтом терминации (и кооперативно в области промотора) и интенсивность связывания с промотором *LSP1*, подчеркнуты. Скорость ортолога NEP принята равной 500 нт/с. Затем указаны уровни транскрипции генов (относительно нулевого момента – времени оплодотворения икры): модельные (mod) и опытные (exp) значения вместе с относительными отклонениями последних в процентах (dev), вычисленные по формуле (4), для трёх лягушек в последовательные моменты времени.

час	***mTERF***	***LSP1***	ND1			COX2		
Frog1			mod	exp	dev	mod	exp	dev
0	0.0157	0.0034	1.0	1.0		1.0	1.0	
5	0.0448	0.0089	1.0	1.1	-12	0.9	0.8	+14
10	0.0872	0.0157	1.2	1.3	-5	1.1	1.1	+1
14	0.0793	0.0173	1.7	2.3	-26	1.6	1.6	-3
16	0.0960	0.0209	2.0	2.9	-31	1.7	1.4	+24
18	0.0542	0.0157	2.1	3.2	-34	1.9	1.7	+14
20	0.0655	0.0157	1.8	3.0	-41	1.6	1.4	+13
23	0.0721	0.0492	9.4	9.7	-4	7.6	5.1	+49
48	0.0542	0.0872	29.3	26.6	+10	26.2	13.4	+96
96	0.0407	0.0960	48.1	48.7	-1	45.3	20.9	+117
час	***mTERF***	***LSP1***	ND1			COX2		
Frog2			mod	exp	dev	mod	exp	dev
0	0.0089	0.0041	1.0	1.0		1.0	1.0	
6	0.0045	0.0023	1.2	1.3	-8	1.2	1.0	+22
9	0.0073	0.0045	1.3	1.5	-14	1.3	1.3	-1
20	0.0157	0.0157	3.8	4.6	-17	3.7	3.7	+1
30	0.0157	0.0230	7.2	7.2	0	7.1	6.8	+4
48	0.0407	0.1056	20.5	19.5	+5	19.7	19.7	0
7 дней	0.0041	0.0073	6.5	6.1	+7	6.6	8.0	-18
час	***mTERF***	***LSP1***	16S			ND6		
Frog3			mod	exp	dev	mod	exp	dev
0	0.0960	0.0026	1.0	1.0		1.0	1.0	
5	0.0407	0.0050	2.2	2.2	+0.9	2.2	2.2	0.0
14	0.0230	0.0081	5.0	5.0	0.0	4.5	4.5	-0.2
20	0.0038	0.0028	5.9	6.0	-1.3	4.0	4.0	+0.5
28	0.0336	0.1056	92.2	92.0	+0.2	25.1	25.0	+0.4
48	0.0143	0.0306	44.1	44.0	+0.2	15.0	15.0	+0.3

Таблица 1.1 – продолжение

час	***mTERF***	***LSP1***	ATP6/8			ND4			ND6			CYTB		
Frog1			mod	exp	dev	mod	exp	dev	mod	exp	dev	mod	exp	dev
0	0.0157	0.0034	1.0	1.0		1.0	1.0		1.0	1.0		1.0	1.0	
5	0.0448	0.0089	0.9	0.9	+1	0.9	2.1	-59	2.4	2.4	-1	0.8	0.7	+19
10	0.0872	0.0157	1.1	0.7	+56	1.0	2.3	-57	4.1	4.0	+2	0.9	0.6	+50
14	0.0793	0.0173	1.5	1.3	+18	1.4	3.0	-53	4.4	4.4	0	1.2	1.2	+3
16	0.0960	0.0209	1.7	1.3	+31	1.5	4.3	-65	5.6	5.8	-4	1.3	1.3	+2
18	0.0542	0.0157	1.9	1.9	+1	1.8	4.5	-60	4.4	4.2	+4	1.6	1.3	+25
20	0.0655	0.0157	1.6	1.8	-12	1.5	4.6	-68	4.2	4.2	0	1.3	1.2	+8
23	0.0721	0.0492	7.4	6.5	+14	6.4	16.1	-60	12.9	12.2	+5	5.3	5.2	+2
48	0.0542	0.0872	26.0	26.1	0	23.8	60.3	-61	18.6	18.6	0	20.2	23.4	-14
96	0.0407	0.0960	45.4	48.3	-6	43.3	104.2	-58	16.7	17.4	-4	38.8	39.3	-1
час	***mTERF***	***LSP1***	ATP6/8			ND4			ND6			CYTB		
Frog2			mod	exp	dev	mod	exp	dev	mod	exp	dev	mod	exp	dev
0	0.0089	0.0041	1.0	1.0		1.0	1.0		1.0	1.0		1.0	1.0	
6	0.0045	0.0023	1.2	1.3	-5	1.2	1.4	-12	0.7	0.7	+6	1.2	1.2	+3
9	0.0073	0.0045	1.3	1.2	+8	1.3	1.6	-19	1.1	1.1	+1	1.3	1.3	-1
20	0.0157	0.0157	3.7	3.7	+1	3.7	3.7	0	2.8	2.8	-2	3.6	4.0	-11
30	0.0157	0.0230	7.1	8.1	-13	7.0	6.2	+14	3.7	3.7	0	6.8	8.1	-17
48	0.0407	0.1056	19.6	28.7	-32	19.1	17.7	+8	8.6	8.4	+2	17.3	23.1	-25
7 дней	0.0041	0.0073	6.6	8.5	-22	6.7	4.9	+36	2.4	2.3	+3	6.6	6.6	+1

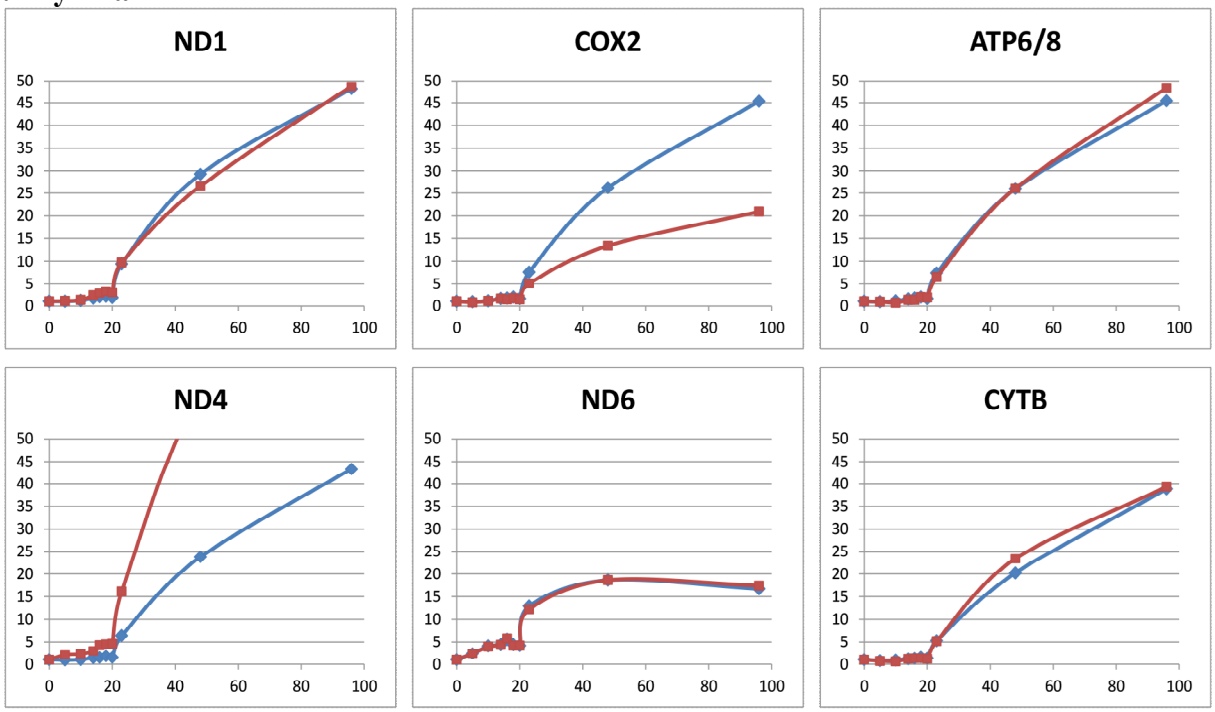
В опыте для крысы отдельно для каждого гена COX1, ATP6/8, COX3, ND4, ND5, CYTB рассматривались отношения количеств мРНК к количеству 16S рРНК. Каждое значение у гипотиреоида вычислялось в процентах от соответствующего отношения у эутиреоида, таблица 0.5. Итак, в опыте определялось отношение $u_j = \frac{(z_j{}^h t_j{}^h)(z_0{}^e t_0{}^e)}{(z_0{}^h t_0{}^h)(z_j{}^e t_j{}^e)}$, где z_j – уровень транскрипции *j*-го гена, кодирующего белок, у гипотиреоида (*h*) и эутиреоида (*e*) в зависимости от верхнего индекса, *j*=1–6, и z_0 – уровень транскрипции 16S рРНК, а t с индексами – соответствующие времена полураспада. Итак, сравнивалось значение $\frac{z_j{}^h z_0{}^e}{z_0{}^h z_j{}^e}$, вычисленное в модели, с опытным значением

$$u_j \cdot \frac{t_0{}^h t_j{}^e}{t_j{}^h t_0{}^e}. \qquad (1a)$$

При этом по отдельности числитель и знаменатель опытного отношения не известны ни у эутиреоида, ни у гипотиреоида.

Другие опытные данные о митохондриях приведены в пункте 3 введения.

лягушка 1



лягушка 2

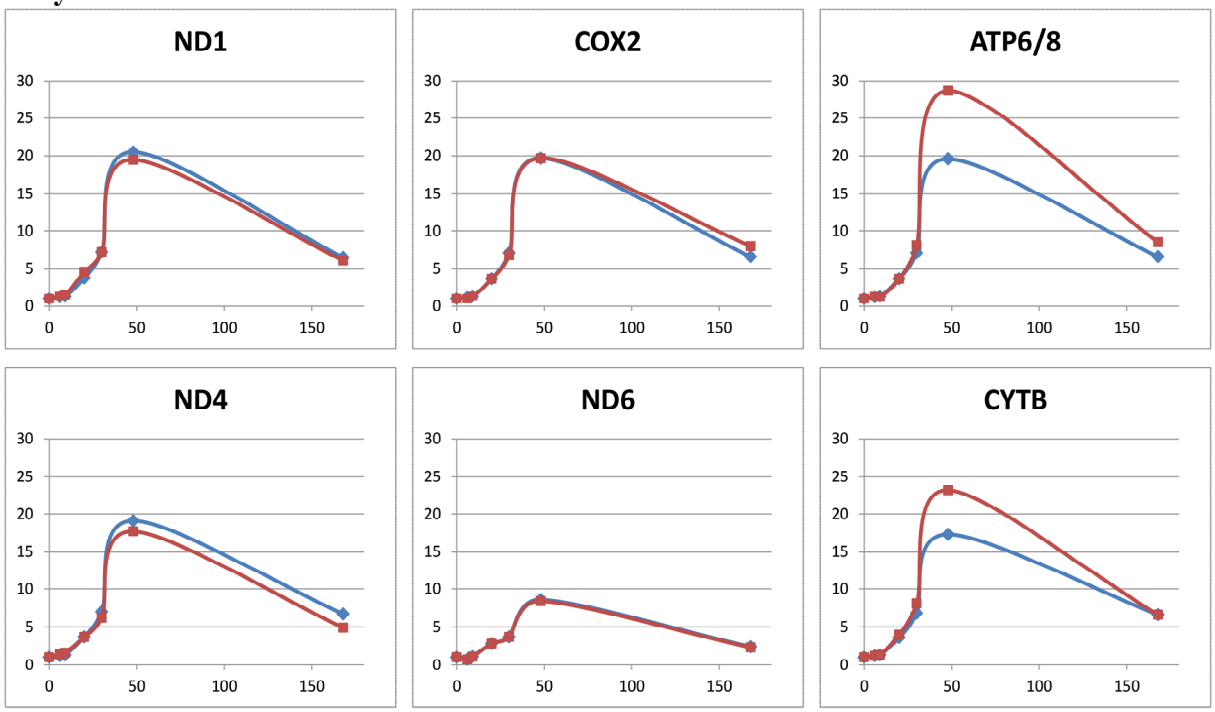


Рисунок 1.2. Графическое представление данных из таблицы 1.1
Показан уровень транскрипции мРНК относительно его значения в нулевой момент – время оплодотворения икры. По оси абсцисс отложено время в часах, по оси ординат – относительное число транскриптов. Все параметры модели, не зависящие от времени, одинаковы для всех лягушек и генов; интенсивности связывания меняются со временем. Линии, помеченные ■, относятся к модели, ♦ – к эксперименту.

Таблица 1.2. Число транскрипций в модели за 9 часов моделируемого времени для 48 часовых эмбрионов лягушек, а также для человека и крысы. Приведено среднее значение ± стандартное отклонение (по 1000 реализаций). В строке “Competition” указан процент РНК-полимераз, сорвавшихся с ДНК в результате столкновения с встречной полимеразой: для тяжёлой цепи (H-цепь) – на участке от начала сайта mTERF до конца tRNA-Thr, на лёгкой цепи (L-цепь) – от начала tRNA-Pro до конца tRNA-Gln.

	Xenopus laevis			*Homo sapiens*		*Rattus norvegicus*	
Ген	Frog 1	Frog 2	Frog 3	WT	MELAS	Эутиреоид	Гипотиреоид
H-цепь:							
tRNA-Phe	3074 ± 271	5003 ± 818	2835 ± 1345	123 ± 27	32 ± 24	1215 ± 51	556 ± 37
12S	3004 ± 276	4931 ± 831	2815 ± 1348	527 ± 24	438 ± 21	2362 ± 46	1090 ± 34
tRNA-Val	2998 ± 277	4926 ± 830	2813 ± 1348	527 ± 24	438 ± 21	2361 ± 46	1090 ± 34
16S	2865 ± 288	4799 ± 855	2779 ± 1351	527 ± 24	438 ± 21	2323 ± 46	1073 ± 35
tRNA-Leu	2012 ± 272	3981 ± 808	2495 ± 1346	22 ± 5	19 ± 5	257 ± 19	86 ± 9
ND1	1934 ± 274	3908 ± 819	2475 ± 1349	22 ± 5	19 ± 5	229 ± 19	75 ± 9
tRNA-Ile	1929 ± 275	3903 ± 820	2474 ± 1349	22 ± 5	19 ± 5	227 ± 19	74 ± 8
tRNA-Met	1920 ± 276	3895 ± 820	2471 ± 1349	22 ± 5	19 ± 5	223 ± 19	72 ± 8
ND2	1849 ± 283	3828 ± 833	2454 ± 1349	22 ± 5	19 ± 5	197 ± 18	62 ± 8
tRNA-Trp	1844 ± 282	3824 ± 834	2453 ± 1350	22 ± 5	19 ± 5	195 ± 18	62 ± 8
COX1	1745 ± 284	3720 ± 856	2426 ± 1352	22 ± 5	18 ± 5	156 ± 16	48 ± 7
tRNA-Asp	1737 ± 284	3713 ± 857	2424 ± 1352	22 ± 5	18 ± 5	153 ± 16	47 ± 7
COX2	1703 ± 285	3676 ± 861	2416 ± 1352	22 ± 5	18 ± 5	141 ± 15	42 ± 7
tRNA-Lys	1700 ± 286	3673 ± 861	2415 ± 1353	22 ± 5	18 ± 5	140 ± 15	42 ± 7
ATP6/8	1665 ± 289	3638 ± 866	2405 ± 1353	21 ± 5	18 ± 5	125 ± 15	37 ± 7
COX3	1634 ± 290	3604 ± 868	2397 ± 1354	21 ± 5	18 ± 5	114 ± 14	34 ± 6
tRNA-Gly	1631 ± 290	3601 ± 868	2396 ± 1354	21 ± 5	18 ± 5	113 ± 14	33 ± 6
ND3	1614 ± 293	3584 ± 869	2391 ± 1355	21 ± 5	18 ± 5	109 ± 14	32 ± 6
tRNA-Arg	1611 ± 293	3582 ± 869	2390 ± 1354	21 ± 5	18 ± 5	108 ± 14	31 ± 6
ND4	1524 ± 291	3487 ± 878	2364 ± 1357	21 ± 5	18 ± 5	87 ± 12	25 ± 5
tRNA-His	1520 ± 290	3481 ± 879	2363 ± 1357	21 ± 5	18 ± 5	86 ± 12	25 ± 5
tRNA-Ser2	1516 ± 291	3477 ± 880	2362 ± 1357	21 ± 5	18 ± 5	86 ± 13	24 ± 5
tRNA-Leu2	1510 ± 291	3472 ± 879	2361 ± 1358	21 ± 5	18 ± 5	85 ± 13	24 ± 5
ND5	1400 ± 288	3323 ± 892	2319 ± 1365	21 ± 5	18 ± 5	67 ± 10	19 ± 5
CYTB	1273 ± 282	3120 ± 886	2259 ± 1369	21 ± 5	18 ± 5	53 ± 10	15 ± 4
tRNA-Thr	1267 ± 281	3111 ± 887	2255 ± 1369	21 ± 5	18 ± 5	53 ± 10	15 ± 4
Competition	37%	22%	10%	5%	5%	79%	83%
L-цепь:							
tRNA-Pro	1605 ± 143	1585 ± 221	741 ± 151	35 ± 7	36 ± 5	1236 ± 34	1248 ± 40
tRNA-Glu	1505 ± 150	1421 ± 229	693 ± 152	35 ± 7	36 ± 5	1227 ± 34	1245 ± 40
ND6	1469 ± 153	1368 ± 233	677 ± 151	35 ± 7	36 ± 5	1222 ± 34	1244 ± 40
tRNA-Ser	1124 ± 150	972 ± 214	572 ± 144	34 ± 7	35 ± 5	1133 ± 36	1215 ± 40
tRNA-Tyr	1040 ± 151	883 ± 198	549 ± 144	34 ± 7	35 ± 5	1099 ± 37	1203 ± 41
tRNA-Cys	1036 ± 149	879 ± 196	548 ± 144	34 ± 7	35 ± 5	1098 ± 37	1202 ± 41
tRNA-Asn	1030 ± 149	871 ± 194	547 ± 144	34 ± 7	35 ± 5	1095 ± 37	1202 ± 41
tRNA-Ala	1026 ± 151	867 ± 194	546 ± 143	34 ± 7	35 ± 5	1094 ± 37	1201 ± 41
tRNA-Gln	941 ± 147	788 ± 184	525 ± 140	34 ± 7	35 ± 5	1062 ± 37	1188 ± 40
Competition	41%	50%	29%	3%	3%	14%	5%

Таблица 1.3. Результаты для человека, полученные в модели и в опыте: здоровый и с MELAS-болезнью. Все результаты приводятся для скорости РНК-полимеразы 500 нт/с и тех значений *p*, *q*, которые найдены для лягушек. Параметры решения выделены полужирным. Указаны относительные уровни транскрипции в модели и в опыте. Отличие опыта и модельного результата везде, кроме CYTB, в пределах опытной ошибки. Для здорового человека по сравнению с мутантом изменились: интенсивности, *HSP1* убывает в 7.75 раза, *mTERF* убывает в 1.21 раза, и уровни транскрипции генов, tRNA-Phe убывает в 3.8 раза, 12S и 16S убывают в 1.2 раза, tRNA-Leu и tRNA-Lys убывают в 1.2 раза.

Параметры решения для здорового человека						Уровень транскрипции относительно гена ND1 в модели (вверху) и в опыте (внизу). Для ND1 в опыте 1.00±0.04.						
LSP	***HSP1***	***HSP2***	***mTERF***	*R*	L_1n	ND2	COX1	COX2	ATP6/8	ND3	ND5	CYTB
0.0031	0.0031	0.0126	0.6456	23.955	1.945	1.00	1.00	1.00	0.96	0.96	0.96	0.96
В опыте для этих генов:						1.40 ±0.34	1.04 ±1.23	1.72 ±1.23	0.91 ±0.78	1.04 ±0.16	1.86 ±1.09	2.31 ±1.06
Отклонение модели от опыта в процентах:						-29	-4	-42	+5	-4	-48	-58
Параметры решения при MELAS-болезни						Изменение уровня транскрипции в модели						
						Phe	12S	Val	16S	Leu	Lys	CYTB
0.0031	0.0004	0.0126	0.5336	24.333		3.84	1.20	1.20	1.20	1.16	1.22	1.17

Отметим, что при сравнении уровней транскрипции генов в модели и опыте вопрос о выборе функционала не вполне ясен. Мы использовали естественный функционал:

$$L_1n = \sum_{ji} \frac{\left|x_{ji} - y_{ji}\right|}{\max\{x_{ji}, y_{ji}\}}, \qquad (2)$$

где x_{ji} и y_{ji} – сравниваемые наборы относительных уровней транскрипции соответственно в опыте и в модели, *j* пробегает имена рассматриваемых генов и *i* – рассматриваемые моменты времени. Если моменты времени отсутствуют, то индекс *i* опускается. В опыте рассматриваются три лягушки, и возникает вопрос о сравнении результатов сразу для всех них. Поэтому использовалось обобщение метрики (2):

$$L_1n(\text{total}) = \sum_{k=1}^{3}\left(\frac{1}{n_k \cdot s}\sum_{ji}\frac{\left|x_{ji} - y_{ji}\right|}{\max\{x_{ji}, y_{ji}\}}\right), \qquad (3)$$

где n_k – «размерность» данных, которыми мы располагаем для каждой из лягушек, $s = \sum_{k=1}^{3}\frac{1}{n_k}$. Эти размерности соответственно равны $n_1 = 54$ (девять моментов времени и шесть генов), $n_2 = 36$ (шесть моментов времени и шесть генов), $n_3 = 10$ (пять моментов времени и два гена).

Рассматривались также другие функционалы:

$$L_2 n = \sqrt{\sum_{ji}\left(\frac{x_{ji}-y_{ji}}{\max\{x_{ji},y_{ji}\}}\right)^2};\ L_1 = \sum_{ji}\left|x_{ji}-y_{ji}\right|;\ L_2 = \sqrt{\sum_{ji}\left(x_{ji}-y_{ji}\right)^2};$$

$$\cos\varphi = \frac{\sum_{ji}\left(x_{ji}y_{ji}\right)}{\|\bar{x}\|\cdot\|\bar{y}\|} \to \max\text{, где } \|\bar{x}\| = \sqrt{\sum_{ji}\left(x_{ji}\right)^2},\ \|\bar{y}\| = \sqrt{\sum_{ji}\left(y_{ji}\right)^2};$$

$$\tilde{\chi}^2 = \sum_{ji}\frac{\left(a_{ji}-b_{ji}\right)^2}{a_{ji}}\text{, где } a_{ji} = \frac{x_{ji}}{\sum_{ji}x_{ji}},\ b_{ji} = \frac{y_{ji}}{\sum_{ji}y_{ji}};$$

$$\chi^2 = \sum_{ji}\frac{\eta_{ji}\left(a_{ji}-b_{ji}\right)^2}{b_{ji}}\text{, где } \eta_{ji} = \exp\left(-\Delta x_{ji}/x_{ji}\right);$$

$$\tilde{S} = \sqrt{\sum_{ji}\left(\frac{x_{ji}}{a}+\frac{y_{ji}}{b}\right)^2}\text{, где } a = \sum_{ji}x_{ji},\ b = \sum_{ji}y_{ji};$$

$$S = \sqrt{\sum_{ji}\left(\frac{x_{ji}}{a}+\frac{y_{ji}}{b}\right)^2}\text{, где } a = \sum_{ji}\frac{x_{ji}}{\eta_{ji}},\ b = \sum_{ji}\frac{y_{ji}}{\eta_{ji}},\ \eta_{ji} = \exp\left(-\Delta x_{ij}/x_{ij}\right).$$

Все функционалы дали примерно одинаковые решения, поэтому приводятся результаты только для функционалов (2)–(3).

4.2. Данные о пластидах

В эксперименте [45] для каждого из трёх генов 1-го локуса подсчитывалось отношение MT/WT уровня транскрипции после нокаута *sig*4 (в числителе) к его уровню в диком типе (в знаменателе). В эксперименте [47] измерялось отношение уровней транскрипции каждого из генов 2-го локуса до и после теплового шока (21°C – нормальная температура и 40°C – температура шока). Таким образом, в этом случае место мутанта занимает клетка после теплового шока. Эти данные приведены в таблицах 1.4 и 1.5.

Отметим, что положения некоторых промоторов были определены нами на основе множественного выравнивания лидерных областей.

Для третьего локуса после нокаутов *sig3* или *sig3* у *Arabidopsis thaliana* происходили сложные изменения уровней транскрипции генов [45]; точнее, были экспериментально измерены отношения уровней транскрипции до и после нокаута, как и для первого локуса. Моделирование показало, что никакие значения интенсивностей связывания с промоторами не приводят к согласию с экспериментом. Это привело к мысли, что здесь действует какой-то неизвестный фактор. И действительно, модель предсказала два фактора прерывания элонгации – терминаторы, которые затем были подтверждены для

каждого из терминаторов выравниванием соответствующих участков ДНК, оказавшихся палиндромами с одинаковой длиной 44; на рисунке 1.1*c* они помечены буквами T1 и T2. Существование палиндрома T1 у небольшого числа видов отмечено в обзоре [65]. Обсуждение этих палиндромов приводится в следующем пункте. Каждый терминатор имеет свою условную вероятность терминации транскрипции, которые были определены при моделировании наряду с интенсивностями связывания промоторов.

Таблица 1.4. Сравнение изменений уровней транскрипции генов в эксперименте и в модели для локусов 1 и 2

Ген	**Эксперимент**	**Модель**
Локус 1 (*Arabidopsis thaliana*)		
ycf1	0.73 ± 0.04	0.76 ± 0.01
ndhF	0.43 ± 0.10	0.47 ± 0.19
*rpl*32	1.52 ± 0.06	1.55 ± 0.02
Локус 2 (*Hordeum vulgare*)		
*rpl*23– *rpl*2	2.15/2.69	2.64 ± 0.02
psbA	0.53/0.55	0.54 ± 0.04

Таблица 1.5. Сравнение изменений уровней транскрипции генов (в строках) в опытах по нокауту генов *sig3* и *sig4* и в модели для третьего локуса

Ген	**Нокаут** *sig3*	**Модель** (*sig3*)	**Нокаут** *sig4*	**Модель** (*sig4*)
psbB	1.02 ± 0.36	1.27 ± 0.12	0.69 ± 0.19	0.84 ± 0.11
psbT	0.98 ± 0.25	1.30 ± 0.12	0.96 ± 0.15	0.85 ± 0.11
psbN	0.49 ± 0.46	0.41 ± 0.12	1.03 ± 0.02	1.02 ± 0.19
psbH	1.31 ± 0.05	1.28 ± 0.12	1.01 ± 0.08	0.83 ± 0.11
petB	0.91 ± 0.15	1.09 ± 0.11	0.87 ± 0.29	0.83 ± 0.11
petD	0.92 ± 0.09	0.89 ± 0.10	0.81 ± 0.21	0.81 ± 0.11
rpoA	0.94 ± 0.14	0.82 ± 0.20	0.79 ± 0.11	1.01 ± 0.14
rps11	0.92 ± 0.33	0.90 ± 0.21	0.98 ± 0.31	1.01 ± 0.13
rpl36	0.88 ± 0.11	1.03 ± 0.21	1.54 ± 0.62	1.08 ± 0.18
rps8	1.11 ± 0.04	1.03 ± 0.21	0.83 ± 0.15	1.08 ± 0.18
rpl14	1.04 ± 0.15	1.03 ± 0.21	1.11 ± 0.02	1.08 ± 0.18
rpl16	1.09 ± 0.03	1.03 ± 0.21	1.18 ± 0.03	1.08 ± 0.18
rps3	1.24 ± 0.26	1.03 ± 0.21	1.25 ± 0.02	1.08 ± 0.18
rpl22	1.09 ± 0.13	1.03 ± 0.21	1.20 ± 0.12	1.08 ± 0.18
rps19	1.15 ± 0.50	1.03 ± 0.21	0.96 ± 0.07	1.08 ± 0.17
rpl2	0.94 ± 0.15	1.03 ± 0.21	0.95 ± 0.06	1.08 ± 0.17
rpl23	1.05 ± 0.04	1.06 ± 0.20	1.35 ± 0.33	1.10 ± 0.17

Другие экспериментальные данные о пластидах приведены в пункте 3 введения.

5. Оценка согласия с опытом

Подробно рассмотрим моделирование в случае митохондрий. Методика моделирования для пластид аналогичная. Распределения переменных u_j, t_0, t_j не известны из опыта. Это не позволяет оценить доверительный интервал опытных значений (1) и (1а) на основе теоретико-вероятностных методов, которые обычно применяются для сравнения предсказаний с опытными значениями. Однако вместо этого можно использовать абсолютные погрешности. Пусть Δ – абсолютная погрешность значения *b* для выражений (1) или (1а), а *a* — значение в модели для того же выражения, тогда можно проверить утверждение: «*a* принадлежит интервалу $b \pm \Delta$».

Погрешность Δ значения выражений (1) и (1а) тривиально оценивается с помощью одной из двух обычно используемых формул [66]. Первая – погрешность суммы равна $\Delta(x \pm y) = \sqrt{\Delta(x)^2 + \Delta(y)^2}$, если погрешности слагаемых статистически независимы; иначе $\Delta(x \pm y) \le \Delta(x) + \Delta(y)$. Вторая – погрешность произведения $x \cdot y$ или отношения x / y равна $\Delta(x \circ y) = (x \circ y) \cdot \sqrt{(\Delta(x)/x)^2 + (\Delta(y)/y)^2}$, если погрешности членов статистически независимы; иначе $\Delta(x \circ y) \le (x \circ y) \cdot (\Delta(x)/x + \Delta(y)/y)$, символ $\circ$ обозначает операцию умножения или деления. Таким образом, либо предполагается статистическая независимость и используются равенства, либо применяются неравенства, и тогда возникает неопределённость. К счастью, оба случая дают близкие результаты.

Предсказания модели укладываются в интервалы погрешностей $b \pm 1.3\Delta$ и $b \pm 2.4\Delta$ (таблицы 1.3, 1.6) в предположении, что погрешности составляющих статистически независимы.

Важно, что модельные и опытные решения, приведённые в таблицах 1.1, 1.3, 1.6, не значимо отличаются между собой и в ещё одном отношении. Согласие между результатами *a* и *b*, полученными соответственно в модели и в опыте, можно вычислять в процентах:

$$100 \cdot (a - b) / b \,. \qquad (4)$$

Эта величина имеет знак, указывающий на убывание или возрастание *a* по сравнению с *b*. В опытах по измерению уровней транскрипции значимым обычно считается различие более, чем в два раза, т.е. незначимым считается отклонения от –50% до +100%, [47]. Практически все модельные результаты в этом смысле не значимо отличаются от опытных результатов, таблицы 1.1, 1.3, 1.6.

Таблица 1.6. Результаты для крыс, полученные в модели и в опыте: эутиреоид и гипотиреоид. Параметры модели выделены полужирным. Все результаты приводятся для скорости РНК-полимеразы 500 нт/с и тех значений *p*, *q*, которые найдены для лягушек. Слева: значения параметров у эутиреоида (вверху) и у гипотиреоида (внизу). Справа: сравнение результатов модели (вверху) и опытных данных (внизу). Обозначение: *HSP = HSP1+HSP2*.

LSP	***HSP***	***mTERF***	*R*	L_1n	Отношение уровней транскрипции у гипотиреоида к эутиреоиду в модели (вверху) и в опыте (внизу, вычислено)					
					COX1	ATP6/8	COX3	ND4	ND5	CYTB
0.1056	0.0721	0.9453	30.605	1.736	0.666	0.641	0.646	0.622	0.614	0.613
0.1056	0.0336	0.9453	30.637		0.61 ±1.02	0.33 ±0.42	0.33 ±0.42	0.61 ±1.02	0.78 ±0.96	0.35 ±0.39
Отклонение модели от опыта в процентах:					+9	+94	+96	+2	-21	+75

6. Методика моделирования

6.1. Обоснование модели

Важно отметить: подстановка в нашу модель значений параметров, непосредственно найденных из опыта, и результат решения обратной задачи с помощью модели, приводят к одинаковым по порядку результатам. Конечно, первый подход применим в немногих случаях и обычно даёт более грубый результат по сравнению со вторым подходом. Поясним это на примере интенсивности связывания PEP в первом локусе пластид. Из приведённых выше опытных данных получены интенсивности связывания: с *ycf1*–33/34 равная 0.09 $с^{-1}$ и с *ndhF*–320 равная 0.15 $с^{-1}$. Решение в модели обратной задачи дали значения интенсивностей связывания: 0.037 $с^{-1}$ для первого промотора и 0.093 $с^{-1}$ для второго. Опытные значения близки по порядку к результатам решения обратной задачи.

В случае экспериментов с нокаутом σ-субъединицы РНК-полимеразы бактериального типа экспрессия других генов в ядре, по-видимому, не меняется, так как такая РНК-полимераза работает только в пластиде, где кодируется её кор-фермент. Концентрация кор-фермента не меняется, поскольку его мРНК транскрибируются РНК-полимеразами фагового типа, не зависящими от σ-субъединиц. В пластидах цветковых растений не кодируются никакие транскрипционные факторы. В опытах с тепловым шоком исследуются изолированные пластиды, когда кодируемые в ядре РНК-полимеразы фагового типа и транскрипционные факторы уже не поступают в пластиду и эксперимент идёт быстрее, чем разложения этих белков. MELAS-мутация происходит в митохондрии, где не кодируются ни РНК-полимеразы, ни факторы. При рассмотрении

митохондрии эмбрионов лягушки явным образом учитывается концентрация транскрипционного фактора mtTFA. В случае изменения концентрации гормонов щитовидной железы вопрос о его влиянии более сложен, мы исходим из простейшего предположения (об отсутствии влияния), которое уже даёт хорошее согласие с экспериментом.

Для контроля проверялось: при удалении данных об экспрессии одного из генов (если остаётся достаточно данных) решение меняется незначительно. Также строилось множество субоптимальных биологически осмысленных решений, при которых уровни транскрипции попадают в интервалы опытных значений. Всё это множество близко к указанному ниже решению (данные не приводятся). В общем виде исследование зависимости решения от исходных данных и их погрешностей не проводилось.

Как уже отмечалось, основным аргументом в пользу предложенной модели является правильность общих положений о протекании транскрипции в сочетании с тем, что модель согласуется практически со всеми биологическими данными о транскрипции в пластидах и митохондриях.

6.2. Случай митохондрий

Вообще говоря, скорость элонгации РНК-полимеразы фагового типа можно включить в число неизвестных параметров и подвергать варьированию. Из-за огромного объёма вычислений были проверены лишь два значения её скорости: 200 нт/с и 500 нт/с. Эта скорость вряд ли ниже 200 нт/с, но значение выше 500 нт/с возможно (см. пункт 3). В работе модель состоит в том, что скорость элонгации полимеразы фагового типа составляет 500 нт/с.

Обозначим *HSP* суммарную интенсивность попыток связывания с промоторами HSP1 и HSP2, т.е. *HSP=HSP1+HSP2*.

Численные данные об относительных уровнях РНК получены в опытах: для лягушек – из [27], для здорового человека в части рРНК – из [33] и более точные данные в части мРНК – из [34], для MELAS-мутации человека – из [31], для крыс – из [22]. Данные для удобства читателя воспроизведены в таблицах 1.1, 0.4 и 0.5.

Стабилизация всех уровней транскрипции генов в модели происходит за время, меньшее 9 часов моделируемого времени. В митохондриях клеток печени здорового человека это время превышает время полураспада каждой рРНК, [33] и каждой мРНК [34], оно превышает время полураспада этих РНК у обеих крыс [22] и продолжительность клеточного цикла у эмбрионов лягушек [67].

Для рассматриваемых организмов (лягушки, человека и крысы) принимались следующие *общие ограничения* на искомое решение.

1) Уровни транскрипции всех генов строго положительны. Это условие заменяется на более сильное: каждый ген транскрибируется не менее двух раз за время полураспада РНК, если оно известно.

2) Все параметры неотрицательны; p, q находятся в интервале от 0 до 1 и $q \le p$, так как в опыте твёрдо установлено, что протекание по лёгкой цепи значительно меньше, чем протекание по тяжёлой цепи, т.е. $0 < q \le p < 1$. Интенсивности *LSP* и *HSP* меняются в интервале от 0.002 до 0.1 ($с^{-1}$), интенсивность *mTERF* – от 0.002 до 1 ($с^{-1}$), так как вне этих ограничений в модели уровни транскрипции не стабилизируются за время, даже значительно превосходящее 9 часов. Заметим, что в модели при увеличении интенсивности связывания *mTERF* выше единицы уровни транскрипции всех генов не изменяются, так как в этом случае ещё бóльшая интенсивность попыток не приводит к большему числу успешных связываний.

Для трёх лягушек использовался функционал $L_1 n(\text{total})$ из (3). Искалось решение – точка его глобального минимума, по которой определялись параметры p и q. Так же были найдены значения параметров интенсивностей *mTERF* и *LSP1* связывания с соответствующими сайтами в каждый из 10 (1-я лягушка), 7 (2-я лягушка) и 6 (3-я лягушка) моментов времени. Поэтому всего функционал $L_1 n(\text{total})$ зависел от 48 переменных. Результаты минимизации этого функционала приведены в таблицах 1.1, 1.7.

Таблица 1.7. Характеристики протекания mTERF-терминатора по тяжёлой и лёгкой цепям. Приведены минимальные значения функционала $L_1 n$ для каждой из трёх лягушек (Frog1, Frog2, Frog3) и функционала $L_1 n(\text{total})$. Для сравнения приведены эти же уровни для скорости РНК-полимеразы 200 нт/с.

Скорость (нт/с)	p	q	$L_1 n$ **(Frog1)**	$L_1 n$ **(Frog2)**	$L_1 n$ **(Frog3)**	$L_1 n$ **(total)**
500	0.0164	0.0056	11.243	3.193	0.043	2.098
200	0.2165	0.0015	10.844	3.240	0.309	2.235

Полученные для лягушек значения p и q брались *без изменения для человека*, здорового и больного, а также *крыс*, эутиреоида и гипотиреоида, так как во всех перечисленных случаях белок mTERF и сайт его связывания на мтДНК высококонсервативны [32].

Для здорового человека рассматривался функционал $L_1 n$ из (2). По его точке глобального минимума определялись интенсивности попыток связывания РНК-полимераз с промоторами LSP, HSP1, HSP2 и регуляторного белка-терминатора mTERF с его сайтом.

Для человека кроме указанных выше *общих ограничений* принимались *специальные ограничения*.

3) Отличие уровней транскрипции генов tRNA-Leu и tRNA-Lyz между больным и здоровым человеком составляет не более 20% у первого и не более 50% у второго [31].

4) $LSP = LSP_-$ и $HSP2 = HSP2_-$ (где LSP_- и $HSP2_-$ – интенсивности связывания промоторов *LSP* и *HSP2* в присутствии MELAS-мутации), поскольку MELAS-мутация не приводит к существенным изменениям уровней инициации транскрипции с этих промоторов. Однако известно её влияние для промотора HSP1, [19].

5) $mTERF > mTERF_-$ и $HSP1 > HSP1_-$ (где $mTERF_-$ и $HSP1_-$ определяются аналогично обозначениям выше), так как при кооперативном связывании белка mTERF, играющего роли терминатора и одновременно активатора промотора HSP1, после мутации оба сайта уменьшают эффективность. Действительно, абсолютная величина энергии связи комплекса терминатора mTERF·ДНК монотонно возрастает с ростом времени полураспада этого комплекса, которое после мутации уменьшается в 7–10 раз. Эта абсолютная величина монотонно возрастает с ростом интенсивности связывания терминатора с сайтом [31]. К сожалению, мы не знаем явного вида этих зависимостей.

6) $1.16 < \text{РНК}/\text{РНК}_- < 1.22$, где в числителе для здорового человека подсчитывается сумма по всем генам (для которых известны времена полураспада, таких оказалось восемь) слагаемых вида $l \cdot t \cdot z$ (длина гена из таблицы 0.4, умноженная на время полураспада соответствующей РНК и умноженная на уровень транскрипции гена), а в знаменателе – аналогичная сумма для больного человека. Из опыта известны отношения количеств суммарной РНК к суммарной ДНК, [31]. Отсюда вычисляются нижняя и верхняя оценки в условии 6, при этом количества ДНК, используемые для нормировки, сокращаются. Времена полураспада РНК для больного человека не известны, и мы принимали для них те же значения, что и для здорового человека; этот вопрос обсуждается в пункте 10 («Обсуждение результатов о митохондриях»).

7) Отношение R уровней транскрипции гена 12S к гену COX2 больше 16.9. Действительно, в [33] количества РНК генов 12S и COX2 равны 12600 и 225, табл. 2 в [33], а верхняя и, соответственно, нижняя границы времён полураспада этих генов равны 146 и 44 минуты, табл. 1, эксперимент 4 в [33], что даёт нижнюю границу для отношения $R>17$. Верхняя граница для R получается, если в качестве времён полураспада взять данные из табл. 1, эксперимент 3 в [33], тогда аналогично получим $R<27$. Ген COX2 был выбран для сравнения, так как время полураспада соответствующей мРНК устойчиво в различных опытах, включая воздействия антибиотиков [34]. Аналогичные вычисления

для других генов дают следующие границы: 19<*R*<37 для ATP6/8, 17<*R*<27 для COX2, 20<*R*<25 для COX3, 17<*R*<25 для CYTB, 41<*R*<52 для COX1. Поэтому мы принимаем границы 17<*R*<37. Только для COX1 наблюдаются другие границы, что, вероятно, связано с экспериментальной ошибкой, так как в области этих генов, соседних и на одной цепи ДНК, нет промоторов и терминаторов. Поэтому трудно представить себе механизм, который бы выделял среди них ген COX1.

Итак, здесь минимизировался функционал $L_1 n$ с восемью переменными и вышеперечисленными ограничениями. Результаты приведены в таблице 1.3.

Для крысы минимизировался тот же функционал $L_1 n$ с шестью переменными: параметрами *LSP*, *HSP=HSP1+HSP2*, *mTERF* соответственно для эу- и гипотиреоидов. Кроме общих принимались следующие *специальные ограничения*.

8) $LSP = LSP_-$ и $HSP2 = HSP2_-$ (обозначения такие же, как выше), равенства можно объяснить малым изменением метилирования в областях соответствующих промоторов [22].

9) Для эутиреоида указанная выше величина *R* лежит в пределах 17<*R*<60. Нижняя граница принята равной таковой у здорового человека, верхняя – у мыши [68].

7. Компьютерная реализация модели

Для полноты описания приведём краткие сведения о компьютерной реализации нашей модели, полученной Л.И. Рубановым. Модель реализована в виде программы на языке C++ в двух вариантах (с интерфейсом командной строки и с графическим пользовательским интерфейсом), которые доступны для загрузки на условиях открытого лицензионного соглашения GNU GPL v3, [69]. В основном программа описана в [55]. Она реализует автомат, управляемый событиями и осуществляющий имитационное моделирование большой совокупности взаимодействующих стохастических и детерминированных процессов, развивающихся в моделируемом времени на фиксированном локусе ДНК. Связывание РНК-полимеразы с каждым промотором моделируется стохастическим процессом. После связывания элонгация РНК-полимеразы моделируется детерминированным процессом. В модели происходят многочисленные коллизии: (i) РНК-полимераза или регуляторный белок пытаются связаться с сайтом, который хотя бы частично занят; (ii) вторичная структура пытается образоваться на сайте, который частично занят; (iii) две *встречные* РНК-полимеразы пытаются занять один и тот же нуклеотид. Сценарии разрешения таких коллизий и вероятностные характеристики исходов являются параметрами программы, которые пользователь может легко задавать и варьировать.

События в модели обрабатываются в хронологическом порядке, для чего все возможные события выстраиваются в сложно организованную систему частично упорядоченных очередей. Быстродействие программы в значительной степени определяется скоростью обслуживания этих очередей.

В пластидах изучалось взаимодействие РНК-полимераз в пределах коротких локусов, вырезанных из длинного пластома (например, локусов с длинами от 4321 до 16583 п.н.). В митохондриях моделировалась транскрипция на всей митохондриальной ДНК длиной до 18 т.п.н. Здесь возникает принципиально новое явление: РНК-полимеразы могут не покидать локус и транскрибировать кольцевую ДНК несколько раз, продолжая элонгацию вплоть до возникновения коллизии. Это приводит к значительному росту числа одновременно моделируемых процессов.

Другой существенный аспект моделирования в митохондриях – транскрипцию осуществляют только РНК-полимеразы фагового типа, скорость элонгации которых экспериментально неизвестна, но, по-видимому, выше, чем у полимераз бактериального типа. В пластидах транскрипцию осуществляют РНК-полимеразы фагового и бактериального типов, но это обстоятельство не играет заметной роли, так как более быстрая полимераза всё равно не может обогнать более медленную и не влияет на неё. Моделирование в случае митохондрий проводилось для скоростей элонгации на порядок более высоких, чем в случае пластид. Это привело к увеличению частоты обращения к очереди событий и росту её длины.

В пластидах принималось, что столкновение РНК-полимеразы с любым белковым фактором или вторичной структурой безусловно приводило к терминации транскрипции. При моделировании в митохондриях рассматривается новый класс объектов – белковые терминаторы с протеканием в обе стороны. Характеристики протекания имеют смысл вероятности и являются параметрами терминатора.

Программная реализация моделирует одиночную траекторию в пространстве возможных событий, вдоль которой вычисляются уровни транскрипции всех генов. При одних и тех же параметрах модели выполняется усреднение этих уровней по многим траекториям. Вычисления могут эффективно выполняться параллельно на высокопроизводительном кластере, поддерживающем среду MPI. Приведённые ниже результаты получены на кластере MVS-100K в Межведомственном Суперкомпьютерном Центре РАН, [70] с использованием 2048 процессоров.

Обратная задача решалась методом многокритериальной оптимизации. Поверхность отклика, например для функционала (3), имеет сложную форму с ярко выраженными «водоразделами» и многочисленными локальными минимумами, что не позволя-

ет применить обычные методы локальной оптимизации, скажем на основе метода градиентного спуска. В таких ситуациях для сокращения перебора обычно используются эвристики. Особенность нашей задачи позволила применить следующую эффективную процедуру.

Промоторы на обеих цепях рассматриваемых мтДНК сосредоточены в компактной области, не содержащей генов кроме tRNA-Phe у человека и крысы. Поэтому из этой области выходит два встречных потока РНК-полимераз, которые конкурируют в основном вне этой области, т.е. там, где расположены практически все гены. Если какой-то ген, например на тяжёлой цепи, не транскрибируется, то это значит, что поток полимераз по лёгкой цепи слишком сильный и полностью блокирует поток, инициируемый промоторами на тяжёлой цепи. Поскольку общие условия требуют, чтобы все гены имели ненулевой уровень транскрипции, не имеет смысла дальнейшее увеличение интенсивности связывания промоторов на лёгкой цепи, так как тогда заведомо не найдется подходящего решения. Это позволяет сильно ограничить перебор интенсивностей связывания промоторов. А именно, для каждого фиксированного набора значений прочих параметров интенсивности связывания с промоторами варьируются в каждую сторону лишь до тех пор, пока полностью не прекратится транскрипция какого-либо гена. Эту стратегию оптимизации можно назвать «активным поиском».

8. Результаты о митохондриях

Для скорости РНК-полимеразы 500 нт/с модель дала по одному решению для каждого организма – трёх лягушек, человека (здорового и больного), и крысы (эу- и гипотиреоидной).

Для лягушек уровни протекания mTERF-терминатора (т.е. доля РНК-полимераз, проходящих через связанный mTERF) оказались следующими: p=0.0164 по тяжёлой цепи и q=0.0056 по лёгкой цепи, указывая на трёхкратную поляризацию терминатора. В остальных случаях они брались фиксированными: такими же, как у лягушек.

Для лягушек интенсивности связывания с промотором LSP1 в основном возрастают со временем, таблица 1.1 и рисунок 1.3. Согласие результатов модели и опытных данных по уровням транскрипции во всех случаях очень высокое: лишь по формуле (4) для 1-й лягушки и генов COX2 и ND4 имеется значимое различие модельного и опытного значений. А именно, только для момента времени 96 часов и гена COX2 различие несколько превышает 100% (в сторону увеличения); для гена ND4 во все моменты времени – несколько превышает –50% (в сторону уменьшения), таблица 1.1.

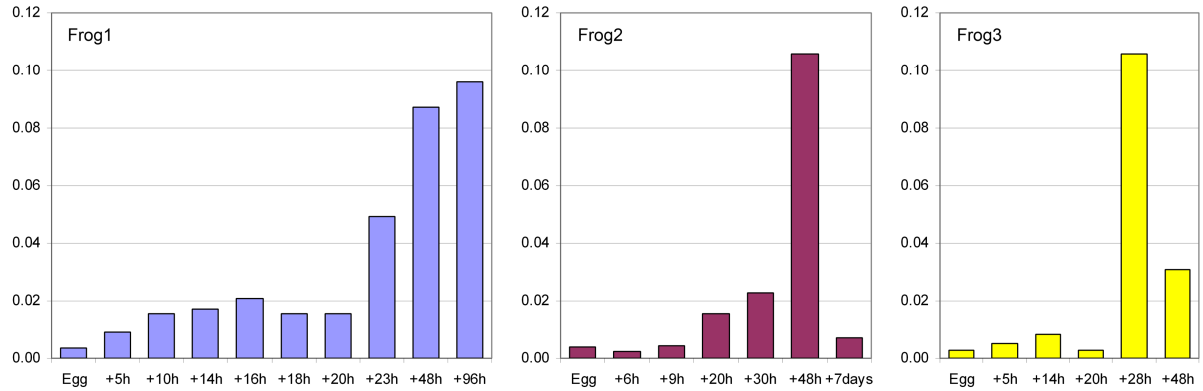


Рисунок 1.3. Графики зависимости от времени интенсивностей связывания промотора LSP1 у эмбрионов трёх лягушек

У здорового человека интенсивности связывания с сайтами *LSP*, *HSP1*, *HSP2* и *mTERF* соответственно равны 0.0031, 0.0031, 0.0126, 0.6456, таблица 1.3. У человека с MELAS-болезнью интенсивности *HSP1* и *mTERF* падают соответственно до 0.0004 и 0.5336, уменьшаясь, таким образом, в 7.75 и 1.21 раза. Отношение *R* уровней транскрипции гена 12S к гену COX2 равно 24, отношение РНК/РНК$_-$ взвешенных суммарных количеств РНК у здорового и больного человека равно 1.18. Уровни транскрипции tRNA-Phe и 16S рРНК падают у больного человека соответственно в 3.84 и 1.2 раза. Уровни транскрипции генов tRNA-Leu и tRNA-Lys уменьшаются в 1.2 раза, т.е. в пределах отклонений, известных из опыта. Минимум функционала при всех ограничениях отличается на 2.4% от его минимума только при общих ограничениях.

Согласование результатов модели и опытных данных по уровням транскрипции у здорового человека находится в пределах опытной погрешности для всех генов, кроме *CYTB*, для которого погрешность, полученная в опыте, превышена в модели на 29%. По формуле (4) для здорового человека это различие только для гена *CYTB* незначительно превышает –50% (в сторону уменьшения), таблица 1.3. Мы вернёмся к случаю *CYTB* в пункте 10.

У крысы интенсивности связывания с сайтом *LSP*, сумма *HSP=HSP1+HSP2* и с сайтом *mTERF* соответственно равны 0.1056, 0.0721 и 0.9453 у эутиреоида; у гипотиреоида *HSP* убывает до 0.0336, таблица 1.6. Отношение *R* уровней транскрипции гена 12S к гену COX2 равно 30.605 у эутиреоида и незначительно возрастает до 30.637 у гипотиреоида. Согласование результатов модели и опытных данных по уровням транскрипции для пары эутиреоид–гипотиреоид находится в пределах опытной погрешности. По формуле (4) различие незначимое, таблица 1.6.

Подробно решения и сравнение полученных результатов с опытными данными приведены в таблицах 1.1–1.3, 1.6. Ещё раз заметим, что практически все результаты находятся в пределах опытной погрешности.

9. Результаты о пластидах

Для 1-го локуса согласие с экспериментом, т.е. отношение sig4/WT уровня транскрипции после нокаута по *sig4* (в числителе) к его уровню в диком типе (в знаменателе), а также отношения уровней транскрипции с разных промоторов хорошо воспроизводятся в нашей модели при следующих численных значениях интенсивностей связывания холофермента или полимеразы фагового типа с промоторами: N1=0.003, N2=0.054, P1=0.010/0.037, P2=0.050/0.093. Здесь и далее *обозначение интенсивности* связывания совпадает с обозначением промотора, а для PEP-промоторов указывается два значения: первое – значение интенсивности связывания холофермента в случае *нокаута*, второе – в случае *дикого типа*. Размерность каждой интенсивности связывания – обратные секунды. В таблице 1.4 для этого локуса при нокауте *sig4* сравниваются изменения уровней транскрипции генов в эксперименте и в модели. Отношения уровней транскрипции гена *ycf1* с промоторов N2 и P1 в опыте и модели равно 1.7. Такое отношение не оценивалось в модели для промоторов N1 и N2, однако интенсивность связывания в опыте с N1 в 20 раз меньше, чем с N2, что согласуется с найденным решением. При нокауте RpoTp (когда N2=0) и P1=0.12 модель предсказывает 2.5-кратное увеличение эффективности PEP-промотора *ycf1*–33/34 по сравнению с диким типом, что хорошо согласуется с данными хроматограмм [48]. Теоретически это может быть оправдано отсутствием после нокаута нелинейного взаимодействия, основанного на одно- и трёхмерной диффузии между РНК-полимеразами, инициировавшими (до нокаута) транскрипцию с NEP-промотора N2 и PEP-промотора P1. Значения, полученные в модели, не выходят за пределы опытной погрешности (таблица 1.4), что свидетельствует о хорошем согласии с биологической реальностью.

Для 2-го локуса модель дала хорошее согласие с экспериментальными данными при значениях интенсивностей связывания холофермента P0=0.2, P1=0.9, P2=0.3, P3=0.1. В таблице 1.4 для этого локуса при увеличении температуры сравниваются изменения уровней транскрипции генов в эксперименте и модели. Видно, что разницы находятся в пределах опытных погрешностей. Уровень экспрессии гена *rps16* (вторая область) возрастает после теплового шока как в опыте, так и в модели. Прирост экспрессии несколько выше, чем предсказанный в [48].

Для 3-го локуса моделирование показало, что никакие значения интенсивностей связывания с промоторами не приводят к согласию с экспериментом без дополнительного предположения о структуре локуса, которое, однако, получило подтверждение (см. пункт 11). С помощью модели были исследованы разные гипотезы о терминации транскрипции различными факторами, которые могут присутствовать в локусе, включая крест-шпильки. Наилучшее согласование результатов моделирования с наблюдаемыми в опыте значениями уровней транскрипции генов было достигнуто в присутствии двух гипотетических терминаторов. Таким образом, модель предсказала два фактора терминации транскрипции – терминаторы, которые были подтверждены анализом выравнивания соответствующих участков ДНК. Эти терминаторы обозначены как T1 и T2 на рисунке 1.1*c* и представляют собой палиндромы длиной 44 нт, рисунок 1.4.

T1 TTAACGTAATCAGCCTCCAAATATTTGGAGGCTGATTACGTTAA

T2 GTATCTAGGGAGTAGTCATTTCCAAATGAATTCTCCCTAGATAC

Рисунок 1.4. Два потенциальных терминатора T1 и T2 в локусе 3
Комплементарные нуклеотиды выделены одиночным и двойным подчёркиванием.
При одинаковой длине в 44 нуклеотида терминаторы существенно различаются по составу.
Терминатор T2 – несовершенный палиндром с тремя не комплементарными парами.

Консервативность палиндрома T1 и его роль описана в обзоре [65] для небольшого числа других видов. Каждый терминатор – T1 и T2 – характеризуется собственной вероятностью (также обозначаемой T1 и T2) терминации транскрипции, которые были определены при моделировании наряду с интенсивностями связывания с промоторами. Эти вероятности оказались равны: T1=0.25, T2=0.25. Были предсказаны следующие значения интенсивности связывания: P1=0.555/0.867/1.355 (для нокаута *sig3*, *sig4* и дикого типа) , P2=0.075/0.227/0.284, P3=0.116/0.146/0.182, N=0.116. Полученные в модели отношения уровней транскрипции хорошо согласуются с результатами опытов с нокаутом *sig3*, а также независимого исследования нокаута *sig4* в *Arabidopsis thaliana*, [45], где уровни транскрипции измерялись до и после нокаута, как в локусе 1. В таблице 1.5 для этого локуса сравниваются изменения при нокауте *sig3* и *sig4* уровней транскрипции генов в эксперименте и в модели; видно, что данные близки. В частности, уровень транскрипции гена *psbB* – около 417 транскрипций в час (выше, чем у других генов), что хорошо согласуется с тем, что он кодирует основной апопротеин второй фотосистемы и, следовательно, должен интенсивно транскрибироваться.

10. Обсуждение результатов о митохондриях

Высокая степень эволюционной консервативности белка mTERF и сайта его связывания [32] позволяет переносить оценки параметров *p* и *q* (вероятностей протекания mTERF-терминатора в обе стороны) на многие другие виды, по крайней мере на хордовые. Однако другие параметры приходится переносить с осторожностью. Например, у мыши известен терминатор D-TERM в 5'-лидерной области гена tRNA-Phe между промоторами LSP и HSP1, [68], не описанный у крысы и человека. Даже у близких видов участки ДНК в области этого терминатора не выравниваются, рисунок 1.5.

Mus musculus	ACCAAAACTCTAATCATACTCTATTACGCAATAAACATTAACAA	16299
Rattus norvegicus	GCCTACCCT---CAGAAAATTCCACATACACCAAA---------	16313
Homo sapiens	GCTAACCCCATACCCCGAACCAACCAAACCCCAAAGACA-----	577

Рисунок 1.5. 5'-Лидерные области гена tRNA-Phe
В первой строке подчёркнут специфический терминатор D-TERM. Указаны координаты правых концов последовательностей ДНК, которые относительно tRNA-Phe имеют координату –1. Области не выравниваются.

Предсказания нашей модели хорошо согласуются с известными из опытов уровнями транскрипции в митохондриях лягушек, человека и крысы, включая болезненные состояния, такие как MELAS-мутацию у человека и понижение уровня гормона щитовидной железы у крысы.

В модели предсказаны значения интенсивностей связывания РНК-полимераз с промоторами, характеристики mTERF-терминатора транскрипции и абсолютные значения (таблица 1.2) уровней транскрипции для всех митохондриальных генов, в то время как из опыта известны только относительные значения уровней транскрипции для части этих генов. В пользу предложенной модели говорит и то, что все терминаторы важны для предсказания транскрипции. Исключение любого терминатора из моделирования приводит к неадекватным предсказаниям, например к транскрипции генов лишь на одной цепи ДНК.

Модель даёт лучшее согласие с опытом при скорости РНК-полимеразы 500 нт/с, чем при 200 нт/с; в частности, в этом случае практически все отклонения результатов модели от опытных значений попадают в пределы опытной погрешности. Вопрос об оценке скорости полимеразы, при которой достигается наилучшее согласие с опытом, требует дальнейших исследований как на основе нашей модели, так и опытным путём.

Из опыта известен монотонный рост концентрации транскрипционного фактора mtTFA на ранних этапах развития эмбриона лягушки [26]. Этот фактор является активатором для всех промоторов, поэтому следует ожидать роста интенсивностей связывания

со всеми промоторами, что мы и наблюдаем в модели с высокой точностью, столбец LSP1 в таблице 1.1. Однако нет опытных оснований, и мы не наблюдаем подобной монотонности для параметра *mTERF*.

В литературе имеется биоинформатическое предсказание: у млекопитающих интенсивность *HSP1* промотора на тяжёлой цепи значительно превосходит интенсивность *HSP2* промотора на той же цепи [13]. Это не соответствует результатам в нашей модели: у нас *HSP2* в 4 раза больше, чем *HSP1*.

У человека полученные абсолютные значения уровней транскрипции белок-кодирующих генов оказались неожиданно маленькими. Транскрипция происходит примерно раз в 15–26 минут в зависимости от гена, таблица 1.2, тогда как время транскрипции всей цепи ДНК составляет 33 секунды при скорости элонгации РНК-полимеразы 500 нт/с. Столь редкая транскрипция качественно согласуется с оценками абсолютных количеств мРНК, полученными в [33].

Более того, у человека в модели вдоль лёгкой и тяжёлой цепей уровни транскрипции генов, кодирующих белки, практически не меняются. Это показывает, что РНК-полимеразы с промоторов HSP1 и HSP2, которые прошли через mTERF-терминатор, практически не испытывают столкновений. То же самое верно для полимераз с промотора LSP, которые прошли первый терминатор, обусловленный G-квадруплексом (тетрамером). Иными словами, эти полимеразы почти не ощущают встречных потоков. Практически отсутствует конкуренция между РНК-полимеразами, связавшимися и свободными, за доступ к промотору: при высокой скорости полимеразы и малых интенсивностях сайт любого промотора освобождается задолго до того, как случается следующая попытка связывания с ним. Вероятность перекрытия промотора РНК-полимеразой, пошедшей на второй круг, также невелика: за 9 часов моделируемого времени только 1 ± 1 РНК-полимераз идут на второй круг по лёгкой цепи и 23 ± 6 – по тяжёлой цепи (как обычно, здесь указаны среднее значение ± несмещенная оценка среднеквадратичного отклонения при $n = 1000$ траекторий).

С эволюционной точки зрения такой низкий уровень конкуренции РНК-полимераз в митохондриях человека мог бы оправдываться тем, что при их столкновении может повреждаться ДНК. Согласно оценкам из [55], в пластидах столкновения РНК-полимераз бактериального типа происходят гораздо чаще и скорость таких полимераз значительно ниже. Это можно связать с тем, что изначально митохондрии, произошедшие от α-протеобактерий, имели РНК-полимеразы бактериального типа, которые были утрачены позже [71], а скорость РНК-полимераз фагового типа значительно

выше. Поэтому предпочтение РНК-полимераз фагового типа может быть связано с уменьшением риска разрыва ДНК при столкновениях РНК-полимераз.

Однако у лягушки и крысы конкуренция РНК-полимераз носит более заметный характер (см. таблицу 1.2), что может быть связано с существенными различиями во временах полураспада соответствующих РНК.

Влияние MELAS-мутации проявляется в модели заметным, в 1.21 раза уменьшением интенсивности связывания терминатора mTERF с его сайтом на ДНК и значительным, в 7.75 раза уменьшением интенсивности промотора HSP1. Это сопровождается уменьшением уровней транскрипции tRNA-Phe в 3.84 раза и рРНК в 1.2 раза. Каков механизм влияния MELAS-мутации на фенотип? Можно думать о двух факторах: уменьшение как уровня транскрипции рРНК, так и уровня фенилаланиновой тРНК.

Сначала обсудим возможный механизм влияния *первого фактора*. Для интенсивно экспрессируемых генов рибосомы плотно заполняют мРНК, предотвращая образование вторичной структуры на ней и, тем самым, защищая её от внешних факторов, таких как разрезание и модификация. Изменение уровней транскрипции рРНК, которое предсказывается нашей моделью, может приводить к возникновению открытых окон на мРНК, что, в свою очередь, приводит к разрушению мРНК и существенному уменьшению количества белков.

Элементарно-вероятностные соображения приводят к следующей формуле для времени τ полураспада любой РНК:

$$\tau = \frac{1}{\mu}(1 + d\lambda)\exp(w\lambda)\ln 2, \qquad (5)$$

где $\lambda = \dfrac{\nu N}{1+\alpha N}$ – интенсивность попыток связывания рибосомы с её сайтом связывания, α – параметр в этой зависимости Микаэлиса – Ментен (насыщение по λ происходит при большом N и равно $\dfrac{\nu}{\alpha}$) и ν – удельная интенсивность при малых N, где N – количество рибосом в митохондрии здорового человека. Далее, w – отношение линейного размера h РНКазы вдоль РНК к скорости V элонгации рибосомы (V =15 кодонов в секунду, $h = Vw = 15w$), d – отношение размера h_1 рибосомы вдоль РНК к той же скорости V элонгации рибосомы (h_1=10 кодонов, $h_1 = Vd$), μ – интенсивность взаимодействия РНКазы с определённым сайтом на мРНК, приводящего к распаду РНК. Здесь в качестве причины распада рассматривается только действие РНКазы, хотя аналогично можно рассмотреть действие и других факторов. Формулы (5)–(7) остаются верными и для пластид, бактерий, архей.

Только ν, α и μ зависят от последовательности РНК, N зависит от экспрессии других генов и, в особенности, от рибосомных генов.

У больного человека время полураспада τ' аналогично выражается через N' – количество рибосом в его митохондрии. Отсюда:

$$\frac{1+d\lambda}{1+d\lambda'}\exp[w(\lambda-\lambda')]=\frac{\tau}{\tau'}. \qquad (6)$$

Из опыта мы ожидаем, что τ/τ' находится в интервале от 1.5 до 3, [31]. Модель позволяет получить значения N и N' как абсолютные количества 12S или 16S рРНК, а w можно оценить в пределах от 2/15 до 4/3 секунд; ν и α не известны и зависят от сайта связывания рибосомы, т.е. в общем случае имеют вид ν_j и α_j, где j пробегает различные РНК. Можно составить систему из уравнений (5)–(6) для разных РНК и выразить ν_j, α_j через w.

Для нас принципиально, что формулы (5)–(6) показывают: маленькое изменение абсолютного количества N рибосом очень значительно меняет время полураспада РНК, а следовательно, – количество соответствующего белка. Это может служить одним из объяснений резкого изменения фенотипа при MELAS-мутации. А именно, хотя в условии (6) мы предполагали, что времена полураспада рРНК и большинства мРНК у здорового и больного человека близки, что подтверждается найденным решением, для объяснения фенотипа больного человека достаточно предполагать, что лишь у одной (возможно, короткой) мРНК время полураспада значительно уменьшается. Тогда отношение в условии (6) мало меняется и остаётся в указанных там пределах, тогда как время полураспада одной (или немногих) мРНК резко изменяется.

Назовём *окном* участок между соседними рибосомами, связанными с РНК, шириной не менее $h=Vw$ – линейного вдоль РНК размера РНКазы. Интенсивность распада любой мРНК в результате взаимодействия с РНКазой равна

$$\frac{\mu}{1+d\lambda}\cdot\exp(-\lambda w). \qquad (7)$$

Формулы (5)–(7) выводится в [72].

Коснёмся возможного механизма влияния *второго фактора* на фенотип больного человека: уменьшение уровня фенилаланиновой тРНК уменьшает экспрессию белок-кодирующих генов и одновременно увеличивает ширину окна между соседними рибосомами на полисоме.

Превышение уровня транскрипции CYTB над уровнями вышележащих генов, которое заявляется в опытных данных, не может быть получено в рамках текущей модели. Однако в опытах с блокированием рибосомы время полураспада мРНК CYTB не

устойчиво [34]. Это позволяет предположить систематическую ошибку при определении этого времени и, следовательно, уровня транскрипции CYTB. Небольшое расхождение (около 6%) между результатами модели и опыта у человека для гена ND2 можно объяснить той же неустойчивостью в определении времени полураспада [34].

В результате моделирования соответствующие эу- и гипотиреоидной крысам интенсивности связывания mTERF, а также промотора LSP оказались равными. В эксперименте метилирование сайта связывания mTERF не изменяется, а промотора LSP – изменяется незначительно, рисунок 1.6*a*.

В модели суммарная интенсивность HSP=HSP1+HSP2 инициации транскрипции с промоторов HSP1 и HSP2 в 2.15 раза меньше у гипотиреоида, чем у эутиреоида, т.е. меняется существенно. В опыте: метилирование области HSP1 меняется существенно, а HSP2 – незначительно, рисунок 1.6*b*. Таким образом, в обоих случаях изменение метилирование согласуется с изменением интенсивностей связывания.

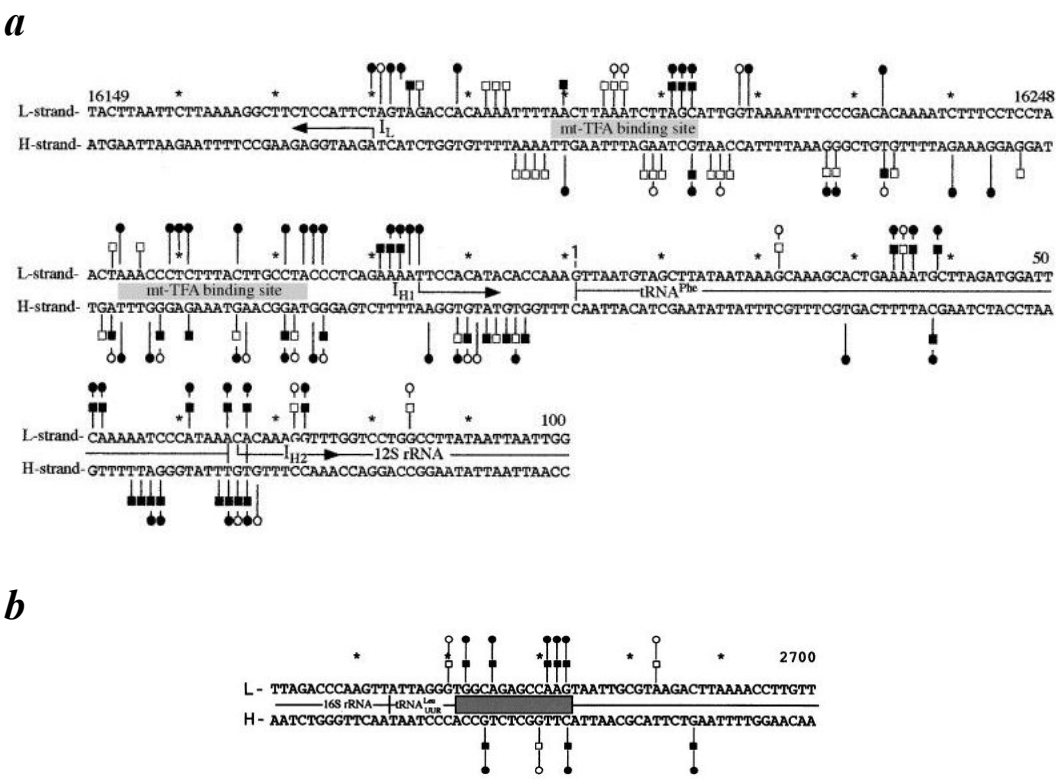


Рисунок 1.6. Сравнение метилирования у гипо- и эутиреоидной крыс.
Данные взяты из [22]. Показаны два локуса митохондриальной ДНК: a – область инициации транскрипции; b –область терминации транскрипции на mTERF, темно-серым показан сайт связывания mTERF. Промоторы: IL – LSP, IH1 – HSP1, IH2 – HSP2. Гиперметилированные позиции показаны светлыми знаками, защищённые позиции – чёрными знаками, квадратик для эутиреоида и кружок для гипотиреоида.

11. Обсуждение результатов о пластидах

Важно заметить, что для всех локусов различие между результатами экспериментов по нокауту и тепловому шоку и соответствующими результатами нашей модели не превышает разброса результатов самого эксперимента (таблицы 1.4, 1.5), также как и с данными хроматограмм из [48]. Это говорит о том, что интенсивности связывания РНК-полимераз с промоторами, предсказанные моделью, хорошо согласуются со всеми доступными опытными данными. В частности, для PEP-промоторов оценка интенсивностей связывания, полученная из экспериментальных данных, не связанных с нокаутами или тепловым шоком, даёт значения, близкие к полученным в модели. Значения других использованных в модели параметров также не расходятся с известными биологическими данными, хотя последние носят в основном косвенный характер.

В первом локусе для гена *ycf1* отношения уровней транскрипции с каждого из трёх его промоторов известны из опыта [48] и близки к отношениям уровней транскрипции, полученным в нашей модели.

Во втором локусе для гена *psbA* уровень транскрипции один из самых высоких [48], и это наблюдается в нашей модели. Боксы промотора перед геном *trnI* (одинаковые перед обеими копиями гена) отличаются от консенсуса лишь в слабо консервативных позициях, а перед –10-боксом вместо оптимальных нуклеотидов TG расположены CG. Промотор перед *psbA* практически не отличается от консенсуса. Это позволяет думать, что промоторы перед *trnI* и перед *psbA* обеспечивают близкие уровни инициации транскрипции этих генов, что мы и наблюдаем в модели. Ген *rpl16* не подвержен конкуренции, после него расположены интенсивно транскрибируемые гены на той же цепи, например ген *rpoA* α-субъединицы РНК-полимеразы. И, соответственно этому, в модели при повышении температуры мы наблюдаем рост транскрипции гена *rpl16*, хорошо объяснимый этим повышением. В целом модель согласуется с экспериментально установленными изменениями уровней транскрипции при тепловом шоке в изолированных хлоропластах, когда они не подвергаются воздействию белков, кодируемых в ядре.

В третьем локусе моделью предсказано существование терминатора T1 между генами *psbT* и *psbN*, непосредственно примыкающего к 3'-концу гена *psbN*, в хлоропласте *Arabidopsis thaliana*. Множественное выравнивание соответствующих участков подтверждает его наличие у широкой группы пластид растений и водорослей, см. также [65]. В хлоропластах растений группы eurosids II, включающей *Arabidopsis thaliana*, это консервативный палиндром длиной в 44 пары оснований с консенсусом TTGAMGTAATCAGCCTCCMAATATTKGGAGGCTGATTACKTCAA, рисунок 1.7.

Рисунок 1.7. Множественное выравнивание потенциального терминатора транскрипции T1 генов *psbT* и *psbN* в хлоропластах группы eurosids II

Межгенная область показана полностью (у *A. thaliana* её положение – 74184..74248). Стрелками и фоновой заливкой обозначены комплементарные плечи палиндрома. Звёздочками отмечены консервативные позиции, чёрточки обозначают пропуски. Условные обозначения видов: *A.c. – Aethionema cordifolium* (NC_009265.1), *A.g. – Aethionema grandiflorum* (NC_009266.1), *A.t. – Arabidopsis thaliana* (NC_000932.1), *D.n. – Draba nemorosa* (NC_009272.1), *B.v. – Barbarea verna* (NC_009269.1), *C.w. – Crucihimalaya wallichii* (NC_009271.1), *A.h. – Arabis hirsuta* (NC_009268.1), *C.b. – Capsella bursa-pastoris* (NC_009270.1), *N.o. – Nasturtium officinale* (NC_009275.1), *L.m. – Lobularia maritima* (NC_009274.1), *L.v. – Lepidium virginicum* (NC_009273.1), *O.p. – Olimarabidopsis pumila* (NC_009267.1), *C.p. – Carica papaya* (NC_010323.1), *C.s. – Citrus sinensis* (NC_008334.1), *G.h. – Gossypium hirsutum* (NC_007944.1).

Можно предположить, что этот палиндром образует крест-шпильку на ДНК, которая выполняет роль терминатора транскрипции. Аналогичный механизм терминации транскрипции в трейлерных областях интенсивно транскрибируемых генов ранее предсказан нами у Actinobacteria, [56]. Однако не исключено, что такой палиндром может служить и местом кооперативного связывания белка с ДНК. Терминатор T1 не может быть высокоэффективным, так как за ним идут активно экспрессируемые гены *petB* и *petD* (в составе полицистронной мРНК от *psbB* до *petD*), и действительно в нашей модели интенсивность терминации на нём мала.

В том же локусе аналогичная ситуация наблюдается с терминатором T2: участок ДНК между генами *petD* и *rpoA* в хлоропласте из *Arabidopsis thaliana* содержит палиндром с длиной 44 и координатами 77719..77762 (весь межгенный участок 77673..77900, согласно аннотации NC_000932 из базы данных GenBank) и тремя не комплементарными парами («несовершенный» палиндром). Нами получено выравнивание для T2 того же качества, что для T1. Оба терминатора могут влиять на транскрипцию в обоих направлениях. Несмотря на совпадение длин гипотетических терминаторов T1 и T2, их нуклеотидный состав существенно различается, рисунок 1.4. Последнее обстоятельство согласуется с гипотезой об образовании крест-шпильки на ДНК, для чего нуклеотидный состав не имеет существенного значения, но важна комплементарность плеч.

Наша модель позволяет предположить механизм клеточного ответа на нокаут и тепловой шок (детально описанный выше), а также некоторые механизмы регуляции экспрессии генов, основанные на конкуренции РНК-полимераз.

12. Заключение

Предложено количественное описание взаимодействия РНК-полимераз в процессах инициации и элонгации транскрипции. Показано, что оно согласуется практически со всеми опытными данными, относящимися к пластидам растений и водорослей, включая: изменения уровней транскрипции генов после нокаутов σ-субъединиц РНК-полимераз и теплового шока изолированных пластид; относительные количества РНК и времена их полураспада в митохондриях лягушек, человека здорового и с MELAS-мутацией, крысы здоровой и с пониженным уровнем тиреоидного гормона.

Предсказаны характеристики транскрипции в митохондриях хордовых животных: доли РНК-полимераз, завершающих транскрипцию на mTERF-зависимом терминаторе в одном и другом направлениях (поляризация); интенсивность связывания регуляторного белка mTERF с сайтом терминации на ДНК; интенсивности инициации транскрипции на промоторах в пластидах растений и в митохондриях лягушки, человека, включая MELAS-мутацию, крысы, включая гипотиреоида. Предсказаны значения уровней транскрипции всех генов, в то время как в опытах известны лишь их относительные количества и только для некоторых генов.

Предположен механизм влияния на фенотип MELAS-мутации: понижение количеств фенилаланиновой и валиновой тРНК, рРНК и, главное, резкого изменения времени полураспада некоторых мРНК.

Подтверждена корреляция между изменением метилирования сайта связывания mTERF и трёх промоторов, характерным для перехода от эутиреоида к гипотиреоиду, с одной стороны, и изменением интенсивностей связывания белка mTERF и инициаций транскрипции, с другой.

ГЛАВА 2. СЕМЕЙСТВА БЕЛКОВ, КОДИРУЕМЫХ В ПЛАСТИДАХ

1. Введение и постановка задачи

Понятие ортологичности двух белков (или кодирующих их генов) ещё не получило окончательной формализации; возможно, что таковая зависит от таксономической группы. Понятие ортологичности и соответствующие базы данных ортологичных генов/белков играют важную роль в биоинформатических исследованиях. В математической постановке поиск ортологичных генов/белков может быть описан как выделение кластеров в графе, вершинам которого соответствуют рассматриваемые гены/белки. Практически все методы кластеризации основаны на приписывании рёбрам этого графа весов («длин») с последующим выделением в нём в том или ином смысле «тесно связных компонент», иными словами кластеров, в процессе некоторой кластеризации графа.

Вес ребра отражает сходство аминокислотных последовательностей при различных способах парного выравнивания, сходство взаимного расположения интронов в генах, сходство в расположении доменов белков, порядок генов на хромосоме (локальную синтению) и т.д. В данной работе при вычислении весов рёбер рассматривались глобальное выравнивание белков (алгоритм Нидлмана – Вунша) и локальное выравнивание с помощью BLAST. Эти два варианта дают в основном сходные белковые семейства, по крайней мере, на наших данных; ниже приводятся результаты, соответствующие глобальному выравниванию. Отметим вариант нашего алгоритма, который учитывает локальную синтению генов на хромосоме; он был применён к различным множествам хордовых и учитывал положение гена на хромосоме или контиге и ортологичность его соседей (не включено в диссертацию).

Рассматривались весьма разнообразные процессы кластеризации: от специально организованного разбиения остовного дерева исходного графа (в предложенном нами и подробно описанном ниже алгоритме ClusterZSL) до оценок времени случайного блуждания в исходном графе (алгоритм OrthoMCL). Второй алгоритм характеризуется тем, что блуждание в кластере должно быть долгим, а переход в другой кластер – редким [73]. Все эти процессы эвристические, сравнение алгоритмов остаётся неформализованной задачей, особенно неопределённой в отсутствии стандартного набора данных для тестирования. В описании OrthoMCL прямо говорится, что даже вопрос о его сходимости трудно обсуждать хотя бы на уровне гипотезы; сходимость алгоритма ClusterZSL очевидна.

Алгоритм ClusterZSL принципиально отличается от обычно применяемых методов, включая OrthoMCL, и тем, что не требует нахождения взаимно наилучших хитов. Это понятие вызывает трудности: в двух геномах может не быть такой пары генов или, наоборот, может быть много таких пар; особенно если рассматривать и почти наилучшие хиты, которые как раз могут быть истинными ортологами. Алгоритм ClusterZSL, в том числе, минимизирует число паралогов (гомологичных белков из одного вида), что не входит в целевой функционал других обычных методов, по крайней мере явно; это условие на семейство ортологичных генов представляется важным.

Алгоритм ClusterZSL (с разработанной автором одноименной программой) имеет сложность не более n^2 с точностью до постоянного множителя. Алгоритм OrthoMCL (Markov Clustering algorithm) использует умножение матриц, сложность этой операции с точностью до мультипликативной константы равна n^{ω}, где для алгоритма Гаусса $\omega = 3$, для алгоритма Штрассена $\omega = \log_2 7 \approx 2.81$, [74]. Известен алгоритм, у которого $\omega \approx 2.37$, однако он даёт выигрыш только на матрицах очень большого порядка [75] и на практике не применяется. Дополнительную трудность представляет оценка числа итераций (включая число матричных умножений) в алгоритме OrthoMCL и проблема его сходимости. В практическом применении OrthoMCL, по-видимому, требует существенно большего времени работы, чем алгоритм ClusterZSL, по крайней мере, на наших данных.

Также сравним алгоритм ClusterZSL с алгоритмом, используемым в базе данных Ensembl. Последний начинает работу, по сути, с того же остовного дерева, что и первый. Но затем алгоритм в Ensembl существенно использует множественное выравнивание белков, приписанных листьям дерева. Время, которое требуется на построение остовного дерева, конечно, одинаковое в обоих алгоритмах, но последующий поиск множественного выравнивания заведомо экспоненциальный по сложности вычислений, если находить оптимальное выравнивание [76]. Алгоритм в Ensembl строит это выравнивание с помощью алгоритмов M-Coffee, [77] или, для больших данных, Mafft, [78]. Оба последних алгоритма чисто эвристические, без гарантии достижения минимума соответствующего функционала. Алгоритм ClusterZSL не использует множественного выравнивания.

Упомянём ещё один метод кластеризации при поиске ортологов, применявшийся нами в другой работе (не включено в диссертацию). Когда размеры кластеров заранее известны, например при поиске белков многокомпонентной системы, у которой размер кластера одной из компонент известен, использовалось выделение наиболее плотного кластера фиксированного размера с помощью алгоритма из работ [79, 80].

Итак, под *кластеризацией белков* понимается разделение множества белков на *кластеры* – семейства, разделяющие паралоги, – сходных (по аминокислотной последовательности) белков. Такая кластеризация позволяет, в частности: уточнять аннотации белков; выполнять поиск семейства белков по филогенетическому профилю; определять белки, уникальные для таксономической группы; судить о работоспособности белковых комплексов, состоящих из нескольких субъединиц, например РНК-полимераз бактериального типа, об эволюции видов и т.д.

Математически задача кластеризации данного множества белков состоит в построении такого разбиения этого множества на кластеры, что в один кластер попадают похожие белки, а паралоги входят в кластеры как можно реже. Полученные с помощью нашего алгоритма семейства белков включают биологически мотивированные паралоги, большинство из которых – точные или почти точные копии друг друга.

В этой главе описывается оригинальный алгоритм решения задачи кластеризации и рассматриваются результаты его применения к белкам, кодируемым в пластидах из трёх обширных групп, кратко описанных в пунктах 1.1–1.3. Для этих групп результат решения задачи кластеризации организован в базу данных, позволяющую, в частности, находить белки (кластеры) по заданному филогенетическому профилю.

Филогенетическим профилем называется разбиение данного множества видов на три части в соответствии с наличием у вида некоторого признака (фиксированного белка, сайта связывания на ДНК или какого-то фенотипического признака и т.д.): первая часть состоит из видов, обладающих данным признаком, вторая часть – из видов, не обладающих им, третья часть – из видов, относительно которых не известно, обладают ли они искомым признаком. Важной задачей является нахождение кластера, соответствующего данному филогенетическому профилю, т.е. содержащего белки из видов первой (и, возможно, третьей) части, но не содержащий белков из видов второй части. Частный случай такой задачи – нахождение кластеров, специфичных для некоторой таксономической группы.

1.1. Пластиды родофитной ветви

Багрянки и виды с пластидами, родственными пластидам багрянок, образуют *родофитную ветвь* в эволюционном дереве пластид [81]. Список рассмотренных пластомов из этой ветви приведён в таблице 2.1. В частности, большой интерес представляют диатомовые водоросли, для которых доступны пять пластомов и два полных ядерных генома. Вместе с ними мы рассмотрели два представителя надтипа Alveolata: *Durinskia baltica* (NC_014287.1) и *Kryptoperidinium foliaceum* (NC_014267.1), у которых

пластомы полностью секвенированы и близки к пластому *Phaeodactylum tricornutum*, [82]. Таблица 2.1 также характеризует полученные семейства белков – результат кластеризации.

Таблица 2.1. Пластомы родофитной ветви. В первом столбце указан номер пластома по базе данных NCBI, во втором – вид, к которому принадлежит пластом; в третьем – число пластомных белков в этом виде, в четвёртом и пятом – количество семейств (кластеров), содержащих хотя бы один белок из данного вида, с общим числом белков строго бóльшим 1 («не-синглетоны») и равным 1 («синглетоны») соответственно.

Пластом	Вид	Белков	(>1)	(1)
NC_012898.1	*Aureococcus anophagefferens*	105	105	0
NC_012903.1	*Aureoumbra lagunensis*	110	110	0
NC_011395.1	*Babesia bovis T2Bo*	32	25	5
NC_014340.1	*Chromera velia*	80	46	31
NC_014345.1	*Chromerida sp. RM11*	81	68	6
NC_013703.1	*Cryptomonas paramecium*	82	78	4
NC_004799.1	*Cyanidioschyzon merolae strain 10D*	207	179	28
NC_001840.1	*Cyanidium caldarium*	197	185	11
NC_014287.1	*Durinskia baltica*	129	128	0
NC_013498.1	*Ectocarpus siliculosus*	148	139	5
NC_004823.1	*Eimeria tenella strain Penn State*	28	27	1
NC_007288.1	*Emiliania huxleyi*	119	117	2
NC_015403.1	*Fistulifera sp. JPCC DA0580*	135	128	4
NC_006137.1	*Gracilaria tenuistipitata var. liui*	203	193	10
NC_000926.1	*Guillardia theta*	147	143	4
NC_010772.1	*Heterosigma akashiwo*	156	138	4
NC_014267.1	*Kryptoperidinium foliaceum*	139	130	9
NC_001713.1	*Odontella sinensis*	140	132	5
NC_008588.1	*Phaeodactylum tricornutum*	132	130	0
NC_000925.1	*Porphyra purpurea*	209	208	1
NC_007932.1	*Porphyra yezoensis*	209	206	3
NC_009573.1	*Rhodomonas salina*	146	142	4
NC_014808.1	*Thalassiosira oceanica CCMP1005*	142	126	1
NC_008589.1	*Thalassiosira pseudonana*	141	127	0
NC_007758.1	*Theileria parva strain Muguga*	44	34	5
NC_001799.1	*Toxoplasma gondii RH*	26	26	0
NC_011600.1	*Vaucheria litorea*	139	139	0

Родофитная ветвь пластид включает апикопласты многих споровиков – органеллы, похожие на пластиды багрянок, но имеющие сильно редуцированный геном. Изучение споровиков особенно важно, поскольку они вызывают заболевания человека и

животных. В частности, *Theileria* и *Babesia* переносятся иксодовыми клещами [83] и вызывают заболевания крупного рогатого скота: *B. bigemina* и *B. bovis* – бабезиоз крупного рогатого скота, *Th. annulata* – тейлериоз крупного рогатого скота, *Th. parva* – лихорадку Восточного Берега; *Eimeria tenella* – эймериоз кур; *Toxoplasma gondii* – токсоплазмоз кошек и человека; различные виды рода *Plasmodium* вызывают малярию у людей (*Pl. falciparum*), грызунов и других животных. Геномы *B. bovis* и *Th. parva* чрезвычайно близки между собой [84]. Обзор особенностей и функций апикопластов приведён в [85]. Отметим, что некоторые споровики, например *Cryptosporidium parvum*, не имеют апикопластов [86].

Исследование разнообразных процессов, связанных с апикопластами, позволит понять их роль в передаче инфекции и в механизмах действия лекарственных средств на апикопласт. Поскольку в апикопласте трансляция и обычно транскрипция имеют бактериальную природу, именно апикопласты являются главной мишенью антибиотиков, не оказывающих прямого воздействия на экспрессию ядерных и митохондриальных генов. Отсюда видно значение проблемы исследования механизмов регуляции и эволюции этих процессов у апикопластов. Некоторые результаты на эту тему содержатся в [87, 88].

Поскольку многие белки, достигающие пластид, кодируются в ядре, исследование пластид не может ограничиваться только пластомами. Нужно сопоставлять данные о белках, кодируемых в ядре, с данными о генах и регуляторных областях в пластоме. Особую роль играют субъединицы РНК-полимераз бактериального типа и РНК-полимеразы фагового типа, гомологичные РНК-полимеразам бактериофага T7, [14, 17], и кодируемые в ядре, которые обеспечивают транскрипцию в пластидах и митохондриях [89].

Кластеризация белков, кодируемых в пластидах, приводит к новой базе данных, в частности, удобной для исследования споровиков – возбудителей многих протозойных инфекций.

В пункте 2.2 этой главы обсуждается полученная нами кластеризация белков, кодируемых в пластомах родофитной ветви. Поиск кластеров по филогенетическому профилю белка, основанный на соответствующей базе данных, доступен на веб-странице [90]. С помощью этой базы данных найдены белки, специфичные для пластомов небольших таксономических групп водорослей и простейших, а также проведён поиск и анализ РНК-полимераз в ядерных геномах споровиков и, в частности, поиск σ-субъединиц РНК-полимераз бактериального типа и РНК-полимераз фагового типа у видов надтипа Alveolata.

1.2. Пластиды хлорофитной ветви

Хлорофитная ветвь состоит из рано отделившихся ветвей зелёных водорослей, включая таксономическую группу Chlorophyta, [91, 92] и видов с родственными пластидами, полученными в результате вторичного эндосимбиоза от видов из Chlorophyta. Это – *Euglena gracilis, E. longa* (из отдела Euglenozoa) и *Bigelowiella natans* (из группы Rhizaria). Таксономическая группа Chlorophyta делится на классы Chlorophyceae (роды *Floydiella*, *Schizomeris*, *Stigeoclonium*, *Chlamydomonas*, *Oedogonium*, *Scenedesmus*), Mamiellophyceae (роды *Micromonas* и *Ostreococcus*), Prasinophyceae (роды *Monomastix*, *Nephroselmis*, *Pycnococcus*, *Pyramimonas*), Trebouxiophyceae (роды *Chlorella*, *Parachlorella*, *Coccomyxa*, *Leptosira*, *Helicosporidium*), Ulvophyceae (*Bryopsis* и *Pseudendoclonium*) и род *Oltmannsiellopsis*, являющийся, вероятно, рано отделившейся ветвью класса Ulvophyceae. Более точное деление внутри класса Chlorophyceae обсуждается в [93], а внутри класса Trebouxiophyceae – в [92]. Заметим, что многие виды из класса Trebouxiophyceae (Требуксиевые), входящего в состав Chlorophyta, являются обычными симбионтами лишайников и простейших, включая инфузорию *Paramecium bursaria* и амёб *Amoeba borokensis* и *A. amazonas*, [94].

Два вида простейших – *Euglena gracilis*, [95] и *E. longa*, [96, 97], – являются ближайшими друг к другу представителями отдела Euglenozoa, [98, 99], хотя они значительно отличаются друг от друга. *E. gracilis* является фотосинтезирующим видом со смешанным типом питания и имеет светочувствительные стигмы, характерные для большинства видов этого отдела. Напротив, *E. longa* не способна к фотосинтезу. Это обусловило значительную редукцию её пластома. Известно, что пластиды, наряду с митохондриями, являются местом для независимого от света синтеза многих веществ и присутствуют у многих видов, лишённых фотосинтеза. Некоторые ортологичные белки *E. longa* и *E. gracilis* хорошо выравниваются, хотя заметно отличаются друг от друга. В частности, на выравнивании рибосомных белков из пластид *E. longa* и *E. gracilis* доли одинаковых аминокислотных остатков составляют: для L2 – 68%, для L20 – 44%, для L22 – 42%, для L23 – 49%, для S19 – 52%, [96]. Состав светособирающих пигментов *E. gracilis* и выравнивание белков, кодируемых в пластидах, показывают родство пластид *Euglena* spp. и пластид зелёных водорослей. Более того, пластиды *Euglena* spp. ближе к пластидам водорослей из классов Chlorophyceae и Trebouxiophyceae, чем к таковым из класса Prasinophyceae, [100]. Однако состав пластома, положение интронов и взаимное расположение генов на хромосоме в пластомах *E. gracilis* и других водорослей значительно различаются, что затрудняет определение непосредственного донора её пластид. Происхождение пластид *E. longa* также остаётся не вполне ясным.

Состав пигментов *Bigelowiella natans* соответствует таковому у зелёных водорослей, и многие белки, кодируемые в пластиде, хорошо выравниваются с белками пластид зелёных водорослей. Напомним также, что вторичное происхождение пластид *B. natans* от зелёных водорослей непосредственно подтверждается наличием нуклеоморфа, остатка от ядра водоросли [101]. Напротив, у *Euglena* spp. нуклеоморф отсутствует.

Филогенетические профили некоторых консервативных пластомных генов из Chlorophyta получены в [91], однако там отсутствуют данные о наличии многих белков (обычно их функция неизвестна).

В пункте 2.3 этой главы обсуждается полученная нами кластеризация белков, кодируемых в пластомах хлорофитной ветви, указанных в таблице 2.2. Поиск кластеров по филогенетическому профилю белка, основанный на соответствующей базе данных, доступен на веб-странице [102].

1.3. Пластиды цветковых растений

С помощью алгоритма и программы ClusterZSL получена кластеризация пластомных белков однодольных растений (пункт 2.4) и более широкой группы – цветковых растений (пункт 2.5).

Однодольные (Liliopsidae) произошли от примитивных травянистых двудольных (в основном, травянистые растения, реже – пальмы). Класс однодольные включает в себя 4 подкласса, 19 порядков, около 70 семейств, свыше 65 тысяч видов.

Список рассматриваемых пластомов однодольных растений (вместе с характеристикой результат кластеризации) приведён в таблице 2.3.

2. Результаты

Описывается алгоритм ClusterZSL, затем он применяется к пластидам родофитной и хлорофитной ветвей, цветковым и однодольным растениям, затем приводятся и обсуждаются полученные на его основе результаты. Таким образом, изложение ведётся по отдельности для каждой из этих групп пластид (см. выше пункты 1.1–1.3).

2.1. Алгоритм кластеризации

Опишем оригинальный алгоритм ClusterZSL кластеризации множества белков. Дано множество белков (последовательностей в соответствующем алфавите). Требуется построить кластеризацию (т.е. разбиение этого множества на попарно непересекающиеся подмножества), так чтобы в каждый кластер, максимальный по размеру, попадали сходные по последовательности белки из разных пластомов, а белки из одного пластома

входили в кластер только в случае, если их сходство больше сходства между белками из разных пластомов, входящими в тот же кластер. Например, белки PsaA и PsaB, хотя имеют близкие последовательности и функционируют вместе в составе первой фотосистемы, не заменяют друг друга и отнесены нашим алгоритмом в разные кластеры. Заметим, что традиционные алгоритмы кластеризации не применимы для решения такой задачи, в том числе, потому, что не учитывают распределения белков по разным видам (у нас – пластомам).

Таблица 2.2. Пластомы хлорофитной ветви. Обозначения, как в таблице 2.1.

Пластом	**Вид**	**Белков**	**(>1)**	**(1)**
NC_008408.1	*Bigelowiella natans*	61	57	0
NC_013359.1	*Bryopsis hypnoides*	69	68	1
NC_005353.1	*Chlamydomonas reinhardtii*	69	65	2
NC_015359.1	*Chlorella variabilis*	80	80	0
NC_001865.1	*Chlorella vulgaris*	174	94	78
NC_015084.1	*Coccomyxa sp. C-169*	80	78	2
NC_001603.2	*Euglena gracilis*	67	61	5
NC_002652.1	*Euglena longa*	46	37	6
NC_014346.1	*Floydiella terrestris*	74	69	4
NC_008100.1	*Helicosporidium sp.* ex *Simulium jonesi*	26	25	1
NC_009681.1	*Leptosira terrestris*	88	82	4
NC_012568.1	*Micromonas pusilla CCMP1545*	27	26	0
NC_012575.1	*Micromonas sp. RCC299*	57	56	0
NC_012101.1	*Monomastix sp. OKE-1*	82	70	10
NC_000927.1	*Nephroselmis olivacea*	155	111	13
NC_011031.1	*Oedogonium cardiacum*	99	80	1
NC_008099.1	*Oltmannsiellopsis viridis*	93	80	3
NC_008289.1	*Ostreococcus tauri*	61	58	2
NC_012978.1	*Parachlorella kessleri*	84	81	0
NC_008114.1	*Pseudendoclonium akinetum*	105	84	17
NC_012097.1	*Pycnococcus provasolii*	68	66	2
NC_012099.1	*Pyramimonas parkeae*	94	83	4
NC_008101.1	*Scenedesmus obliquus*	77	73	1
NC_015645.1	*Schizomeris leibleinii*	77	72	4
NC_008372.1	*Stigeoclonium helveticum*	79	72	6

Таблица 2.3. Пластомы однодольных растений. Обозначения, как в таблице 2.1.

Пластом	Вид	Белков	(>1)	(1)
NC_015820.1	*Acidosasa purpurea*	82	76	0
NC_010093.1	*Acorus americanus*	84	78	0
NC_007407.1	*Acorus calamus*	84	78	0
NC_008591.1	*Agrostis stolonifera*	85	77	0
NC_014062.1	*Anomochloa marantoidea*	85	78	0
NC_015830.1	*Bambusa emeiensis*	84	77	0
NC_012927.1	*Bambusa oldhamii*	82	76	0
NC_011032.1	*Brachypodium distachyon*	81	74	0
NC_013273.1	*Coix lacryma-jobi*	104	87	0
NC_013088.1	*Dendrocalamus latiflorus*	85	78	0
NC_009601.1	*Dioscorea elephantipes*	84	78	0
NC_015831.1	*Ferrocalamus rimosivaginus*	84	77	0
NC_011713.2	*Festuca arundinacea*	80	73	0
NC_008590.1	*Hordeum vulgare* subsp. vulgare	83	76	0
NC_015803.1	*Indocalamus longiauritus*	82	76	0
NC_010109.1	*Lemna minor*	85	78	0
NC_009950.1	*Lolium perenne*	84	77	0
NC_014056.1	*Oncidium Gower Ramsey*	74	68	0
NC_005973.1	*Oryza nivara*	119	95	1
NC_008155.1	*Oryza sativa* Indica Group	64	59	0
NC_001320.1	*Oryza sativa* Japonica Group	108	92	1
NC_015990.1	*Panicum virgatum*	85	77	0
NC_007499.1	*Phalaenopsis aphrodite* subsp. formosana	95	73	15
NC_013991.2	*Phoenix dactylifera*	95	80	1
NC_015817.1	*Phyllostachys edulis*	84	77	0
NC_015826.1	*Phyllostachys nigra* var. henonis	84	77	0
NC_014874.1	*Rhizanthella gardneri*	23	20	0
NC_006084.1	*Saccharum hybrid cultivar* NCo 310	117	92	2
NC_005878.2	*Saccharum hybrid cultivar* SP80-3280	97	82	0
NC_008602.1	*Sorghum bicolor*	84	76	0
NC_015891.1	*Spirodela polyrhiza*	83	76	0
NC_002762.1	*Triticum aestivum*	83	76	0
NC_013823.1	*Typha latifolia*	86	79	0
NC_015899.1	*Wolffia australiana*	83	77	0
NC_015894.1	*Wolffiella lingulata*	83	77	0
NC_001666.2	*Zea mays*	111	92	0

Кластеры формируются измельчением, начиная с единственного кластера, содержащего все данные белки. Кластер может включать далёкие белки, если при измельчении они не попали в разные кластеры. Такой подход полезен при рассмотрении далёких видов и их белков, которые произошли от одного предкового белка и сохранили общую функцию, когда сходство этих белков меньше сходства между паралогами). Общий план работы алгоритма показан на рисунке 2.1.

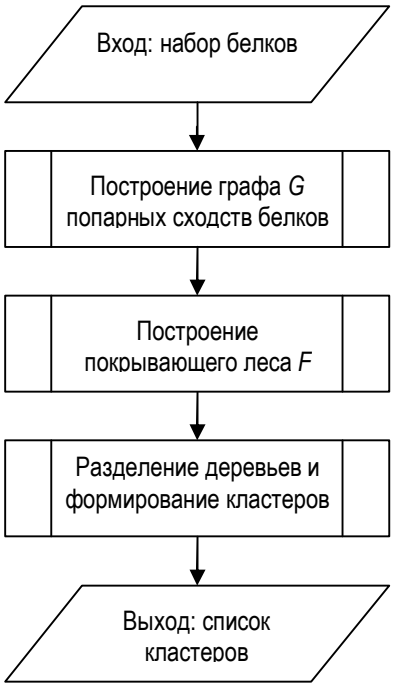


Рисунок 2.1. Общий план алгоритма кластеризации

Пусть задан набор пластомов S_i и для каждого пластома перечислены его белки P_{ij}. Для всех пар белков (P_{ij}, P_{kl}) из всех пар пластомов вычисляется характеристика сходства $s_0(P_{ij}, P_{kl})$ белков как качество оптимального глобального выравнивания их последовательностей; при этом само парное выравнивание не используется (и не вычисляется). Эта характеристика вычисляется стандартным алгоритмом Нидлмана – Вунша [103], в котором в качестве меры сходства последовательностей, включающих делеции, используется сумма соответствующих элементов матрицы BLOSUM62, [104]. После этого по следующей формуле вычисляется нормированное сходство белков: $s(P_{ij}, P_{kl}) = 2s_0(P_{ij}, P_{kl})(s_0(P_{ij}, P_{ij}) + s_0(P_{kl}, P_{kl}))^{-1}$.

Рассматривается полный неориентированный граф G_0 с множеством вершин $\{P_{ij}\}$, в котором каждому ребру (P_{ij}, P_{kl}) приписано значение $s(P_{ij}, P_{kl})$, которое называется *весом* этого ребра; рёбра соединяют различные вершины, т.е. петли отсутствуют. По G_0 строится разреженный граф G, включающий лишь рёбра (P_{ij}, P_{kl}), удовлетворяющие условиям: $s(P_{ij}, P_{kl}) = \max_m s(P_{im}, P_{kl}) = \max_m s(P_{ij}, P_{km})$ и $s(P_{ij}, P_{kl}) \geq L$, где максимумы берутся по всем белкам из соответствующих пластомов, *i*-го и *k*-го, а L – параметр алгоритма (по умолчанию равный нулю). Если $i = k$, то предполагается ещё условие $m \neq l$ и второе равенство не учитывается.

Для полученного графа *G* алгоритм процедурой Крускала строит лес *F* (ациклический подграф, компоненты связности которого – деревья), покрывающий *G* (рисунок 2.2). Сумма весов всех рёбер леса называется его *весом*. Итак, в *G* перебираются рёбра в порядке убывания их веса (при совпадении весов сначала выбираются рёбра, соединя-

ющие белки одного пластома), которые объявляются рёбрами строящегося леса F, если добавление к F очередного ребра из G не приводит к появлению в F цикла. В результате F не содержит циклов, т.е. является лесом, и включает все вершины из G. Вес полученного леса максимален по сравнению с любым другим лесом в G.

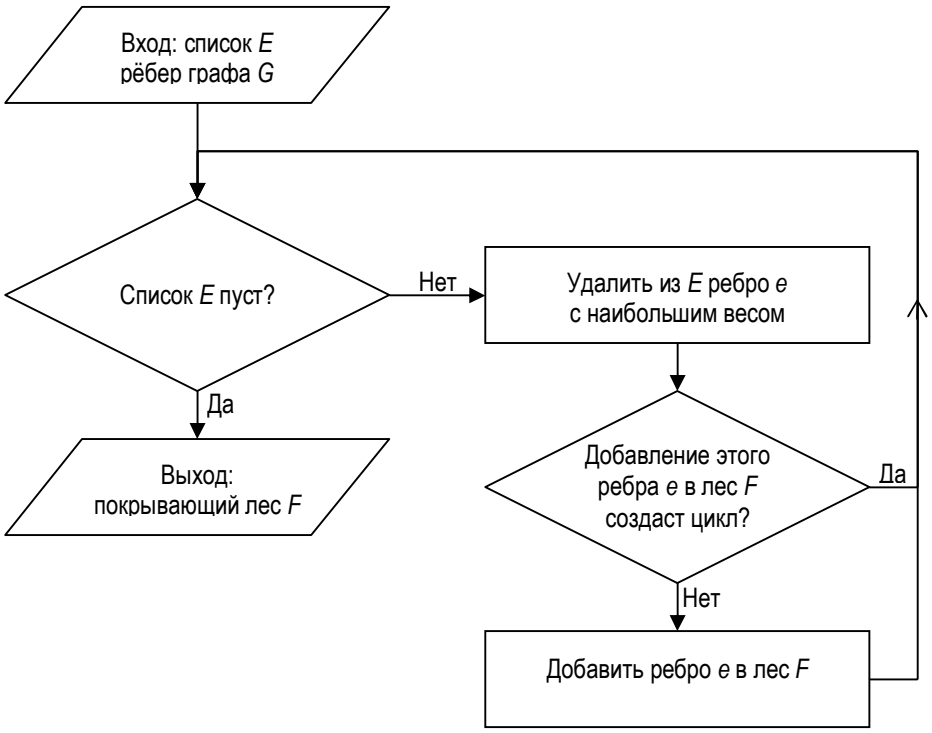


Рисунок 2.2. Схема алгоритма построения накрывающего леса
В начале список E содержит все рёбра графа G, а лес F – все вершины графа G. В результате: список E пуст, а лес F накрывает все вершины графа G и его вес максимальный.

Затем к лесу F применяется следующая процедура разделения деревьев (рисунок 2.3), которая строит набор C искомых белковых кластеров. Пусть T –дерево из F и e_0 – ребро в T с минимальным в T весом s_0. Если $s_0 < H$, где H – параметр алгоритма, и T не удовлетворяет сформулированному ниже *критерию сохранения дерева*, то T заменяется в F на два новых дерева T' и T'' путём удаления из T ребра e_0; в противном случае (т.е. критерий выполнен или $s_0 \geq H$) дерево T перемещается из F в список C.

Критерий сохранения дерева T состоит в выполнении 3-х условий (рисунок 2.4):
(1) $|T| \leq pn$, где $|T|$ – число вершин в дереве T, n – число видов, p – параметр алгоритма;
(2) ребро $e_0=(P_{mq}, P_{kl})$ соединяет белки P_{mq} и P_{kl}, у которых $m \neq k$;
(3) любая пара вершин P_{mq} и P_{ml} дерева T, соответствующих белкам из одного пластома, соединена в T путём, состоящим из вершин, соответствующих белкам из того же пластома (т.е. подграфы в T, состоящие из вершин одного пластома, связны).

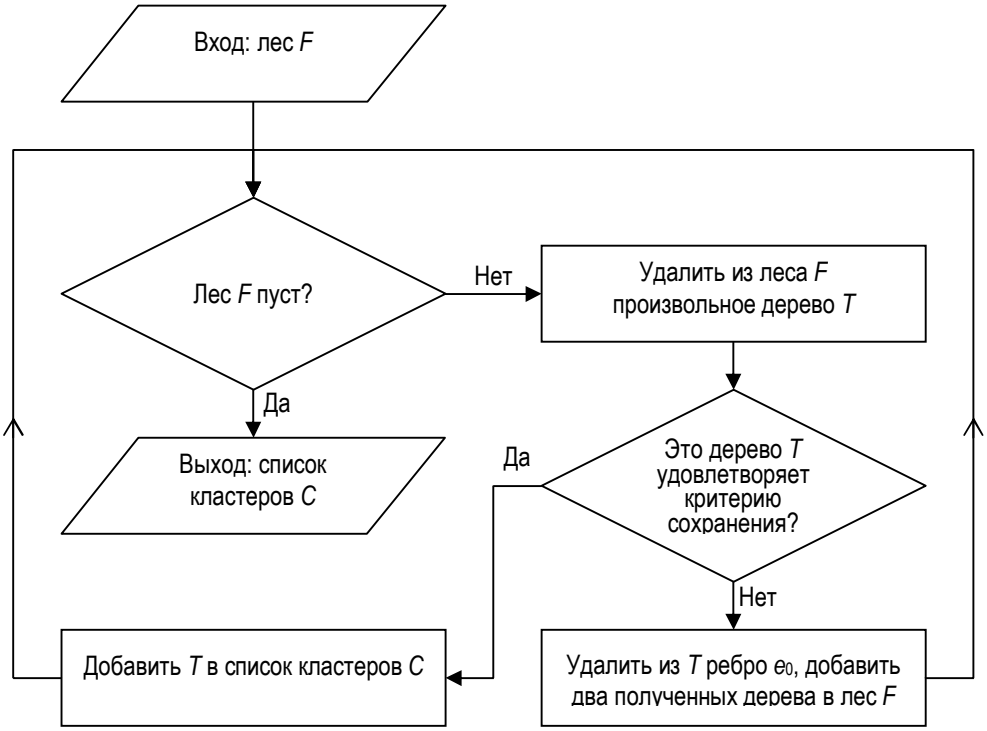


Рисунок 2.3. Схема алгоритма разделения леса и формирования кластеров
Вначале лес *F* содержит накрывающие *G* деревья, а список кластеров *C* пуст. В результате лес *F* пуст, а список *C* содержит набор искомых кластеров.

Если в *F* ещё остались деревья, то рассматривается следующее дерево *T* из *F*, иначе алгоритм завершает работу. Полученный в результате набор деревьев *C* представляет собой кластеры исходных белков: *один кластер состоит из последовательностей, приписанных всем вершинам одного дерева*. Конец описания алгоритма.

Предложение 1. Для любых белков P_0 и P_n, если в графе *G* существует путь от P_0 к P_n с весами рёбер не меньше *H*, то алгоритм помещает P_0 и P_n в один кластер.

Доказательство. Для $n=1$ утверждение справедливо, т.к. по условию разделения алгоритм никогда не удаляет из леса рёбра с весом, превышающим H. Пусть утверждение справедливо для $n=k$, т.е. белки P_0 и P_k принадлежат одному кластеру, и выполнено условие утверждения для $n=k+1$, т.е., в частности, $s(P_k,P_{k+1})>H$. Тогда, поскольку алгоритм никогда не удаляет из дерева рёбра с весом, превышающим H, ребро (P_k,P_{k+1}) будет сохранено, т.е. белки P_k и P_{k+1} попадут в один кластер, а значит и белки P_0 и P_n попадут в один и тот же кластер. Таким образом, утверждение доказано по индукции для любого натурального n. □

Предложение 2. Пусть C_1 и C_2 – две кластеризации одного множества белков при значениях H_1 и H_2 параметра *H*, соответственно. Если $H_1 > H_2$, то $C_1 = C_2$ или C_1 – измельчение C_2.

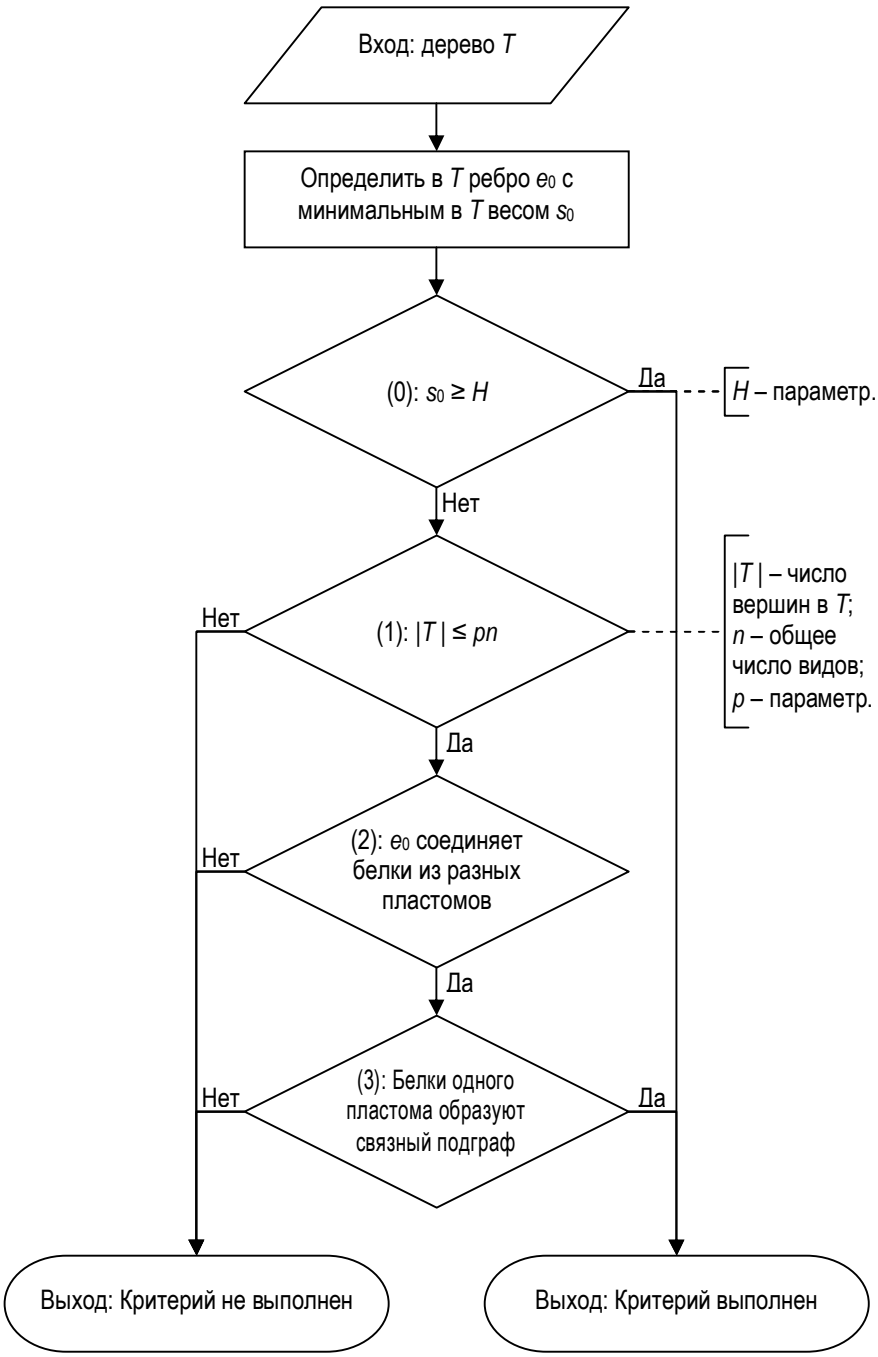


Рисунок 2.4. Схема проверки критерия сохранения дерева

Доказательство. По построению кластеризации параметр H влияет только на принятие решения об удалении некоторых рёбер в ходе выполнения процедуры разделения, т.е., в частности, покрывающий лес, строящийся алгоритмом для данного набора белков, не зависит от H. При удалении каждого ребра из леса одно дерево (компонента связности, которой принадлежит удаляемое ребро) заменяется на два. Таким образом, при увеличении значения H каждое дерево-кластер либо останется неизменным, либо разделится на два или более кластеров, что и требовалось доказать. □

Следствие 1. Условие: указаны наборы белков, элементы которых должны находиться в разных кластерах. Существует не более одного максимального по включению интервала, для которого выполняется: при любом значении параметра H из интервала алгоритм выдаёт кластеризацию, удовлетворяющую условию, и никакие два её кластера нельзя объединить с сохранением условия. □

Следствие 2. Условие: указаны наборы белков, ни один набор не должен разделяться кластерами. Существует максимальный по включению интервал, для которого выполняется: при любом значении параметра H из интервала алгоритм выдаёт кластеризацию, удовлетворяющую условию, и ни один кластер нельзя разбить на меньшие с сохранением условия. □

В обоих следствиях границы интервалов – рациональные числа (или бесконечность), которые вычисляются алгоритмически. Число из пересечения этих интервалов бралось в качестве значения параметра H, своего для каждой филогенетической группы. Например, у цветковых растений это пересечение – узкая окрестность, включающая выбранное значение $H = 0.5$.

Пример работы алгоритма

В примере кластеризуются девять коротких белков, кодируемых в трёх пластомах: NC_000925 (*Porphyra purpurea*), NC_000926 (*Guillardia theta*), NC_000927 (*Nephroselmis olivacea*). А именно, из каждого пластома взято по три коротких белка:

NP_053804.1: photosystem_I subunit IX, *Porphyra purpurea*

MNNNFTKYLSTAPVIGVLWMTFTAGFIIELNRFFPDVLYFYL;

NP_054005.1: photosystem_I subunit XII, *Porphyra purpurea*

MIDDSQIFVALLFALVSAVLAIRLGKELYQ;

NP_053866.1: ribosomal protein S18, *Porphyra purpurea*

MAVYRKKISPIKPTEAVDYKDIDLLRKFITEQGKILPKRSTGLTSKQQKKLTKAIKQARILSLLPFLNKD;

NP_050719.1: photosystem_I subunit VIII, *Guillardia theta*

MTAAYLPSILVPIIGIIFPGLTMAFAFIYIEQDQIN;

NP_050713.1: photosystem_I subunit IX, *Guillardia theta*

MDNNFLKYLSTAPVLLTIWLSFTAALVIEANRFYPDMLYFPI;

NP_050701.1: photosystem_I subunit XII, *Guillardia theta*

MISDTQIFVALILALFSFVLAIRLGTSLY;

NP_050833.1: photosystem_I subunit VIII, *Nephroselmis olivacea*

MVTSFLPSLFVPLVGLVFPAVAMASLFLYIEKDEIA;

NP_050847.1: photosystem_I subunit IX, *Nephroselmis olivacea*

MKDFTTYLSTAPVLAAVWFGFLAGLLIEINRFFPDALSFSFV;

NP_050819.1: ribosomal protein L36, *Nephroselmis olivacea*

MKVRPSVRKICDKCCLIRRHRKLLVICSNPKHKQRQG.

Обозначим эти белки в указанном порядке как: 1:1, 1:2, 1:3; 2:1, 2:2, 2:3; 3:1, 3:2, 3:3. Таким образом, пара *n*:*m* обозначает *m*-й белок из *n*-го пластома.

Значения сходства s_0 всех пар белков приведены в таблице 2.4*a*. Значения нормированного сходства *s* всех пар белков приведены в таблице 2.4*b*. Значения в таблице округлены до двух значащих цифр. Нормированные сходства указаны в процентах.

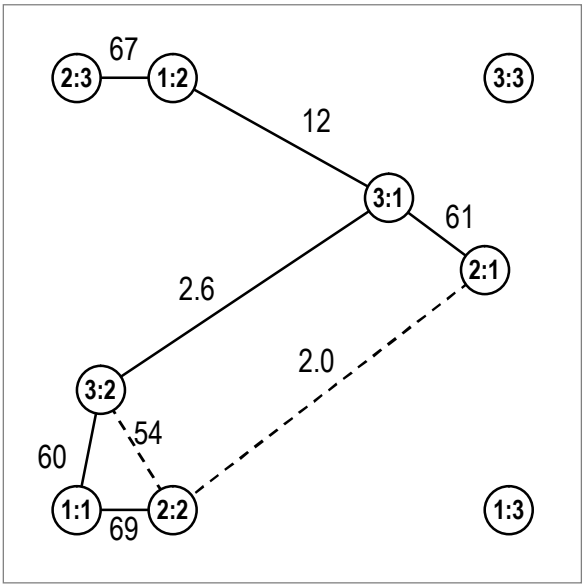
В таблице 2.4*c* после отбрасывания рёбер в графе *G* по второму условию разряжения остается 15 чисел. После отбрасывания рёбер в графе *G* по первому условию разряжения остаются 8 чисел, отмеченных в таблице полужирным шрифтом. Сам граф показан на рисунке 2.5*a*.

Граф имеет три компоненты связности: две, состоящие из изолированных вершин 1:3, 3:3, и одну, содержащую все остальные вершины. Первым двум компонентам соответствуют тривиальные накрывающие деревья (из одной вершины), для которых выполнен критерий сохранения, так что они образуют два одноэлементных кластера. Рассмотрим нетривиальную компоненту связности. Для неё имеется одно накрывающее дерево *T*, которое получается, удалением из неё рёбер, показанных на рисунке 2.5*a* пунктиром. Пусть параметр *p* равен двум. Тогда *T* не удовлетворяет первому условию сохранения. (Если $p = 3$, то *T* не удовлетворяет второму условию сохранения.) В *T* удаляется ребро 3:1–3:2. Получается набор из двух деревьев, показанный на рисунке 2.5*b*. Дерево с четырьмя вершинами не удовлетворяет третьему условию сохранения. Ребро 1:2–3:1 с минимальным весом удаляется. Получается набор из трёх деревьев, показанный на рисунке 2.5*c*. Полученные деревья удовлетворяют всем условиям сохранения. Алгоритм завершает работу.

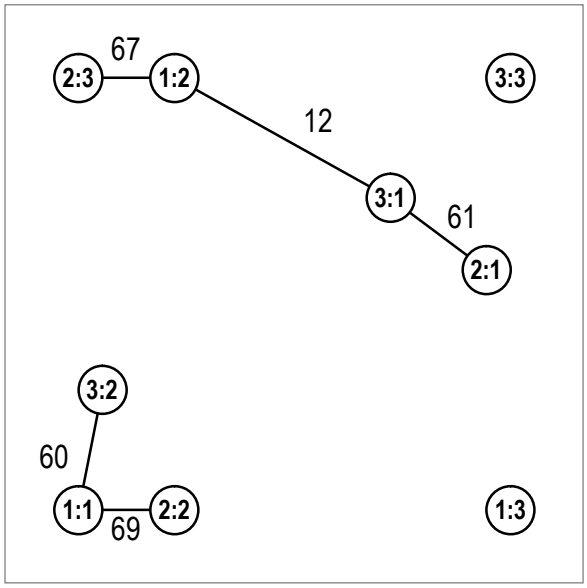
В результате по пяти деревьям получены следующие пять белковых кластеров: кластер 1 (1:1, 2:2, 3:2): {photosystem_I subunit IX, *Porphyra purpurea*, photosystem_I subunit IX, *Guillardia theta*, photosystem_I subunit IX, *Nephroselmis olivacea*}; кластер 2 (1:2, 2:3): {photosystem_I subunit XII, *Porphyra purpurea*, photosystem_I subunit XII, *Guillardia theta*}; кластер 3 (2:1, 3:1): {photosystem_I, subunit VIII *Guillardia theta*, photosystem_I subunit VIII, *Nephroselmis olivacea*}; и два одноэлементных кластера: 4 (1:3): {ribosomal protein S18, *Porphyra purpurea*} и 5 (3:3): {ribosomal protein L36, *Nephroselmis olivacea*}.

Таблица 2.4. Значения сходства s_0 и s пар белков; разрежение графа сходств

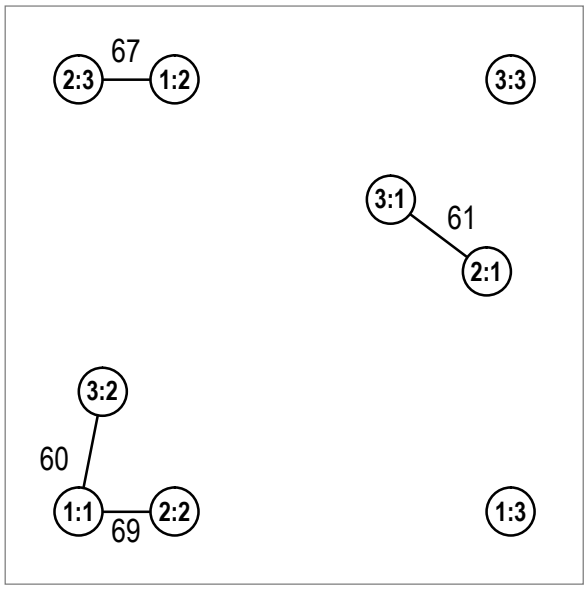
***a*) значения сходства s_0 пар белков**									
s_0	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1	225	-9	-60	11	153	1	7	131	-36
1:2	-9	139	-97	0	1	91	18	-6	-12
1:3	-60	-97	345	-66	-69	-101	-77	-54	-57
2:1	11	0	-66	180	4	-3	108	5	-17
2:2	153	1	-69	4	219	-4	12	118	-21
2:3	1	91	-101	-3	-4	134	8	-1	-27
3:1	7	18	-77	108	12	8	174	5	-22
3:2	131	-6	-54	5	118	-1	5	215	-27
3:3	-36	-12	-57	-17	-21	-27	-22	-27	203
***b*) значения нормированного сходства s пар белков**									
s	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1	100	-4.9	-21	5.4	69	0.6	3.5	60	-17
1:2	-4.9	100	-40	0.0	0.6	67	12	-3.4	-7.0
1:3	-21	-40	100	-25	-25	-42	-30	-19	-21
2:1	5.4	0.0	-25	100	2.0	-1.9	61	2.5	-8.9
2:2	69	0.6	-25	2.0	100	-2.3	6.1	54	-10
2:3	0.6	67	-42	-1.9	-2.3	100	5.2	-0.6	-16
3:1	3.5	12	-30	61	6.1	5.2	100	2.6	-12
3:2	60	-3.4	-19	2.5	54	-0.6	2.6	100	-13
3:3	-17	-7.0	-21	-8.9	-10	-16	-12	-13	100
***c*) граф G определяется полужирными значениями**									
G	1:1	1:2	1:3	2:1	2:2	2:3	3:1	3:2	3:3
1:1				5.4	**69**	0.6	3.5	**60**	
1:2					0.6	**67**	**12**		
1:3									
2:1					**2.0**		**61**	2.5	
2:2							6.1	**54**	
2:3							5.2		
3:1								**2.6**	
3:2									
3:3									



a) исходный граф



b) граф после 1-го шага разделения



c) граф после 2-го шага разделения

Рисунок 2.5. Граф *G* в процессе разделения деревьев

2.2. Кластеризация белков родофитной ветви пластид

Пластомы, указанные в таблице 2.1, получены из базы данных NCBI. В их числе – пластомы недавно секвенированных диатомовых водорослей [105, 106]. Некоторые фрагменты ядерных геномов *Eimeria tenella* и *Neospora caninum Liverpool* получены из базы данных Sanger Institute [107]. Счёт проводился при значениях параметров $H = 0.7$, $p = 2$, $L = 0$; полученные результаты сохраняются, если параметры остаются в преде-

лах: $0.6 \le H \le 0.7$, $1 \le p \le \infty$ и $-\infty \le L \le 0.05$. В целом параметры подобраны так, чтобы полученные кластеры хорошо согласовывались с доступными биологическими сведениями о семействах белков.

Развернутое статистически значимое исследование влияния параметров не проводилось, в том числе потому, что точный вид кластеров не известен. Несколько слов о влиянии параметров: при $p < 1$ кластеры максимального размера распадаются; при больших значениях p (даже при $p = +\infty$), т.е. без учёта условия (1) сохранения дерева, результаты не меняются, но время счёта увеличивается. Если значение L превышает 0.05, то с его ростом число рёбер в графе G быстро уменьшается, а число компонент связности в нём быстро возрастает, при этом кластеры, деревья которых содержат ребро с маленьким весом, распадаются. При $H \le 0.55$ некоторые кластеры объединяются, а при $H \ge 0.75$ – распадаются.

В некоторых редких случаях на основе биологической информации пришлось объединять или разделять кластеры. Например, кластер L-субъединиц протохлорофиллидредуктазы ChlL был выделен из большего кластера, сформированного алгоритмом и включающего белки, заведомо не относящиеся к синтезу хлорофилла и не сопровождаемые N-субъединицами. Выделение основано на эволюции генов *chlL* и *chlN*, как и *chlB*, кодирующих субъединицы независимой от света протохлорофиллидредуктазы, которая описана в работе [108]. Так же выделены ещё два кластера, один из них составили фрагменты β"-субъединицы РНК-полимеразы бактериального типа у Piroplasmida (*Babesia bovis* и *Theileria parva*), а другой – киназы из водорослей *Rhodomonas salina* и *Heterosigma akashiwo*.

Результаты кластеризации представлены в базе данных, доступной через веб-интерфейс [90], обеспечивающий ряд функций, среди которых отметим поиск белка (кластера) по заданному филогенетическому профилю.

Для контроля наших результатов и построения филогенетических деревьев, например при исследовании РНК-полимераз, использовался пакет программ MEGA 5, [109]. Поиск субъединиц РНК-полимераз выполнялся программой BLAST, [110], соответствующее значение E-value обозначается ниже E.

2.2.1. Характеристика кластеров пластомных белков родофитной ветви

Мы рассмотрели многочисленные таксономические группы родофитной ветви, охватывающие все её виды и представленные в базе данных GenBank, NCBI (на 01.10.2011), см. таблицу 2.1. Рассмотрено 3426 белков, из них образовано 260 кластеров, содержащих строго больше одного белка («не-синглетоны»), и 143 одноэлемент-

ных кластера («синглетоны»). Последние в совокупности содержат только 4% от числа всех белков, каждый из 11 не-синглетонов состоит из паралогичных белков. Подавляющее большинстве кластеров (359) не содержат паралогов, 44 кластера содержат их. Распределение кластеров в зависимости от числа представленных в них видов показано на рисунке 2.6.

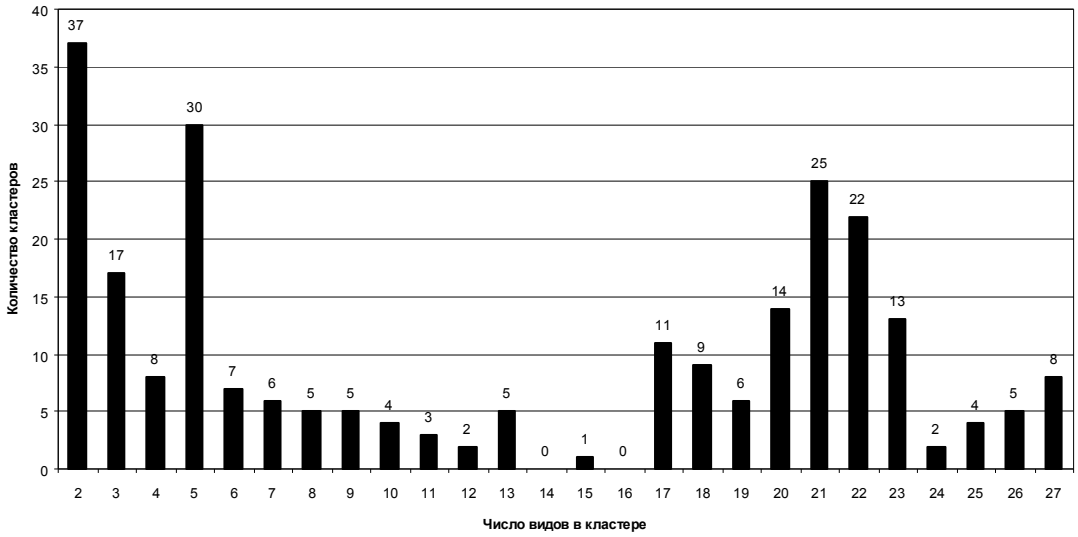


Рисунок 2.6. Распределение числа кластеров родофитной ветви пластид в зависимости от числа представленных в них видов

Белки, общие для пластомов всех рассмотренных видов, составляют 8 кластеров: рибосомные белки S2, S12, L2, L6, L14 и L16, фактор элонгации Tu и β-субъединица РНК-полимеразы бактериального типа. Рибосомный белок S19 определён у всех рассмотренных видов, кроме споровика *Babesia bovis*.

Для нескольких таксономических групп удалось выделить белки, которые *характеризуют эту группу* («филогенетические подписи»), т.е. кодируются в её пластомах и только в них. А именно, белки, кодируемые в пластидах багрянок (*Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Gracilaria tenuistipitata*, *Porphyra purpurea* и *P. yezoensis*) и отсутствующие в остальных рассмотренных пластомах (т.е. специфичные для багрянок), составляют 24 кластера: третий фактор инициации трансляции, α-, β-, β_{18}-, γ-субъединицы аллофикоцианина, α- и β-субъединицы фикоцианина, два формообразующих белка фикобилисом и связанный с деградацией фикобилисом белок Ycf18, тиоредоксин, белки комплекса ацетил-CoA-карбоксилазы, пренилтрансфераза, ацетилглутаматкиназа, ферредоксин-зависимая глутаматсинтаза, α- и β-субъединицы пируватдегидрогеназы E1, субъединицы антранилатсинтазы, α-субъединица триптофансинтазы и гипотетические консервативные белки.

Не найдено белка, специфичного для криптофитовых водорослей *Cryptomonas paramecium*, *Guillardia theta* и *Rhodomonas salina*; как и для Chromerida (*Alveolata* sp. CCMP3155 и *Chromera velia*).

Белки, специфичные для споровиков группы Piroplasmida (*Babesia bovis*, *Theileria parva*), составили 5 кластеров: два из них – слабые гомологи рибосомных белков, ещё два – молекулярные шапероны, гомологичные ClpC (YP_002290851.1, XP_762692.1, YP_002290850.1, XP_762693.1) и фрагменты β"-субъединицы РНК-полимеразы бактериального типа (YP_002290845.1, XP_762712.1).

Группа “Diatoms и Dinotoms” содержит *Durinskia baltica*, *Kryptoperidinium foliaceum*, *Fistulifera* sp. JPCC DA0580, *Odontella sinensis*, *Phaeodactylum tricornutum*, *Thalassiosira oceanica*, *Thalassiosira pseudonana*. Среди них 5 пластомов диатомовых водорослей: *Fistulifera* sp. JPCC DA0580, *P. tricornutum*, *O. sinensis, T. oceanica* и *T. pseudonana*. Пластиды *D. baltica* и *K. foliaceum* близки к пластидам *P. tricornutum*. Специфичными для этой группы оказались два кластера: один содержит гомологи белка Ycf88, другой – по два паралога, гомологичных белку Ycf89, из каждого вида этой группы.

Некоторые кластеры получили дополнительное обоснование при исследовании 5'-лидерных областей соответствующих генов. А именно, найдены консервативные участки в некодирующих областях пластомов перечисленных видов из этой группы, включая ещё недавно секвенированный пластом *Synedra acus* (NC_016731). Большое число пластомов в выравнивании позволяет говорить о достоверном выделении консервативных участков в некодирующих областях геномов. Для пар ортологичных генов, позиционно сцепленных хотя бы у 7-ми из 8-ми видов, были проведены дополнительные выравнивания лидерных областей. В хлоропластах диатомовых водорослей консервативные участки в составе длинных лидерных областей, в целом неконсервативных, имеются перед генами *rps20, ycf12, atpA, atpB, atpG, psaB, psaL, psbA, psbE, psbI, psbK, psbN, psbV, psbZ, rbcS, trnG, petF*. Из них только ген *petF*, кодирующий ферредоксин, отсутствует в пластоме *T. oceanica*; и был перенесён в ядро.

2.2.2. Поиск РНК-полимераз в ядерных геномах споровиков

У штаммов *Toxoplasma gondii* ME49 (XP_002367014.1), *T. gondii* VEG (EEE31947.1), *T. gondii* GT1 (EEE23737.1) и у *Neospora caninum* (CBZ55882.1) найдено по одной копии РНК-полимеразы фагового типа (номера указаны в скобках). У штаммов *T. gondii* ME49 и VEG белки совпадают, у штамма GT1 белок содержит замены аминокислотных остатков в нескольких позициях и вставку, занимающую позиции от 347 до 354. У *Eimeria tenella* не удалось определить РНК-полимеразу фагового типа.

Гомологи РНК-полимераз фагового типа найдены у многих споровиков, не являющихся кокцидиями: у *Plasmodium berghei* (XP_676913.1), *Pl. falciparum* 3D7 (XP_001347935.1), *Pl. knowlesi* H (XP_002259256.1), *Pl. vivax* SaI-1 (XP_001615369.1), *Pl. yoelii* 17XNL (XP_727223.1), *Pl. chabaudi* (XP_739650.1), *Babesia bovis* (XP_001611431.1), *Theileria annulata* (XP_953797.1), *Th. parva* (XP_766496.1). Дерево РНК-полимераз фагового типа показано на рисунке 2.7. Однако ортологичный белок не найден у кокцидии *Cryptosporidium parvum*, которая в отличие от многих споровиков не имеет пластид.

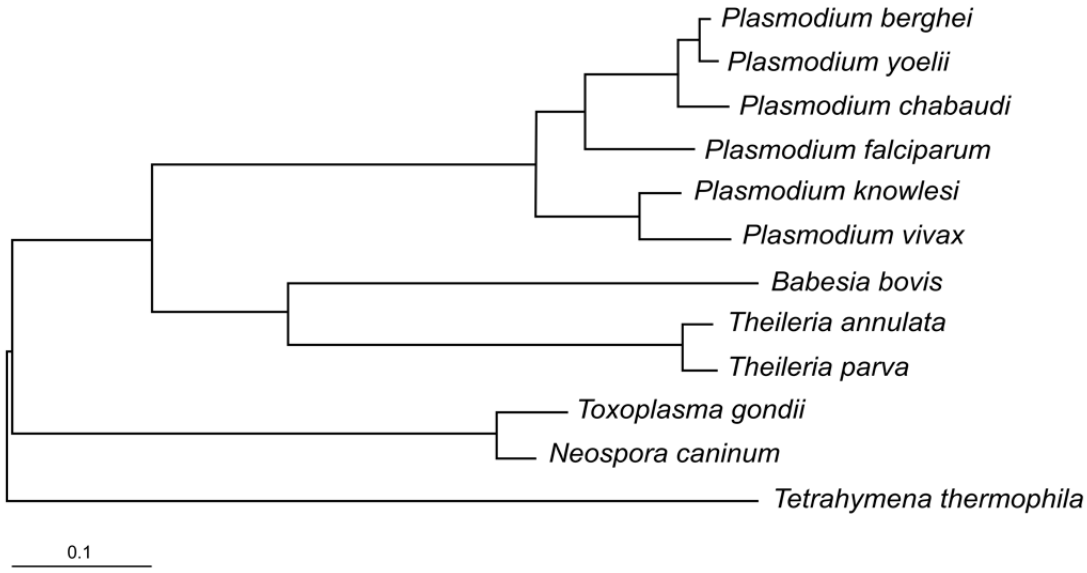


Рисунок 2.7. Дерево РНК-полимераз фагового типа у простейших надтипа Alveolata

В ядерном геноме *Toxoplasma gondii* обнаружен только один ген, кодирующий σ-субъединицу РНК-полимеразы бактериального типа. Её длина – 1002 аминокислотных остатка у штаммов ME49 и GT1, 1001 – у штамма VEG. Ниже рассматривается белок XP_002367841.1 штамма ME49. В ядерном геноме *Neospora caninum* ген CBZ51366.1 кодирует σ-субъединицу РНК-полимеразы длиной 1206 аминокислотных остатков. У *T. gondii* и *N. caninum* С-концы σ-субъединиц РНК-полимераз чрезвычайно близки друг к другу, но не имеют существенного сходства с σ-субъединицами диатомовых водорослей *Phaeodactylum tricornutum* CCAP 1055/1 и *Thalassiosira pseudonana* CCMP1335, золотистой водоросли *Aureococcus anophagefferens*, криптофитовых водорослей *Guillardia theta* и *Hemiselmis andersenii*. σ-Субъединицы, ближайшие к этим σ-субъединицам кокцидий, найдены у цианобактерий *Cyanothece* sp. PCC 7822 (YP_003885480.1), *Microcoleus chthonoplastes* PCC 7420 (ZP_05024793.1), *Acaryochloris marina* MBIC11017 (YP_001519047.1) и у δ-протеобактерии *Desulfarculus baarsii* DSM 2075

(YP_003809216.1). Бактериальные ортологи имеют длины от 260 до 363 аминокислотных остатков. У всех видов хорошо выравниваются С-концы второго региона, весь третий регион и N-концы четвёртого региона σ-субъединиц РНК-полимераз. По всей длине четвёртый регион выравнивается у *T. gondii, N. caninum* и *D. baarsii*.

Также ортологи σ-субъединиц РНК-полимеразы найдены у простейших из отряда Haemosporida: *Plasmodium berghei* (XM_669238.1), *Pl. falciparum* 3D7 (XP_966194.1), *Pl. knowlesi* H (XM_002261430.1), *Pl. vivax* SaI-1 (XP_001616222.1), *Pl. yoelii* 17XNL (XP_724777.1), *Pl. chabaudi* (XM_739944.1). В каждом из них отсутствуют другие σ-субъединицы. Не удалось определить σ-субъединицы РНК-полимеразы у видов из отряда Piroplasmida: *Theileria parva, Th. annulata, Babesia bovis*. Дерево σ-субъединиц показано на рисунке 2.8.

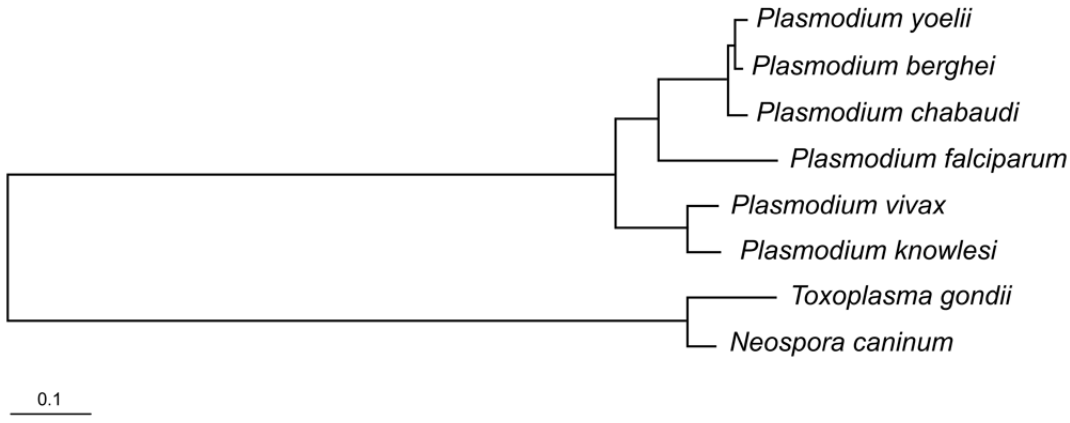


Рисунок 2.8. Дерево σ-субъединиц РНК-полимераз у споровиков

Особенностью пластомов споровиков является отсутствие у них α-субъединиц РНК-полимераз бактериального типа. Рассмотрено три вида кокцидий: *Eimeria tenella*, *Toxoplasma gondii* и *Neospora caninum*.

Данные о *T. gondii* и об обсуждаемых водорослях и бактериях доступны в базе данных NCBI. У *T. gondii* ME49 α-субъединица кодируется в ядре, соответствующий белок XP_002367289.1 имеет 836 аминокислотных остатков. У этого белка имеется отличие в одной позиции между штаммами *T. gondii* ME49 и GT1. В ядерном геноме *E. tenella* обнаружена близкая (E=1.1×10^{-71}) α-субъединица, для которой определены фрагменты четырёх экзонов на контиге dev_EIMER_contig_00028796 с координатами соответственно 5283..5453, 5682..6167, 6576..6785 и 7273..7965. В ядерном геноме *N. caninum* обнаружена близкая (E=9.9×10^{-288}) α-субъединица, для которой определены два экзона на контиге Contig892 с координатами соответственно 45655..47412 и 47940..48611.

2.2.3. Обсуждение результатов кластеризации для родофитной ветви

Белок NP_045121.1 у *Cyanidium caldarium* входит в кластер, содержащий белки YP_537023.1 из *Porphyra yezoensis* и NP_053952.1 из *P. purpurea*. Эти белки имеют относительно короткий консервативный домен, характерный для транскрипционного фактора NtcA (Ycf28). Белок NP_849012.1 из *Cyanidioschyzon merolae* является гомологом для NtcA, однако он не вошёл в кластер NtcA из-за значительного отличия, в том числе в наиболее консервативном домене фактора. Ещё меньшее сходство в соответствующем домене – у NtcA и его гомолога у *Gracilaria tenuistipitata*. Эволюционно это изменение связано с переносом в ядро или потерей гена *glnB* из пластома, транскрипция которого регулируется фактором NtcA у багрянок *Porphyra* spp. и *Cyanidium caldarium*, [111].

Пластом *Gracilaria tenuistipitata* содержит гены *leuC* и *leuD*, кодирующие большую (YP_063540.1) и малую (YP_063541.1) субъединицы 3-изопропилмалатдегидрогеназы, которые отсутствуют в других рассмотренных пластомах. Как отмечается в [112], это свидетельствует о раннем разделении таксономических групп Florideophyceae (включающей *G. tenuistipitata*) и Bangiophyceae в составе отдела багрянок.

Особенностью пластомов споровиков является отсутствие в них α-субъединиц РНК-полимераз бактериального типа, однако их гомологи найдены в ядерных геномах большинства споровиков.

Наличие у диатомовых водорослей и близких к ним третичных эндосимбионтов общих белков, отсутствующих в пластидах других видов, позволяет предположить раннее обособление диатомовых водорослей от других представителей родофитной ветви.

Неконсервативность большинства субъединиц РНК-полимеразы бактериального типа у Piroplasmida позволяет сомневаться в работоспособности этого фермента. Эта гипотеза подтверждается тем, что в их ядерных геномах не удалось определить σ-субъединицу. Можно предположить, что у Piroplasmida транскрипция всего пластома осуществляется исключительно РНК-полимеразами фагового типа, что означает неэффективность в борьбе с Piroplasmida антибиотиков, ингибирующих РНК-полимеразы бактериального типа. Напротив, такие антибиотики могут быть применены против *Plasmodium* spp, *Toxoplasma gondii* и *Neospora caninum*.

Дерево σ-субъединиц РНК-полимераз бактериального типа у споровиков, исключая виды из Piroplasmida, хорошо согласуется как с деревом видов, так и с деревом РНК-полимераз фагового типа. Наличие не более одной σ-субъединицы РНК-полимеразы у споровиков указывает на незначительную роль регуляции пластома на уровне транскрипции. Вероятно, здесь наибольшее значение имеет регуляция на уровне трансляции или процессинга, что подтверждается наблюдениями [87].

РНК-полимеразы фагового типа у видов рода *Plasmodium* хорошо выравниваются между собой и образуют кладу на дереве белков. Также эти полимеразы формируют отдельные клады у Piroplasmida и Coccidia. Однако РНК-полимеразы Coccidia существенно отличаются от ортологичных белков у других споровиков. Напротив, РНК-полимеразы фагового типа у кокцидий близки к ортологичным белкам тетрахимены, не имеющей пластид. Можно предположить, что у кокцидий РНК-полимеразы фагового типа не играют роли в транскрипции пластома. Наши данные не выявили значительного разнообразия РНК-полимераз фагового типа у простейших. Вероятно, РНК-полимеразы фагового типа у споровиков имеют древнее происхождение и не связаны с приобретением пластид. Напротив, у высших растений наблюдается большое разнообразие РНК-полимераз фагового типа, которые нацелены на различные органеллы [88, 89].

2.3. Кластеризация белков хлорофитной ветви пластид

Пластомы 25-ти видов получены из базы данных NCBI и перечислены в таблице 2.2. Для контроля результатов и построения филогенетических деревьев использовались пакет программ MEGA 5, [109] и база данных Pfam, [113].

Веб-интерфейс [102] обеспечивает для хлорофитной ветви функциональность, описанную выше (пункт 2.2) для родофитной ветви.

2.3.1. Характеристика кластеров пластомных белков хлорофитной ветви

Кластеризация охватывает 1992 белка, из которых сформированы 166 одноэлементных кластеров («синглетонов») и 156 кластеров, включающих строго более одного белка («не-синглетонов»); среди не-синглетонов 87 содержат не более одного белка из каждого вида, 68 содержат пары белков из одного вида, один кластер содержит тройки белков из одного вида. Детали показаны в таблице 2.2. Распределение кластеров по числу представленных в них видов показано на рисунке 2.9. Из них 13 кластеров имеют представителей в каждом виде группы Chlorophyta (из таблицы 2.2). Более того, представители этих 13-ти кластеров найдены и у вторичных эндосимбионтов *Euglena gracilis*, *E. longa* и *Bigelowiella natans*. Для каждого из классов Mamiellophyceae, Prasinophyceae, Trebouxiophyceae и Ulvophyceae не найдено ни одного специфичного кластера белков, т.е. такого, что его белки присутствуют в каждом виде данного класса, но отсутствуют в других классах группы Chlorophyta. Среди Chlorophyta только класс Chlorophyceae имеет, и при том ровно один, специфичный кластер белков, вероятно связанных с делением пластид. Это показывает значительную близость пластомов рассмотренных видов из группы Chlorophyta. Здесь ситуация принципиально отличается от той, которую мы видели в отделе Rhodophyta, рассмотренной в пункте 2.2.1.

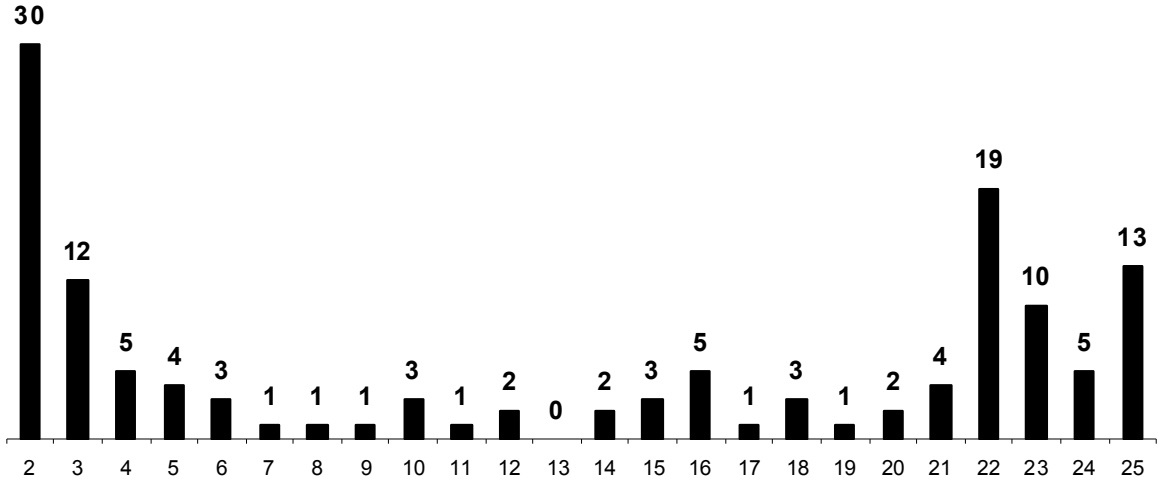


Рисунок 2.9. Распределение числа кластеров хлорофитной ветви пластид в зависимости от числа представленных в них видов

Некоторые алгоритмически полученные кластеры объединялись из биологических соображений. Ниже перечислены все такие случаи.

Три кластера, состоящие из рибосомных белков S3, объединены в один кластер. Эти белки имеют консервативные участки на N- и C-концах, но в середине у некоторых видов произошли длинные неконсервативные вставки; вероятно, это – не имеющие описания интроны с длинами, кратными 3 п.н.

Два кластера, состоящие из рибосомных белков S4, объединены в один.

К большому кластеру, состоящему из α-субъединиц РНК-полимераз бактериального типа (RpoA), добавлены два маленьких кластера гомологичных белков: один состоял из белков двух видов *Euglena* spp., другой – из белков вида *Pycnococcus provasolii*. По Pfam наибольшее значение E-value при сравнении с доменами, характерными для RpoA, наблюдаются у белка из *P. provasolii*, однако ближайший к нему гомолог у *Pyramimonas parkeae*, несомненно, является α-субъединицей РНК-полимеразы.

Объединены в один кластер белки АТФ-связывающей субъединицы протеазы ClpP. На выравнивании видно, что на N-конце белков имеется консервативный участок длиной около 50 аминокислотных остатков. У видов из класса Chlorophyceae присутствует вставка длиной около 300 аа, и вблизи C-конца имеется ещё один консервативный участок длиной около 130 аа. При проверке по базе данных Pfam у всех белков обнаружены домены, характерные для Clp протеазы.

Объединены в один кластер некоторые белки, гомологичные белку Ycf1, для которых удалось построить хорошее множественное выравнивание.

Объединены в один кластер некоторые белки, гомологичные белку FtsH и связанные с делением пластид, для которых удалось построить хорошее множественное выравнивание и которые имеют домен, характерный для АТФаз.

Объединены в один кластер белки, являющиеся фрагментами β-субъединицы РНК-полимеразы бактериального типа (RpoB). Так полученный кластер подробнее описан в конце пункта 2.3.2.

2.3.2. Обсуждение результатов кластеризации для хлорофитной ветви

Всего 29 кластеров имеют представителей в обоих видах *Euglena longa* и *E. gracilis*. Из них только один кластер представлен ровно в двух видах *E. longa* и *E. gracilis*; он состоит из белков NP_074988.1 и NP_041917.1 с неизвестной функцией; 14 кластеров содержат белок из *E. longa*, но не содержат соответствующего белка из *E. gracilis*; 37 кластеров содержат белок из *E. gracilis*, но не содержат белка из *E. longa*. Наиболее часто белки из пластома *E. longa*, отсутствующие в *E. gracilis*, имеют гомологи у водорослей *Chlorella vulgaris* и *Leptosira terrestris* из класса Trebouxiophyceae (Требуксиевые). Наша гипотеза: донором пластид для *E. longa* и *E. gracilis* служит класс Trebouxiophyceae. Это хорошо согласуется с высокой частотой вхождения водорослей именно этого класса в состав симбиотических сообществ, включая лишайники и некоторые простейшие [94]. Асимметрия (14<37) между их пластомами связана с заметно меньшим размером пластома *E. longa* по сравнению с *E. gracilis*.

Выравнивание с помощью MUSCLE, [114] белков каждого из 13-ти кластеров, имеющих представителей в каждом из рассмотренных видов (таблица 2.2), и последующее построение деревьев белков методом Maximum Likelihood показало, что в 11-ти из 13-ти кластеров ближайшим к белку из *Euglena longa* является белок из *E. gracilis*, а в двух других случаях – из *Helicosporidium* sp. ex *Simulium jonesii*. При этом в одном из этих двух случаев три вида *Euglena* spp. и *Helicosporidium* sp. образуют отдельную кладу. Всего в 4-х из 13-ти кластеров три вида *Euglena* spp. и *Helicosporidium* sp. образуют отдельную кладу. В остальных случаях положение белков из *Euglena* spp. относительно белков других родов меняется хаотически, что не позволяет сделать уверенное предположение о доноре пластид для *Euglena* spp., однако принадлежность *Helicosporidium* к классу Trebouxiophyceae согласуется с нашей гипотезой о происхождении пластид *Euglena* spp. от вида из этого класса.

Наконец, опишем содержание кластера β-субъединиц РНК-полимераз бактериального типа (RpoB). Здесь оказалось полезным учесть архитектуру белка RpoB у разных видов. У этих белков с помощью базы данных Pfam было определено взаимное

расположение доменов. Ниже используется следующая нумерация доменов из базы данных Pfam: 1-й – RNA_pol_Rpb2_1, 2-й – RNA_pol_Rpb2_2, 3-й – RNA_pol_Rpb2_3, домены 4 и 5 объединены в рассматриваемых видах в один домен RNA_pol_Rpb2_45, 6-й – RNA_pol_Rpb2_6, и 7-й – RNA_pol_Rpb2_7.

У вида *Coccomyxa* sp. C-169 в белке RpoB (YP_004222037.1) домен 1 имеет две копии. В двух видах *Leptosira terrestris* (YP_001382217.1) и *Oedogonium cardiacum* (YP_002000391.1) домен 6 имеет по две копии, расположенные в указанных белках рядом. В пяти видах – *Chlamydomonas reinhardtii* (NP_958398.1), *Oltmannsiellopsis viridis* (YP_635874.1), *Pseudendoclonium akinetum* (YP_636174.1), *Scenedesmus obliquus* (YP_635950.1) и *Stigeoclonium helveticum* (YP_764419.1) – домен 2 имеет по две копии в указанных белках. В 13-ти видах – *Pyramimonas parkeae* (YP_002600950.1), *Nephroselmis olivacea* (NP_050839.1), *Euglena gracilis* (NP_041951.1), *Oltmannsiellopsis viridis* (YP_635874.1), *Chlorella vulgaris* (NP_045893.1), *Ostreococcus tauri* (YP_717229.1), *Micromonas* sp. RCC299 (YP_002808641.1), *Pseudendoclonium akinetum* (YP_636174.1), *Micromonas pusilla* CCMP1545 (YP_002808499.1), *Chlorella variabilis* (YP_004347774.1), *Bigelowiella natans* (YP_778610.1), *Parachlorella kessleri* (YP_003058290.1), *Coccomyxa* sp. C-169 (YP_004222037.1) – все домены представлены во всех белках из кластера. К этим 13-ти видам примыкают *Euglena longa* (NP_074962.1), у которой потерян только домен 2, и *Monomastix* sp. OKE-1 (YP_002601004.1), у которой потеряны только домены 4 и 5.

Ещё у семи видов β-субъединица РНК-полимеразы (RpoB) разделилась на два белка, один из которых включает только домены 6 и 7, а другой – только остальные домены с 1 по 5. Это – *Chlamydomonas reinhardtii* (NP_958398.1, NP_958397.1), *Scenedesmus obliquus* (YP_635950.1, YP_635949.1), *Schizomeris leibleinii* (YP_004581337.1, YP_004581350.1), *Stigeoclonium helveticum* (YP_764419.1, YP_764412.1), *Floydiella terrestris* (YP_003795481.1, YP_003795537.1), *Leptosira terrestris* (YP_001382216.1, YP_001382217.1) и *Oedogonium cardiacum* (YP_002000410.1, YP_002000391.1). Особняком стоят ещё два вида: *Helicosporidium sp.* ex *Simulium jonesii* (YP_635922.1) с доменами 3, 6, 7 и *Bryopsis hypnoides* (YP_003227091.1) с доменами 1, 2, 3 у белка RpoB. У *Pycnococcus provasolii* белок RpoB не был определён. Отметим, что у *P. provasolii* размеры белков RpoC1 и RpoC2, кодирующих β'- и β''-субъединицы РНК-полимеразы, значительно больше, чем у близкого вида *Pyramimonas parkeae*.

2.3.3. Дополнительное исследование кластеров CysA и CysT

В пластидах Viridiplantae ген *cysT* присутствует у зелёных водорослей группы Chlorophyta: *Bryopsis hypnoides*, *Nephroselmis olivacea*, *Pycnococcus provasolii*, *Chlorella variabilis*, *Chlorella vulgaris*, *Coccomyxa subellipsoidea* C-169, *Helicosporidium sp. ex Simulium jonesii*, *Leptosira terrestris*, *Parachlorella kessleri*; зелёных водорослей группы Streptophyta: *Chlorokybus atmophyticus*, *Mesostigma viride*, *Zygnema circumcarinatum*; мохообразных: *Anthoceros formosae*, *Marchantia polymorpha*, псевдогены – у *Aneura mirabilis* и *Ptilidium pulcherrimum*.

Белки CysT консервативны почти по всей длине (кроме короткого N-концевого участка) и представляют собой трансмембранный домен ABC-транспортёра. Однако у *Bryopsis hypnoides* и *Leptosira terrestris* они укорочены на C-конце. Ортологичные белки с хорошим выравниванием имеются у цианобактерий.

Во всех перечисленных видах Viridiplantae, кроме *Helicosporidium* sp. и *Pycnococcus provasolii*, в пластидах наряду с геном *cysT* присутствует также ген *cysA*. У *Marchantia polymorpha* ортологичный *cysA* ген имеет необычное имя *mbpX*. Многие виды, близкие к перечисленным, не имеют генов *cysA* или *cysT*. Удивительно, что эти гены или псевдогены сохранились у мохообразных, хотя отсутствуют у многих высокоорганизованных водорослей, близких к наземным растениям: *Chaetosphaeridium globosum*, *Chara vulgaris*, *Staurastrum punctulatum*. Также они отсутствуют в пластомах мха *Physcomitrella patens* и всех сосудистых растений. Среди зелёных водорослей эти гены чаще встречаются в классе Trebouxiophyceae (роды *Chlorella*, *Coccomyxa*, *Helicosporidium*, *Leptosira*, *Parachlorella*).

Белки CysA пластид хорошо выравниваются с ортологичными белками цианобактерий. Белок CysA у всех Viridiplantae имеет сильно консервативный N-концевой домен, характерный для АТФ-связывающей кассеты ABC-транспортёров. У всех рассмотренных видов группы Chlorophyta, за исключением *Nephroselmis olivacea*, этот белок укорочен на C-конце. Напротив, у видов группы Streptophyta, у *Nephroselmis olivacea* и у цианобактерий присутствует консервативный C-концевой домен. У *Mesostigma viride* и *Chlorokybus atmophyticus* этот домен имеет гомологию с доменом TOBE, вероятно связанным с распознаванием сульфата. Согласно базе данных Pfam 26.0 значения E-value для этого домена составляют 0.0017 для *M. viride* и 0.00007 для *Ch. atmophyticus*. У других белков из пластид сходство домена меньше, но на выравнивании прослеживается много консервативных позиций.

В большинстве случаев в 5'-лидерной области рассматриваемого гена расположен один или два кандидата в промоторы бактериального типа. Исключением является

ген *cysA* у *Anthoceros formosae*, перед которым расположены три потенциальных промотора близкого качества. Единственный кандидат в промоторы перед геном *cysA* у *Chlorella vulgaris* имеет необычный -35 бокс AAGAAA. Однако перед этим геном у *Ch. variabilis* определён хороший потенциальный промотор с TG-расширением -10 бокса. Не удалось определить промоторы перед обоими генами *cysA* и *cysT* у видов *Nephroselmis olivacea*, *Pycnococcus provasolii*, *Bryopsis hypnoides*, *Leptosira terrestris*, *Aneura mirabilis* и *Ptilidium pulcherrimum*; а также перед геном *cysA* у *Chlorokybus atmophyticus* и перед геном *cysT* у *Zygnema circumcarinatum*. Возможно, в этих случаях гены транскрибируются вместе с предыдущими генами или с помощью РНК-полимеразы фагового типа.

У многих видов вблизи промоторов перед генами *cysT* и *cysA* найден консервативный однобоксовый мотив с консенсусом TAAWATGATT, иногда повторяющийся дважды или даже трижды. Консенсус получен по двум генам, 28 сайтам из 9 видов: *Coccomyxa subellipsoidea*, *Chlorella variabilis*, *Chlorella vulgaris*, *Helicosporidium*, *Parachlorella kessleri*, *Mesostigma viride*, *Chlorokybus atmophyticus*, *Zygnema circumcarinatum*, *Anthoceros formosae*. У *C. subellipsoidea* C-169 перед промотором гена *cysA* расположен двукратный повтор последовательности с небольшими вариациями, включающей на 5'-конце предсказанный мотив, но с отклонением от консенсуса. Не удалось определить мотив вблизи промотора у *Chlorokybus atmophyticus* и печеночника *Marchantia polymorpha*. Отклонения от консенсуса часто одинаковы для разных сайтов внутри одного вида, что может отражать изменчивость транскрипционного фактора.

У большинства видов мотив расположен выше -35 бокса промотора или перекрывает его. В случае промоторов перед геном *cysA* из *Zygnema circumcarinatum* и *Anthoceros formosae* мотивы расположены между боксами промоторов или перекрывают -10 бокс промотора.

Позиционная сцепленность с промотором позволяет предположить, что найденный мотив является сайтом связывания транскрипционного фактора. Изменчивость расстояния между мотивом и промотором, а также близость боксов промотора к консенсусу, говорит в пользу того, что это – сайт связывания репрессора, а не активатора транскрипции. Повтор мотива характерен для кооперативного связывания нескольких экземпляров транскрипционного фактора, что может компенсировать отличие сайтов от консенсуса мотива, например у *Coccomyxa subellipsoidea*.

Изложенные в данном пункте сведения служат, в частности, независимым подтверждением корректности полученных алгоритмически кластеров CysA и CysT.

2.4. Кластеризация пластомных белков однодольных растений

Пластомы 36-ти видов однодольных получены из базы данных NCBI и перечислены в таблице 2.3. Для контроля результатов использовались пакет программ MEGA 5, [109], и база данных Pfam, [113].

Веб-интерфейс [115] обеспечивает для однодольных растений функциональность, описанную выше (пункт 2.2) для родофитной ветви.

Немного алгоритмически полученных кластеров были объединены из биологических соображений. Ниже перечислены все такие случаи.

Основной кластер PetG (cytochrome b6/f complex subunit V) объединён с кластером, состоящим из двух белков: YP_654227.1 (из *Oryza sativa* Indica Group) и YP_358627.1 (из *Phalaenopsis aphrodite* subsp. formosana).

Основной кластер RpL23 (ribosomal protein L23) пополнен двумя белками, образовавшими одноэлементные кластеры: YP_874745.1 (из *Agrostis stolonifera*) и YP_899416.1 (из *Sorghum bicolor*).

Основной кластер RpL2 (ribosomal protein L2) объединён с кластером, состоящим из двух паралогов: YP_654244.1 и YP_654261.1 (из *Oryza sativa* Indica Group).

Наилучших результатов удалось достичь при следующих значениях параметров: p=2, L=0, H=0.5. При указанных значениях и после трёх вышеописанных объединений образуется 105 неодноэлементных кластеров и 20 одноэлементных. Из неодноэлементных кластеров 71 содержит не более одного белка из каждого вида, 30 содержат пары белков из некоторых видов, 2 содержат тройки белка из некоторых видов и 2 содержат 4 белка из одного вида.

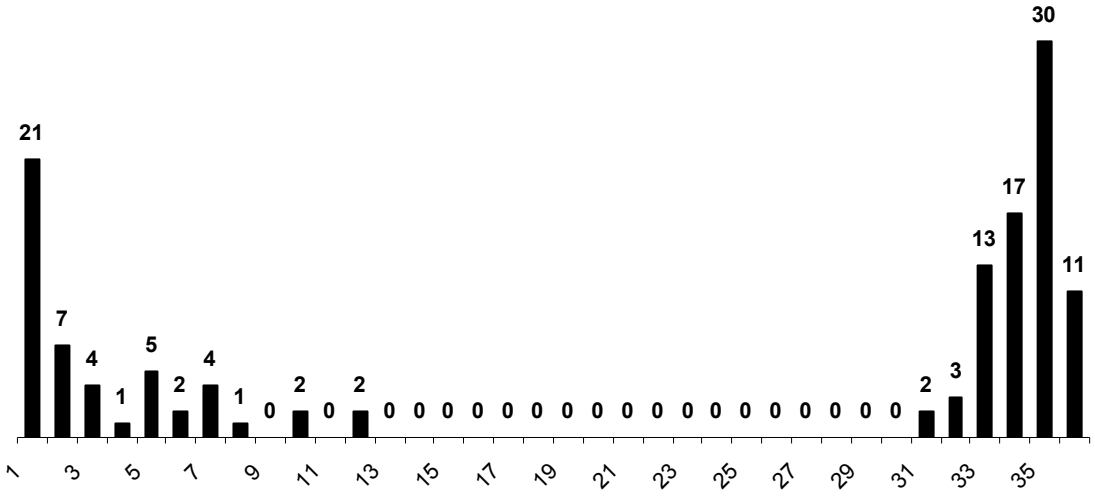


Рисунок 2.10. Распределение числа кластеров белков однодольных в зависимости от числа представленных в них видов

Распределение размеров кластеров (рисунок 2.10) заметно отличается от случаев родофитной и хлорофитной ветвей: в 29-ти (неодноэлементных) кластерах представлены от одного до 12-ти видов; нет кластеров, в которых представлены от 13-ти до 30-ти видов; в 76-ти кластерах представлены от 31-го до 36-ти видов (границы диапазонов везде включены). Максимум распределения – 30 кластеров с 35-ю видами. В таблице 2.5 перечислены все кластеры, за исключением образованных гипотетическими белками с неизвестной функцией.

2.5. Кластеризация пластомных белков цветковых растений

Описанная в пункте 2.4 кластеризация была расширена на белки всех доступных на момент исследования (конец 2012 года) в базе данных GenBank полных пластомов цветковых растений (186 видов).

В трёх случаях алгоритмически полученные кластеры были объединены из биологических соображений: белок YP_003934083.1 из *Geranium palmatum*, составлявший единичный кластер, был добавлен к кластеру AccD; белок YP_654227.1 из *Oryza sativa* Indica Group – к кластеру PetG; белки YP_874745.1 из *Agrostis stolonifera* и YP_899416.1 из *Sorghum bicolor* – к кластеру Rpl23.

Веб-интерфейс [116] обеспечивает для цветковых растений функциональность, описанную выше (пункт 2.2) для родофитной ветви.

Для контроля результатов использовались пакет программ MEGA 5, [109] и база данных Pfam, [113].

Кластеризация охватывает 15 507 белков, включает 165 кластеров, из них 122 содержат белки из двух и более различных пластид. Среди таких кластеров 39 содержат не более одного белка из каждого вида, 78 – содержат пары белков из одного вида, но не более двух белков из каждого вида, и 5 – содержат более двух белков из одного вида, но не более четырёх белков из каждого вида.

Размер кластера понимается как число различных видов, представленных в нём. Из 122-х кластеров, включающих белки из разных видов, 38 (31%) имеют размер меньше десяти, 12 (10%) имеют размер от 10-ти до 170-ти, и 72 (59%) имеют размер более 170-ти (т.е. охватывают более 90% исходных видов). Чаще других встречаются кластеры с размером 182 и 183 (по 15 кластеров каждого размера). Более трети неединичных кластеров имеют размер больше 180-ти, т.е. каждый из них содержит белки из более чем 97% рассмотренных видов. Распределение числа кластеров в зависимости от их размера n на рисунке 2.11.

Таблица 2.5. Перечень аннотированных кластеров пластидных белков однодольных растений. В таблицу не включены белки с неизвестной функцией. Кластер обозначается именем белка, который в него входит. Здесь разным белкам соответствуют разные кластеры.

Белок	Описание
AccD	acetyl-CoA carboxylase beta subunit
AtpA	ATP synthase CF1 alpha subunit
AtpB	ATP synthase CF1 beta subunit
AtpE	ATP synthase CF1 epsilon subunit
AtpF	ATP synthase CF0 B subunit
AtpH	ATP synthase CF0 C subunit
Atpl	ATP synthase CF0 A subunit
CcsA	cytochrome c biogenesis protein
CemA	envelope membrane protein
ClpP	ATP-dependent Clp protease proteolytic subunit
InfA	translation initiation factor 1
MatK	maturase K
NdhA	NADH dehydrogenase subunit 1
NdhB	NADH dehydrogenase subunit 2
NdhC	NADH dehydrogenase subunit 3
NdhD	NADH dehydrogenase subunit 4
NdhE	NADH dehydrogenase subunit 4L
NdhF	NADH dehydrogenase subunit 5
NdhG	NADH dehydrogenase subunit 6
NdhH	NADH dehydrogenase subunit 7
NdhJ	NADH dehydrogenase subunit J
NdhK	NADH dehydrogenase subunit K
NdhL	NADH dehydrogenase subunit I
PetA	cytochrome f
PetB	cytochrome b6
PetD	cytochrome b6/f complex subunit IV
PetG	cytochrome b6/f complex subunit V
PetL	cytochrome b6/f complex subunit VI
PetN	cytochrome b6/f complex subunit VIII
PsaA	photosystem I P700 chlorophyll a apoprotein A1
PsaB	photosystem I P700 chlorophyll a apoprotein A2
PsaC	photosystem I subunit VII
PsaJ	photosystem I subunit IX
Psal	photosystem I subunit VIII
PsbA	photosystem II protein D1
PsbB	photosystem II 47 kDa protein
PsbC	photosystem II 44 kDa protein
PsbD	photosystem II protein D2
PsbE	photosystem II protein V

Белок	Описание
PsbF	photosystem II protein VI
PsbH	photosystem II protein H
PsbJ	photosystem II protein J
PsbK	photosystem II protein K
Psbl	photosystem II protein I
PsbL	photosystem II protein L
PsbM	photosystem II protein M
PsbN	photosystem II protein N
PsbT	photosystem II protein T
PsbZ	photosystem II protein Z
RbcL	ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
RpL2	ribosomal protein L2
RpL14	ribosomal protein L14
RpL16	ribosomal protein L16
RpL20	ribosomal protein L20
RpL22	ribosomal protein L22
RpL23	ribosomal protein L23
RpL32	ribosomal protein L32
RpL33	ribosomal protein L33
RpL36	ribosomal protein L36
RpoA	RNA polymerase alpha subunit
RpoB	RNA polymerase beta subunit
RpoC1	RNA polymerase beta' subunit
RpoC2	RNA polymerase beta" chain
RpS2	ribosomal protein S2
RpS3	ribosomal protein S3
RpS4	ribosomal protein S4
RpS7	ribosomal protein S7
RpS8	ribosomal protein S8
RpS11	ribosomal protein S11
RpS12	ribosomal protein S12
RpS14	ribosomal protein S14
RpS15	ribosomal protein S15
RpS16	ribosomal protein S16
RpS18	ribosomal protein S18
RpS19	ribosomal protein S19
Ycf1	protein Ycf1
Ycf2	protein Ycf2
Ycf3	photosystem I assembly protein Ycf3
Ycf4	photosystem I assembly protein Ycf4

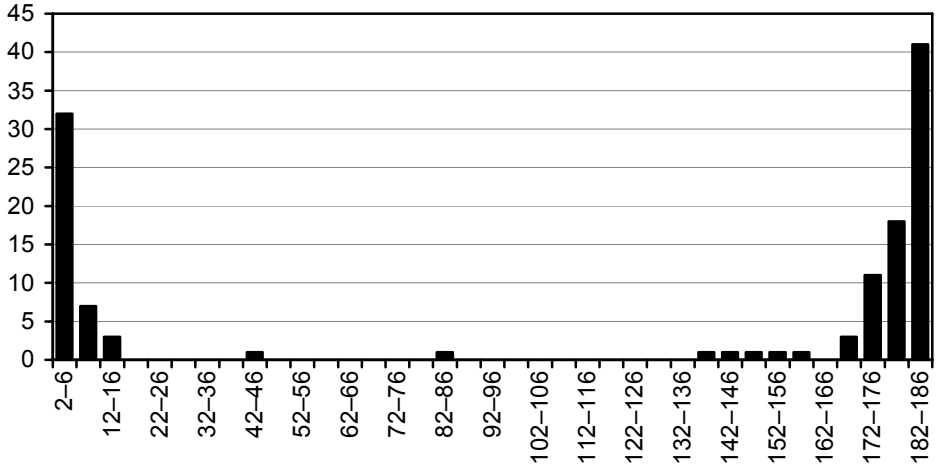


Рисунок 2.11. Распределение кластеров белков цветковых растений в зависимости от числа представленных в них видов

ГЛАВА 3. СОПРЯЖЕНИЕ ТРАНСЛЯЦИИ И ПРОЦЕССИНГА мРНК В ПЛАСТИДАХ

1. Введение и постановка задачи

Поиск сайтов связывания и регуляторных вторичных структур мРНК в не транслируемых участках – обширная область исследований, в качестве примеров отметим работы [117, 118]. Мы рассмотрим такой поиск в важном, но очень специальном случае сопряжения трансляции и процессинга мРНК в пластидах. Точнее, мы хотим ответить на вопрос: как может осуществляться задержка инициации трансляции до завершения процессинга.

В хлоропластах регуляция экспрессии генов может быть основана на связывании сайтов мРНК, кодируемой в пластидах, с белками, кодируемыми в ядре. Такая регуляция экспериментально установлена у нескольких водорослей и растений [119–121] и предсказана у большого их числа [122].

В хлоропластах трансляционный механизм близок к таковому у бактерий. В частности, элонгация рибосомы происходит непосредственно вслед за РНК-полимеразой, если только связыванию рибосомы не мешает какой-то механизм. Многие белок-кодирующие гены хлоропластов содержат интроны или нуждаются в редактировании. Рибосома не должна достигать интрона, причём в рассматриваемых генах хлоропластов первый экзон короткий, поэтому их трансляция не может начинаться сразу вслед за транскрипцией. В редких случаях задержка начала трансляции обеспечивается редактированием инициирующего кодона: AUG получается из ACG, [120]. Однако, например, пластом печёночника *Marchantia polymorpha* хорошо изучен и в нём отсутствует редактирование каких-либо мРНК, [123]. Это указывает на наличие других механизмов задержки инициации трансляции до завершения сплайсинга или редактирования. Такие механизмы были указаны в [122]: это сайты связывания белковых факторов или шпильки, которые обеспечивают перекрывание сайта связывания рибосомы. Поиску таких сайтов и шпилек посвящена эта короткая глава.

Для поиска мотива использовался стандартный алгоритмы MEME, [124] и оригинальный алгоритм поиска клики в многодольном графе, названный Clique. Второй из них является развитием алгоритма из [125], которое описано ниже. Напомним, что многодольным называется граф, множество вершин которого разбито на непересекающиеся подмножества, называемые долями, и ребра соединяют вершины только из разных долей. Алгоритм MEME при использовании параметров по умолчанию не определил мо-

тивы, предсказанные алгоритмом Clique, у видов, находящихся за пределами цветковых растений. Мотивы, предсказанные алгоритмом Clique, показаны на рисунках 3.1–3.7; сайты этих мотивов, соответствующие некоторым видам, приведены в таблице 3.2. Отрицательный результат, связанный с MEME, можно объяснить отсутствием среди его параметров числа видов, которые должны быть охвачены мотивом. Отметим, что эти алгоритмы основаны на совершенно разных принципах. В частности, в алгоритме Clique имеется параметр – размер клики, т.е. число сайтов (из разных последовательностей) в искомом мотиве. Не исключено, что вычисления можно организовать таким образом, что алгоритм MEME также найдёт мотивы, указанные на рисунках 3.1–3.7. Сравнение алгоритмов для поиска мотива не входило в задачу диссертанта.

Опишем упомянутое улучшение в алгоритме Clique; в остальном этот алгоритм описан в [125]. Фиксируем длину участка k и определим исходный многодольный граф, в котором ищутся клики заданного размера. В нём каждая доля соответствует одной из данных нуклеотидных последовательностей, а каждая вершина доли – участку этой последовательности с длиной k; любые две вершины из разных долей соединяются ребром, помеченным числом – сходством участков, приписанных вершинам ребра. Сходство отражает консервативность участков и устанавливается с учётом их GC-состава. А именно, *сходство участков* полагается равным сумме по позициям сходств соответствующих пар нуклеотидов в них, а сходство отдельных пар определяется таблицей 3.1. В ней p – средняя доля вхождений G или C в геномах всех данных последовательностей, тогда аналогичная средняя доля вхождений A или T равна $1–p$.

Таблица 3.1. Сходства пар нуклеотидов

	A	**C**	**G**	**T**
A	1	½	½	p
C	½	1	$1–p$	½
G	½	$1–p$	1	½
T	p	½	½	1

Следующее простое предложение относится к выбору сходства между парами нуклеотидных остатков в таблице 3.1: если p мало, то сходство A и T мало, а сходство C и G велико; если p велико, то наоборот. Действительно, редкое событие несёт больше информации.

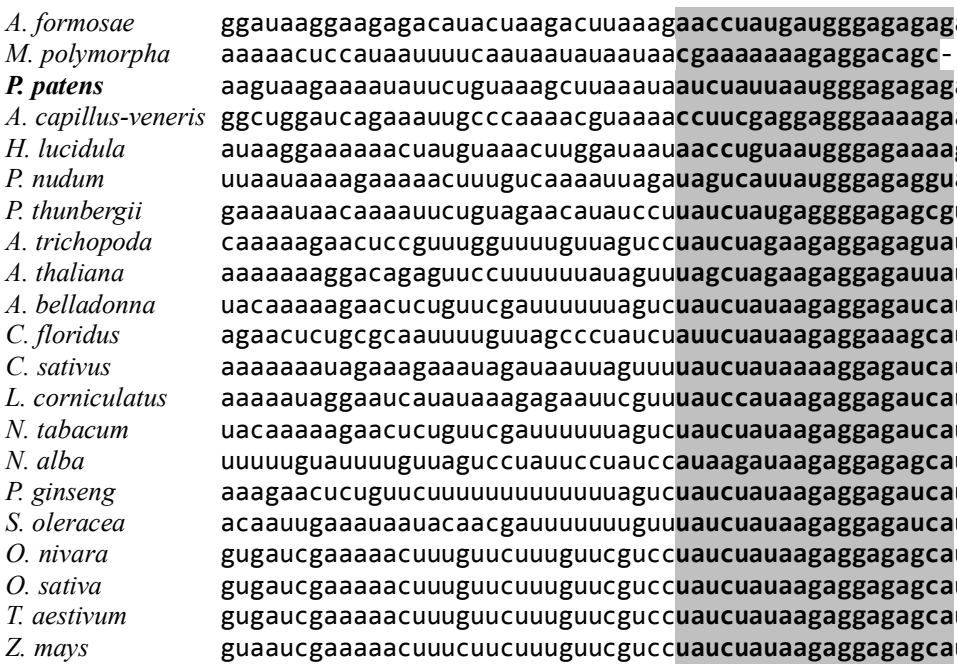


Рисунок 3.1. Выравнивание 5'-нетранслируемых участков перед геном *atpF*

Здесь и на следующих пяти рисунках серым показаны консервативные столбцы, которые образуют мотив

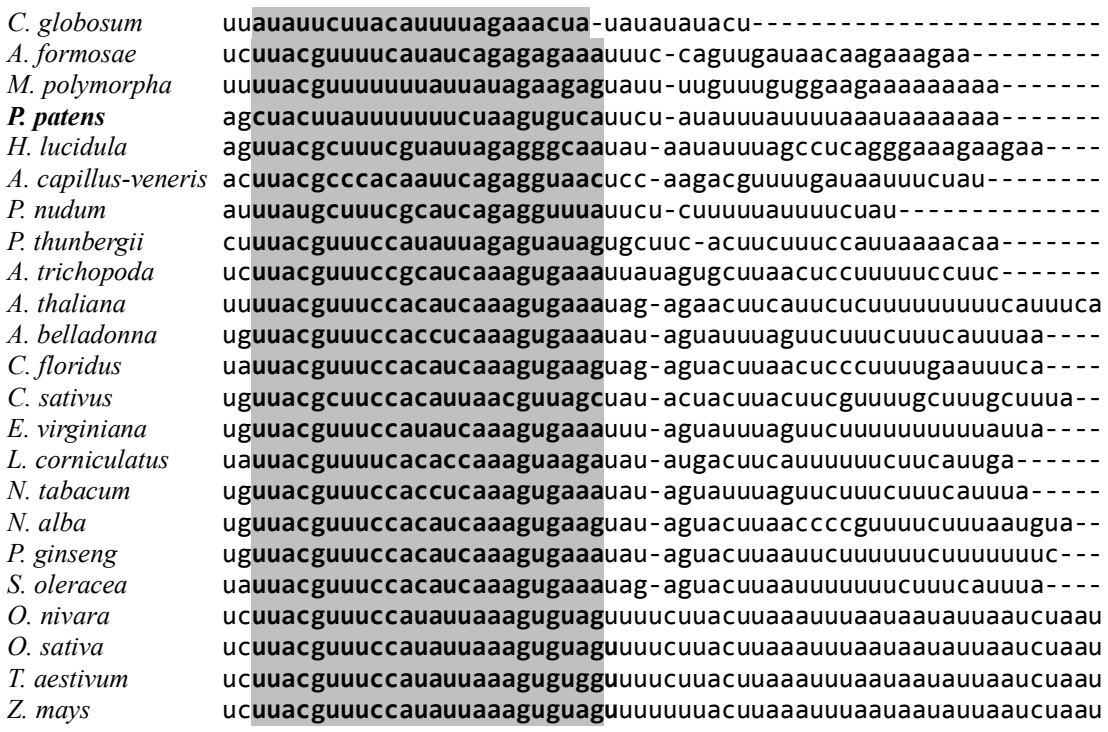


Рисунок 3.2. Выравнивание 5'-нетранслируемых участков перед геном *clpP*

Рисунок 3.3. Выравнивание 5'-нетранслируемых участков перед геном *petB*

Рисунок 3.4. Выравнивание 5'-нетранслируемых участков перед геном *psaA*

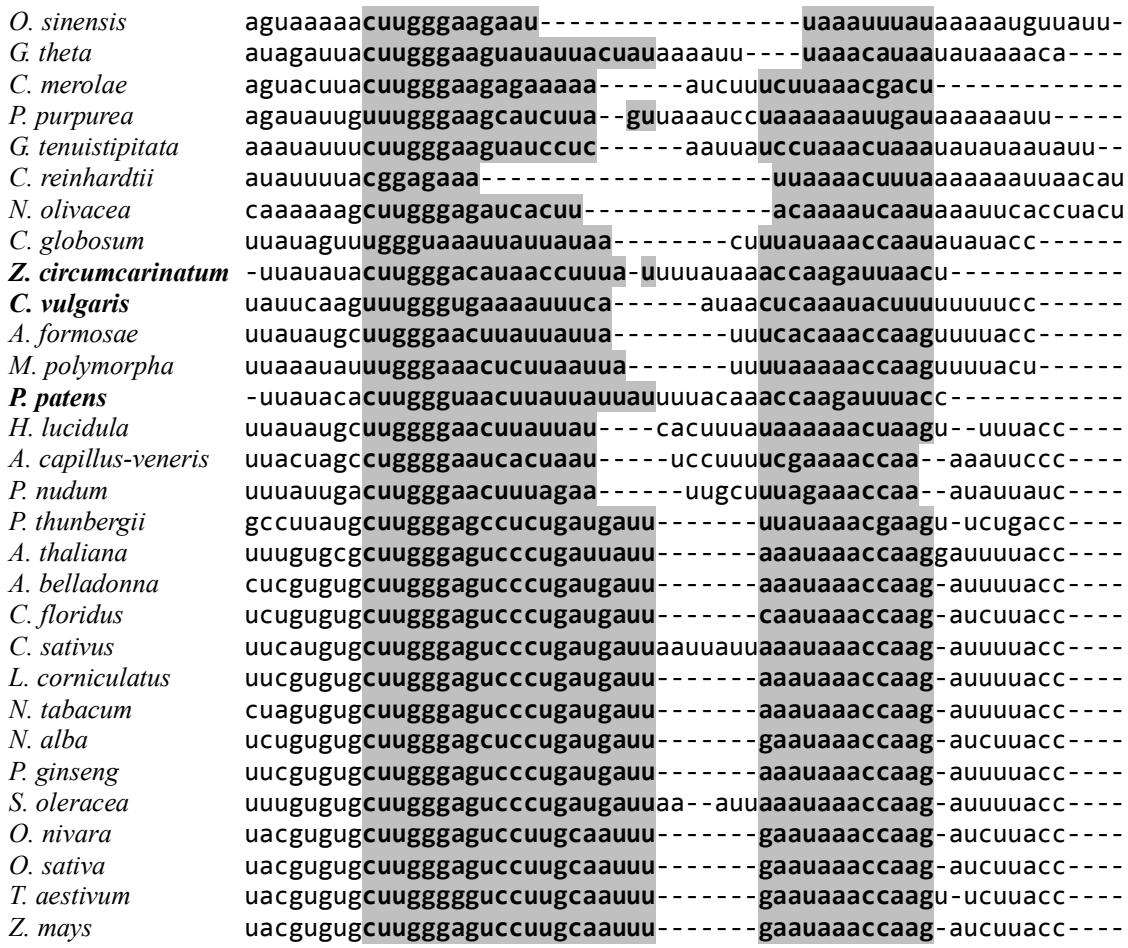


Рисунок 3.5. Выравнивание 5'-нетранслируемых участков перед геном *psbA*

Рисунок 3.6. Выравнивание 5'-нетранслируемых участков перед геном *psbB*

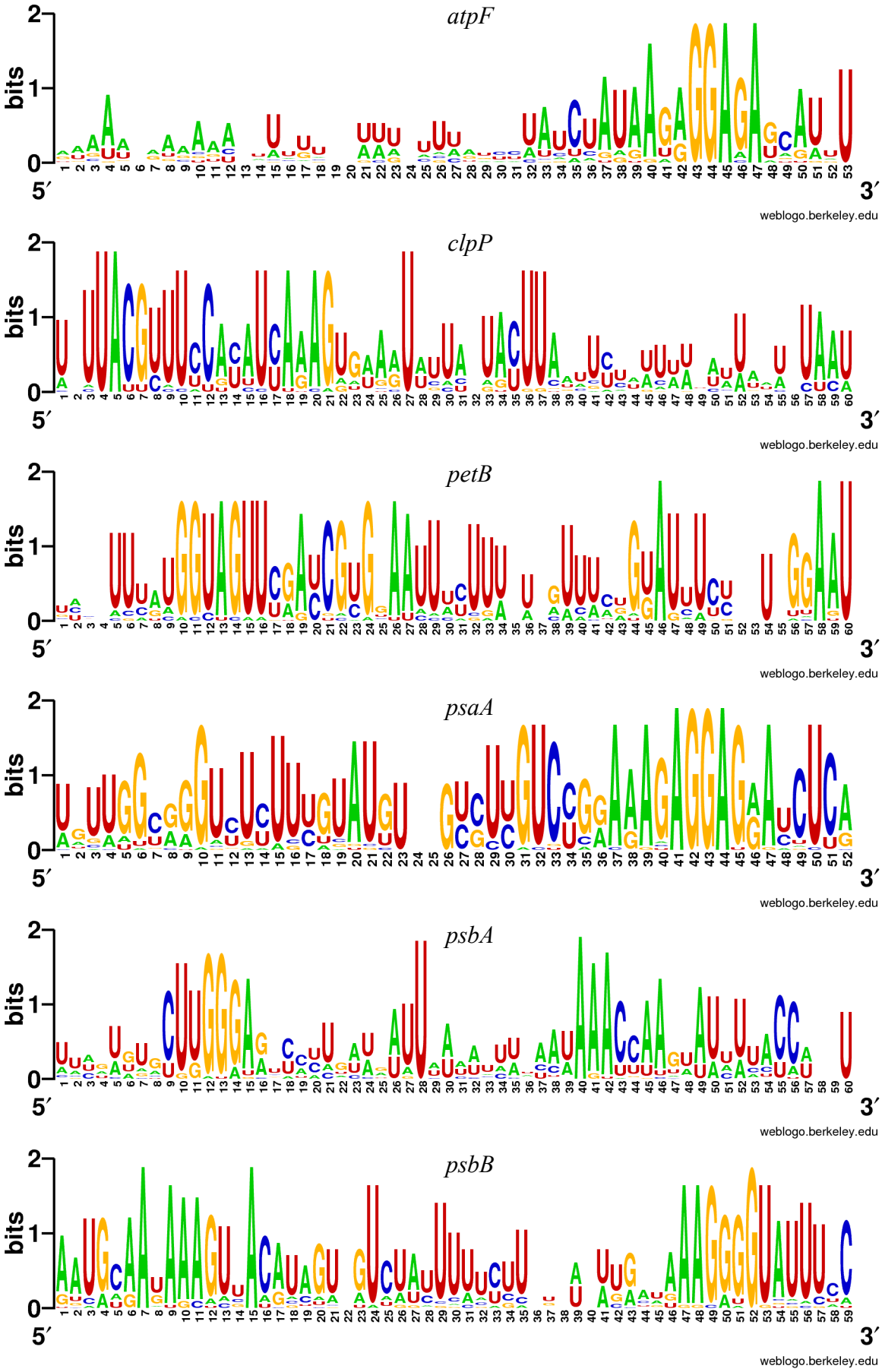


Рисунок 3.7. Диаграмы LOGO выравниваний, приведённых на рисунках 3.1–3.6

Предложение. Пусть $0 \le p < \frac{1}{2}$ и даны два случайных участка одинаковой длины в алфавите {A,C,G,T}, в которые буквы G и C входят с вероятностью $p/2$, а буквы A и T – с вероятностью $(1-p)/2$. Тогда в любой позиции этих участков вероятность появления пары {A,T} строго больше вероятности появления пары {G,C}. Если $\frac{1}{2} < p \le 1$, то выполнено противоположное неравенство.

Доказательство. Вероятность появления нуклеотида G или C в каждой позиции равна $p/2$. Поскольку участки независимы, вероятность появления в данной позиции G в первом участке и C во втором равна $p^2/4$. А вероятность появления пары {G, C} равна $p^2/2$. Аналогично, вероятность появления пары {A, T} равна $(1-p)^2/2$. Остаётся заметить, что при $p < \frac{1}{2}$ последнее выражение $(1-p)^2/2 = \frac{1}{2} - p + p^2/2 > p^2/2$ □

2. Материалы и методы

Геномы хлоропластов получены из базы данных GenBank (NCBI). В качестве исходного набора последовательностей были взяты 5'-нетранслируемые области перед генами *atpF* (субъединица АТФ-синтазы), *petB* (цитохром b6), *clpP* (протеолитическая субъединица АТФ-зависимой протеазы Clp), *psaA* (P700 фотосистемы I), *psbA* (белок D1 фотосистемы II) и *psbB* (P680 фотосистемы II), *accD* (бета субъединица ацелил-CoA карбоксилазы) и *atpH* (субъединица АТФ-синтазы) пластид у следующих 34 видов: *Cyanidioschyzon merolae, Cyanidium caldarium, Gracilaria tenuistipitata*, *Guillardia theta*, *Nephroselmis olivacea*, *Odontella sinensis*, *Porphyra purpurea*, *Chlamydomonas reinhardtii, Chaetosphaeridium globosum, Chara vulgaris, Mesostigma viride, Zygnema circumcarinatum, Anthoceros formosae, Adiantum capillus-veneris, Huperzia lucidula, Marchantia polymorpha, Physcomitrella patens, Psilotum nudum, Pinus thunbergii, Amborella trichopoda, Arabidopsis thaliana, Atropa belladonna, Calycanthus floridus, Cucumis sativus, Epifagus virginiana, Lotus corniculatus, Nicotiana tabacum, Nymphaea alba, Panax ginseng, Spinacia oleracea, Oryza nivara, Oryza sativa, Triticum aestivum, Zea mays*. Заметим, что *Epifagus virginiana* не является фотосинтезирующим видом: гены фотосистем в его пластидах отсутствуют.

Для поиска консервативных сайтов использовались программы MEME и Clique. Для определения вторичной структуры РНК и вычисления её энергии использовалась программа RNAstructure, являющаяся обновлением программы, описанной в [126]. Для контроля использовалась оригинальная программа, которая учитывает и вторичные структуры с псевдоузлами; однако она не привела к новым результатам на этих данных и потому здесь не описывается.

3. Результаты

Результаты приводятся только для видов *Chara vulgaris*, *Zygnema circumcarinatum*, *Physcomitrella patens*, так как для остальных видов они в основном предсказаны уже в [122]. Найденные нами выравнивания (рисунки 3.1–3.6) указали на сайты у трёх упомянутых видов *C. vulgaris*, *Z. circumcarinatum*, *P. patens* перед шестью генами: *atpF*, *clpP*, *petB*, *psaA*, *psbA*, *psbB*; эти сайты указаны в таблице 3.2. В таблице 3.3 приведены сведения о наличие сайтов и интронов, включая результаты из [122]. Эти сайты представляют интерес, так как, по крайней мере в части случаев, их присутствие коррелирует с наличием интронов. Вопрос об их функциональной роли в других случаях остаётся открытым.

Таблица 3.2. Консервативные сайты перед генами *atpF*, *clpP*, *petB*, *psaA*, *psbA*, *psbB* у *Chara vulgaris*, *Zygnema circumcarinatum* и *Physcomitrella patens*

***atpF*:**	
P. patens	aaguaagaaaauauucuguaaagcuuaaauaaucuau**uaa**ug**ggagag**agaau
***clpP*:**	
P. patens	agc**uac**u**u**au**u**uuu**u**u**uc**u**aagug**ucauucuauauuuauuuuaaauaaaaaaa
***petB*:**	
Z. circumcarinatum	uca**uuu**uac**ggu**ucaauugcgcaauu**auu**a---------cauuggagaauguuucaau--
C. vulgaris	aacu**uua**a**ugg**c**agu**u**u**ag**ucgug**aauaaaucaaauu-aaaauggagaaggauucguaau
P. patens	uuuacuaaa**ggu**a**gu**u**u**a**aucgug**uaa**uca---auua**auu**aaaggauuua**ugga**uu----
***psaA*:**	
C. vulgaris	au**u**auuucuag**c**aaa**uuu**u**u**cuuuug**c**u**c**g**uc**uaaagac**aggagaa**u**cuc**g
Ph.p.	**u**a**uugg**u**ggg**u**uuu**cc-**uaug**----c**u**c**guc**ugaagaga**gga**gaa**c**c**uc**g
***psbA*:**	
Z. circumcarinatum	**uu**auauacu**ugg**acau**a**accuu**u**a-uuuu**auaaaccaag**au**u**aacu-------
C. vulgaris	**u**auucaag**uuugg**u**g**aaaa**u**uuca------au**aac**ucaaauac**uu**uuuuuuucc
P. patens	**uu**auacacu**ugg**u**a**acuua**uu**a**u**uauuuua**caaaccaag**auu**uacc-------
***psbB*:**	
C. vulgaris	**a**ua**gcaagaaagu**c**a**uaaaa**u**aucaac**uugu**cua**u**gacaaaagg**u**g**u**cauuu**c**
P. patens	**aaugcaa**aaaaguuacauagucuuuaauucucuu-**uga**gaaagggguauuucc

Мы видим, что для трёх генов *atpF*, *clpP* и *petB* имеется высокая корреляция между присутствием интронов в гене и наличием найденных сайтов связывания. Для других трёх генов *psaA*, *psbA* и *psbB* корреляция отсутствует.

Затем были рассмотрены гены *accD* и *atpH*. У растений видов *Anthoceros formosae*, *Adiantum capillus-veneris*, *Huperzia lucidula*, *Psilotum nudum*, *Pinus thunbergii* в

лидерных областях этих генов консервативного сайта обнаружить не удалось. Поэтому предположено, что задержка, необходимая для выполнения редактирования, обеспечивается у них *неконсервативной* шпилечной структурой мРНК. Для её нахождения вычислялась свободная энергия шпилек (в ккал/моль) и определялась сама шпилька с наименьшей энергией на участке мРНК длиной в 40 нуклеотидов перед инициирующими кодонами генов *accD* и *atpH*.

Таблица 3.3. Распределение консервативных сайтов связывания белка перед шестью генами у хлоропластов всех перечисленных выше видов. Обозначения: в заголовках столбцов 3–8 указаны имена шести генов, внутри этих столбцов знак «+» означает наличие сайта, знак «–» – его отсутствие, знак «s» – соответствующий ген содержит интроны.

Отдел	Вид	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Bacillariophyta	*Odontella sinensis*	-	-	-	+	+	-
Cryptophyta	*Guillardia theta*	-	-	-	+	+	-
Rhodophyta	*Cyanidioschyzon merolae*	-	-	-	-	+	-
	Cyanidium caldarium	-	-	-	-	-	-
	Porphyra purpurea	-	-	-	+	+	+
	Gracilaria tenuistipitata	-	-	-	-	+	-
Chlorophyta	*Chlamydomonas reinhardtii*	-	-	-	-s	+s	-
	Nephroselmis olivacea	-	-	-	+	+	+
Streptophyta, водоросли	*Chaetosphaeridium globosum*	-	+s	-s	+	+	+
	Chara vulgaris	-s	-s	+s	+	+	+
	Mesostigma viride	-	-	-	+	-	-
	Zygnema circumcarinatum	-	-	+	-	+	-
Anthocerophyta	*Anthoceros formosae*	+s	+s	+s	+	+	+
Bryophyta	*Physcomitrella patens*	+s	+s	+s	+	+	+
Hepatophyta	*Marchantia polymorpha*	+s	+s	+s	+	+	+
Lycopodiophyta	*Huperzia lucidula*	+s	+s	+s	+	+	+
Pteridophyta	*Adiantum capillus-veneris*	+s	+s	-s	+	+	+
Psilophyta	*Psilotum nudum*	+s	+s	+s	+	+	+
Pinophyta	*Pinus thunbergii*	+s	+	+s	+	+	+
Magnoliophyta	разные	+s	+s	+s	+	+	+

В результате в 5'-нетранслируемых областях транскриптов этих генов у *Anthoceros formosae* и *Adiantum capillus-veneris* обнаружены шпильки большой длины с низкой энергией, перекрывающие сайт связывания рибосомы. Как известно, мРНК именно этих генов у этих растений редактируются [123]. Энергии найденных шпилек приведены в таблице 3.4, а сами шпильки показаны на рисунке 3.8.

Дополнительно к этому поиску длина участка, на котором искалась шпилька, варьировалась вплоть до 70 нуклеотидов; в результате новые шпильки с низкой энергией, перекрывающие сайт связывания рибосомы, не обнаружены.

Таблица 3.4. Наличие мощной шпильки перед генами *accD* и *atpH*. Обозначения: в столбцах 2 и 4 приведена свободная энергия для каждой из найденных шпилек в ккал/моль на участке мРНК длиной в 40 нуклеотидов перед инициирующими кодонами генов *accD* и *atpH*. В столбцах 3 и 5 знак «+» показывает, что мРНК в соответствующем виде подвергается редактированию, знак «–» – не подвергается.

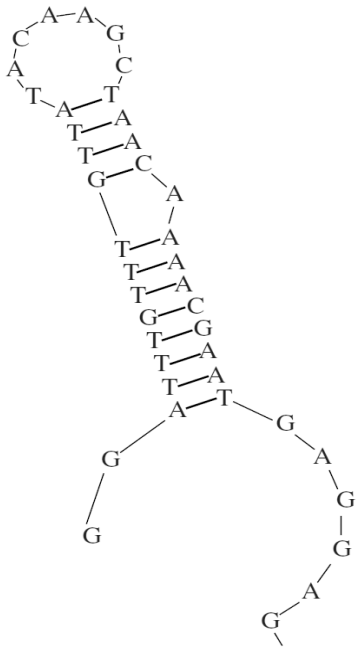
Вид	Ген *accD*		Ген *atpH*	
1	2	3	4	5
Anthoceros formosae	-7.0	+	-5.1	+
Adiantum capillus-veneris	-7.2	+	-5.2	+
Huperzia lucidula	-4.8	–	-2.9	–
Psilotum nudum	-0.8	–	-2.9	–
Pinus thunbergii	-3.6	–	-2.8	–

4. Обсуждение

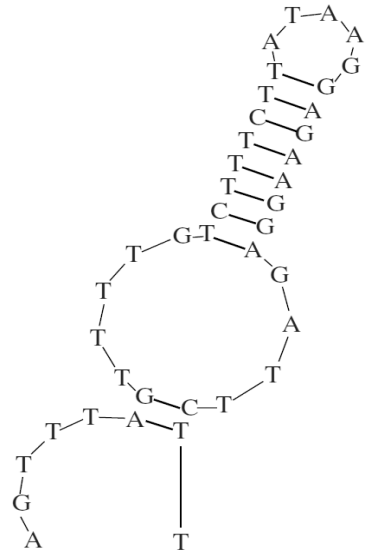
В пластидах растений родов *Anthoceros* и *Adiantum* мРНК *accD* и *atpH* редактируются [123]. Происходит значительное изменение кодонов этих генов, так что правильная трансляция до завершения редактирования невозможна.

У *Anthoceros formosae*, *Adiantum capillus-veneris*, *Huperzia lucidula*, *Psilotum nudum*, *Pinus thunbergii* ген, непосредственно предшествующий *accD*, кодирует тРНК; лидерные области этого гена примерно одинаковой длины. Это указывает на отсутствие крупномасштабных перестроек хромосомы перед этим геном у рассмотренных видов, что делает их естественной группой для изучения механизма задержки редактирования мРНК *accD*. Напротив, у цветковых растений непосредственный предшественник *accD* – ген *rbcL*, последний кодирует белок, а не тРНК. Такая перестройка хромосомы затрудняет сопоставление 5'-лидерных областей *accD* у пяти упомянутых видов (таблица 3.4), и у цветковых растений. Поэтому в связи с регуляцией *accD* (и также *atpH*) последние не рассматривались.

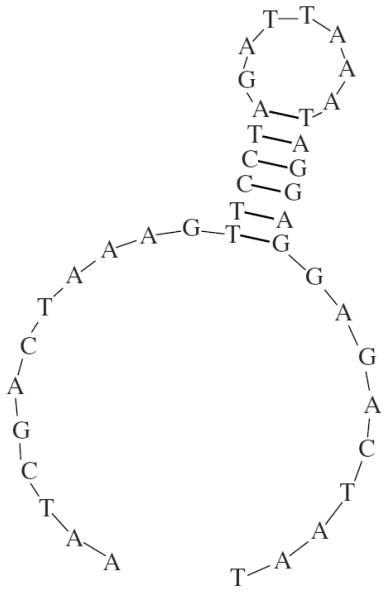
В 5'-нетранслируемых областях мРНК *accD* и *petH* у *Anthoceros* и *Adiantum* присутствуют длинные шпильки с низкой энергией, и одновременно эти мРНК у этих родов редактируются. Низкая энергия этих шпилек обеспечивает их стабильность в течение длительного времени, и гипотеза состоит в том, что они препятствуют началу трансляции до завершения редактирования. В таблице 3.4 указаны свободные энергии на том же участке мРНК у секвенированных представителей всех пяти родов.



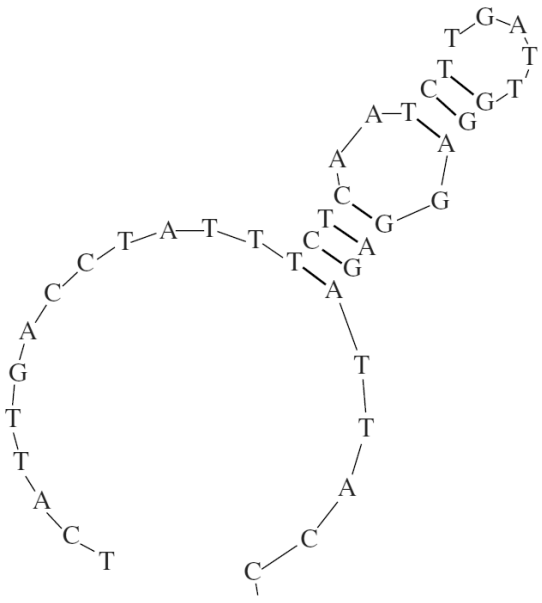
a) *accD*, *Adiantum capillus-veneris*



b) *accD*, *Anthoceros formosae*



c) *atpH*, *Adiantum capillus-veneris*



d) *atpH*, *Anthoceros formosae*

Рисунок 3.8. Вторичные структуры мРНК перед генами *accD* и *atpH* у *Adiantum capillus-veneris* и *Anthoceros formosae*, предположительно перекрывающие сайт связывания рибосомы

У *Huperzia lucidula* редактирование этих двух генов отсутствует, и одновременно, хотя у неё имеется шпилька с достаточно низкой энергией, сайт связывания рибосомы (RBS) перед геном *accD* находится в петле этой шпильки, а все другие шпильки на этом участке имеют значительно большие значения энергии: большие чем –1.7 ккал/моль. Для оставшихся двух строк таблицы 3.4 шпильки с низкой энергией на этом участке отсутствуют и одновременно эти мРНК не редактируются. Найденные шпильки

у *Anthoceros* и *Adiantum* в 5'-нетранслируемых областях мРНК *accD* и *atpH* в пластидах не консервативны; они привязаны к редактированию этих мРНК – событию, которое редко встречается у близких видов [123].

Вывод о перекрывании RBS сайтом связывания регуляторного белка основан на том, что консервативный участок перед геном *atpF* имеет значительно большую протяжённость и включает AG-богатый мотив, характерный для RBS. Предполагаемый регуляторный белок, взаимодействуя с рибосомой, препятствует инициации трансляции, обеспечивает её задержку до завершения сплайсинга.

Перед геном *petB* отсутствует типичный RBS, но консервативная спираль РНК, может быть связана с процессингом 5'-лидерной области мРНК. В рассмотренных видах ген *petB* имеет интроны, если и только если (за одним исключением) присутствуют консервативные участки (сайты или спирали), что позволяет предположить: этот процессинг обеспечивает задержку начала трансляции до завершения сплайсинга. Исключение составляет *Adiantum*, у которого отсутствует консервативный участок, но тогда задержка инициации трансляции может объясняться редактированием инициирующего кодона мРНК.

Трансляционная регуляция гена *psbA* экспериментально изучена у *Chlamydomonas reinhardtii*, где транскрипция происходит конститутивно, в то время как трансляция активируется на свету белком 47 кДа, который образует комплекс с другими белками, непосредственно не связанными с мРНК, [119]. Этот комплекс разрушается в темноте. Можно думать, что найденный нами сайт связывает этот комплекс ортологичных белков.

Консервативные участки в 5'-лидерных областях генов *psbA* и *psaA* найдены почти перед всеми их ортологами, в том числе не содержащими интронов. Это указывает: найденная нами регуляция этих генов возникла до появления интронов. Можно предположить, что интроны в этих генах возникли потому, что ранее сформировалась задержка трансляции и в результате не было препятствий для протекания сплайсинга.

Консервативные участки в 5'-нетранслируемых областях мРНК *petB*, *clpP*, *psbA* и *psaA* содержат шпильки, окружённые консервативными нуклеотидами, что характерно для многих регуляторных систем у бактерий [127]. Отметим, что для гена *ycf3* это не так: он содержит интроны, имеет длинную 5'-лидерную область, но в ней консервативный участок отсутствует.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Altenhoff A.M., Dessimoz C. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods // *PLoS Computational Biology*. 2009. Vol. 5, № 1. e1000262.

2 Waterhouse R.M., Zdobnov E.M., Tegenfeldt F., Li J., Kriventseva E.V. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs // *Nucleic Acids Research*. 2013. Vol. 41. P. 358–365.

3 Электронный ресурс http://orthomcl.cbil.upenn.edu/.

4 Электронный ресурс http://www.genedb.org/

5 Электронный ресурс http://roundup.hms.harvard.edu/browse/.

6 Электронный ресурс http://inparanoid.sbc.su.se/.

7 Электронный ресурс http://www.omabrowser.org/.

8 Электронный ресурс http://eggnog.embl.de/.

9 Электронный ресурс http://www.ncbi.nlm.nih.gov/COG/.

10 Kang D., Kim S.H., Hamasaki N. Mitochondrial transcription factor A (TFAM): roles in maintenance of mtDNA and cellular functions // *Mitochondrion*. 2007. Vol. 7. P. 39–44.

11 Bogenhagen D.F. Interaction of mtTFB and mtRNA polymerase at core promoters for transcription of *Xenopus laevis* mtDNA // *The Journal of Biological Chemistry*. 1996. Vol. 271. P. 12036–12041.

12 De Virgilio C., Pousis C., Bruno S., Gadaleta G. New isoforms of human mitochondrial transcription factor A detected in normal and tumoral cells // *Mitochondrion*. 2001. Vol. 11. P. 287–295.

13 Asin-Cayuela J., Gustafsson C.M. Mitochondrial transcription and its regulation in mammalian cells // *Trends in Biochemical Sciences*. 2007. Vol. 32. P. 111–117.

14 Ma N., McAllister W.T. In a head-on collision, two RNA polymerases approaching one another on the same DNA may pass by one another // *Journal of Molecular Biology*. 2009. Vol. 391. P. 808–812.

15 Liere K., Maliga P. In vitro characterization of the tobacco *rpoB* promoter reveals a core sequence motif conserved between phage-type plastid and plant mitochondrial promoters // *EMBO Journal*. 1999. Vol. 18. P. 249–257.

16 Datta K., Johnson N.P., Hippel P.H. Mapping the conformation of the nucleic acid framework of the T7 RNA polymerase elongation complex in solution using low-energy CD and fluorescence spectroscopy // *Journal of Molecular Biology*. 2006. Vol. 360. P. 800–813.

17 Jeruzalmi D., Steitz T.A. Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme // *EMBO Journal*. 1998. Vol. 17. P. 4101–4113.

18 Chang D.D., Clayton D.A. Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA // *Cell*. 1984. Vol. 36. P. 635–643.

19 Martin M., Cho J., Cesare A.J., Griffith J.D., Attardi G. Termination factor-mediated DNA loop between termination and initiation sites drives mitochondrial rRNA synthesis // *Cell*. 2005. Vol. 123. P. 1227–1240.

20 Pham X.H., Farge G., Shi Y., Gaspari M., Gustafsson C.M., Falkenberg M. Conserved sequence Box II directs transcription termination and primer formation in mitochondria // *The Journal of Biological Chemistry*. 2006. Vol. 281. P. 24647–24652.

21 Bogenhagen D.F., Applegate E.F., Yoza B.K. Identification of a promoter for transcription of the heavy strand of human mtDNA: In vitro transcription and deletion mutagenesis // *Cell*. 1984. Vol.36. P. 1105–1113.

22 Enríquez J.A., Fernández-Silva P., Garrido-Pérez N., López-Pérez M.J., Pérez-Martos A., Montoya J. Direct regulation of mitochondrial RNA synthesis by thyroid hormone // *Molecular and Cellular Biology*. 1999. Vol. 19. P. 657–670.

23 Bogenhagen D.F., Yoza B.K. Accurate in vitro transcription of Xenopus laevis mitochondrial DNA from two bidirectional promoters // *Molecular and Cellular Biology*. 1986. Vol. 6. P. 2543–2550.

24 Bogenhagen D.F., Yoza B.K., Cairns S.S. Identification of initiation sites for transcription of *Xenopus laevis* mitochondrial DNA // *The Journal of Biological Chemistry*. 1986. Vol. 261. P. 8488–8494.

25 Bogenhagen D.F., Romanelli M.F. Template sequences required for transcription of *Xenopus laevis* mitochondrial DNA from two bidirectional promoters // *Molecular and Cellular Biology*. 1988. Vol. 8. P. 2917–2924.

26 Shen E.L., Bogenhagen D.F. Developmentally-regulated packaging of mitochondrial DNA by the HMG-box protein mtTFA during *Xenopus oogenesis* // *Nucleic Acids Research*. 2001. Vol. 29. P. 2822–2828.

27 Ammini C.V., Hauswirth W.W. Mitochondrial gene expression is regulated at the level of transcription during early embryogenesis of *Xenopus laevis* // *The Journal of Biological Chemistry*. 1999. Vol. 274. P. 6265–6271.

28 Shock L.S., Thakkar P.V., Peterson E.J., Moran R.G., Taylor S.M. DNA methyltransferase 1, cytosine methylation, and cytosine hydroxymethylation in mammalian mitochondria // *Proc. Natl. Acad. Sci. U.S.A.* 2011. Vol. 108. P. 3630–3635.

29 Wanrooij P.H., Uhler J.P., Simonsson T., Falkenberg M., Gustafsson C.M. G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation // *Proc. Nat. Acad. Sci. U.S.A.* 2010. Vol. 107. P. 16072–16077.

30 Bogenhagen D.F., Morvillo M.V. Mapping light strand transcripts near the origin of replication of *Xenopus laevis*; mitochondrial DNA // *Nucleic Acids Research*. 1990. Vol. 18. P. 6377–6383.

31 Chomyn A., Martinuzzi A., Yoneda M., Daga A., Hurko O. *et al.* MELAS mutation in mtDNA binding site for transcription termination factor causes defects in protein synthesis and in respiration but no change in levels of upstream and downstream mature transcripts // *Proc. Nat. Acad. Sci. U.S.A.* 1992. Vol. 89. P. 4221–4225.

32 Valverde J.R., Marco R., Garesse R. A conserved heptamer motif for ribosomal RNA transcription termination in animal mitochondria // *Proc. Natl. Acad. Sci. U.S.A.* 1994. Vol. 91. P. 5368–5371.

33 Gelfand R., Attardi G. Synthesis and turnover of mitochondrial ribonucleic acid in HeLa cells: the mature ribosomal and messenger ribonucleic acid species are metabolically unstable // *Molecular and Cellular Biology*. 1981. Vol. 1. P. 497–511.

34 Piechota J., Tomecki R., Gewartowski K., Szczęsny R., Dmochowska A. *et al.* Differential stability of mitochondrial mRNA in HeLa cells // *Acta Biochimica Polonica*. 2006. Vol. 3. P. 157–168.

35 Селиверстов А.В., Лысенко Е.А., Любецкий В.А. Быстрая эволюция промоторов пластомных генов *ndhF* у цветковых растений // *Физиология растений*. 2009. Т. 56. С. 926–934.

36 Lysenko E.A. Plant sigma factors and their role in plastid transcription // *Plant Cell Reports*. 2007. Vol. 26. P. 845-859.

37 Lyubetsky V.A., Rubanov L.I., Seliverstov A.V. Lack of conservation of bacterial type promoters in plastids of Streptophyta // *Biology Direct*. 2010. Vol. 5, P. 34.

38 Миронов А.А., Кистлер А.Э. Теоретический анализ кинетики образования вторичной структуры РНК в процессе транскрипции и трансляции. Учёт дефектных спиралей // *Молекулярная биология*. 1985. Т. 19. С. 1350–1357.

39 Danilova L.V., Pervouchine D.D., Favorov A.V., Mironov A.A. RNAKINETICS: A web server that models secondary structure kinetics of an elongating RNA // *Journal of Bioinformatics and Computational Biology*. 2006. Vol. 4, № 2. P. 589–596.

40 Lyubetsky V.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling classic attenuation regulation of gene expression in bacteria // *Journal of Bioinformatics and Computational Biology*. 2007. Vol. 5. P. 155–180.

41 Dodd I.B., Shearwin K.E., Sneppen K. Modelling Transcriptional Interference and DNA Looping in Gene Regulation // *Journal of Molecular Biology*. 2007. Vol. 369. P. 1200–1213.

42 Sneppen K., Dodd I.B., Shearwin K.E., Palmer A.C., Schubert R.A., Callen B.P., Egan J.B. A Mathematical Model for Transcriptional Interference by RNA Polymerase Traffic in Escherichia coli // *Journal of Molecular Biology*. 2005. Vol. 346. P. 399–409.

43 Palmer A.C., Ahlgren-Berg A., Egan J.B., Dodd I.B., Shearwin K.E. Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter // *Molecular Cell*. 2009. Vol. 34, № 5. P. 545–555.

44 Elias-Arnanz M., Salas M. Resolution of head-on collisions between the transcription machinery and bacteriophage Φ29 DNA polymerase is dependent on RNA polymerase translocation // *The EMBO Journal*. 1999. Vol. 18, № 20. P. 5675–5682.

45 Favory J.-J., Kobayshi M., Tanaka K., Peltier G., Kreis M., Valay J.-G., Lerbs-Mache S. Specific function of a plastid sigma factor for *ndhF* gene transcription // *Nucleic Acids Research*. 2005. Vol. 33. P. 5991–5999.

46 Zghidi W., Merendino L., Cottet A., Mache R., Lerbs-Mache S. Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the *psbN* gene in plastids // *Nucleic Acids Research*. 2007. Vol. 35. P. 455–464.

47 Зубо Я.О., Лысенко Е.А., Алейникова А.Ю., Кузнецов В.В., Пшибытко Н.Л. Изменение транскрипционной активности генов пластома ячменя в условиях теплового шока // *Физиология растений*. 2008. Т. 55. С. 323–331.

48 Swiatecka-Hagenbruch M., Emanuel C., Hedtke B., Liere K., Börner T. Impaired function of the phage-type RNA polymerase RpoTp in transcription of chloroplast genes is compensated by a second phage-type RNA polymerase // *Nucleic Acids Research*. 2008. Vol. 36. P. 785–792.

49 Homann A., Link G. DNA-binding and transcription characteristics of three cloned sigma factors from mustard (*Sinapis alba L.*) suggest overlapping and distinct roles in plastid gene expression // *European Journal of Biochemistry*. 2003. Vol. 270. P. 1288–1300.

50 Swiatecka-Hagenbruch M., Liere K., Börner T. High diversity of plastidial promoters in *Arabidopsis thaliana* // *Molecular Genetics and Genomics*. 2007. Vol. 277. P. 725–734.

51 Westhoff P., Herrmann R.G. Complex RNA maturation in chloroplasts. The *psbB* operon from spinach // *European Journal of Biochemistry*. 1988. Vol. 171. P. 551–564.

52 Hoffer P.H., Christopher D.A. Structure and blue-light-responsive transcription of a chloroplast *psbD* promoter from *Arabidopsis thaliana* // *Plant Physiology*. 1997. Vol. 115. P. 213–222.

53 Электронный ресурс http://www.ncbi.nlm.nih.gov/genbank/.

54 Ahn J., Kraynov V.S., Zhong X., Werneburg B.G., Tsai M.D. DNA polymerase beta: effects of gapped DNA substrates on dNTP specificity, fidelity, processivity and conformational changes // *Biochemical Journal*. 1998. Vol. 331. P. 79–87.

55 Lyubetsky V.A., Zverkov O.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase competition: the effect of σ-subunit knockout and heat shock on gene transcription level // *Biology Direct*. 2011. Vol. 6, Iss. 3. P. 1–16.

56 Любецкая Е.В., Селиверстов А.В., Любецкий В.А. У актинобактерий число длинных шпилек в межгенных трейлерных областях велико по сравнению с другими областями генома // *Молекулярная биология*. 2007. Т. 41, № 4. С. 739–742.

57 Abbondanzieri E.A., Shaevitz J.W., Block S.M. Picocalorimetry of transcription by RNA polymerase // *Biophysical Journal: Biophysical Letters*. 2005. Vol. 89. P. 61–63.

58 Ryals J., Little R., Bremer H. Temperature dependence of RNA synthesis parameters in *Escherichia coli* // *Journal of Bacteriology*. 1982. Vol. 151. P. 879–887.

59 Gotta S.L., Miller O.L., French S.L. rRNA transcription rate in *Escherichia coli* // *Journal of Bacteriology*. 1991. Vol. 173, P. 6647–6649.

60 Ederth J., Artsimovitch I., Isaksson L.A., Landick R. The downstream DNA jaw of bacterial RNA polymerase facilitates both transcriptional initiation and pausing // *The Journal of Biological Chemistry*. 2002. Vol. 277. P. 37456–37463.

61 Johnson R.S., Strausbauch M., Cooper R., Register J.K. Rapid kinetic analysis of transcription elongation by *Escherichia coli* RNA polymerase // *Journal of Molecular Biology*. 2008. Vol. 381, P. 1106–1113.

62 Северинов К.В. Структурно-функциональные исследования взаимодействий ДНК-зависимой РНК-полимеразы бактерий с промоторами // Диссертация в форме научного доклада на соискание учёной степени доктора биологических наук. Москва, 2006. 43 с.

63 Neidhardt F.C., Curtiss R.I., Gross C.A., Ingraham J.L., Lin E.C.C. *et al. Escherichia coli and Salmonella typhimurium: cellular and molecular biology*. Vol. 1 // American Society for Microbiology, Washington, D.C, 1987.

64 Seliverstov A.V., Lysenko E.A., Lyubetsky V.A. Rapid evolution of promoters in Magnoliophyta chloroplasts // *Proceedings of Computational Phylogenetics and Molecular Systematics: CPMS'2007*. Moscow: KMK Scientific Press, 2007. P. 286–292.

65 Quandt D., Müller K., Huttunen S. Characterisation of the chloroplast DNA *psbT-H* region and the influence of dyad symmetrical elements on phylogenetic reconstructions // *Plant Biology (Stuttgart)*. 2003. Vol. 5. P. 400–410.

66 Тейлор Дж.Р. *Введение в теорию ошибок* // М.: Мир, 1985. 272 стр. (Taylor J.R. *An introduction to error analysis* // 1982. Univ. Science Books Mill Valley, Calif.)

67 Cooper G.M. *The Cell: A Molecular Approach*. 2nd edition // Sunderland: Sinauer Associates, 2000.

68 Camasamudram V., Fang J.-K., Avadhani N.G. Transcription termination at the mouse mitochondrial H-strand promoter distal site requires an A/T rich sequence motif and sequence specific DNA binding proteins // *European Journal of Biochemistry*. 2003. Vol. 270. P. 1128–1140.

69 Электронный ресурс http://lab6.iitp.ru/ru/rivals/.

70 Электронный ресурс http://www.jscc.ru/.

71 Singer M., Berg P. *Genes & Genomes* // MillValley: University Science Books, 1991.

72 Lyubetsky V.A., Zverkov O.A., Pirogov S.A., Rubanov L.I., Seliverstov A.V. Modeling RNA polymerase interaction in mitochondria of chordates // *Biology Direct*. 2012. Vol. 7, Iss. 26. P. 1–16.

73 van Dongen S., Abreu-Goodger C. Using MCL to extract clusters from networks // *Methods in Molecular Biology*. 2012. Vol. 804. P. 281–295.

74 Strassen V. Gaussian elimination is not optimal // *Numerische Mathematik*. 1969. Vol. 13. P. 354–356.

75 Coppersmith D., Winograd S. Matrix multiplication via arithmetic progressions // *Journal of Symbolic Computation*. 1990. Vol. 9. P. 251–280.

76 Vilella A.J., Severin J., Ureta-Vidal A., Heng L., Durbin R., Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates // *Genome Research*. 2009. Vol. 19, № 2. P. 327–335.

77 Wallace I.M., O'Sullivan O., Higgins D.G., Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee // *Nucleic Acid Research*. 2006. Vol. 34, № 6. P. 1692– 1699.

78 Katoh K., Standley D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability // *Molecular Biology and Evolution*. 2013. Vol. 30, № 4. P. 772–780.

79 Galashov A.E., Kel'manov A.V. A 2-approximate algorithm to solve one problem of the family of disjoint vector subsets // *Automation and Remote Control*. 2014. Vol. 75, № 4, P. 595–606.

80 Кельманов А.В., Романченко С.М. FPTAS для одной задачи поиска подмножества векторов // *Дискретный анализ и исследование операций*. 2014. Т. 21, № 3. С. 41–52.

81 Lemieux C., Otis C., Turmel M. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies // *BMC Biology*. 2007. Vol. 5, № 2. P. 1–17.

82 Imanian B., Pombert J.-F., Keeling P.J. The complete plastid genomes of the two 'Dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum* // *PLoS ONE*. 2010. Vol. 5, № 5. e10711.

83 Балашов Ю.С. *Иксодовые клещи – паразиты и переносчики инфекций*. СПб.: Наука, 1998.

84 Brayton K.A., Lau A.O.T., Herndon D.R., Hannick L., Kappmeyer L.S. *et al.* Genome sequence of *Babesia bovis* and comparative analysis of *Apicomplexan Hemoprotozoa* // *PLoS Pathogens*. 2007. Vol. 3. e148.

85 Wilson R.J.M., Rangachari K., Saldanha J.W., Rickman L., Buxton R.S., Eccleston J.F. Parasite plastids: maintenance and functions // *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2003. Vol. 358, P. 155–164.

86 Zhu G., Marchewka M.J., Keithly J.S. *Cryptosporidium parvum* appears to lack a plastid genome // *Microbiology*. 2000. Vol. 146. P. 315–321.

87 Садовская Т.А., Селиверстов А.В. Анализ 5'-лидерных областей некоторых генов пластид у простейших типа Apicomplexa и у красных водорослей // *Молекулярная биология*. 2009. Т. 43, № 4. С. 599–604.

88 Селиверстов А.В., Любецкий В.А. Эволюция РНК-полимераз и их промоторов в пластидах // *Юбилейная конференция 50лет ИППИ РАН*. Москва. 2011. С. 58–62.

89 Kühn K., Bohne1 A.-V., Liere K., Weihe A., Börner T. *Arabidopsis* phage-type RNA polymerases: accurate in vitro transcription of organellar genes // *The Plant Cell*. 2007. Vol. 19. P. 959–971.

90 Электронный ресурс http://lab6.iitp.ru/ppc/redline/.

91 Lü F., Xü W., Tian C., Wang G., Niu J., Pan G., Hu S. The *Bryopsis hypnoides* plastid genome: multimeric forms and complete nucleotide sequence // *PLoS ONE*. 2001. Vol. 6, № 2. e14663.

92 Turmel M., Otis C., Lemieux C. The chloroplast genomes of the green algae *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitaria* reveal a shared ancestry between the Pedinomonadales and Chlorellales // *Molecular Biology and Evolution*. 2009. Vol. 26, № 10. P. 2317–2331.

93 Brouard J.S., Otis C., Lemieux C., Turmel M. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae // *Genome Biology and Evolution*. 2010. Vol. 2. P. 240–256.

94 Карлов А.С. Взаимодействие зоохлорелл с новым потенциальным хозяином – крупными свободно живущими амёбами // *Цитология*. 1992. Т. 34, № 4. С. 73.

95 Hallick R.B., Hong L., Drager R.G., Favreau M.R., Monfort A. *et al.* Complete sequence of *Euglena gracilis* chloroplast DNA // *Nucleic Acids Research*. 1993. Vol. 21, № 15. P. 3537–3544.

96 Gockel G., Baier S., Hachtel W. Plastid ribosomal protein genes from the nonphotosynthetic flagellate *Astasia longa* // *Plant Physiology*. 1994. Vol. 105. P. 1443–1444.

97 Gockel G., Hachtel W. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa* // *Protist*. 2000. Vol. 151, № 4. P. 347–351.

98 Linton E.W., Karnkowska-Ishikawa A., Kim J.I., Shin W., Bennett M.S. *et al.* Reconstructing euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of Euglenaria gen. nov. (Euglenophyta) // *Protist*. 2010. Vol. 161, № 4. P. 603–619.

99 Brosnan S., Shin W., Kjer K.M., Triemer R.E. Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA // *International Journal of Systematic and Evolutionary Microbiology*. 2003. Vol. 53, № 4. P. 1175–1186.

100 Turmel M., Otis C., Lemieux C. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: Insights into the architecture of ancestral chloroplast genomes // *Proc. Natl. Acad. Sci. U.S.A.* 1999. Vol. 96, P. 10248–10253.

101 Gilson P.R., Su V., Slamovits C.H., Reith M.E., Keeling P.J., McFadden G.I. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus // *Proc. Natl. Acad. Sci. U.S.A.* 2006. Vol. 103, № 25. P. 9566–9571.

102 Электронный ресурс http://lab6.iitp.ru/ppc/chlorophyta/.

103 Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // *Journal of Molecular Biology*. 1970. Vol. 48, № 3. P. 443–453.

104 Электронный ресурс http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt.

105 Lommer M., Roy A.-S., Schilhabel M., Schreiber S., Rosenstiel P., LaRoche J. Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation // *BMC Genomics*. 2010. Vol. 11. P. 718.

106 Tanaka T., Fukuda Y., Yoshino T., Maeda Y., Muto M. *et al.* High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine pennate diatom, *Fistulifera* sp. strain JPCC DA0580 // *Photosynthesis Research*. 2011. Vol. 109, № 1–3, P. 223–229.

107 Электронный ресурс http://www.sanger.ac.uk/

108 Fong A., Archibald J.M. Evolutionary dynamics of light-independent protochlorophyllide oxidoreductase genes in the secondary plastids of Cryptophyte algae // *Eukaryotic Cell*. 2008. Vol. 7, № 3. P. 550–553.

109 Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods // *Molecular Biology and Evolution*. 2011. Vol. 28. P 2731–2739.

110 Электронный ресурс http://blast.ncbi.nlm.nih.gov/.

111 Лопатовская К.В., Селиверстов А.В., Любецкий В.А. Регулоны NtcA и NtcB у цианобактерий и хлоропластов водорослей отдела Rhodophyta // *Молекулярная биология*. 2011. Т. 45, № 3. С. 570–574.

112 Hagopian J.C., Reis M., Kitajima J.P., Bhattacharya D., Oliveira M.C. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids // *Journal of Molecular Evolution*. 2004. Vol. 59, № 4. P. 464–477.

113 Finn R.D., Mistry J., Tate J., Coggill P., Heger A. *et al.* The Pfam protein families database // *Nucleic acids research*. 2010. Vol. 38, Database issue D211-D222.

114 Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput // *Nucleic Acids Research*. 2004. Vol. 32, № 5. P. 1792–1797.

115 Электронный ресурс http://lab6.iitp.ru/ppc/liliopsida/.

116 Электронный ресурс http://lab6.iitp.ru/ppc/magnoliophyta/.

117 Bauer A.L., Hlavacek W.S., Unkefer P.J., Mu F. Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites // *PLoS Computational Biology*. 2010. Vol. 6, № 11. e1001007.

118 Sun E.I., Rodionov D.A. Computational analysis of riboswitch-based regulation // *Biochimica et Biophysica Acta – Gene Regulatory Mechanisms*. 2014. pii: S1874–9399(14)00032-7.

119 Hauser C.R., Gillham N.W., Boynton J.E. Translation regulation of chloroplast genes // *The Journal of Biological Chemistry*. 1996. Vol. 271. P. 1486–1497.

120 Zerges W. Translation in chloroplasts // *Biochimie*. 2000. Vol. 82. P. 583–601.

121 Nickelsen J. Chloroplast RNA binding proteins // *Current Genetics*. 2003. Vol. 43. P. 392–399.

122 Seliverstov A.V., Lyubetsky V.A. Translation regulation of intron-containing genes in chloroplasts // *Journal of Bioinformatics and Computational Biology*. 2006. Vol. 4. P. 783–792.

123 Wolf P.G., Rowe C.A., Hasebe M. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris* // *Gene*. 2004. Vol. 339. P. 89–97.

124 Bailey T.L., Williams N., Misleh C., Li W.W. MEME: discovering and analyzing DNA and protein sequence motifs // *Nucleic Acids Research*. 2006. Vol. 34: W369–W373.

125 Любецкий В.А., Селиверстов А.В. Некоторые алгоритмы, связанные с конечными группами // *Информационные процессы*. 2003. Т. 3, № 1. С. 39–46.

126 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction // *Nucleic Acids Research*. 2003. Vol. 31. P. 3406–3415.

127 Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria // *BMC Microbiology*. 2005. Vol. 5, № 54. P. 54.