Позиционная связь генов пластомов растений и водорослей

О.А. Зверков, А.В. Селиверстов, В.А. Любецкий
ИППИ РАН
lyubetsk@iitp.ru

Аннотация

Проведён полный анализ позиционной связи генов в пластомах растений и водорослей. Анализ основан на сопоставлении аннотаций генов в банке данных GenBank. Предположен новый вид аттенюаторной регуляции, который показан на примере пары генов ycf33 и ilvB и основан на сопряжённой трансляции этих генов; в ней ycf33 играет роль гена лидерного пептида для гена ilvB. Обсуждаются роли генов ycf12 в формировании железосероцентров и ycf34 в синтезе аминокислот.

1. Введение

Пластиды многих видов делятся на две большие группы, обозначаемые ниже GreenLine и RedLine. Первая группа объединяет виды из таксонов Viridiplantae, Euglenozoa и Chlorarachniophyceae. Во вторую группу входят весьма различные виды, как по внешнему виду, так и по таксономической принадлежности: багрянки, бурые, жёлто-зелёные, золотистые, золотисто-бурые (Raphidophyceae), диатомовые и криптофитовые водоросли; гаптофитовые (Haptophyceae) и несколько видов споровиков, чьи близкие родственники не имеют пластид. Эта группа объединяет пластомы разной длины, что связано с потерей в них генов (с полной потерей или с переносом в ядро), а иногда связано со значительным сжатием межгенных участков. Пластиды внутри каждой из этих групп имеют самостоятельное происхождение от цианобактерий [1]. Пластиды у видов *Cyanophora paradoxa* и *Paulinella chromatophora* имеют своё собственное происхождение от цианобактерий [1]; пластом *P. chromatophora* имеет большой размер и является почти полным бактериальным геномом. Эти два вида не относятся к группам GreenLine и RedLine.

Аттенюаторной называется регуляция экспрессии генов, при которой трансляция одного гена (лидерного пептида) существенно влияет на экспрессию другого гена, входящего в состав полицистронной мРНК и лежащего ближе к 3'-концу. Аттенюаторная регуляция может встретиться у любых бактерий и пластид. Хотя до сих пор она не описана у цианобактерий и пластид, она хорошо известна у других бактерий из многих таксономических групп [2].

2. Материалы

Геномы получены из банка данных GenBank. Таксономическая группа GreenLine включает все виды из таксономических групп Viridiplantae, Euglenozoa и Chlorarachniophyceae. Из них Euglenozoa содержит два вида: *Euglena gracilis*, *E. longa* (или *Astasia longa*); Chlorarachniophyceae содержит один вид: *Bigelowiella natans*. Таксономическая группа RedLine включает следующие 22 вида, для которых геномы пластид секвенированы полностью: *Porphyra purpurea*, *P. yezoensis*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Gracilaria tenuistipitata*, *Ectocarpus siliculosus*, *Fucus vesiculosus*, *Vaucheria litorea*, *Heterosigma akashiwo*, *Aureococcus anophagefferens*, *Aureoumbra lagunensis*, *Guillardia theta*, *Rhodomonas salina*, *Cryptomonas paramecium*, *Emiliania huxleyi*, *Phaeodactylum tricornutum*, *Odontella sinensis*, *Thalassiosira pseudonana*, *Eimeria tenella*, *Toxoplasma gondii*, *Babesia bovis*, *Theileria parva*. Рассмотрен также короткий фрагмент ДНК водоросли *Porphyra umbilicalis*, содержащий гены *argB*, *ycf33* и *ilvB*.

3. Методы

Сравнительный анализ аннотаций геномных последовательностей проводился разработанным нами простым алгоритмом поиска пар генов, сохраняющих соседнее расположение в геномах различных организмов. Этот алгоритм на вход получает набор полностью аннотированных геномов в форматах банков данных GenBank NCBI, www.ncbi.nlm.nih.gov/genbank, и EMBL, www.embl.org. Выходом алгоритма являются таб-

лицы частоты встречаемости соседних пар генов в геномах, а также координаты такой пары генов в каждом геноме. Алгоритм выдаёт также списки некодирующих областей, лежащих между парами одноимённых генов.

Приведём краткое описание алгоритма. Выделяются существенные для задачи фрагменты аннотации гена в геноме, формируется упорядоченная последовательность генов и составляется список пар соседних генов. В качестве существенных фрагментов аннотации используются записи *CDS* (последовательность, кодирующая белок), *tRNA*, *rRNA* (гены РНК) и *exon* (экзон). Если запись в аннотации содержит сведения о нескольких кодирующих последовательностях, то для каждой из них создаются новые записи, которые рассматриваются независимо. Поскольку один и тот же локус ДНК часто снабжается несколькими записями, расположенными в разных местах аннотации генома, применяется дополнительный проход по сформированному списку записей (каждая из них соответствует одному гену без интронов или экзону) с объединением тех из них, которые имеют совпадающие координаты начала и конца. При этом взаимодополняющая информация в окончательной записи для каждого экзона объединяется: например, порядковый номер экзона обычно указывается в записи типа *exon*, продукт гена указывается в других записях; координаты, название гена и его принадлежность к одной из цепей ДНК обычно дублируются в разных исходных записях. Затем гены упорядочиваются по возрастанию координаты их левого конца, и к каждому из них применяется следующая процедура поиска соседей. Поскольку гены в алгоритме идентифицируются по именам, приходится исключать из рассмотрения гены, в аннотации которых имя не указано.

Обозначим текущий рассматриваемый ген g_c. Для каждого гена g_p из списка «предшествующих» (в начале обработки каждого генома он пуст): пара (g_p, g_c) добавляется в список пар, и, если координата правого конца g_p меньше координаты левого конца g_c (т.е. g_c отделяет g_p от последующих генов), g_c удаляется из списка «предшествующих». Ген g_c добавляется в список «предшествующих», и начинается обработка следующего гена в очереди.

После завершения обработки всех геномов для каждой отобранной пары генов формируется слово, служащее ключом для отнесения её к группе гомологичных пар генов из разных геномов. Ключ состоит из специальным образом унифицированных имён генов данной пары, снабжённых номерами экзонов (если они есть), и записанных в лексикографическом порядке. Все пары генов с совпадающим ключом собираются в группы. Для каждого ключа подсчитывается суммарное число геномов, в которых встречается соответствующая пара соседних генов. Межгенные промежутки выдаются с учётом порядка следования генов, так как в различных геномах гены могут находиться на разных цепях ДНК.

4. Результаты

4.1. Общий анализ

Обычно позиционно сцепленные гены в пластидах составляют опероны, кодирующие субъединицы одного фермента, типичные примеры *rbcLS* и *chlLN*; или кодируют рибосомные белки, факторы элонгации, субъединицы РНК-полимеразы, различные компоненты фотосистем. В каждой из групп GreenLine и RedLine имеется много сцепленных пар, как характерных для обеих групп, так и специфичных для каждой группы.

Следующие пары генов позиционно сцеплены более чем в 75% геномов каждой из групп GreenLine и RedLine: *psbF+psbL*, *psbJ+psbL*, *atpB+atpE*, *psbB+psbT*, *psbE+psbF*, *psbH+psbN*, *psbN+psbT*, *rpoA+rps11*, *atpH+atpI*, *psaA+psaB*, *psbC+psbD*, *rpl33+rps18*, *rpoC2+rps2*, *rpl22+rps3*, *psaB+rps14*, *rpl22+rps19*.

Следующие пары генов позиционно сцеплены более чем в 75% геномов из RedLine, но относительно редко сцеплены в геномах из GreenLine: *rps12+rps7*, *rps12+rps7*, *rps7+tufA*, *rpl2+rps19*, *rpl16+rps3*, *rpl2+rpl23*, *rpl5+rps8*, *rpoC1+rpoC2*, *rpoB+rpoC1*, *petB+petD*.

Следующие пары генов позиционно сцеплены более чем в 75% геномов из RedLine, но не сцеплены в геномах из GreenLine: *atpA+atpD*, *atpG+atpH*, *rbcL+rbcS*, *rpl31+rps12*, *rpl31+rps9*, *rpl6+rps8*, *rps10+tufA*, *rps11+rps13*, *rps5+secY*, *atpD+atpF*, *atpF+atpG*, *rpl1+rpl11*, *rpl13+rps9*, *rpl14+rpl24*, *rpl14+rps17*, *rpl24+rpl5*, *rpl36+rps13*, *rpl36+secY*, *petA+tatC*, *psaF+psaJ*, *rpl13+rpoA*, *rpl20+rpl35*, *ycf16+ycf24*.

Следующие пары генов позиционно сцеплены в наборах от 50% до 75% геномов из RedLine, но не сцеплены в геномах из GreenLine: *rpl1+rpl12*, *rpl16+rpl29*, *rpl18+rpl6*, *rpl18+rps5*, *rpl21+rpl27*, *rpl23+rpl4*, *rpl3+rpl4*, *atpB+ycf3*, *chlI+psaM*, *petG+psbK*, *rpl29+rps17*, *rpoB+rps20*, *rpl34+ycf46*, *dnaK+rpl3*, *petL+psaL*, *rpl34+secA*, *rps16+ycf65*, *rps2+tsf*, *atpE+tatC*, *atpI+tsf*, *petL+ycf4*, *petM+petN*.

Следующие пары генов позиционно сцеплены в наборах от 30% до 50% геномов из RedLine, но не сцеплены в геномах из GreenLine: *psaI+psbJ*, *ftsH+psaE*, *ilvB+ycf33*, *rpl11+trnW*, *cbbX+rbcS*, *ftrB+psaI*, *petJ+psbV*, *psaD+trnS*, *rpl33+rps20*, *rps16+ycf19*, *oleG+natA*, *rps18+ycf3*, *trnG+ycf4*, *clpC+rpl19*, *dnaB+trnF*, *ftrB+ycf12*, *petG+rps14*, *psaE+psbH*, *trnC+trnL*.

Наибольшее число одних и тех же в группе RedLine пар сцепленных генов наблюдается у багрянок *Porphyra spp.* и *Gracilaria tenuistipitata*, у бурых водорослей *Ectocarpus siliculosus* и *Fucus vesiculosus*, а также у *Heterosigma akashiwo*.

В качестве примеров опишем подробнее три пары позиционно сцепленных генов из RedLine.

4.2. Пара *ycf33 + ilvB*

У многих видов из RedLine перед геном *ilvB* расположен ген *ycf33*. Среди них водоросли родов Cyanidioschyzon, Gracilaria, Porphyra, Guillardia, Rhodomonas, Aureococcus, Aureoumbra, Ectocarpus, Fucus, Vaucheria. Здесь нами предположена регуляция, которая будет описана в разделе 5.2 Обсуждения.

4.3. Пара *ycf12 + ftrB*

Ген *ftrB* (или *ftrC*) обнаружен у багрянок (*Porphyra purpurea*, *P. yezoensis*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Gracilaria tenuistipitata*), *Heterosigma akashiwo* (CCMP452 и NIES293), бурых (*Ectocarpus siliculosus*, *Fucus vesiculosus*), криптофитовых (*Guillardia theta*, *Rhodomonas salina*) и жёлто-зелёных (*Vaucheria litorea*) водорослей. В то же время *ftrB* отсутствует в пластидах *Emiliania huxleyi*, диатомовых (*Phaeodactylum tricornutum*, *Odontella sinensis*, *Thalassiosira pseudonana*) и золотистых (*Aureococcus anophagefferens*, *Aureoumbra lagunensis*) водорослей и у нефотосинтезирующей криптофитовой водоросли *Cryptomonas paramecium*. Ген *ycf12* кодирует консервативный белок неизвестной функции длиной от 31 до 36 аминокислотных остатков. Ген *ycf12* присутствует в пластидах многих водорослей и наземных растений, включая мохообразные, папоротникообразные, гнетовые и саговниковые, но отсутствует в пластидах хвойных и цветковых растений, у простейших группы Rhizaria (*Bigelowiella natans* и *Paulinella chromatophora*), у споровиков (Eimeria, Toxoplasma, Babesia, Theileria), а также в сильно редуцированных пластидах нефотосинтезирующих видов, чьи родственники обладают этим геном: у криптофитовой водоросли *Cryptomonas paramecium*, у эвгленовой водоросли *Astasia longa*, у зелёной паразитирующей водоросли *Helicosporidium sp. ex Simulium jonesii*. Положение гена *ycf12* на хромосоме значительно отличается даже у близких видов. Однако у большинства видов, чьи пластиды содержат ген *ftrB* (за исключением двух видов из семейства Cyanidiaceae), гены *ycf12* и *ftrB* позиционно сцеплены, причём транскрибирующие их РНК-полимеразы конкурируют друг с другом. У *Cyanidium caldarium* и *Cyanidioschyzon merolae* ген *ftrC* удалён от *ycf12*, но и здесь ген *ycf12* позиционно сцеплен с опероном *ycf24–ycf16*. Причём опять РНК-полимеразы конкурируют друг с другом.

4.4. Пара *ycf34 + ilvH*

В пластидах ген *ycf34* обнаружен лишь у шести видов водорослей: *Porphyra purpurea*, *P. yezoensis*, *Gracilaria tenuistipitata*, *Heterosigma akashiwo*, *Ectocarpus siliculosus* и *Fucus vesiculosus*. Во всех случаях ген *ycf34* расположен на небольшом расстоянии от гена *ilvH*. У *Heterosigma* всего 8 пн разделяют гены *ycf34* и *ilvH*. Транскрибирующие эти гены РНК-полимеразы конкурируют друг с другом, двигаясь по противоположным цепям ДНК навстречу друг другу.

5. Обсуждение

5.1. Общий анализ

Анализ позиционной сцепленности генов подтвердил таксономическое разделение групп GreenLine и RedLine и позволил выявить в них много сцепленных пар одноимённых генов, специфических для каждой группы.

Хотя пластиды из видов, принадлежащих группе RedLine, имеют общее происхождение [1], сама эта группа не является монофилетической. После исключения отдела багрянок она становится частью монофилетической группы Chromalveolata [3], которая объединяет большое число видов, как имеющих, так и не имеющих пластид, например, *Cryptosporidium parvum* и *Phytophthora spp.* Собственно RedLine и состоит из багрянок и видов из Chromalveolata, имеющих пластиды.

Позиционная сцепленность генов в пластидах багрянок часто нарушается в семействе Cyanidiaceae (*C. merolae*, *C. caldarium*), что коррелирует с обособленностью Cyanidiaceae в дереве белков пластид [1; 4].

Отмечалось значительное сходство белков пластид *E. siliculosus*, *F. vesiculosus* и *H. akashiwo* [4]. Однако вопреки этому сходству белков пластид этих родов и диатомовых водорослей, позиционная сцепленность генов в пластидах диатомовых водорослей часто нарушается. Это связано с потерей или переносом генов из пластид в ядро. Примером является перенос гена *ycf34* в ядерный геном у *Phaeodactylum tricornutum* и *Thalassiosira pseudonana*, см. пункт 5.4.

Отсутвие позиционной сцепленности у *Aureococcus anophagefferens*, *Aureoumbra lagunensis* и *Emiliania huxleyi* также связано с потерей из их пластомов некоторых генов. С другой стороны, даже в

сильно редуцированных пластомах нефотосинтезирующих видов *Cryptomonas paramecium*, *Eimeria tenella*, *Toxoplasma gondii*, *Babesia bovis*, *Theileria parva* присутствуют пары позиционно сцепленных генов, которые остаются таковыми у большинства видов.

5.2. Случай *ycf33* + *ilvB*

Ген *ilvB* кодирует большую субъединицу фермента, важного для синтеза разветвлённых аминокислот. Поскольку ген *ycf33* содержит много кодонов разветвлённых аминокислот, он может играть роль лидерного гена, вовлечённого в разновидность аттенюаторной регуляции на уровне трансляции гена *ilvB*. В этом случае скорость движения рибосом по открытой рамке считывания *ycf33* должна влиять на вторичную структуру вблизи области связывания рибосомы, транслирующей *ilvB*.

Метиониновый кодон на С-конце гена *ycf33* абсолютно консервативен. У *P. umbilicalis* он находится непосредственно перед инициирующим кодоном, в остальных случаях после него идут от одного до трёх кодонов разветвлённых аминокислот. Этот метиониновый кодон также служит инициирующим кодоном очень короткого лидерного пептида; обозначим старт кодон *ycf33* буквами MN, а этот дополнительный метиониновый кодон – буквами ML. Анализ вторичных структур РНК у *Aureococcus anophagefferens* показывает, что если рибосома остановилась на кодонах разветвлённых аминокислот ниже ML, то вторичная структура РНК не закрывает сайт связывания рибосомы (ШДСГ) перед *ilvB*, что согласуется с гипотезой о регуляции. Механизм такой регуляции можно представить себе следующим образом. Если разветвлённых аминокислот мало, то ML занято рибосомой, вторичная структура не закрывает ШДСГ, с которой идёт трансляция *ilvB*. Если разветвлённых аминокислот много, то по мРНК гена *ycf33* постоянно идут рибосомы, которые не задерживаются на регуляторных кодонах и освобождают мРНК после достижения стоп-кодона. Поэтому каждая из этих рибосом разрушает вторичную структуру вблизи ШДСГ лишь на короткое время, а в остальное время ШДСГ закрыта вторичной структурой, препятствующей инициации трансляции *ilvB*. В любом случае при большой концентрации аминокислот ШДСГ в основном закрыта: рибосомой со стоп кодона *ycf33* или вторичной структурой около ШДСГ.

Отметим, что регуляция инициации трансляции, опосредованная скоростью трансляции лидерного пептида, предсказана нами для гена *leuA* у многих актинобактерий [5; 2].

У некоторых видов гены *ycf33* и *ilvB* независимы друг от друга. У *Heterosigma akashiwo* и *Paulinella chromatophora* гены *ycf33* и *ilvB* расположены далеко друг от друга. У *Cyanophora paradoxa* и у диатомовых водорослей *Odontella sinensis*, *Phaeodactylum tricornutum*, *Thalassiosira pseudonana* в пластоме ген *ilvB* отсутствует, хотя там имеется ген *ycf33*. Напротив, в пластоме у нефотосинтезирующей *Cryptomonas paramecium* присутствует ген *ilvB*, но там нет гена *ycf33*. Заметим, что здесь важную роль могут играть другие потенциальные лидерные пептиды для *Cryptomonas paramecium* и *Heterosigma akashiwo*, найденные нами.

5.3. Случай *ycf12* + *ftrB*

Ген *ftrB* (или *ftrC*) кодирует каталитическую цепь ферредоксин-тиоредоксин редуктазы [6], содержащей железосероцентр типа [Fe4-S4]. Гены *ycf24* и *ycf16* кодируют белки SufB и SufC, участвующие в формировании железосероцентров [7].

Позиционная сцепленность гена *ycf12* позволяет предположить, что он кодирует белок, непосредственно связанный с формированием железосероцентров (гены *ycf24* и *ycf16*) или регулирующий экспрессию белков с железосероцентрами (ген *ftrB*). Действительно, у бактерий и хлоропластов транскрипция и трансляция происходят одновременно. Поэтому транскрипционные факторы бактерий нередко кодируются вблизи сайтов их связывания с ДНК, что ускоряет достижение фактором своего сайта. Такая сцепленность регулируемого гена и гена фактора полезна и для регуляторных белков, связывающих РНК. Заметим, что такая сцепленность имеет значение только для факторов, реагирующих на концентрацию вещества в цитоплазме, и бесполезна для белков двухкомпонентных систем передачи сигнала от мембраны, поскольку в последнем случае фактор сначала достигает рецептора на мембране, а после модификации возвращается на ДНК.

5.4. Случай *ycf34* + *ilvH*

Ген *ilvH* кодирует малую субъединицу фермента, важного для синтеза разветвлённых аминокислот. Роль гена *ycf34* неизвестна. Его ортологи присутствуют в геномах многих цианобактерий и в ядерных геномах диатомовых водорослей (на хромосоме 1 у *Phaeodactylum tricornutum* CCAP 1055/1 и хромосоме 4 у *Thalassiosira pseudonana* CCMP1335), но там нет позиционной сцепленности *ycf34* с *ilvH*. На N-конце белков Ycf34, кодируемых генами *ycf34*, присутствуют четыре абсолютно консервативные позиции, занятые остатками цистеина, рис. 1. Исключение составляет *Ph. tricornutum*, у которой соответствующий участок расположен дальше от N-конца. Наличие консервативных ос-

Вид		Домен
P.purpurea	1	MCICINCNHIAECSTYHLIESRHK
P.yezoensis	1	MCICINCNHITKCNTYHLIESQHK
G.tenuistipitata	1	MCICVNCRHVQDCMTYYMIQKQHD
H.akashiwo	1	MCICLNCEYIDFCQQYHFIEKKHI
E.siliculosus	1	MCICLNCDRINSCHIYFLIEKQHQ
F.vesiculosus	1	MCICLNCERINSCKIFVNIEEKHK
Ph.tricornutum	47	HCICVDCARVTNCAAYHFVETKHE
Th.pseudonana	1	MCICIHCARVTNCQAYHFVEEQHS
P.marinus	1	MCICVNCQWVDRCKAYHAVEKQHG
A.marina	1	MCICINCQFVDQCITYHAVETQHL
		_***__*_____*_____*_

Рисунок 1. Выравнивание доменов белков Ycf34, которые могут входить в состав железосероцентра, у водорослей и у цианобактерий *Prochlorococcus marinus* CCMP1375 и *Acaryochloris marina* MBIC11017.
Число обозначает номер первого аминокислотного остатка домена от N-конца белка. Звёздочка отмечает консервативные позиции.

татков цистеина позволяет предположить, что белок Ycf34 содержит железосероцентр. Позиционная сцепленность генов в хлоропластах и наличие железосероцентра позволяют думать, что белок Ycf34 вовлечён в синтез аминокислот. Это подтверждается и тем, что поиск доменов в базе Pfam обнаруживает гомологию коротких фрагментов белка Ycf34 и фермента MetW синтеза метионина. Например, у *Ectocarpus siliculosus* этот фрагмент NFFSKQGIYLL расположен от 45 до 55 позиций. Однако малая длина гомологичного фрагмента не позволяет сделать определённый вывод о роли Ycf34.

Список литературы

[1] C. Lemieux, C. Otis, M. Turmel “A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atomphyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies” *BMC Biology*, 2007, 5(2), pp. 1–17.

[2] K.V. Lopatovskaya, A.V. Seliverstov, V.A. Lyubetsky “Attenuation regulation of the amino acid and aminoacyl-tRNA biosynthesis operons in bacteria: a comparative genomic analysis” *Molecular Biology*, 2010, 44(1), pp. 128–139.

[3] C. Martens, K. Vandepoele, Y. Van de Peer “Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species” *Proc. Natl. Acad. Sci. USA*, 2008, 105(9), pp. 3427–3432.

[4] G. Le Corguille, G. Pearson, M. Valente, C. Viegas, B. Gschloessl, E. Corre, X. Bailly, A. Peters, C. Jubin, B. Vacherie, M. Cock, C. Leblanc “Chloroplast genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids” *BMC Evolutionary Biology*, 2009, 9(253), pp. 1–14.

[5] A.V.Seliverstov, H.Putzer, M.S.Gelfand, V.A.Lyubetsky “Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria” *BMC Microbiology*, 2005, 5(54), p. 14.

[6] S. Dai, C. Schwendtmayer, P. Schurmann, S. Ramaswamy, H. Eklund “Redox signaling in chloroplasts: cleavage of disulfides by an iron-sulfur cluster” *Science*, 2000, 287, pp. 655–658.

[7] J.F. Eccleston, A. Petrovic, C.T. Davis, K. Rangachari, R.J.M. Wilson (Iain) “The kinetic mechanism of the SufC ATPase” *The Journal of Biologiczl Chemistry*, 2006, 281, pp. 8371–8378.