Информационные процессы, Том 5, № 4, 2005, стр. 265–270.

Поиск консервативных участков в лидерных областях генов в случае известного дерева видов[1]

В.А. Любецкий, А.В. Селиверстов

Институт проблем передачи информации РАН,
127994, Россия, Москва,
Большой Каретный переулок, 19,
e-mail: lyubetsk@iitp.ru, slvstv@iitp.ru
Поступила в редколлегию 22.08.2005

Аннотация—Предложен алгоритм поиска консервативных нетранслируемых участков в мРНК с учетом дерева видов. В частности, найдены сигналы у генов, имеющих интроны и траскрибируемых в хлоропластах растений из Embryophyta. У многих хлоропластов гены белков имеют интроны. Поэтому трансляция не может происходить сразу после транскрипции, как это имеет место у бактерий. В тоже время аппарат трансляции у хлоропластов очень близок к таковому у бактерий. В статье описан потенциальный сигнал, который мог бы обеспечивать такую задержку начала трансляции гена с интронами.

1. ВВЕДЕНИЕ

У многих хлоропластов гены белков имеют интроны, или белок транслируется с мРНК, образованной транс-сплайсингом из разных транскриптов. Поэтому трансляция не может происходить сразу после транскрипции, и, соответственно, рибосома не может двигаться вдоль мРНК сразу после РНК-полимеразы, как это имеет место у бактерий. В тоже время аппарат трансляции у хлоропластов очень близок к таковому у бактерий (не требуется кэпирование $5'$ конца мРНК, как у эукариотов, строение рибосом близкое, трансляция мРНК ингибируется хлорамфениколом, как и у бактерий, и т.д.).

В редких случаях время, необходимое для сплайсинга, могло бы обеспечиваться за счет редактирования мРНК: например, за счет преобразования кодона ACG в старт кодон AUG, [1]. Известно, что у *Chlamydomonas reinhardtii* трансляция многих генов (без интронов) регулируется белками, связывающими $5'$ нетранслируемые области мРНК, [2], [3], [4].

В [5] высказано предположение, что начало трансляции гена с интронами некоторым механизмом задерживается до завершения сплайсинга. В нашей заметке описан потенциальный сигнал, который мог бы обеспечивать такую задержку начала трансляции гена с интронами. В заметке также предлагается модификация алгоритма, описанного в [6] и [7]. Эта модификация тестировалась и показала свою эффективность, в частности, на примере поиска сигналов, т.е. консервативных $5'$ нетранслируемых участков в мРНК, у генов *atpF* и *petB*, имеющих интроны и траскрибируемых в хлоропластах растений из Embryophyta. Эти растения перечислены в таблице 1.

[1] Работа частично поддержана грантом МНТЦ 2766.

Таксон	Вид	Рибосомальные белки
Embryophyta	*Anthoceros formosae*	L2 L16 S12 S16
	Marchantia polymorpha (печеночник)	L16 S12
- Tracheophyta	*Adiantum capillus-veneris* (папоротник)	L2 L16 S12 S16
	Huperzia lucidula (плаун)	L2 L16 S12 S16
	Psilotum nudum (папоротник)	L2 L16 S12
- - Spermatophyta	*Pinus koraiensis*	L2 L16 S12
	Pinus thunbergii (чёрная сосна)	L2 L16 S12
- - - Magnoliophyta	*Amborella trichopoda*	L2 L16 S12 S16
	Arabidopsis thaliana	L2 L16 S12 S16
	Atropa belladonna	L2 L16 S12 S16
	Calycanthus floridus	L2 L16 S12 S16
	Cucumis sativus (огурец)	L2 S12 S16
	Epifagus virginiana	L2 L12 L16 S12
	Lotus corniculatus (ледвянец рогатый)	L2 L16 S12 S16
	Nicotiana tabacum	L2 L16 S12 S16
	Nymphaea alba (кувшинка белая)	L2 L16 S12 S16
	Panax ginseng (женьшень обыкновенный)	L2 L16 S12 S16
	Spinacia oleracea (шпинат)	L16 S12 S16
	Oryza nivara (рис)	L2 L16 S12 S16
	Oryza sativa (рис)	L2 L16 S12 S16
	Triticum aestivum (пшеница)	L2 L16 S12 S16
	Zea mays (кукуруза)	L2 L16 S12 S16

Таблица 1. Рибосомальные белки хлоропластов растений из Embryophyta, гены которых подлежат сплайсингу. Знак "-" указывает на принадлежность к подтаксону предыдущего таксона.

2. МАТЕРИАЛЫ И МЕТОДЫ

Геномы хлоропластов взяты из ГенБанка.

В качестве набора исходных последовательностей брались лидерные области перед ортологичными генами. В общем случае можно искать сигнал – консервативные участки в таком наборе с помощью алгоритма поиска клики в многодольном графе, который описан в [6] и [7]. Но когда исходные последовательности принадлежат геномам, для которых известно дерево видов, этот алгоритм может быть удачно усовершенствован, как это описано ниже. При тестировании так усовершенствованного алгоритма был получен результат, показанный в таблицах 2 и 3.

Исходный набор последовательностей в алфавите {A, U, C, G} разбивается на группы, состоящие из близкородственных по дереву видов последовательностей, и в каждой группе по отдельности ищется сигнал – набор попарно близких сайтов фиксированной длины n (которая возрастала до тех пор, пока сигнал не теряется в заметной части последовательностей из такой группы). Таким образом, для каждой филогенетической группы находится один или несколько сигналов. Каждый сигнал определяет свою весовую матрицу – матрицу размера $4 \times n$, в которой указаны частоты встречаемости каждой буквы в каждом столбце сигнала. Затем отбираются цепочки таких матриц, у которых имеются попарно близкие подматрицы одного (по возможности, большого) размера $4 \times m$, и эта цепочка определяет набор последовательностей уже из разных филогенетических групп и соответствующий этому набору общий сигнал. Этот шаг алгоритма выполняется с помощью того же исходного алгоритма поиска клики в многодольном графе.

Наконец, выполняется сборка односайтовых сигналов в многосайтовые, которые уже, вообще говоря, разделены неконсервативными участками. И затем строится множественное выравнивание, согласованное с найденными сигналами и с исходным деревом видов.

3. РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

3.1. АТФ синтаза

Ген *atpF*, кодирующий одну из субъединиц АТФ синтазы, имеет интроны в генах хлоропластов из Embryophyta (кроме *Epifagus virginiana* – не фотосинтезирующего паразита). У *Adiantum capillus-veneris* старт кодон AUG получается из кодона ACG при редактировании мРНК. В остальных случаях старт кодоном является RUG, где R – один из нуклеотидов A или G. В лидерных областях гена *atpF* у видов из Embryophyta найдены по два консервативных участка, консенсусы которых соответственно равны AAUNAAAAAA и UAUCUAUAAGAGGAGANNA, второй участок расположен правее и вблизи старта трансляции гена, он перекрывает предполагаемый сайт связывания рибосомы, таблица 2. Здесь N означает произвольный нуклеотид. Между этими консервативными участками расположены AU богатые неконсервативные участки с длинами от 19 до 61 нуклеотидов, неконсервативные даже в Magnoliophyta. В таблице 2 показано соответствующее множественное выравнивание $5'$ лидерных областей, непосредственно примыкающих к старт кодону гена. Вероятно, некоторый белок *предотвращает* инициацию трансляции, перекрывая область связывания рибосомы, см. дискуссию.

Anthoceros formosa	ggaaugaauaauagcuuaagcuuuccuauuuuaaguaggauuggau
Marchantia polymorpha	uaaaugaaaaaaacuguugaguaaacaaaaaacucca*********
Adiantum capillus-veneris	uuaauaauuaaugugucaauucuugcuaggcuggauca********
Huperzia lucidula	auaaggaaaaaacuauguaaacuu**********************
Psilotum nudum	uuaauaaaagaaaaacuuugucaa**********************
Pinus thunbergii	caaauaagaaaauaacaaaauucugua********************
Pinus koraiensis	aggaucaaaaaagaaaagaacaaaaugaug*****************
Amborella trichopoda	uaaauaaaaauaagauaugggugaagugauacaaaaaga******
Arabidopsis thaliana	agaauaaaaaaaaaaggacagaguuccu*******************
Atropa belladonna	auaauaaauaaaaagaggggcgaagugcuacaaaaaga********
Calycanthus floridus	ucaaucaaaaaaaaggggggcgaagugauacaaaaagaacu****
Cucumis sativus	aaaauaaaaaaaauagaaagaaauaga********************
Lotus corniculatus	caaaaaaaaaaauaaagaaaaaaucuacaaaaauagga********
Nicotiana tabacum	agaauaaauaaaaagaggggcgaagugcuacaaaaaga********
Nymphaea alba	gugauacaaaaaaggaaaucuuuuuuuguauuuug**********
Panax ginseng	uauauaaaaaaagaacaggggcaagguuauacaaaaagaac*****
Spinacia oleracea	ucaauaaaaaaaauaaaauucauuuagaaguagcaaacaauu****
Oryza nivara	gcauuaaaaaaaccgaucaaaaagggcgagcgaaguaagugau***
Oryza sativa	gcauuaaaaaaaccgaucaaaaagggcgagcgaaguaaaguga***
Triticum aestivum	ccgaucaaaaaagggcgagcgaaguaagugaucgaaa*********
Zea mays	ccgaucaaaaaaggacgagcgaaguaaguaaucgaaa*********
консенсус	**AAUNAAAAAA********************************

Таблица 2. Выравнивание $5'$ лидерных областей перед инициирующим кодоном гена *atpF* у видов из Embryophyta. Здесь N обозначает произвольный нуклеотид, а * нуклеотид или делецию.

Anthoceros formosa	aaggaagagacauacuaagacuuaaagaaccuaugaugggagagagagu
Marchantia polymorpha	******uaauuuucaauaauauaauaacgaaaaaaagaggacagc****
Adiantum capillus-veneris	******gaaauugcccaaaacguaaaaccuucgaggagggaaaagaau*
Huperzia lucidula	**********************ggauaauaaccuguaaugggagaaaagu*
Psilotum nudum	**********************aauuagauagucauuaugggagagguauu
Pinus thunbergii	****************gaacauauccuuaucuaugaggggagagcgu**
Pinus koraiensis	**************uagaacauauccuuaucuaugaggggcgagcau**
Amborella trichopoda	****acuccguuugguuuuguuaguccuaucuagaagaggagaguau**
Arabidopsis thaliana	****************uuuuuauaguuuagcuagaagaggagauuau**
Atropa belladonna	******acucuguucgauuuuuuagucuaucuauaagaggagaucau**
Calycanthus floridus	**cugcgcaauuuuguuagcccuaucuauucuauaagaggaaagcau**
Cucumis sativus	****************uaauuaguuuuaucuauaaaaggagaucau**
Lotus corniculatus	******aucauauaaagagaauucguuuauccauaagaggagaucau**
Nicotiana tabacum	******acucuguucgauuuuuuagucuaucuauaagaggagaucau**
Nymphaea alba	*********uuaguccuauuccuauccauaagauaagaggagagcau**
Panax ginseng	***ucuguucuuuuuuuuuuuuuagucuaucuauaagaggagaucau**
Spinacia oleracea	**gaaauaauacaacgauuuuuuuguuuaucuauaagaggagaucau**
Oryza nivara	*cgaaaaacuuuguucuuuguucguccuaucuauaagaggagagcau**
Oryza sativa	ucgaaaaacuuuguucuuuguucguccuaucuauaagaggagagcau**
Triticum aestivum	******aacuuuguucuuuguucguccuaucuauaagaggagagcau**
Zea mays	******aacuuucuucuuuguucguccuaucuauaagaggagagcau**
консенсус	**************************UAUCUAUAAGAGGAGANNA***

Продолжение таблицы 2.

3.2. Цитохром b6

В 5′ лидерной области гена *petB* (цитохром b6) хлоропластов у всех видов из Embryophyta, кроме *Adiantum capillus-veneris* и *Epifagus virginiana*, найден консервативный участок, примыкающий к старт кодону AUG с консенсусом равным GGTAGTTCGAYCGYGGAATT*YTTT*** GTTTNNGTATTTYYGGAAT, таблица 3. Здесь Y означает один из нуклеотидов C или U, и N произвольный нуклеотид, а * возможная вставка не более чем одного нуклеотида. Здесь обнаружена консервативная спираль, показанная подчеркиванием в множественном выравнивании в таблице 3.

Перед геном *petB* на расстоянии 3–10 нуклеотидов от старта трансляции этого гена нет высоко консервативного AG-богатого участка, характерного для сайта связывания рибосомы. Найденный нами консервативный участок имеет большую длину и перед ним расположен слабо консервативный участок. Поэтому можно думать, что наш консервативный участок перекрывает сайт связывания рибосомы. И тогда некоторый белок, связываясь с этим консервативным участком в 5′ лидерной области гена *petB* с учетом вторичной структуры, в отличии от предыдущего случая активирует инициацию трансляции.

У *Epifagus virginiana* ортолог гена *petB* отсутствует. У *Adiantum capillus-veneris* соответствующая 5′ лидерная область не выравнивается с другими видами из Embryophyta. У всех видов из Embryophyta ген *petB* содержит интроны.

Anthoceros formosa	****uuuucccagug*gugguaguuuaaucgugcaacuacugaaaaaaaaaaggauuuuugaaau
Marchantia polymorpha	ucauuuuuuuaaauuuu*agguaguuuaauuguguaauua*uuaa**auucaaggauuu*uugaau
Huperzia lucidula	auugauccuucccuuu*ugguaguuuaaucguguaauu*cuga***aucaaaggaucuuuagaau
Psilotum nudum	*ucauaaaaaaagac*gaggcaguugaaucacgcaaauuauua***auuuaaugauguuuguaau
Pinus thunbergii	*agcuuaucuuguuc**cacuaguuugaucguguaauuacuuuu**cucuaaggauuuuuggaau
Pinus koraiensis	*agcuuaucuuguuc**caauaguuugaucguguaauuacuuuu**cucuaaggauuuuuggaau
Amborella trichopoda	uuggguuucuagguu*a*gguaguucgaccgugcaauuccuuu***guuucgguauuuccggaau
Arabidopsis thaliana	***ccuauucuccuu**ugguaguucgaccgcgaaauuuuuuucugcauuguauauuuccggaau
Atropa belladonna	*cauucuauuuucuuu*ugguaguucgaucguggaauuucuuu***guuucuguauuuccggagu
Calycanthus floridus	*uuuucuaggccauuc*ugguaguucgaccguggaauuccguu***guuucgguauuuccggaau
Cucumis sativus	*uuagccuacucuuuuuugguaguucgaucguggaauuuauuu****uuucuguauuuccggaau
Lotus corniculatus	*cauuccuuuuuuauu*ugguaguucgaucguggaacuucuuu***guuucuguauuuccggaau
Nicotiana tabacum	*cauucuauuuucuuu*ugguaguucgaucguggaauuucuuu***guuucuguauuuccggagu
Nymphaea alba	ucaucucauucuguu**ugguaguucgaccgcggaauucuuuu***guuucgguauuuccggaau
Panax ginseng	*cagcccauucuauuu*ugguaguucgaccgcggaauuucuuu***guuucuguauuucuggaau
Spinacia oleracea	*uauuucuaucccuuu*ugguaguucgaucgcggaauuucuuu***cuuucuguauuuccggaau
Oryza nivara	*cauuucuagacauuc*ugguaguucgaccguggaauu*uuuug**guuucgguaucucuggaau
Oryza sativa	*cauuucuagacauuc*ugguaguucgaccguggaauu*uuuug**guuucgguaucucuggaau
Triticum aestivum	*cauuucuagauauu*augguaguucgaccgcggaauuuuuuu***guuucgguaucucuggaau
Zea mays	acauuucuagacauuc*ugguaguucgaccguggauuu*uuuu***guuuugguaucucuggaau
консенсус	*****YY******UYY*UGGUAGUUCGAYCGYGGAAUU*YUUU***GUUUNNGUAUUUYYGGAAU

Таблица 3. Выравнивание 5′ лидерных областей перед инициирующим кодоном гена *petB* у видов из Embryophyta. Здесь N обозначает произвольный нуклеотид, а * нуклеотид или делецию.

4. ДИСКУССИЯ

Найденные сигналы – консервативные участки перед генами *atpF* и *petB*, вероятно, связаны с задержкой инициации трансляции до завершения сплайсинга. Сейчас нет достаточных оснований, чтобы решить, основана ли эта регуляция на связывании белка с мРНК. Перед геном *petB* консервативный участок не содержит очевидного сайта связывания рибосомы, но имеется консервативная шпилька, что позволяет предположить: после транскрипции происходит образование вторичной структуры в 5′ лидерной области гена и туда присоединяется белок, активирующий инициацию трансляции.

Интроны обнаружены в генах, кодирующих различные белки, экспрессируемые в хлоропластах как растений, так и некоторых одноклеточных эукариотов (*Chlamydomonas reinhardtii*, *Euglena gracilis*). Часто они встречаются в генах белков фотосистем, некоторых субъединиц РНК-полимеразы и NADH-дегидрогеназы, и в генах рибосомальных белков. Количества генов с интронами в хлоропластах примерно одинаковы у всех растений, за исключением *Pinus* spp. и *Epifagus virginiana*, где таких генов меньше.

Хлоропласты у всех видов из Embryophyta, а также у *Chaetosphaeridium globosum* и *Euglena gracilis* содержат интроны в генах белков большой и малой субъединиц рибосомы, таблица 1. Можно думать, что интроны в хлоропластах играют своеобразную регуляторную роль. Если рибосом достаточно много, то рибосомы, транслирующие начальный фрагмент транскрибированной РНК, предотвращают или изменяют сплайсинг, препятствуя образованию новых рибосом. Если рибосом мало, то мРНК длительное время остается открытой для “правильного” сплайсинга и, возможно, редактирования.

СПИСОК ЛИТЕРАТУРЫ

1. Sugiura M., Hirose T., Sugita M. Evolution and mechanism of translation in chloroplasts, *Annu Rev Genet*, 1998, 32, стр. 437–59.

2. Hauser C.R., Gillham N.W., Boynton J.E. Translation regulation of chloroplast genes, *The Journal of Biological Chemistry*, 1996, 271: 3, стр. 1486–1497.

3. Nickelsen J. Chloroplast RNA binding proteins, *Curr Genet*, 2003, 43, стр. 392–399.

4. Herrin D.L., Nickelsen J. Chloroplast RNA processing and stability, *Photosynthesis Research*, 2004, 82, стр. 301–314.

5. Zegers W. Translation in chloroplasts, *Biochimie*, 2000, 82, стр. 583–601.

6. Любецкий В.А., Селиверстов А.В. Некоторые алгоритмы, связанные с конечными группами, *Информационные процессы*, 2003, 3, 1, стр. 39–46.

7. Lyubetsky V.A., Seliverstov A.V. Note on Cliques and Alignments, *Информационные процессы*, 2004, 4, 3, стр. 241–246.