

ГЕНЕТИКА

УДК 575.852

**ИЗУЧЕНИЕ ВСТАВОК ПРЯМЫХ ПОВТОРОВ
В МИКРОЭВОЛЮЦИИ МИТОХОНДРИЙ И ПЛАСТИД РАСТЕНИЙ
НА ОСНОВЕ КЛАСТЕРИЗАЦИИ БЕЛКОВ****О.А. Зверков*, Л.Ю. Русин, А.В. Селиверстов*, В.А. Любецкий****(лаборатория геносистематики; e-mail: zverkov@iitp.ru)*

На основе кластеризации пластидных белков трех обширных групп растений и водорослей исследованы вставки точных прямых повторов и длины повторяемых участков в пластомах и митохондриях семенных растений. Предположено, что в ходе эволюции некодирующих участков ДНК прямые повторы часто возникают одномоментно, возможно, в результате репликативных ошибок, состоящих в удвоении участка ДНК. Обсуждается роль таких дупликаций в эволюции пластома.

Ключевые слова: вставки точных прямых повторов, пластиды, митохондрии, семенные растения, микроэволюция некодирующих участков ДНК, кластеризация пластидных белков.

Введение и постановка задачи. Широко рассматривается задача поиска в одном геноме *неточных* повторов некоторого участка ДНК (“слова”). Часто поиск выполняется с помощью программы REPuter [1]. В [2] эта программа применена к пластоми *Phoenix dactylifera* и в нем найдены 11 прямых и обратных неточных повторов с длиной повторяемого слова 30 п.н. В [3] с помощью REPuter найдены прямые неточные повторы с длиной повторяемого слова от 10 до 100 п.н. в пластомах пяти видов рода *Oenothera*. В [2] рассматриваются неточные повторы в двух видах рода *Silene*. При секвенировании новых пластов среди первичной информации о них обычно ищутся и неточные повторы слов [4–6]. Можно привести много примеров такого типа результатов, в которых рассматриваются единичные геномы, в частности пластома, и повторяемые слова не короткие (иначе возникает огромное перепредсказание), а повторы неточные (часто они имеют 90% точность).

В [7] на основе большого числа недавно секвенированных пластов рассматривалась эволюция большого инвертированного повтора, что не связано с изучением прямых точных повторов слов, которые, как показывают, в частности, наши результаты, имеют сравнительно небольшую длину. Работа [8] посвящена полиморфизмам в почти совпадающих геномах митохондрий некоторых близких растений родов *Oryza* и *Brassica*; в ней отмечено, что ранее рассматривался и вид *Beta vulgaris*. В нашей заметке сопоставлены более далекие друг от друга митохондрии видов из больших таксономических групп.

Существенное обстоятельство, побуждающее к исследованию *произвольной длины точных повторов* некоторого слова, состоит в том, что модель независимых однонуклеотидных замен и вставок, обсуждаемая, например, в [9], не позволяет объяснить их в случае слова с длиной четыре и более п.н. в некодирующих областях генома. Хотя такие вставки часто встречаются, например в хлоропластах.

В [9, 10] приведен перечень основных эволюционных событий в некодирующих областях хлоропластов и отмечена высокая частота вставок (точных) прямых повторов слов и особенно многократной вставки одного нуклеотида. Однако в [9, 10] рассматривались только короткие специальные участки геномов, например интроны первой группы гена *trnL* у *asterids*. Такие вставки использовались для построения и классификации деревьев видов [11].

Мы рассмотрим задачу поиска и подсчета числа вставок (точных) прямых повторов любой длины в некодирующих областях, включая интроны, в пластах и митохондриях видов семенных растений. Подчеркнем, что эта задача не сводится к поиску повторов в отдельных последовательностях и требует существенно *множественного выравнивания последовательностей*, например ради возможности отличить вставку от серии независимых нуклеотидных замен. Чтобы пояснить разницу между повторами и вставками повторов, мы дадим следующее определение. Вставка прямого повтора наблюдается как множественное выравнивание нуклеотидных последовательностей,

*Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН, г. Москва.

TGACTCTTTCAAG----ATTCAATTCATCTCT
 TGAATCTTGAAAG-----ATTCAATTCATCTCT
 TGGATCTTGAAAG-----ATTCAATTCATCTCT
 TGGATCTTGAAAG----ATTCAATTCATCTCT
 TGGATCTTGAAAGAAAGATTCAATTCATCTCT

Рис. 1. Множественное выравнивание участков 5'-лидерных областей гена *psbM*. Повторяемое слово φ выделено полужирным

в котором имеются две строки: одна (“образец”) содержит точный прямой повтор вида $\varphi\varphi$ некоторого слова φ , а в другой строке это $\varphi\varphi$ выравнивается на $\varphi-$ или $-\varphi$, где “-” означает делецию всех букв в φ ; при этом для образца и каждой последовательности выполняется: $\varphi\varphi$ выравнивается на один из четырех типов участков $\varphi\varphi$, $\varphi-$, $-\varphi$ или $--$. На рис. 1 приведено иллюстративное выравнивание с повтором слова длины 4 п.н., оно является частью выравнивания на рис. 2. На рис. 2 приведен пример неточного повтора, который произошел у *Keteleeria davidiana* в результате, вероятно, сначала вставки точного повтора слова длины 11 п.н., а затем замены в одной из его позиций.

Объекты и методы. В части пластид нами проведен широкомасштабный поиск вставок прямых по-

второв в следующих семействах, которые охватывают почти все семейства семенных растений, включающих хотя бы два вида с полностью секвенированным пластомом. После названия вида указан номер пласта в базе GenBank. **Acoraceae:** *Acorus calamus* (NC_007407), *Acorus americanus* (NC_010093). **Asteraceae:** *Lactuca sativa* (NC_007578), *Helianthus annuus* (NC_007977), *Guizotia abyssinica* (NC_010601), *Parthenium argentatum* (NC_013553). **Brassicaceae:** *Arabidopsis thaliana* (NC_000932), *Aethionema cordifolium* (NC_009265), *Aethionema grandiflorum* (NC_009266), *Olimarabidopsis pumila* (NC_009267), *Arabis hirsuta* (NC_009268), *Barbarea verna* (NC_009269), *Capsella bursa-pastoris* (NC_009270), *Crucihimalaya wallichii* (NC_009271), *Draba nemorosa* (NC_009272), *Lepidium virginicum* (NC_009273), *Lobularia maritima* (NC_009274), *Nasturtium officinale* (NC_009275). **Convolvulaceae:** *Cuscuta gronovii* (NC_009765), *Cuscuta reflexa* (NC_009766), *Cuscuta obtusiflora* (NC_009949), *Cuscuta exaltata* (NC_009963), *Ipomoea purpurea* (NC_009808). **Fabaceae:** *Lotus japonicus* (NC_002694), *Medicago truncatula* (NC_003119), *Glycine max* (NC_007942), *Phaseolus vulgaris* (NC_009259), *Cicer arietinum* (NC_011163), *Trifolium subterraneum* (NC_011828), *Vigna radiata*

Ca TTAGTTAGATTTTGTATTTCGAACCAATTTTATAAATT-----ACATCTCTTTTCA
 Cd TTAGTTAGATTCGGTTACTCGAACCGATTCTATTAATT-----CAATCACATCTATTTTCA
 Kd TTAGTTATATTCCATTACTCGAACCGATTCTATTCATTTTCTATTAAATTCATCATATCTATTTTCA
 Ps TTAGTTAGATTCGGTTACTCGAACCGATTCTATCAATT-----ACATCTATTTTCA
 Pc TTAGTTAGATTCGGTTACTCGAACCAATTCTATCAATG-----ACATAGATCTATTTTCA
 Pg TTAGTTAGATTCGGTTACTCGAACCGATTCCATAGATA-----AAATACATCTATTTTCA
 Pkr TTAGTTAGATTCGGTTACTCGAACCGATTCCATAGATA-----AAATACATCTATTTTCA
 Pk TTAGTTAGATTCGGTTACTCGAACCGATTCCATAGATA-----AAATACATCTATTTTCA
 Pt TTAGTTAGATTCGGTTACTCGAACCAATTCTATCAATG-----ACATACATCTATTTTCA
 Cj TCAGATTGATCCTATTGATGGAATT-ACTCCATGGATT-----CAATCCATTTTTT

Ca TGACTCTTTCAAG----ATTCAATTCATCTCT-----ATGAGATCAAAATTCGAGCTATTTT
 Cd TGAATCTTGAAAG-----ATTCAATTCATCTCT-----ATGAGATCAAAATCTCGAGCTATTTT
 Kd TGGATCTTGAAAG-----ATTCAATTCATCTCT-----ATGAGATAAAATCTCGAGCTATTTT
 Ps TGGATCTTGAAAG----ATTCAATTCATCTCT-----ATGAAATTAAATCTCGAGCTATTTT
 Pc TGGATCTTGAAAGAAAGATTCAATTCATCTCTATGAGATCAAAATGAGATCAAAATGAGATCAAAATCTCGAGCTATTTT
 Pg TGGATCTTGAAAG----ATTCAATTCATCTCT-----ATGAGATAAAATCTCGAGCTATTTT
 Pkr TGGATCTTGAAAG----ATTCAATTCATCTCT-----ATGAGATAAAATCTCGAGCTATTTT
 Pk TGGATCTTGAAAG----ATTCAATTCATCTCT-----ATGAGATAAAATCTCGAGCTATTTT
 Pt TGGATCTTGAAAGAAAGATTCAATTCATCTCTATGAGATCAAAATGAGATCAAAATGAGATCAAAATCTCGAGCTATTTT
 Cj TTTGAC-----AAAAAATGAGATACTCT-----ATGAGATCAAAATCTCGAGCTATTTT

Ca T-GAACGAAGTAAAAATAAGGAGATC
 Cd TAGAACGAAGTAAAAATAAGGAGATC
 Kd T-AAACGAAGTAAAAATCAGGGGATC
 Ps TTTAACGAAGTAAAGATCAGGAGATC
 Pc TTGAACAAAGTTAAAAATAAGGAGATC
 Pg TGGAACAAAGTGAAAATCAGGAGATC
 Pkr TGGAACAAAGTGAAAATCAGGAGATC
 Pk TGGAACAAAGTGAAAATCAGGAGATC
 Pt TTGAACAAAGTTAAAAATAAGGAGATC
 Cj T----AAAACGAAGGGAAAATCAATC

Рис. 2. Множественное выравнивание участков 5'-лидерных областей гена *psbM*, в которых возникли трехкратные прямые повторы слов длиной 11 п.н. и двукратные повторы слов длиной 4 п.н. Подчеркнуты повторы, а в случае *K. davidiana* подчеркнуто и продолжение повтора, которое отличается от повторяемого слова в одной позиции, — выделено серым цветом. Полужирным выделены повторяемые слова φ . Обозначения видов: Ca — *C. argyrophylla*, Cd — *C. deodara*, Kd — *K. davidiana*, Ps — *P. sitchensis*, Pc — *P. contorta*, Pg — *P. gerardiana*, Pkr — *P. krempfii*, Pk — *P. koraiensis*, Pt — *P. thunbergii*, Cj — *C. japonica*

(NC_013843), *Pisum sativum* (NC_014057), *Lathyrus sativus* (NC_014063). **Geraniaceae**: *Pelargonium* × *hortorum* (NC_008454), *Erodium texanum* (NC_014569), *Geranium palmatum* (NC_014573), *Monsonia speciosa* (NC_014582). **Malvaceae**: *Gossypium hirsutum* (NC_007944), *Gossypium barbadense* (NC_008641). **Myrtaceae**: *Eucalyptus globulus* (NC_008115), *Eucalyptus grandis* (NC_014570). **Nymphaeaceae**: *Nymphaea alba* (NC_006050), *Nuphar advena* (NC_008788). **Oleaceae**: *Jasminum nudiflorum* (NC_008407), *Olea europaea* (NC_013707). **Onagraceae**: *Oenothera elata* (NC_002693), *O. argillicola* (NC_010358), *O. glazioviana* (NC_010360), *O. biennis* (NC_010361), *O. parviflora* (NC_010362). **Pinaceae**: *Pinus thunbergii* (NC_001631), *Pinus koraiensis* (NC_004677), *Picea sitchensis* (NC_011152), *Pinus contorta* (NC_011153), *Pinus gerardiana* (NC_011154), *Pinus krempfii* (NC_011155), *Keteleeria davidiana* (NC_011930), *Cedrus deodara* (NC_014575), *Cathaya argyrophylla* (NC_014589). **Poaceae** (BEP): *Oryza sativa Japonica Group* (NC_001320), *Triticum aestivum* (NC_002762), *O. nivara* (NC_005973), *O. sativa Indica Group* (NC_008155), *Hordeum vulgare* (NC_008590), *Agrostis stolonifera* (NC_008591), *Lolium perenne* (NC_009950), *Brachypodium distachyon* (NC_011032), *Festuca arundinacea* (NC_011713), *Bambusa oldhamii* (NC_012927), *Dendrocalamus latiflorus* (NC_013088). **Poaceae** (PACCAD): *Zea mays* (NC_001666), *Saccharum hybrid SP80-3280* (NC_005878), *Saccharum officinarum* (NC_006084), *Sorghum bicolor* (NC_008602), *Coix lacryma-jobi* (NC_013273). **Ranunculaceae**: *Ranunculus macranthus* (NC_008796), *Megaleranthis saniculifolia* (NC_012615). **Saliceae**: *Populus alba* (NC_008235), *Populus trichocarpa* (NC_009143). **Solanaceae**: *Nicotiana tabacum* (NC_001879), *Atropa belladonna* (NC_004561), *Nicotiana sylvestris* (NC_007500), *Nicotiana tomentosiformis* (NC_007602), *Solanum lycopersicum* (NC_007898), *Solanum bulbocastanum* (NC_007943), *Solanum tuberosum* (NC_008096). Также рассмотрен пластом *Cryptomeria japonica* (NC_010548) из семейства **Cupressaceae**, который относительно плохо выравнивается с пластами других видов и на рис. 2 играет роль внешней группы.

В части митохондрий вместо семейства **Brassicaceae** рассмотрены порядки **Brassicales**: *Brassica napus* (NC_008285), *Arabidopsis thaliana* (NC_001284), *Carica papaya* (NC_012116); **Caryophyllales**: *Beta vulgaris* subsp. *vulgaris* (NC_002511), *Beta vulgaris* subsp. *maritima* (NC_015099). А также рассмотрены митохондрии видов семейства **Poaceae**: *Triticum aestivum* (NC_007579), *Sorghum bicolor* (NC_008360), *Oryza sativa Indica Group* (NC_007886), *Oryza sativa Japonica Group* (NC_011033), *Oryza rufipogon* (NC_013816), *Zea mays* subsp. *mays* (NC_007982), *Zea perennis* (NC_008331), *Zea mays* subsp. *parviglumis* (NC_008332), *Zea luxurians* (NC_008333), *Tripsacum dactyloides* (NC_008362).

Для решения указанной задачи нами использовалась следующая база данных, представляющая и самостоятельный интерес. База содержит результаты кластеризации (разбиения на семейства родственных

по последовательности) пластидных белков из трех обширных групп: родофитной и хлорофитной ветвей и однодольных растений. Для ее построения разработан оригинальный алгоритм кластеризации.

Непосредственно вставки прямых повторов искались следующим образом. Для каждой пары видов искались пары позиционно сцепленных генов или экзонов из близких видов, при этом очень короткие неконсервативные гены игнорировались. Затем собранные вместе некодирующие области между такими генами или экзонами из всех видов одного семейства выравнивались, на выравнивании определялись вставки прямых повторов. Для этого использовались программы [12–14].

Результаты. Получена база данных, отражающая кластеризацию белков, кодируемых в пластидах указанных групп растений [15]. Поиск кластеров по филогенетическому профилю белка на ее основе доступен по адресу <http://lab6.iitp.ru/ppc/>. Там же расположено описание упомянутого алгоритма и инструкции по его применению. Коротко напомним смысл параметров алгоритма. Параметр H — верхний порог сходства белков из разных кластеров (белки со сходством выше H не разделяются). Параметр L — нижний порог сходства (белки со сходством ниже L считаются не обладающими сходством). Параметр p — верхний порог размера кластера относительно числа видов N (кластеры с числом белков больше $p \times N$ всегда разделяются, если в них есть ребро с весом ниже H). Приведем результаты для однодольных растений, для которых наилучших результатов удалось достичь при значениях параметров: $p = 2$, $L = 0$, $H = 0,41$. Полученные при этих значениях кластеры были исправлены вручную: в кластер PetG добавлены белки YP_654227.1 из *Oryza sativa Indica Group* и YP_358627.1 из *Phalaenopsis aphrodite*; в кластер RpL23 добавлены белки YP_874745.1 из *Agrostis stolonifera* и YP_899416.1 из *Sorghum bicolor*; в кластер RpL2 добавлены паралоги YP_654244.1 и YP_654261.1 из *Oryza sativa Indica Group*. Так образовано 105 неоднородных кластеров и 20 одноэлементных. Из неоднородных кластеров 71 содержит не более одного белка из одного вида, 30 содержат пары белков из одного вида, 2 содержат по три белка из одного вида и 2 содержат по 4 белка из одного вида. В 29 (неоднородных) кластерах представлены от одного до 12 видов; нет кластеров, в которых представлены от 13 до 30 видов; в 76 кластерах представлены от 31 до 36 видов (границы диапазонов везде включены). Максимум распределения — 30 кластеров с 35 видами.

В результате получены следующие распределения числа вставок прямых повторов в зависимости от длины слова φ в семействах цветковых; после двоеточия используется запись “длина слова φ : число найденных вставок с этой длиной”. В семействе **Acoraceae**: 1: 9. В семействе **Asteraceae**: 1: 153, 2: 10, 3: 5, 4: 9, 5: 27, 6: 31, 7: 9, 8: 4, 9: 1, 10: 5, 11: 1, 12: 2, 13: 1, 17: 2, 18: 2, 21: 2, 22: 3, 23: 1. В семействе **Brassicaceae**: 1: 1373, 2: 82, 3: 26, 4: 31, 5: 66, 6: 57, 7: 32, 8: 16, 9: 5, 10: 3, 11: 3,

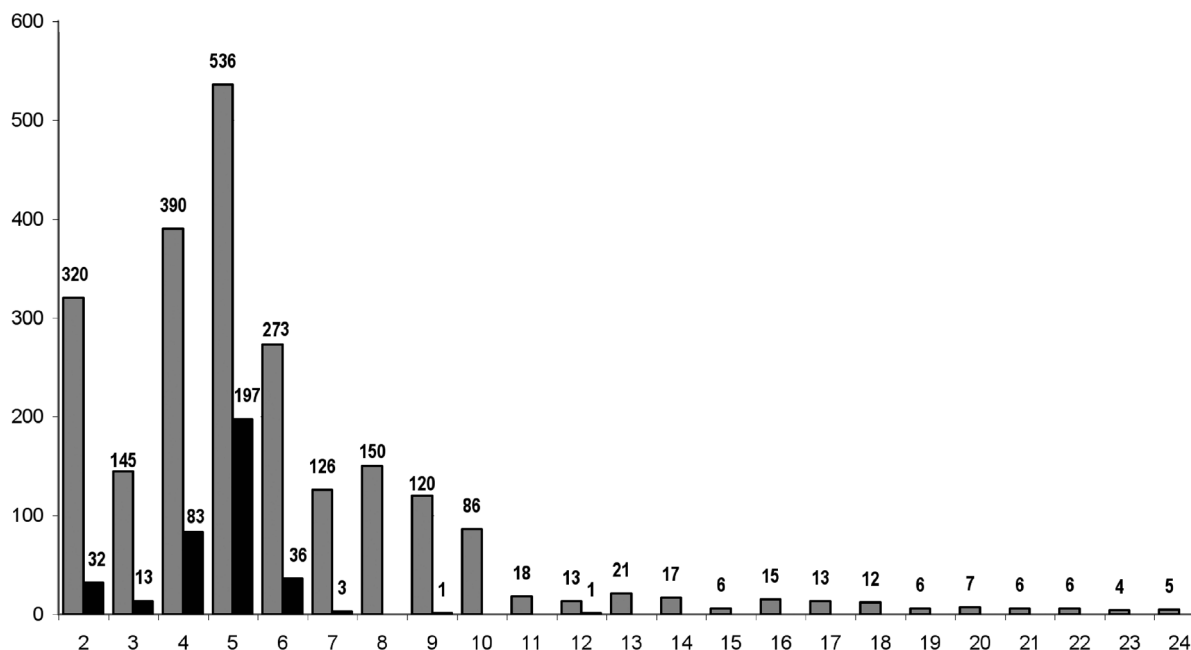


Рис. 3. Число всех найденных вставок прямых повторов в зависимости от длины повторяемого слова, результаты показаны для длин от 2 до 24 п.н. Черные столбики показывают числа вставок прямых повторов в пластидах, серые в митохондриях

12: 1, 13: 6, 15: 1, 16: 1, 17: 2, 20: 3, 22: 1, 23: 1. В семействе Fabaceae: 1: 783, 2: 84, 3: 41, 4: 91, 5: 71, 6: 25, 7: 10, 8: 3, 10: 1, 12: 2, 13: 1, 14: 1, 16: 2. В семействе Geraniaceae: 1: 186, 2: 14, 3: 6, 4: 25, 5: 34, 6: 21, 7: 9, 8: 2, 9: 1. В семействе Convolvulaceae: 1: 168, 2: 8, 3: 4, 4: 6, 5: 4, 6: 7, 7: 3, 8: 2, 15: 1, 18: 1, 19: 1. В семействе Malvaceae: 1: 45, 2: 1, 4: 2, 5: 8, 6: 4, 7: 2, 8: 2, 51: 1. В семействе Myrtaceae: 1: 34, 6: 1, 18: 1, 19: 1. В семействе Nymphaeaceae: 1: 58, 2: 8, 3: 3, 4: 22, 5: 32, 6: 8, 7: 5, 9: 2, 11: 1, 14: 2, 24: 1. В семействе Oleaceae: 1: 46, 2: 4, 3: 3, 4: 5, 5: 6, 6: 3, 7: 1, 8: 2, 11: 1, 12: 1, 14: 1, 17: 1. В семействе Onagraceae: 1: 114, 2: 6, 3: 1, 4: 5, 5: 14, 6: 10, 7: 4, 9: 1, 11: 1, 12: 3, 13: 2, 15: 2, 16: 2, 17: 1, 19: 1, 20: 1, 21: 2, 22: 2, 23: 2, 24: 2, 26: 1, 29: 1, 35: 1, 50: 1, 78: 1. В семействе Pinaceae: 1: 378, 2: 41, 3: 38, 4: 104, 5: 99, 6: 23, 7: 6, 8: 4, 9: 4, 10: 2, 11: 3, 12: 3, 13: 2, 16: 1. В кладе ВЕР семейства Rosaceae: 1: 695, 2: 22, 3: 9, 4: 47, 5: 108, 6: 52, 7: 15, 8: 6, 9: 6, 10: 1, 11: 1, 13: 2, 14: 2, 15: 1, 16: 1, 18: 1, 19: 1, 21: 2, 24: 1, 27: 1. В кладе РАССАД семейства Rosaceae: 1: 202, 2: 7, 3: 4, 4: 11, 5: 30, 6: 7, 7: 5, 8: 2, 9: 2, 14: 3, 16: 3, 17: 3, 18: 3, 19: 1, 20: 1, 26: 1. В семействе Ranunculaceae: 1: 111, 2: 13, 3: 3, 4: 12, 5: 22, 6: 7, 7: 5. В семействе Salicaceae: 1: 49, 2: 2, 4: 1, 6: 4, 7: 6, 8: 1, 9: 1, 10: 1, 11: 4, 13: 2, 14: 4, 16: 1, 17: 1, 18: 1, 20: 1. В семействе Solanaceae: 1: 238, 2: 18, 3: 2, 4: 19, 5: 15, 6: 13, 7: 14, 8: 10, 9: 7, 10: 3, 11: 3, 12: 1, 13: 5, 14: 4, 15: 1, 16: 4, 17: 3, 18: 3, 19: 1, 20: 1, 24: 1, 30: 1.

Прямые повторы слов с наибольшей длиной найдены в семействе Onagraceae: 50 п.н. между генами *petA* и *psbJ* и 78 п.н. между генами *atpH* и *atpI*. В семействе Malvaceae найден повтор слова длиной 51 п.н. между генами *petN* и *psbM*. Преобладают вставки слова из одного нуклеотида, их найдено 4642. Среди вставок прямых повторов слов большей длины наиболее

часто встречаются слова длины 5 п.н., что видно из рис. 3. Вставки повторов слов длиной более 24 п.н. наблюдаются очень редко. Это — две вставки при длине слова 26 и по одной вставке при длинах слов 27, 29, 30, 35, 50, 51, 78. При поиске вставок отсутствовало какое-либо ограничение на длину повтора.

Рассмотрим пример вставок прямых повторов в 5'-лидерных областях гена *psbM* в семействе Pinaceae, рис. 2. Здесь имеются троекратные повторы слова длиной 11 п.н. у *K. davidiana* и *P. thunbergii*, которые расположены в разных местах относительно иницирующего кодона *psbM* и различны по нуклеотидному составу, что указывает на их независимое возникновение в эволюции. У *C. argyrophylla*, *C. deodara*, *P. sitchensis*, *P. gerardiana*, *P. krempfii*, *P. koraiensis* перед этим геном отсутствуют повторы слов длины 11 п.н., у *P. contorta* и *P. thunbergii* имеется повтор слова ATGAGATCAAA, который отсутствует у остальных видов. У *K. davidiana* имеется повтор слова TTCTATTTCATT с одним отличием в третьей копии, который отсутствует у двух видов сосен и расположен дальше от начала гена. Также в этой области наблюдаются вставки прямых повторов слова AAAG у *P. contorta* и *P. thunbergii* и слова ATTTC у *C. argyrophylla*, *P. sitchensis* и *Pinus* spp. Можно думать, что 5'-лидерная область гена *psbM* удобна для возникновения повторов участков независимо от их нуклеотидного состава. У близкого вида *C. japonica*, не входящего в семейство Pinaceae, повторы перед *psbM* отсутствуют, что подтверждает возникновение, а не потерю повторов в этом семействе. Отметим, что области, расположенные по соседству от этих повторов, также обладают заметной консервативностью.

При исследовании митохондрий получены следующие распределения числа вставок прямых повторов

в зависимости от длины слова φ . В порядке Brassicales: 1: 159, 2: 20, 3: 7, 4: 12, 5: 16, 6: 5; в семействе Poaceae: 1: 164, 2: 12, 3: 6, 4: 71, 5: 181, 6: 31, 7: 3, 9: 1, 12: 1. Так же как и в пластидах, абсолютный максимум числа вставок прямых повторов приходится на длину 1 повторяемого слова, а локальный максимум — на длину 5.

Обсуждение и выводы. Нами анализировались пластомы и митохондрии близких видов с целью расширить и уточнить представления об эволюции некодирующих участков ДНК. Полученные результаты показывают, что вставки (точных) прямых повторов являются частыми элементарными событиями в микроэволюции относительно коротких некодирующих участков ДНК пластид и митохондрий. Повторяемое слово обычно имеет длину пять. Распределение длин этих слов в пластомах и митохондриях сходно (рис. 3).

СПИСОК ЛИТЕРАТУРЫ

1. Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale // Nuc. Acids Res. 2001. Vol. 29. P. 4633—4642.
2. Yang M., Zhang X. et al. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.) // PLoS ONE. 2010. Vol. 5. 9. e12762.
3. Greiner S., Wang X., Rauwolf U., Silber M.V., Mayer K., Meurer J., Haberer G., Herrmann R.G. The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution // Nuc. Acids Res. 2008. Vol. 36. N 7. P. 2366—2378.
4. Ogihara Y., Terachi T., Sasakuma T. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species // Proc. Natl. Acad. Sci. USA. 1988. Vol. 85. N 22. P. 8573—8577.
5. Cai Z., Guisinger M., Kim H.G., Ruck E., Blazier J.C., McMurtry V., Kuehl J.V., Boore J., Jansen R.K. Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions // J. Mol. Evol. 2008. Vol. 67. N 6. P. 696—704.
6. Timme R.E., Kuehl J.V., Boore J.L., Jansen R.K. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats // Amer. J. Bot. 2007. Vol. 94. N 3. P. 302—312.
7. Moore M.J., Hassan N., Gitzendanner M.A., Bruenn R.A., Croley M., Vandeventer A., Horn J.W., Dhingra A., Brockington S.F., Latvis M., Ramdial J., Alexandre R., Piedrahita A., Xi Z., Davis C.C., Soltis P.S., Soltis D.E. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region // Intern. J. Plant Sci. 2011. Vol. 172. N 4. P. 541—558.
8. Honma Y., Yoshida Y., Terachi T., Toriyama K., Mikami T., Kubo T. Polymorphic minisatellites in the mitochondrial DNAs of *Oryza* and *Brassica* // Current genetics. 2011. Vol. 57. N 4. P. 261—270.
9. Borsch T., Quandt D. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA // Plant Syst. Evol. 2009. Vol. 282. P. 169—199.
10. Kelchner S.A. The evolution of non-coding chloroplast DNA and its application in plant systematics // Ann. Missouri Bot. Garden. 2000. Vol. 87. P. 482—498.
11. Ingvarsson P.K., Ribstein S., Taylor D.R. Molecular evolution of insertions and deletion in the chloroplast genome of *Silene* // Mol. Biol. Evol. 2003. Vol. 20. N 11. P. 1737—1740.
12. Официальный сайт лаборатории Института проблем передачи информации им. А.А. Харкевича РАН (URL: <http://lab6.iitp.ru/ru/repeats> 25.08.2011).
13. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Res. 2004. Vol. 32. P. 1792—1797.
14. Официальный сайт лаборатории Института проблем передачи информации РАН (URL: <http://lab6.iitp.ru/en/treal/> 25.08.2011).
15. Зверков О.А., Селиверстов А.В., Любецкий В.А. Белковые семейства, специфичные для пластомеров небольших таксономических групп водорослей и простейших // Молекулярная биология. 2012. Т. 46. № 5. С. 799—809.
16. Селиверстов А.В., Любецкий В.А. Прямые повторы в некодирующих областях хлоропластов у семенных растений // Тр. 52-й науч. конф. Моск. физ.-тех. ун-та. 2009. Т. 1. № 1. P. 116—117.
17. Зверков О.А., Русин Л.Ю., Селиверстов А.В., Любецкий В.А. Вставки прямых повторов в микроэволюции пластид и митохондрий семенных растений // Информационные процессы. 2012. Т. 12. № 3. С. 191—197.

* * *

Работа выполнена при частичной поддержке Министерства образования и науки РФ (14.740.11.1053, 8858, 8481, 8091, 8823).

Поступила в редакцию
25.08.2011

A STUDY OF DIRECT REPEATS IN MICROEVOLUTION OF PLANT MITOCHONDRIA AND PLASTIDS BASED ON PROTEIN CLUSTERING

O.A. Zverkov, L.Y. Rusin, A.V. Seliverstov, V.A. Lyubetsky

The study focuses on insertions of perfect direct repeats of words of arbitrary length in plastomes and mitochondriomes. The approach is exemplified on seed plants. Plastomes of close species were analyzed to further develop and refine published evidence on the evolution of non-coding DNA. The results suggest that insertions are common elementary events in microevolution of short non-coding DNA regions. The repeated word length is usually 5, the word length distribution is similar between plastomes and mitochondriomes.

Key words: *perfect direct repeat insertions, plastids, mitochondria, Spermatophyta, microevolution, protein clustering.*

Сведения об авторах

Зверков Олег Анатольевич — мл. науч. сотр., Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН. Тел.: 8-495-694-33-38; e-mail: zverkov@iitp.ru.

Русин Леонид Юрьевич — канд. биол. наук, науч. сотр. лаборатории геносистематики биологического факультета МГУ, Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН. Тел.: 8-495-694-33-38, e-mail: rusin@iitp.ru.

Селиверстов Александр Владиславович — канд. физ.-мат. наук, ст. науч. сотр., Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН. Тел.: 8-495-694-33-38, e-mail: slvstv@iitp.ru.

Любецкий Василий Александрович — докт. физ.-мат. наук, зав. лаб., Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН. Тел.: 8-495-694-33-38, 8-910-464-69-17; e-mail: lyubetsk@iitp.ru