N3 Фактическое выполнение

Выполнена классификация и определение типов аттенюаторной регуляции на уровнях транскрипции и инициации трансляции у актинобактерий и альфа-протеобактерий. В частности, предсказан новый тип аттенюаторной регуляции, характеризуемый экстремально малым расстоянием между структурным и лидерных генами, который оказался характерным для актинобактерий и связанным с реинициализацией рибосомы, транслирующей лидерный пептид. Проведён массовый поиск всех типов регуляторных систем у актинобактерий и альфа-протеобактерий.

В 196 геномах актинобактерий предсказаны десятки тысяч лидерных генов с расстоянием между структурным и лидерным генами до 1305 п.н. Среди них наиболее часто встречается случай, когда лидерный ген отделён от структурного гена 10-12 нуклеотидами. Это позволяет предложить гипотезу о реинициализации рибосомы после завершения трансляции лидерного пептида, в результате увеличивается поток рибосом на белок кодирующую область РНК.

Например, для генов *trpE,B,S* биосинтеза триптофана и триптофанил-тРНК у родов *Streptomyces* и *Corynebacterium* обнаружена опероннная и регуляторная организация, значительно отличающаяся от наблюдаемой у *E. coli.* Длинный оперон, включающий *trpE,B*, с общей аттенюаторной регуляцией у *E. coli* распадается, и его части имеют такую же регуляцию; также как и не регулируемый аттенюаторно у *E. coli* ген *trpS*. Эта регуляция с гистидиновыми регуляторными кодонами наблюдается у транскрипционных факторов семейств LysR и TetR, у белка, участвующего в окислительно-восстановительных реакциях цитохрома p450, субъединицы ABC-транспортёров, белков с доменами Helicase_C, DEAD, Phage_integrase и доменами неизвестного назначения UPF0182, Pfam-B_340, Pfam-B_671, Pfam-B_11008, [O.A. Zverkov *et al* 2015. A Search for Genes Encoding Histidine-Containing Leader Peptides in Actinobacteria. *ИТuС'2015*].

Описаны регулоны различных типов аттенюаторной регуляции. Прослежена эволюция аттенюаторной регуляции генов пути синтеза триптофана у актинобактерий и формирование у некоторых коринебактерий такой регуляции одновременно перед двумя операнами, в то время как у других видов регулируется не более одного оперона, включающего гены этого пути. Аналогично для протеобактерий получены следующие пары чисел (расстояние от структурного гена – число генов во всех видах): 10 – 11701, 11 – 11957, 12 – 10123.

Разработана программа поиска генов лидерных пептидов. Существенно расширена развитая нами ранее модель аттенюаторной регуляции за счёт учёта в ней вторичных структур ДНК и РНК, РНК-триплексов, температуры среды обитания бактерии, G-квадруплексов. Например, модель позволила объяснить экспериментально наблюдаемую аттенюаторную регуляцию гена *trpE* у *Streptomyces coelicolor*, как и у *Streptomyces coelicolor*, что невозможно без учёта перечисленных факторов. Модель реализована эффективной параллельной компьютерной программой http://lab6.iitp.ru/ru/rnamodel/. На основе этих двух программ массово определены потенциальные лидерные пептиды и гены, связанные с метаболизмом аминокислот, у актинобактерий и альфа-протеобактерий. На основе второй программы определена эффективность найденных аттенюаторов. Например, у актинобактерий вероятность преждевременной терминации транскрипции (аттенюаторная регуляция) гена *trpE* часто высокая при концентрации c=0 триптофанил-тРНК, затем сразу достигает незначительной величины (минимума) при c<0.05 и затем монотонно возрастает при увеличении c до 0.5 с дальнейшим выходом на горизонтальную асимптоту, близкую к значению при нулевой концентрации.

Определены взаимно ближайшие белки цианобактерий и апикопластов кокцидий. Проведён сравнительный анализ удлинений кодируемых в ядре апикопластных белков, что обеспечивает их транслокацию в апикопласт и регуляцию: дополнительный N-конец у *T. gondii* в среднем в полтора раза длиннее, чем у *N. caninum*, и в два раза длиннее, чем у *P. falciparum.* Точные величины удлинений, как и упомянутые выше взаимно ближайшие

белки, приведены в [A.V. Seliverstov *etal*, 2015, *BioMed Research International*, ID 452958]. Предложена гипотеза о регуляции активности пластид кокцидий, которая основана на посттрансляционной модификации избыточно длинного N-конца белка, направляющегося в апикопласт, что играет ключевую роль в регуляции реактивации кокцидий через реактивацию апикопластов в целом.

Выполнена кластеризация всех доступных (на конец 2015 года) белков, кодируемых в пластидах родофитной ветви. Результаты представлены в общедоступной базе данных по адресу http://lab6.iitp.ru/ppc/redline72. База данных позволяет проводить быстрый поиск кластера (семейства белков) как по фрагменту аминокислотной последовательности белка, так и по его филогенетическому профилю. Обоснована наша гипотеза: транскрипционный фактор Ycf28 в пластидах родофитной ветви регулирует транскрипцию гена *moeB*. Показано, что специфичными для пластид багрянок (включая новые пластомы *Lepidodinium chlorophorum*, *Choreocolax polysiphoniae*, *Vertebrata lanosa*, *Trachydiscus minutus*) являются гены *odpA/pdhA*, *odpB/pdhB*, *trpA*, *trpG*, *tilS/ycf62* и *infC*. Определены потенциальные промоторы ряда генов, включая *moeB*, *ftsH*, на основе нового метода [Гершгорин et al, 2015. *Journal of Communications Technology and Electronics*].

Разработан и реализован алгоритм поиска ультраконсервативных элементов ДНК, основанный на поиске плотных подграфов в многодольном графе. У 22 видов из надтипа Alveolata построены кластеры ультраконсервативных элементов. Подтверждено, что род *Cryptosporidium* не входит в класс кокцидий, а является близким родственником плазмодиев, пироплазмид и *Gregarina niphandrodes*. Подтверждено, что фотосинтезирующие простейшие *Chromera velia* и *Vitrella brassicaformis* близкие родственники. Показано, что в составе типа Apicomplexa кокцидии сохранили большее число элементов, присутствовавших у общего предка надтипа Alveolata.

Получен метод совместного анализа большого числа полных геномов с целью нахождения в них ультраконсервативных элементов – похожих друг на друга (не обязательно тождественных) участков, встречающихся сразу во многих геномах. Задача связана с обработкой больших данных, для ее решения разработан ряд оригинальных параллельных алгоритмов с линейной или близкой к ней сложностью.

Для *Dicyema* sp. на основе 480 тыс. чтений предсказаны транскриптом и протеом, который содержит около 15 тыс. потенциальных белков/доменов. Кроме дициемы, предсказаны протеомы: гнатостомулиды *Austrognathia* sp. и плоских червей катенулид *Catenula lemnae*, *Stenostomum leucops* и *Stenostomum sthenum* и других видов. Кластеризация белков выполнена с использованием OrthoMCL-DB-v5. Реконструкция филогении выполнена методом максимального правдоподобия. Она позволила предположить близость дициемы к гнатостомулидам и синдерматам.

Разработана указанная в задании модель, которая составлена из трёх модулей. Модули с учётом всех перечисленных в задании эволюционных событий представляют: согласование двух филогенетических деревьев – генов/белков и видов, описание как эволюции хромосомных перестроек, так и совместных сценариев эволюции регуляторных систем, генов и видов. Модули реализованы для вычислений на суперкомпьютере и объединены в программно-вычислительный комплекс, их можно найти соответственно по адресам http://lab6.iitp.ru/ru/embed3gl, http://lab6.iitp.ru/ru/hrom_reconstruction/, http://lab6.iitp.ru/ru/super3gl.

В качестве примера использования 1го модуля рассмотрена совместная регуляция транскрипционного фактора Rho и его сайтов связывания у актинобактерий.

В качестве примера использования 2го модуля рассмотрены хромосомные перестройки в пластидах родофитной ветви и в митохондриях споровиков.

В качестве примера совместного использования 1го и 3го модулей рассмотрена эволюция, связанная с синтезом пролина у гамма-протеобактерий. В этом случае построены три вложения в дерево видов: дерева генов *pro*, ответственных за синтез пролина, дерева

доменов транскрипционных факторов и дерева сайтов их связывания. Анализ вложений показал хорошую согласованность сценария совместной эволюции.
Эти модули также использованы в работах следующего п. 3.6.
Опишем принципиально новые особенности каждого модуля. Первый из них проводит согласование двух деревьев; первое дерево не обязательно бинарное, и модуль позволяет проводить его оптимальную бинаризацию; алгоритм имеет кубическую вычислительную сложность [Rusin *et al*. *BioMed Research International* (Current Advances in Molecular Phylogenetics), 2014, ID 642089].
Второй модуль решает три задачи. Набор операций и описание хромосомной структуры допускаются в максимальной общности [Gershgorin *et al*. ITaS 2015, #1570151809]. Первая задача состоит в определении кратчайшей (по суммарной цене) последовательности операций, которая переводит одну хромосомную структуру в другую. Если цены всех операций одинаковы, решение задачи было известно. Мы предложили новый (даже для случая равных цен) точный и линейный по вычислительный сложности алгоритм решения этой задачи в отсутствии равенства цен операций. Вторая и третья задачи – построение эволюционного дерева хромосомных структур и реконструкция хромосомных структур, заданных в листьях, вдоль него. Они решаются также точными и линейными алгоритмами [Горбунов et al. *Молекулярная биология*, 2015, том. 49, №3].
Третий модуль – описание совместной эволюции на основе единого функционала, объединяющего требования к вложениям в дерево видов: дерева регулируемого гена, дерева фактора регуляции, дерева регуляторного сайта и т.д. Она решается точным алгоритмом кубической сложности.
На основе этих работ построено супердерево эубактерий, на его примере наша программа сравнивалась с другими известными программами построения дерева видов и находится наравне с лучшими программами. Построено супердерево пластид и цианобактерий, которое сопоставлено с известными деревьями видов, что позволило выделить основные группы видов, чьи пластиды имеют общее происхождение. А именно, подавляющее большинство пластид принадлежат родофитной ветви, пластиды которой ведут общее происхождение от пластид багрянок, или хлорофитной ветви, пластиды которой также имеют общего предка. Родофитная ветвь включает много вторичных и третичных эндосимбионтов. Хлорофитная ветвь включает все наземные растения, чьи пластиды ведут первичное происхождение от пластид зелёных водорослей, но в её составе значительно меньшее разнообразие вторичных эндосимбионтов. Значительные различия пластид эвгленовых водорослей, имеющих вторичное происхождение от пластид зелёных водорослей, не позволяют уточнить филогенетическое положение донора их пластид. Пластида (cyanelle) у *Cyanophora paradoxa* значительно отличается от пластид как родофитной, так и хлорофитной ветвей.
Хромосомные структуры у большинства видов отделов Rhodophyta (багрянки) и Cryptophyta (криптофитовые водоросли) весьма близки друг другу, что говорит об их происхождении от общего предка. Хотя у *Porphyridium purpureum* произошло очень много хромосомных перестроек, не характерных для других видов этого отдела. Также неожиданно много хромосомных перестроек наблюдается у *Chromera velia* из надтипа Alveolata, чьи пластиды имеют вторичное происхождение от пластид багрянок.
Выделены гены пластид, пригодных для определения таксономической принадлежности видов. Например, специфичными для всех пластид рассмотренных видов того же отдела являются гены *odpA/pdhA, odpB/pdhB, trpA, trpG, tilS/ycf62* и *infC*. Характерной особенностью пластид зелёных водорослей, родственных хламидомонаде, служит интенсивная перестройка хромосомы. Мы предполагаем, что эти перестройки отражаются и на типе регуляции: регуляция на уровне инициации трансляции встречается у этих видов чаще, чем на уровне транскрипции из-за быстрой смены промоторов и разрушения сайтов связывания транскрипционных факторов при хромосомных перестройках. Описаны события, сопровождающие исчезновение фотосинтеза: значительная редукция

генома, исчезновение генов фотосистем. В то же время порядок оставшихся генов на хромосоме во многих случаях меняется незначительно. Например, на дереве хромосомных структур пластид фотосинтезирующие *Guillardia theta*, *Rhodomonas salina* и нефотосинтезирующая *Cryptomonas paramecium* криптофитовые водоросли расположились в одной кладе. Изучение построенного супердерева не позволило сделать окончательные выводы о таксономической принадлежности цианобактерий, от которых произошли пластиды различных групп, из-за значительной редукции геномов цианобактерий, сопровождаемой хромосомными перестройками. Ближайшие к пластидам красных водорослей цианобактерии принадлежат порядкам ностоковые (Nostocales) и хроококковые (Chroococcales).

Данные о хромосомных структурах были сопоставлены с распространением цианобактерий и зелёных водорослей в составе симбионтов: инфузориях, лишайниках. Предсказано: нефотосинтезирующая водоросль *Helicosporidium* sp., паразитирующая на насекомых, ведёт происхождение от требуксиевых водорослей, которые часто выступают симбионтами инфузорий и лишайников.

Построено супердерево хромосомных структур митохондриальной ДНК у споровиков – простейших из группы Alveolata – в которых неоднократно происходила линеаризация и зацикливание хромосом. У животных не встречаются линейные митохондриальные ДНК, обнаруженные у многих споровиков – возбудителей протозойных инфекций, опасных для человека и домашних животных. Это может использоваться для разработки лекарственных препаратов, нацеленных на повреждение митохондрий этих паразитов. Однако быстрая смена типа митохондриальной ДНК не позволяет одновременно воздействовать на весь спектр споровиков.

Построено бинарное дерево пластид семенных растений, для чего политомии разрешались на основе филогенетического распределения вставок прямых повторов в пластидах этих растений. Получено распределение вставок прямых повторов в основных таксономических группах высших растений, и оценка скорости возникновения вставок прямых повторов. В некодирующих областях наиболее частыми событиями оказываются повтор одного нуклеотида или повтор участка длиной пять нуклеотидов. Это может быть использовано для моделирования эволюции некодирующих областей ДНК.

N4 Достигнутые конкретные результаты

В 196 геномах актинобактерий предсказаны десятки тысяч лидерных генов с расстоянием между структурным и лидерным генами до 1305 п.н. Среди них наиболее часто встречается случай, когда лидерный ген отделён от структурного гена 10-12 нуклеотидами. Это позволяет предложить гипотезу о реинициализации рибосомы после завершения трансляции лидерного пептида, в результате увеличивается поток рибосом на белок кодирующую область РНК. Аналогично для протеобактерий получены следующие пары чисел (расстояние от структурного гена – число генов во всех видах): 10 – 11701, 11 – 11957, 12 – 10123.

Проведён сравнительный анализ удлинений кодируемых в ядре апикопластных белков, что обеспечивает их транслокацию в апикопласт и регуляцию: дополнительный N-конец у *T. gondii* в среднем в полтора раза длиннее, чем у *N. caninum*, и в два раза длиннее, чем у *P. falciparum.* Предложена гипотеза о регуляции активности пластид кокцидий, которая основана на посттрансляционной модификации избыточно длинного N-конца белка, направляющегося в апикопласт, что играет ключевую роль в регуляции реактивации кокцидий через реактивацию апикопластов в целом.

Выполнена кластеризация всех доступных белков, кодируемых в пластидах родофитной ветви. Результаты представлены в общедоступной базе данных по адресу http://lab6.iitp.ru/ppc/redline72.

Разработан и реализован алгоритм поиска ультраконсервативных элементов ДНК, основанный на поиске плотных подграфов в многодольном графе.

Для *Dicyema* sp. предсказаны транскриптом и протеом, который содержит около 15 тыс. потенциальных белков/доменов. Кроме дициемы, предсказаны протеомы: гнатостомулиды *Austrognathia* sp. и плоских червей катенулид *Catenula lemnae*, *Stenostomum leucops* и *Stenostomum sthenum* и других видов. Кластеризация белков выполнена с использованием OrthoMCL-DB-v5. Реконструкция филогении выполнена методом максимального правдоподобия. Она позволила предположить близость дициемы к гнатостомулидам и синдерматам.

Разработана указанная комплексная модель: согласования филогенетических деревьев – генов/белков и видов, описания, как эволюции хромосомных перестроек, так и совместных сценариев эволюции регуляторных систем, генов и видов. Модель реализована для вычислений на суперкомпьютере: http://lab6.iitp.ru/ru/chromoggl/, http://lab6.iitp.ru/ru/super3gl.