О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Н.Н. Усанов, Н.Г. Усанов

Симметричные структуры в комплементарных цепях геномных ДНК и возможные алгоритмы их организации

Препринт №

Москва

Ордена Ленина ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
им. М.В. Келдыша
Российской академии наук

Н.Н. Усанов, Н.Г. Усанов

СИММЕТРИЧНЫЕ СТРУКТУРЫ В КОМПЛЕМЕНТАРНЫХ ЦЕПЯХ ГЕНОМНЫХ ДНК И ВОЗМОЖНЫЕ АЛГОРИТМЫ ИХ ОРГАНИЗАЦИИ

Москва
2011

АННОТАЦИЯ

При подсчете олигонуклеотидных ($N < 12$) субпоследовательностей в полных текстах первой и второй цепей ДНК геномов прокариот и эукариот образуются массивы данных с высокими степенями линейной корреляции, вычисленной по Пирсону. Корреляции вырождаются при изменении алгоритма генерации последовательности комплементарной цепи и уменьшении длины обрабатываемых фрагментов ДНК. Высокая симметрия первой и второй цепей геномов обусловливается наличием в геномных ДНК большого числа комплементарных разделенных палиндромов, которые могут рассматриваться в виде совершенных, инвертированных, разнесенных повторов. В компьютерной модели, работающей по алгоритму генерации случайных повторов, а также крупноблочных делеций, поддерживающих постоянство длины трансформируемого *in silico* нуклеотидного текста, показано возникновение обсуждаемой симметрии цепей в матрице псевдослучайной ДНК.

Авторы признательны коллегам из Института биологии Уфимского научного центра РАН, где выполнялась основная часть данной работы, и участникам школы «Будущее прикладной математики» за поддержку и обсуждение.

N.N. Usanov, N.G. Usanov

Symmetrical structures in the complementary chains of the genomic DNA and the possible algorithms of their organization

ABSTRACT

Calculation of substrings ($N < 12$) quantity in the nucleotide sequences of the first and second DNA strands of prokaryotic and eukaryotic genomes form the arrays of data with high degrees of the linear Pearson correlation. Changing the algorithm of complementary strands generation (second from the first) or decreasing the length of genomic DNA fragments being analyzed causes the correlations to extinct. The high degree of symmetry of the first and second strands of genomes is explained by presence of the large number of complementary divided palindromes, which may be considered as the form of perfect, inverted, diverse repeating regions. The appearance of the discussed complementary chains symmetry in the pseudorandom DNA matrix is shown via the computer model, which works on the random repeats generation algorithm and long deletions, which ensure the constancy of the nucleotide text length being transformed *in silico*.

Authors sincerely thanks collaborators from Biology institute of Ufa Scientific Center of RAS where the main part of this work was done and participants of school “Future of Applied Mathematics” for support and discussion.

Работа выполнена при частичной поддержке РФФИ (проект 07-01-00618-а).

1. Введение

Первая непрерывная последовательность ДНК большой длины (1,80 млн.п.н.[1]) бактерии *Haemophilus influenzae* была прочитана в 1995 году [1], и это событие, по существу ознаменовало начало качественно нового этапа в развитии геномики. Последующий прогресс в секвенировании геномов прокариот и эукариот привел к возникновению колоссальных по своему объему баз данных [2], в которые были внесены полные коды нуклеотидных комбинаций, устанавливающих биохимическое, молекулярное, морфологическое и популяционное функционирование сотен организмов с различной клеточной организацией. Вместе с тем, лишь одно накопление новой информации без развития новых методических путей анализа текстов ДНК, постановки новых задач, поиска новых фундаментальных закономерностей не может дать качественного изменения понимания.

Одним из известных способов анализа генетических текстов является частотный подсчет комбинаций нуклеотидов, например, их двойных и тройных сочетаний. Считается, что в кодирующих участках геномной ДНК эти символьные последовательности подчиняются определенным правилам, тогда как в промежутках между генами, там, где нет ничего существенного, частота отдельных олигонуклеотидных «слов» может быть низкой или, наоборот, высокой. Неудивительно, что поиску локальных корреляций в текстах ДНК, равно как и разработке математических методов исследования этой молекулы, написанию компьютерных программ, предназначенных для изучения генома, посвящено очень много работ. Соответствующая информация может быть найдена, например, в книгах Б. Вейера [3] и М.С. Уотермена [4], Д. Гасфилда [5], А.А. Александрова с соавторами [6], изданных на русском языке, а также в многочисленных статьях [7-9], обзорах [10] и диссертациях [11].

Вопреки первоначальным ожиданиям, последовательности нуклеотидов в уже прочитанных протяженных последовательностях геномных ДНК не нашли объяснений, основанных только на механизмах простой мутационной изменчивости, которые должны были бы совершенствовать структуру текстов методом случайных ошибок. Результатом подобных изменений, конечно же, могли бы стать коды, нормально функционирующие на уровне трансляции в РНК, и белки, придающие последним необходимую функциональность. Но случайная мутационная изменчивость никак не объясняет присутствие в текстах геномных ДНК многочисленных элементов «инфраструктуры», обнаруживаемых в виде повторяющихся последовательностей, палиндромов, т.е. буквосочетаний, одинаково читающихся в обоих направлениях, [12] или других симметричных элементов. В этой связи, в настоящее время принято считать, что молекулярная эволюция геномной ДНК является суммой многих процессов [13-14], в которых мутационная изменчивость является лишь частью.

---

[1] Аббревиатура «п.н.» означает: пар нуклеотидов; млн.п.н. – миллионов пар нуклеотидов.

Необходимо отметить еще один момент, и это, в частности, касается методической стороны настоящей работы. При установлении структурных особенностей текстов ДНК последние всегда рассматриваются в виде лидирующей, т.е. непрерывно синтезируемой цепи ДНК, направление которой (5′–3′) совпадает с направлением движения репликационной вилки. И это справедливо на уровне текстов отдельных генов. Между тем, уже составленные многочисленные карты геномов как эукариот, так и прокариот показывают, что трансляция информации с ДНК на РНК, в равной степени может происходить как с первой, так и со второй цепи, считающейся условно реверсной. Более того, абсолютное направление генома нельзя выбрать, исходя из ориентации репликации ДНК. Так, в геноме *E.coli* K12 (NC_000913), имеется один центр инициации репликации oriC, инициирующий двустороннюю репликацию, ориентированную как по часовой стрелке, так и против нее. Оба процесса заканчиваются на противоположном участке генетической карты [15]. Таким образом, выбор главных (первых) цепей геномов, тексты которых имеются в общедоступных базах данных, во многом является условным. Можно привести интересный пример, когда в течение нескольких лет (2000-03 г.г.), геномы двух близкородственных штаммов бактерии *Chlamydophila pneumoniae* CWL029 (NC_000922.1) *Chlamydophila pneumoniae* J138 (NC_002491.1) были представлены в виде записей комплементарных к *Chlamydophila pneumoniae* AR39 (NC_002179.1), текст которой, к тому же, имел иные начальные координаты. Ситуация была исправлена после публикации в NCBI генома четвертого близкородственного штамма *Chlamydophila pneumoniae* TW-183, NC_005043.1 авторы последовательности которой избрали первый вариант написания. Лишь после этого направление и точка отсчета ДНК NC_002179.1 были также скорректированы.

Нам представлялось вероятным, что при статистическом исследовании лишь одного текста первой цепи хромосом, могут быть упущены какие-то интересные моменты, так как протяженные нуклеотидные последовательности несут в себе приблизительно равноценное количество генов, транслируемых в молекулы РНК с обеих цепей ДНК.

Алгоритм получения реверсно-комплементарного текста второй цепи из первой не подразумевает даже его схожесть с исходным вариантом, что не позволяет дать априорных ответов на такие вопросы, как, например, насколько частоты встречаемости отдельных олигонуклеотидных сочетаний, найденных в первой цепи, будут соответствовать таковым, обнаруженным для комплементарной, условно реверсной второй.

1.1. Алгоритм получения реверсно-комплементарного текста ДНК

Предположим, что произвольная гипотетическая последовательность[2] **TTAACCGGGGGGAATGGG,** является фрагментом текста генома. Представленная запись по умолчанию также означает, что соседние нуклеотиды между собой

[2] Синонимы: строка, запись, текст.

соединены в цепи фосфодиэфирной связью, образованной 5'-фосфатной (5'-PO_3) и 3'-гидроксильной (3'-OH) группами. Чтобы подтвердить или акцентировать указанную направленность текста, вышеприведенная запись может иметь следующий вид:

(5')TTAACCGGGGGGAATGGG(3').

Именно в такой форме с направлением **(5')→(3')** строки нуклеотидных последовательностей размещаются в соответствующих базах данных, например, генбанке национального центра биотехнологической информации (NCBI – http://www.ncbi.nlm.nih.gov/Genbank).

Вторая, комплементарная цепь нуклеотидов, строка которой при написании располагается под первой, имеет относительную направленность (3')→(5'), поэтому для ее генерации имеющийся текст первой цепи переписывается справа налево:

(3')GGGTAAGGGGGGCCAATT(5'),

и в нем производится замена нуклеотидов по принципам комплементарности: A→T, G→C, T→A, C→G. Итоговая строка выглядит следующим образом:

(5')CCCATTCCCCCCGGTTAA(3').

Из-за того, что первая и вторая цепи ДНК антипараллельны, второй текст перед совмещением его с первым должен быть еще раз переписан в обратном направлении:

(3')AATTGGCCCCCCTTACCC(5').

Итоговая запись двух антипараллельных комплементарных цепей будет иметь вид:

(5')TTAACCGGGGGGAATGGG(3') первая цепь
(3')AATTGGCCCCCCTTACCC(5') вторая цепь

Для всех манипуляций со второй цепью, например, для поиска сайтов рестрикции (последовательностей нуклеотидов в молекуле ДНК, которые определяют места ее специфического расщепления ферментом рестриктазой), последняя запись должна быть приведена в соответствующее направление (5')→(3'), то есть снова записана в реверсном виде:

(5')CCCATTCCCCCCGGTTAA(3').

1.2. Частотный анализ и составление словарей подпоследовательностей

Выше отмечалось, что одним из направлений анализа текстов ДНК является построение и анализ словарей нуклеотидных последовательностей (субпоследовательностей, подпоследовательностей). Общепринятым считается алгоритм действий, в котором назначается *устройство считывания*, представляющее собой математический инструмент, выделяющий с помощью подвижной *рамки* фрагмент нуклеотидного текста, заносящий его в список и назна-

Таблица. 1. Подсчет трех нуклеотидных сочетаний

Слово	В 1-ой цепи	Во 2-ой цепи
AAC	1	0
AAT	1	0
ACC	1	0
ATG	1	0
ATT	0	1
CAT	0	1
CCA	0	1
CCC	0	5
CCG	1	1
CGG	1	1
GAA	1	0
GGA	1	0
GGG	5	0
GGT	0	1
GTT	0	1
TAA	1	1
TCC	0	1
TGG	1	0
TTA	1	1
TTC	0	1

чающий равным единице специальный показатель – число копий фрагмента в исследуемом генетическом тексте.

Устройство считывания характеризуется размером – числом считываемых за один шаг нуклеотидов, шагом, с которым рамка движется по исследуемому тексту, а также направлением движения. Рамка считывания может включать в себя любое число нуклеотидов: два, три, четыре пять и более, при этом будут получены все возможные варианты триплетов, тетраплетов, пентаплетов и других многобуквенных слов, присутствующих в исследуемой строке ДНК. В случае отсутствия в словаре какого-либо олигонуклеотидного сочетания число его копий назначается равным нулю, но строка не удаляется. Движение осуществляться в любом направлении от начала (конца) текста, например, (5')→(3').

Проиллюстрируем составление списка (словаря) субпоследовательностей нашего фрагмента ДНК, приведенного в разделе 1.1, выполнив частотный анализ одновременно на обеих цепях с рамкой считывания размером три и пять нуклеотидов. Результаты подсчета для первой и второй цепей представим в табл. 1 и 2, соответственно.

```
          ←——
(5')TTAACCGGGGGGAATGGG(3')      первая цепь
(3')AATTGGCCCCCCTTACCC(5')      вторая цепь
              ——→
```

Сравнивая числа подпоследовательностей из 3 и 5 нуклеотидов, из которых складываются первая и вторая цепи случайного фрагмента ДНК, нетрудно

Таблица. 2. Подсчет 5-нуклеотидных сочетаний

Слово	В 1-ой цепи	Во 2-ой цепи	Слово	В 1-ой цепи	Во 2-ой цепи
AACCG	1	0	**CCGGT**	0	1
AATGG	1	0	**CGGGG**	1	0
ACCGG	1	0	**CGGTT**	0	1
ATGGG	1	0	**GAATG**	1	0
ATTCC	0	1	**GGAAT**	1	0
CATTC	0	1	**GGGAA**	1	0
CCATT	0	1	**GGGGA**	1	0
CCCAT	1	0	**GGGGG**	2	0
CCCCC	0	2	**GGTTA**	0	1
CCCCG	0	1	**GTTAA**	0	1
CCCGG	0	1	**TAACC**	1	0
CCGGG	1	0	**TCCCC**	0	1

убедиться в том, что в них одновременно присутствует лишь небольшая часть одинаковых триплетов (CCG, CGG, TAA, TTA) и полностью отсутствуют совпадения на уровне 5-нуклеотидных комбинаций. Проводить поиск схожести на более высоком уровне – 6, 7, 8 и более нуклеотидных слов, в данном случае не имеет смысла.

2. Сравнение словарей первой и второй цепей геномных ДНК

Словари нуклеотидных подпоследовательностей ДНК легко сгенерировать, используя пакет программ утилит[3], написанный нами на языке Паскаль и находящийся по адресу http://www.insilico.ru/programs/frequency.rar. В указанном архиве находятся программы Gclean.exe, Mirror.exe, random.exe, Abr.exe, Am.exe и др., а об их назначении рассказано в описании readme.txt, прилагаемом к пакету.

В качестве объектов исследования будем использовать тексты геномов бактерий или архей (одноклеточных прокариотов, на молекулярном уровне заметно отличающихся как от бактерий, так и от эукариотов), а также отдельные хромосомы эукариотических организмов, являющихся молекулярно-генетическими «стандартами»: *Arabidopsis thaliana* и *Drosophila melanogaster*. В качестве ссылок на тексты геномов показаны номера, под которыми они депонированы в базе данных NCBI (ftp://ftp.ncbi.nih.gov/genomes)

Рис. 1. Общий алгоритм генерации объединенного словаря подпоследовательностей из N-нуклеотидов, обнаруживаемых в первой и второй цепях ДНК методом движущейся рамки.

Из файла[4] хромосомы Chr.2 *Arabidopsis thaliana*, имеющей линейную длину ~19,25 млн.п.н., легко генерируется текст второй комплементарной цепи. По описанному на рис. 1 алгоритму составим словарь всех 3÷8 нуклеотидных подпоследовательностей в текстах первой и второй цепей. Отсортированные наборы данных отобразим графически в координатах номер строки – число олигонуклеотидов в первой и второй цепи ДНК. При этом будут получены соответствия (см. рис. 2), свидетельствующие о высокой степени сходства частот встречаемости. В частности, из по-

[3] Утилиты и программы были написаны для иллюстрации настоящей работы.
[4] ftp://ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana/CHR_II/NC_003071.fna

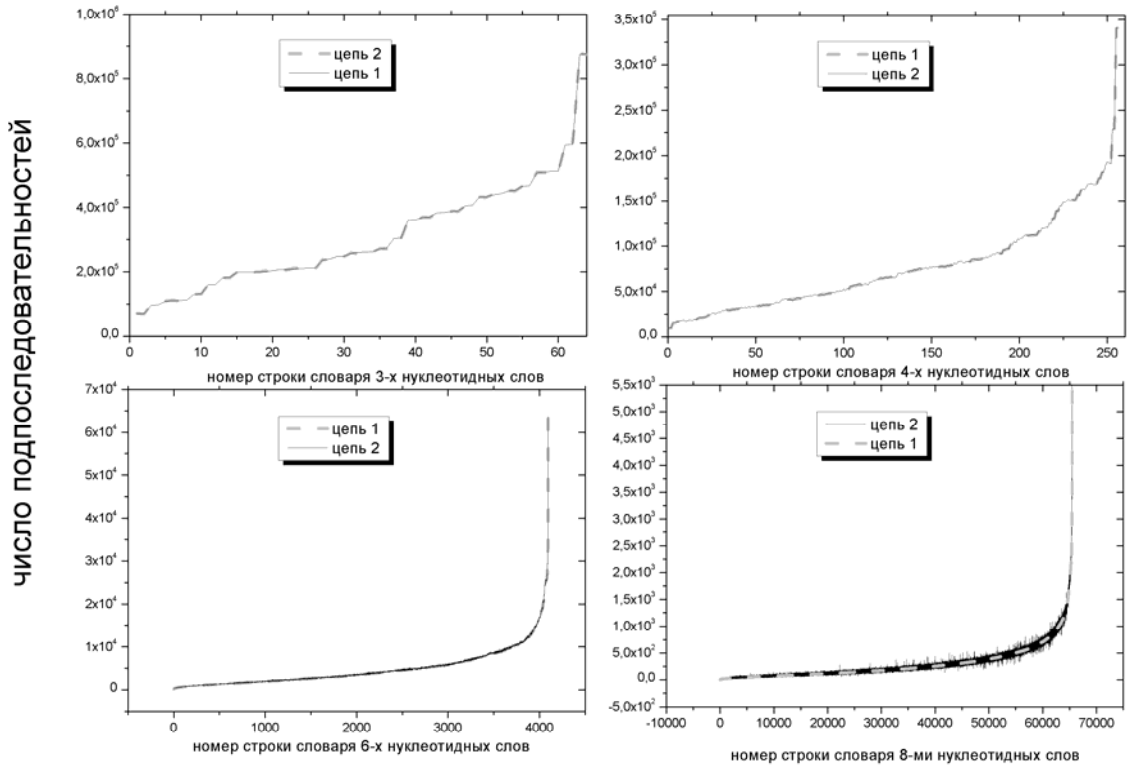

Рис. 2. Наложение частот встречаемости 3-, 4-, 6- и 8-нуклеотидных подпоследовательностей в словарях первой и второй цепей линейной хромосомы Chr.2 эукариота *Arabidopsis thaliana*. Сортировка словарей выполнена по возрастанию встречаемости олигонуклеотидов в словаре первой цепи.

лученных соответствий последует, что частоты встречаемости одних и тех же нуклеотидных комбинаций в текстах первой и второй ДНК цепей хромосомы Chr.2 *Arabidopsis thaliana* настолько близки между собой, что кривые практически налагаются друг на друга. Абсолютное расхождение между значениями числа копий одних и тех же «слов» не превышает 1% для триплетов одного типа или 2÷2,5% для одинаковых шестинуклеотидных комбинаций. При этом число повторений «редко» встречающегося триплета «GCG» в обеих цепях ДНК линейной хромосомы Chr.2 *Arabidopsis thaliana*, близко к 70 тыс., а наиболее часто встречающегося «AAA» лежит в интервале 872÷875 тыс.

Хромосомы *Arabidopsis* представляют собой весьма протяженные тексты, имеющие размеры в десятки млн.п.н. Чтобы убедиться в том, что эффект абсолютной близости частот встречаемости идентичных олигонуклеотидов в обеих цепях не является результатом присутствия каких-либо избыточных мотивов, весьма характерных для эукариот, выполним аналогичные построения с более компактными текстами прокариотических геномов, почти лишенных нефункциональных участков. Используем для этих целей весьма «лаконичные» последовательности генома археи *Nanoarchaeum equitans* (~450 тыс.п.н., NC_005213) и текст хромосомы бактерии *Buchnera aphidicola* str. APS (~640 тыс.п.н., NC_002528). По аналогии с *Arabidopsis thaliana,* составим словари шестинукле-

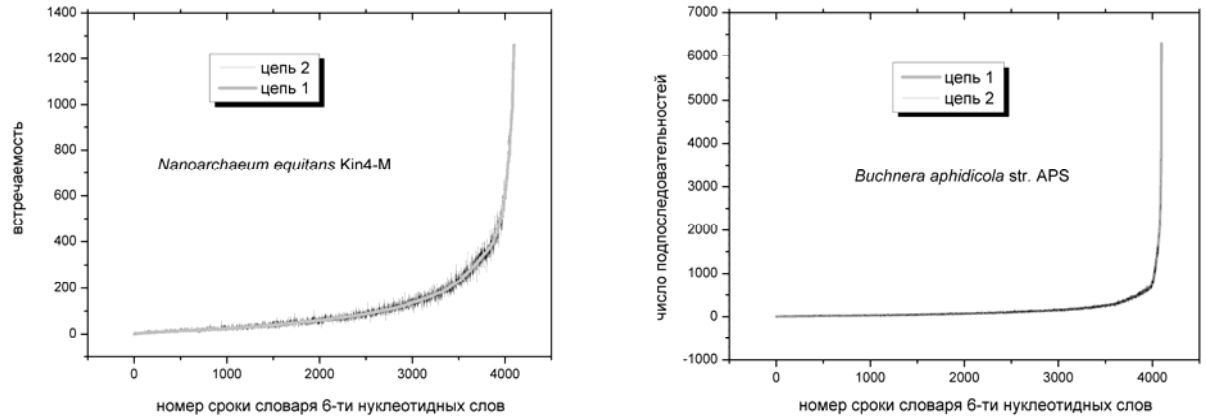

Рис. 3. Наложение частот встречаемости 6-нуклеотидных подпоследовательностей в словарях первой и второй цепях хромосомы прокариот *Nanoarchaeum equitans* и *Buchnera aphidicola* str. APS относящихся к царствам архей и бактерий.

отидных подпоследовательностей 1 и 2 цепей ДНК указанных организмов и отобразим графически массивы натуральных чисел, отсортированных по возрастанию значений одной из цепей (см. рис. 3).

При сравнении рисунков 2 и 3 нетрудно убедиться, что прокариотические геномы организованы сходным образом с эукариотическими. Высокое совпадение частот встречаемости одних и тех же нуклеотидных комбинаций в первой и второй цепях геномной ДНК бактерий и архей качественно проявляется столь же ярко, как и при лингвостатистическом анализе текстов двух цепей ДНК хромосомы 2 эукариота *Arabidopsis thaliana.* Следует подчеркнуть, что бактерии, археи и эукариоты иерархически связаны очень слабо и принадлежат к трем известным царствам живых организмов.

Используя те же эмпирические подходы, попытаемся ответить на вопрос: проявится ли обнаруженный нами эффект при работе с неорганизованными последовательностями ДНК? Для этого при помощи утилиты Randomizer[5] создадим неструктурированный линейный массив из 1 млн. нуклеотидов и, пользуясь прежней методикой, рассчитаем словари 4- и 6-нуклеотидных подпоследовательностей, для которых также построим кривые наложения. Как следует из рис. 4, в этом случае, в отличие от предыдущих примеров, корреляции в явном виде не обнаруживаются.

Проверим, является ли совпадение частотных составов подпоследовательностей 1 и 2 цепей ДНК общим свойством живых систем, будучи связанным, к примеру, с триплетной организацией последовательностей ДНК? В этом случае следует предполагать наличие кросс-соответствий между текстами, например, первых цепей ДНК геномов филогенетически удаленных организмов. Проведем сравнение ДНК двух разных бактерий с близкими размерами геномов. Для нашей цели хорошо подходят тексты ДНК *Aeropyrum pernix* K1 (NC000854.1, размер 1,67 млн.п.н.) и *Helicobacter pylori* 26695 (NC000915.1

[5] Утилита random.exe, основанная на использовании алгоритмов генерации псевдослучайных чисел

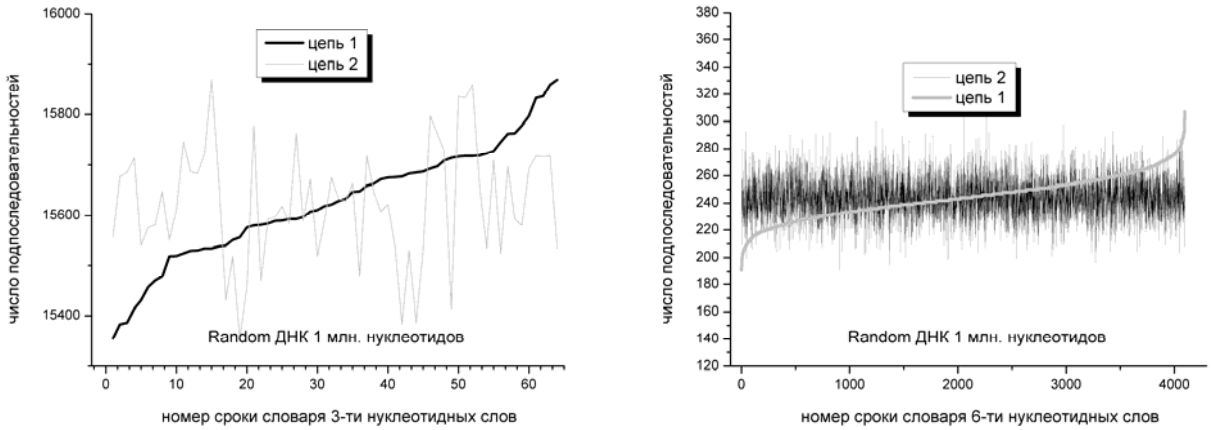

Рис. 4. Наложение частот встречаемости 3- и 6-нуклеотидных подпоследовательностей в отсортированных словарях первой и второй цепей псевдослучайной ДНК из 1 млн. нуклеотидных оснований

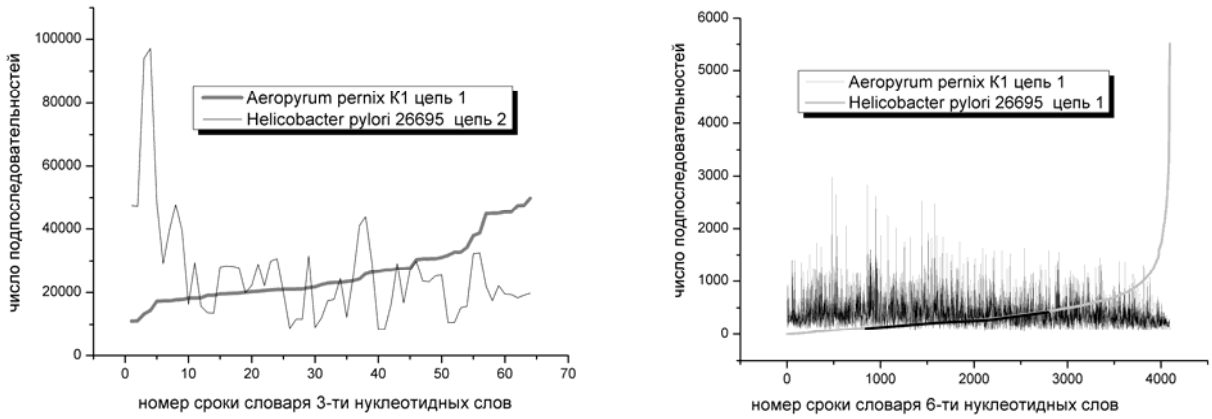

Рис. 5. Кросс-соответствия между однообразно сортированными 3- и 6-нуклеотидными словарями геномов археи *Aeropyrum pernix* K1 и бактерии *Helicobacter pylori* 26695.

размер 1,67 млн.п.н.) Первый микроорганизм принадлежит к царству *Archaea*, второй к царству *Bacteria*. Результаты совмещения рядов встречаемости 6-нуклеотидных подпоследовательностей в первых цепях ДНК этих двух микроорганизмов, представленные на рис. 5, позволяют сделать вывод о существенной разнице в составах их 3- и 6-нуклеотидных словарей.

Из проделанных вычислений можно сделать вывод о том, что геномные последовательности всех существующих царств живых систем: эукариот, архей и бактерий имеют весьма характерную структуру, отличающую природные тексты ДНК от любых искусственных. Данное свойство выражается в наличии олигонуклеотидной симметрии между текстами первой и второй цепей ДНК, легко выявляемой эмпирическими методами. Для простоты дальнейшего изложения обозначим обнаруженное общее свойство геномов живых систем термином «симметрия лингвистической структуры цепей ДНК», или «симметрия олигонуклеотидного состава первой и второй цепей геномной ДНК», или совсем кратко: «лингвистическая симметрия» геномной ДНК.

3. Количественная оценка олигонуклеотидной симметрии двух цепей геномных ДНК

Подсчет числа олигонуклеотидных подпоследовательностей позволяет оценить схожесть частотных словарей текстов первой и второй цепей лишь качественно. Чтобы провести количественное сравнение, следует воспользоваться какими-то другими известными методами, например, расчетом коэффициентов корреляции в массивах полученных значений.

Коэффициент корреляции Пирсона *r*, представляет собой безразмерный индекс, лежащий в интервале от –1,0 до 1,0 включительно, который может быть использован для отражения степени линейной зависимости между двумя множествами численных значений. Он вычисляется по формуле:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)\sigma_x\sigma_y}, \qquad (1)$$

где X_i и Y_i – значения двух переменных, $\overline{X}$ и $\overline{Y}$ – их средние значения, σ_x и σ_y – их стандартные отклонения, а n – количество пар значений.

Считается, что две переменные коррелируют между собой положительно, если между ними существует прямое, однонаправленное соотношение. При однонаправленном соотношении большие значения одной переменной соответствуют большим значениям другой переменной, а малые значения – малым.

Для словесного описания величин коэффициента линейной корреляции Пирсона мы будем использовать следующие соответствия, приведенные в таблице.

Значение коэффициента корреляции *r*	**Уровень корреляции**
$0 < r \le 0{,}2$	Очень слабая
$0{,}2 < r \le 0{,}5$	Слабая
$0{,}5 < r \le 0{,}7$	Средняя
$0{,}7 < r \le 0{,}9$	Сильная
$0{,}9 < r \le 1$	Очень сильная

3.1. Лингвистическая симметрия в хромосомах Arabidopsis thaliana *и* Drosophila melanogaster. *Внутренние и внешние кросс-корреляции*

С помощью утилиты abr.exe рассчитаем коэффициенты внутренней лингвистической корреляции (между первым и вторым текстами одной строки ДНК), а также кросс-корреляции (между первыми текстами двух разных хромосом), проведем некоторые измерения. Интервал длин нуклеотидных подпоследовательностей, в котором мы будем проводить вычисления, обозначим как диапазон корреляций.

Для построения графических зависимостей будем использовать соответствие вычисленных коэффициентов корреляции Пирсона, длинам олигонуклеотидных подпоследовательностей *N*, принятым во внимание и лежащим в диапазоне $N = 3 \div 12$.

С использованием перечисленных критериев схожесть словарей первой и второй цепи всех хромосом *Arabidopsis thaliana* (рис. 6а), будет определяться

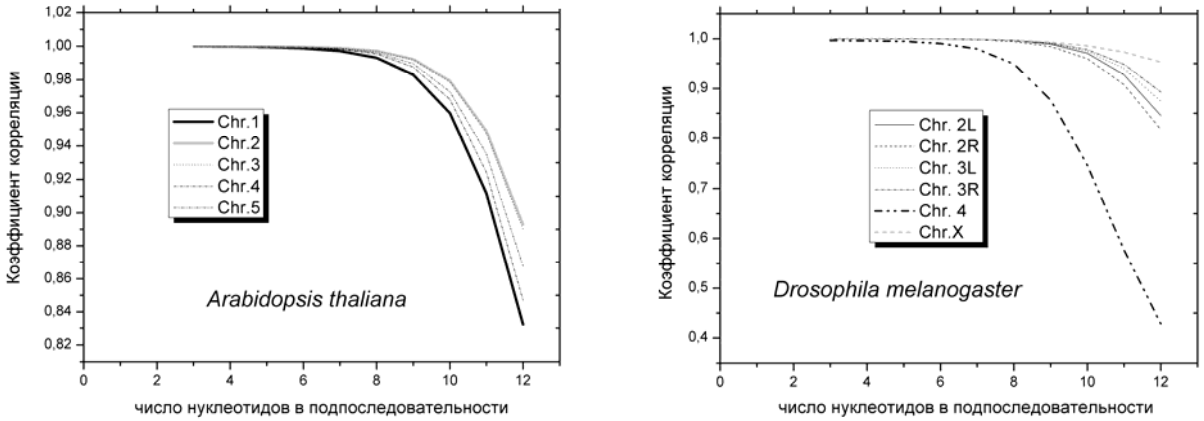

Рис. 6. Демонстрация схожести словарей первой и второй цепей всех хромосом *Arabidopsis thaliana* и *Drosophila melanogaster* 2L,2R,3L,3R,4,X – обозначения хромосом плодовой мушки.

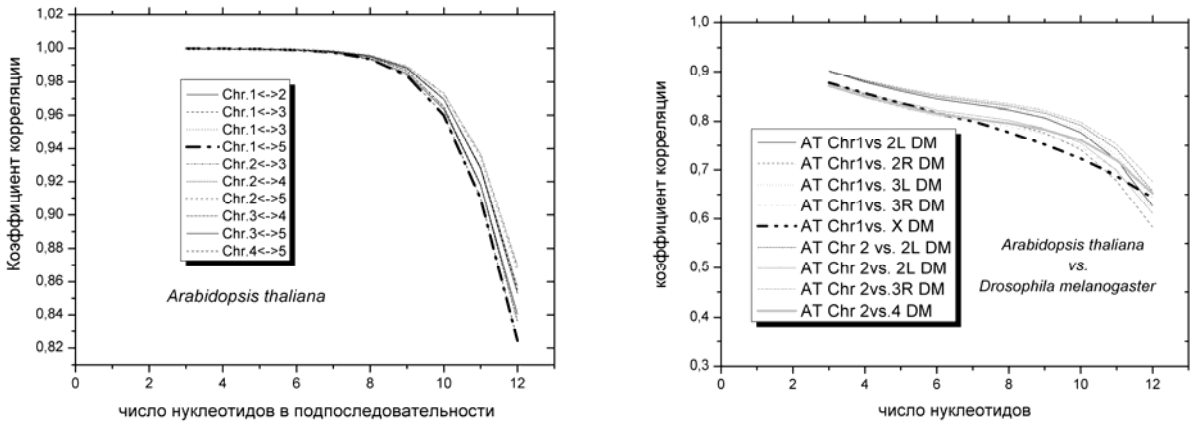

Рис. 7. Внутренние и внешние соответствия нуклеотидных словарей хромосом растения *Arabidopsis thaliana* и плодовой мушки *Drosophila melanogaster*.

сериями кривых (одна кривая на 1 хромосому), представленных на графике. Напомним, что соответствие точек экспериментальных кривых требованию $0{,}9 < r \leq 1$ соответствует критерию «очень сильная корреляция», а $0{,}7 < r \leq 0{,}9$ – сильная корреляция.

Для полноты картины измерим коэффициенты лингвистической корреляции для всех хромосом другого эукариотического организма - известной мухи дрозофилы, являющейся излюбленным объектом классической генетики.

Как видно из правого графика, у *Drosophila melanogaster*, лишь для 4-ой хромосомы лингвистическая симметрия выражена не столь отчетливо, но даже и в этом случае, в диапазоне $N = 8 \div 10$ нуклеотидов коэффициенты корреляций Пирсона подпадают под описание «сильная корреляция» $0{,}7 < r \leq 0{,}9$.

Напрашивается вопрос, обнаруживаются ли кросс-корреляции внутри одного организма между разными хромосомами или это внутреннее свойство индивидуальных молекул ДНК «написанных собственным языком»? Проведя подробный анализ (рис. 7), легко обнаружить эффект высокого лингвистиче-

ского сходство текстов ДНК разных хромосом одного объекта, в данном случае растения *Arabidopsis thaliana.*

Столь ли сильно растение отличается от насекомого? Количественный ответ будет получен в результате расчетов внешних кросс-корреляций между парами текстов хромосомных ДНК этих двух объектов. Результаты проведенного сравнения представлены на рис. 7. Полученные коэффициенты r_i, которые, по всей видимости, определяются степенью дивергенции *Arabidopsis thaliana* и *Drosophila melanogaster*, для словарей $N = 3 \div 10$ нуклеотидов находятся в интервале $0{,}7 < r \leq 0{,}9$, что соответствует критерию сильной корреляции.

3.2. Множественность явления: проверка на прокариотах, вирусах, фагах. Наблюдаемые экстремумы корреляций $\mathbf{r_i}$

Проведем серию измерений для большей выборки прокариотических геномов из 39 бактерий и архей. Сравнение полученных данных позволяет убедиться в достоверности и множественности явления. Можно с уверенностью говорить о том, что лингвистическая симметрия двух цепей ДНК наблюдается практически у всех микроскопических объектов, способных к самостоятельному размножению (рис. 8).

Сравнивая абсолютные значения коэффициентов лингвистической симметрии, полученные для организмов с различным типом клеточной организации, следует отметить несколько большую «сбалансированность» олигонуклеотидного состава текстов хромосом эукариот по сравнению с прокариотами.

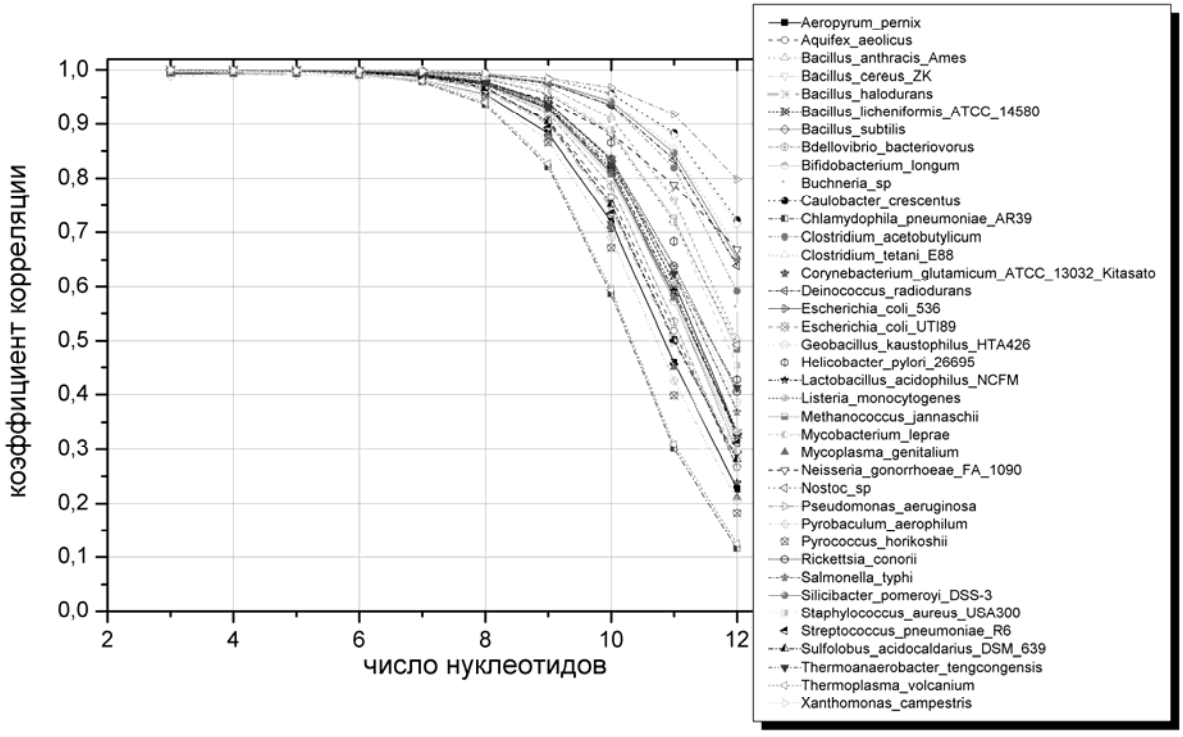

Рис. 8. Сравнение корреляций в текстах первой и второй цепей геномных ДНК некоторых бактерий и архей.

Таблица. 3. Абсолютные значения r_i ДНК некоторых организмов

Организм	r_3	r_4	r_6	r_8	r_{10}	r_{11}	r_{12}
Arabidopsis thaliana Chr.1	0,9999	0,9998	0,9995	0,9988	0,9971	0,9931	0,9829
Drosophila melanogaster Chr.2R	0,9998	0,9997	0,9996	0,9991	0,9978	0,9941	0,9842
Drosophila melanogaster Chr.X	0,9998	0,9998	0,9996	0,9993	0,9985	0,9967	0,9929
*Anaeromyxobacte*r (NC_007760.1)	0,9998	0,9998	0,9996	0,9982	0,9880	0,9684	0,9179
Pseudomonas fluorescens (NC_004129)	0,9993	0,9991	0,9983	0,9938	0,9597	0,8941	0,7522
Buchneria aphidicola (NC_002528.1)	0,9988	0,9977	0,9943	0,9810	0,9007	0,7734	0,5602
Synechococcus sp. (NC_007775.1)	0,9998	0,9996	0,9978	0,9847	0,8840	0,7317	0,522
Idiomarina loihiensis (NC_006512.1)	0,9588	0,9643	0,9632	0,9168	0,6153	0,3477	0,1618

Представление об этом можно получить, в частности, на основании данных, приведенных в табл. 3. Вместе с тем, лингвистическая симметрия цепей геномной ДНК, некоторых прокариотических организмов таких, как *Anaeromyxobacter dehalogenans* 2CP-C (NC_007760), также может быть оценена как «очень высокая» даже на уровне 12 нуклеотидных подпоследовательностей. В целом же, для геномов безъядерных организмов характерна организация с «очень высоким» уровнем комплементарной симметрии лишь до 9-нуклеотидных сочетаний.

В числе исследованных нами микроорганизмов, геномы которых обладают наименьшей симметрий, следует назвать бактерию *Idiomarina loihiensis* L2TR (5.64 млн.п.н., NC_006512.1), значения коэффициентов корреляции которой также приведены в табл. 3. Необычность текста геномной ДНК этой глубоководной морской бактерии, выделенной из термального источника, заключается еще и в том, что максимум корреляций приходится на 5-нуклеотидные сочетания, которые превышают таковые для 4- и даже 3-нуклеотидных сочетаний ($r_3 < r_4 < r_5 > r_6 > r_7 > r_N$).

По аналогии с эукариотами, несложно подтвердить наличие межхромосомного лингвистического сходства в прокариотических объектах, несущих одновременно несколько хромосом, таких, например, как *Burkholderia* sp., *Brucella suis*, *Leptospira interrogans*, и мн. др. Аналогичный результат будет получен при сравнении олигонуклеотидного состава многих плазмид большого размера с составом геномной ДНК организма носителя.

Говоря об универсальности лингвистической симметрии обеих цепей ДНК, являющейся, как мы показали, общим свойством клеточных организмов, нельзя оставить в стороне последний вопрос: распространяется ли это явление на вирусы и фаги[6], относимые с некоторыми оговорками к живым системам?

В табл. 4 приведены данные, свидетельствующие как в пользу наличия эффекта симметрии в ДНК обсуждаемых объектов, так и против него. Абсо-

[6] Те же вирусы, объектом атаки которых являются бактерии

Таблица. 4. Коэффициенты лингвистической корреляции r_i некоторых фагов, вирусов и плазмид

Организм	r_3	r_4	r_5	r_6	r_7
Enterobacteria phage P22 (NC 002371)	0,83531	0,8184	0,76642	0,62622	0,40318
Vaccinia virus (NC 006998)	0,99932	0,99824	0,99486	0,98333	0,95157
Goatpox virus (NC 004003)	0,99599	0,99508	0,99264	0,98719	0,97236
Chlorella virus 1 (U42580)	0,99825	0,99514	0,98769	0,96459	0,90001
Deinococcus plasmid MP1 (NC 000958.1)	0,99085	0,98354	0,97446	0,9532	0,90401
Alteromonas phage PM2 (NC 000867)	0,61726	0,59342	0,50331	0,38713	0,2306

лютные значения коэффициентов корреляции нуклеотидных текстов некоторых вирусов, фагов и плазмид (внехромосомных самовоспроизводящихся генетических элементов бактерий и архей), близки к таковым, обнаруживаемым, например, у прокариот. Вместе с тем, учитывая биологическую множественность вирусных форм, проявляющуюся в существовании одноцепочечных и двухцепочечных ДНК- и РНК-вирусов, мы воздержимся от каких-либо обобщений, утверждающих универсальность обнаруженного эффекта в отношении всех бесклеточных организмов.

4. К ВОПРОСУ О СПЕЦИФИКЕ СТРУКТУРЫ ТЕКСТОВ ДНК С ВЫСОКИМИ КОЭФФИЦИЕНТАМИ ЛИНГВИСТИЧЕСКОЙ КОРРЕЛЯЦИИ

Строение двухцепочечной молекулы ДНК определяет тот факт, что если в первой цепи имеется некая уникальная строка, состоящая из N нуклеотидов, то в записи второй цепи обязательно будет присутствовать ее комплементарная копия. Предположим, что на обеих цепях присутствует две идентичные последовательности, расположенные рядом без промежутков, то есть, крест-накрест. Соответственно, каждая из копий будет иметь комплементарное отражение в противоположной цепи, а в полученной крестообразной структуре будет присутствовать центр точечной симметрии. В итоге, в обеих цепях ДНК будут одновременно находиться четыре варианта одной и той же последовательности, состоящей из половин этой записи. Описываемая структура, изображенная на рисунке, называется комплементарным палиндромом [5] и относится к одной из разновидностей повторяющихся последовательностей геномной ДНК:

```
(5')TTAACCGGGGGGAATGGGCCCATTCCCCCCGGTTAA(3')   первая цепь
(3')AATTGGCCCCCCTTACCCGGGTAAGGGGGGCCAATT(5')   вторая цепь
```

Строка 1: Комплементарный палиндром

Как известно, обычный палиндром читается одинаково в обоих направлениях, имеет симметричную структуру:

```
(5')AATTGGCCCCCCTTACCCCCCATTCCCCCCGGTTAA(3')   первая цепь
(3')TTAACCGGGGGGAATGGGGGGTAAGGGGGGCCAATT(5')   вторая цепь
```

Строка 2: Обычный палиндром

Таблица. 5. Коэффициенты корреляции цепей модельных ДНК

Строка	Коэффициенты корреляции									
	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}
1	1	1	1	1	1	1	1	1	1	1
2	0,01357	0,00252	0,01639	0,00638	0,00183	$4{,}4\cdot10^{-4}$	$1{,}1\cdot10^{-4}$	$3\cdot10^{-5}$	$1\cdot10^{-5}$	0
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
3+4	0,97659	0,96174	0,94404	0,92637	0,90872	0,8922	0,87497	0,85713	0,83871	0,81967
3a+4a+ 3b+4b	0,96277	0,92424	0,84103	0,79383	0,72617	0,6766	0,62491	0,5714	0,51612	0,45901

Напомним, что термин «палиндром» используют не только в молекулярной биологии, но и в более широком смысле, применяя к любым символьным конструкциям. Комплементарные палиндромы специфичны только для ДНК, так как «симметричные» половины слова располагаются на обеих цепях молекулы. Подсчетом олигонуклеотидов в первой и второй цепях нашего фрагмента ДНК (строка 1), будут получены идентичные наборы подпоследовательностей заданной длины, а коэффициенты корреляции Пирсона окажутся равными единице (см. табл. 5).

Отметим, что вычисления r_i, выполненные с обычным палиндромом (первая и вторая цепи строки 2), дадут иной результат и в общем случае корреляции не будет обнаружено.

Картина представляется простой и ясной, если речь идет о присутствии в изучаемом фрагменте ДНК одного комплементарного палиндрома. В случае с протяженными ДНК-текстами геномов, наличие в них конструкции лишь из одного весьма протяженного комплементарного палиндрома, даже вырожденного (то есть со следами мутаций), не представляется вероятным. Можно лишь сделать общее предположение, объясняющее наблюдаемые эффекты лингвистической симметрии, о которых говорилось выше, результатом слияния множества сравнительно коротких реверсных палиндромов в одной молекуле ДНК:

TTAACCGGGGGGAATGGGCCCATTCCCCCCGGTTAAGATCCTAGGCTGGACCGTACGGTCCAGCCTAGGATC 1
AATTGGCCCCCCTTACCCGGGTAAGGGGGGCCAATTCTAGGATCCGACCTGGCATGCCAGGTCGGATCCTAG 2

Строка 3 Строка 4

Легко убедиться, что после смешения четырех половин двух палиндромов корреляции не исчезнут полностью, даже если будет разрушена видимая крестообразная симметрия этих структур, и их комплементарные части окажутся расположенными на удалении друг от друга:

TTAACCGGGGGGAATGGGGATCCTAGGCTGGACCGTCCCATTCCCCCCGGTTAAACGGTCCAGCCTAGGATC 1
AATTGGCCCCCCTTACCCCTAGGATCCGACCTGGCAGGGTAAGGGGGGCCAATTTGCCAGGTCGGATCCTAG 2

Строка 3a Строка 4a Строка 3b Строка 4b

Совокупность пар фрагментов цепей ДНК, представленных на верхнем рисунке, может быть отнесена к разделенным комплементарным палиндромам, которые также могут быть рассмотрены в качестве совершенных (полностью идентичных) разделенных (не стоящих друг за другом) повторяющихся после-

довательностей [5]. Эта структура продолжает сохранять достаточно высокий уровень лингвистической симметрии (см. табл. 5).

С ДНК фрагментом, несущим несколько разделенных комплементарных палиндромов, можно проделать другие манипуляции, например, ввести между фрагментами хаотический текст, но симметрия при этом останется и будет выявлена методом подсчета коэффициентов Пирсона.

На основании выполненных построений и сравнений может быть сделан вывод о том, что в основе наблюдаемой лингвистической симметрии первой и второй цепей геномной ДНК безъядерных и эукариотических организмов лежит присутствие в структуре двух текстов множества смежных и разделенных комплементарных палиндромов. Последние могут быть также рассмотрены в виде множества несвязанных, повторяющихся, разделенных последовательностей.

На примере генома любого конкретного организма можно показать, что в результате искусственного «укорачивания» текстов корреляции в них быстро разрушаются. В качестве такового на рис. 9 показан процесс «распада симметрии» текста генома *Anaeromyxobacter dehalogenans* при вычислении коэффициентов корреляций r_i для фрагментов ДНК этого организма, имеющих различную длину.

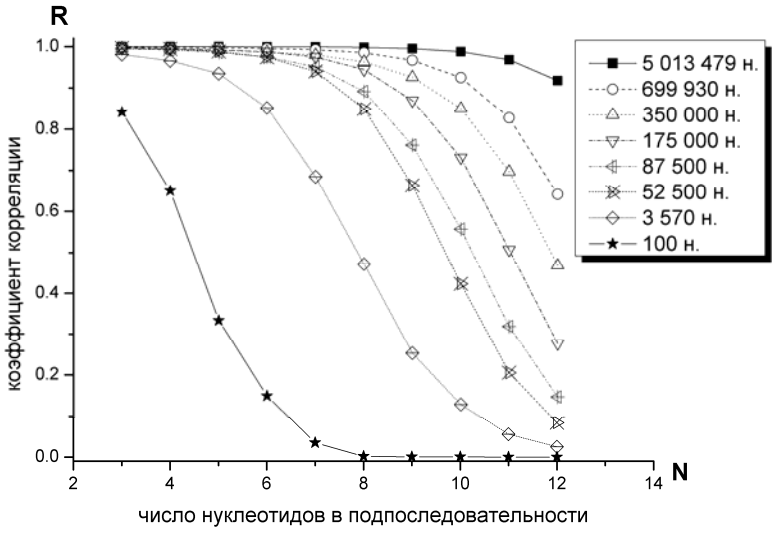

Рис. 9. Разрушение лингвистических корреляций в процессе укорачивания текста кольцевого генома *Anaeromyxobacter dehalogenans* NC_007760 имеющего длину 5,01 млн.п.н.

Существенное снижение корреляций при подсчете 8 нуклеотидных подпоследовательностей наблюдается уже при работе с фрагментом 50 тыс. нуклеотидов, и становятся ощутимым на уровне триплетной организации текста ДНК в случайно отобранном фрагменте из 100 нуклеотидов. Уместно отметить, что из фрагмента текста 50 тыс. нуклеотидов генерируется 99 986 8-нуклеотидных строк, а 100-буквенный текст ДНК содержит в себе 196 триплетов (на двух цепях). Таким образом, уменьшение коэффициентов Пирсона в процессе укорачивания текста достоверно свидетельствует в пользу доминирования в геноме разделенных реверсных палиндромов или просто разнесенных повторяющихся последовательностей, имеющих 100% гомологию (две строки считаются гомологичными, если при наложении, они могут быть выровнены до состояния совпадения части идентичных букв и фраз, имеющихся в каждой из строк).

Если бы комплементарные палиндромы шли друг за другом, укорачивание двойной цепи, приводило бы к существенному изменению коэффициентов кор-

реляции. Когда же куски палиндромов (мотивы) равномерно распределены в виде смеси по всему геному, то искусственное укорачивание текста приводит к исчезновению недостающих пар, соответственно, корреляция быстро снижается.

Следует отметить, что все сказанное выше отнюдь не исключает возможности присутствия в геномных ДНК **смежных** реверсных палиндромов, долю которых еще следует уточнить.

Естественным представляется вопрос, какова доля текста суммы всех реверсных палиндромов во всей последовательности генома? Предположим, что в процессе подсчета *N*-нуклеотидных слов в первой и второй цепях ДНК, каждому сочетанию нуклеотидов *i* соответствуют значения m_i и n_i. Учитывая, что реверсные палиндромы складываются из пар слов, их число для данного сочетания нуклеотидов всегда будет соответствовать меньшему значению из пары.

Иными словами, если *i* строка AAAGCG встречается в первой цепи $m = 123$ раза, а во второй цепи $n = 114$, то количество реверсных палиндромов AAAGCG во всем геноме также будет равно 114.

Сделав выбор меньшего числа из пары значений, равных числу слов в первой и во второй цепях, выполнив эту операцию во всех найденных строках, и рассчитав произведение суммы выбранных меньших величин на число нуклеотидов в рамке *R*, можно узнать приблизительную сумму (количество) всех нуклеотидов D_s, из которых складываются реверсные палиндромы, обнаруживаемые во всем геноме:

$$D_S = \sum\nolimits_i D_i \cdot R = R \cdot \sum\nolimits_i \left(m_i + n_i - \left|m_i - n_i\right|\right)/2 . \qquad (2)$$

Остается определить сумму нуклеотидов во всех разновидностях строк, считанных в момент движения рамки. Очевидно, что величина суммы значительно больше длины генома. Для ширины рамки *R* при длине генома *L* число шагов движения рамки составляет $N = (L–R+1)$, следовательно, суммарное число нуклеотидов составит значение $M = N \cdot R = (L–R+1) \cdot R$. Таким образом, доля комплементарных палиндромов *P* в сумме длины нуклеотидов из всех подпоследовательностей, образующихся в процессе движения рамки, будет равна отношению

$$P = D/\left(L - R + 1\right) \cdot R , \qquad (3)$$

где $D = 0{,}5 \cdot (m+n-|m-n|)$ – число нуклеотидов в реверсных палиндромах, *L* – длина генома, *R* – ширина рамки.

Вычисления, сделанные в качестве примера для ДНК бактерии *Anaeromyxobacter dehalogenans* длиной 5,01 млн.п.н., показывают, что доля комплементарных разделенных палиндромов, половины которых состоят из 6 букв, занимают не менее 98% общей длины. Для реверсных палиндромов, половины которых состоят из семи и восьми нуклеотидов, эти величины составят соответственно 97% и 95%. Величины такого же порядка обнаруживаются со всеми другими испытанными геномами, в частности, с эукариотическими.

5. Изменение алгоритма написания комплементарной цепи ДНК

Во всех предыдущих построениях и измерениях мы пользовались естественным алгоритмом генерации комплементарной цепи ДНК. Не отказываясь от принципов антипараллельности расположения двух цепей ДНК, изменив порядок комплементарности, представим межцепочечные соответствия нуклеотидов, которые не существуют в природе. При этом не предполагается, что соответствие является обратным по отношению к самому себе, как это имеет место в реальности. Однако сохраняется требование сопоставления разным исходным нуклеотидам разных результирующих.

В комбинации из 4 нуклеотидов возможны 4! = 24 варианта таких соответствий, лишь один из которых является правильным: A → T; G → C; T → A; C → G; остальные – неверными. Назовем их *нонсенс-соответствиями*, исходный текст генома – *матричной ДНК*, генерацию текста второй цепи из матричной по природному алгоритму *процессом трансляции*, по 23 искусственным алгоритмам – *нонсенс-трансляциями*. Самым простым для понимания вариантом нонсенс-соответствия является идентичная подстановка: A → A; C → C; G → G; T → T, так как она приводит к образованию текста второй цепи в виде обратно прочтенной строки первой цепи.

Используя в качестве матриц последовательности хромосом различных прокариотических организмов просчитаем коэффициенты лингвистической корреляции между цепями всех «искусственных разновидностей» ДНК.

На первый взгляд кажется очевидным, что изучение ДНК с нонсенс-структурой не имеет смысла и априори не должно выявить никаких лингвистических закономерностей. И в этом, отчасти, можно удостовериться, исследовав кольцевой геном бактерии *Shigella boydii* Sb227, (NC_007613.1) (длина текста 4,51 млн.п.н.) – см. рис. 10. Лишь один текст второй цепи из 24, написанный «правильным» языком (верхняя кривая), будет иметь высокие уровни корреляции с матрицей. Значимых межцепочечных соответствий в «искусственных»

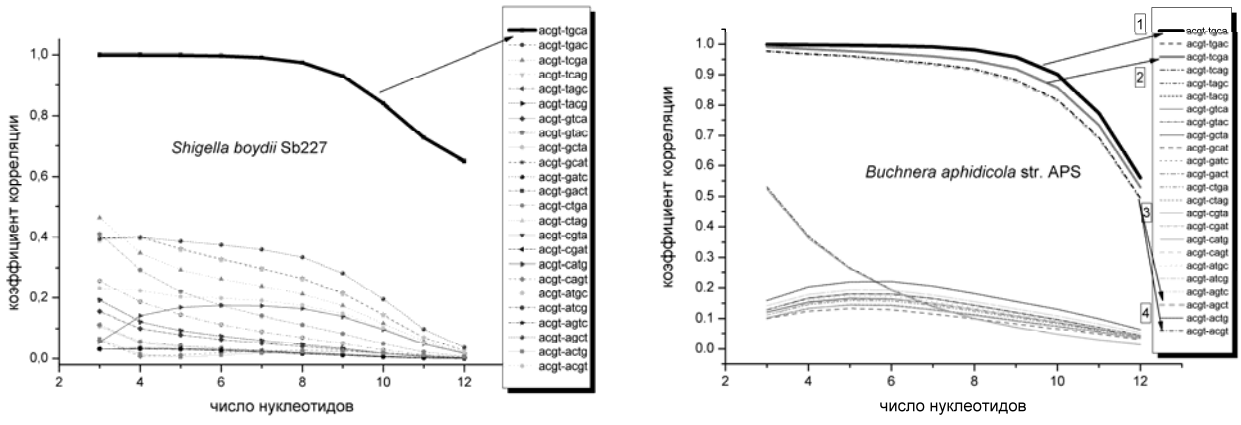

Рис. 10. Испытание нонсенс-алгоритмов написания второй цепи геномов бактерий *Shigella boydii* Sb227 (NC_007613.1) и *Buchnera aphidicola* str. APS, (NC_002528.1).

текстах нонсенс-ДНК, сгенерированных *in silico* на основе геномной матрицы *Shigella boydii*, обнаружено не будет.

Тем более неожиданными покажутся результаты аналогичных манипуляций с достаточно короткой последовательностью бактерии *Buchnera aphidicola* str. APS, (640 тыс. п.н.) NC_002528.1. Как следует из рис. 10, в данном случае обнаруживаются четыре типа высоких корреляций текста матричной ДНК с последовательностями, сгенерированными, в том числе, нонсенс-методами. Наиболее высокие коэффициенты r_i будут обнаруживаться при нормальном варианте трансляции, и несколько меньшие – для трех других результатов нонсенс-трансляции.

Таблица. 6. Значимость алгоритмов «трансляции», приводящих к возникновению значимых лингвистических корреляций в геноме *Buchnera aphidicola* str. APS, (NC_002528.1)

Номер алгоритма														
1			**>**	**2**			**>**	**3**			**>**	**4**		
A	→	T		A	→	T		A	→	A		A	→	A
C	→	G		C	→	C		C	→	C		C	→	C
G	→	C		G	→	G		G	→	G		G	→	G
T	→	A		T	→	A		T	→	T		T	→	T

Объединим в табл. 6 полученные с последовательностью *Buchnera aphidicola* результаты значимости корреляций и типы подстановок нуклеотидов при трансляции, обозначив коррелирующие типы перестановок номерами 1, 2, 3 и 4 в порядке убывания значений коэффициентов Пирсона.

Как уже говорилось, наиболее «правильным» и коррелирующим является первый алгоритм, в котором текст комплементарной цепи был сгенерирован классическим способом.

(5')TTAACCGGGGGGAATGGGCCCATTCCCCCCGGTTAA(3') первая цепь
(3')AATTGGCCCCCCTTACCCGGGTAAGGGGGGCCAATT(5') вторая цепь

Комплементарный палиндром, сгенерированный по методу 1

За ним по значимости следует алгоритм №2, являющийся модификацией алгоритма №1, с той разницей, что в комплементарных блоках, из которых складываются половины палиндрома, произведены замены цитозина (C) на гуанин (G)-при высокой консервативности положения нуклеотидов A и T. Образец нонсенс-текста ДНК, записанный этим методом, также будет обладать крестообразной симметрией, на обнаружение которых, собственно, и направлен описываемый метод исследования:

(5')TTAACCGGGGGGAATGGGCCCATTCCCCCCGGTTAA(3') первая цепь
(3')AATTCCGGGGGGTTAGGGCCCTAACCCCCCGGAATT(5') вторая цепь

Комплементарный палиндром, сгенерированный по методу 2

Таким образом, если в конкретном тексте первой цепи генома ДНК *Buchnera aphidicola* обнаруживается последовательность (5')ACGT(3'), то наряду с присутствием в другой цепи полной копии данного тетрануклеотида также высока вероятность встречи его вырожденного аналога: (5')AGCT(3').

Рассмотрим вопрос, какие комплементарные палиндромы будут возникать при генерации второго текста по алгоритму №3?

(5')TTAACCGGGGGGAATGGGGGGTAAGGGGGGCCAATT(3') первая цепь
(3')TTAACCGGGGGGAATGGGGGGTAAGGGGGGCCAATT(5') вторая цепь

Комплементарный палиндром, сгенерированный по методу 3

Из представленного рисунка видно, что первая цепь этого образца «ДНК» имеет центральную симметрию и одинаково читается слева направо и справа налево, то есть, является классическим (не реверсным) палиндромом. Таким образом, в отношении текста генома *Buchnera aphidicola* можно будет сделать вывод о присутствии в нем большого количества как смежных, так и разделенных классических палиндромов.

Соответствие №4 является разновидностью соответствия №3 с той лишь разницей, что в симметричных блоках, из которых складываются палиндромы, произведена замена цитозина (C) на гуанин (G) (или наоборот). Этот же тип вырождения мы обсуждали, рассматривая корреляции, полученные при использовании второго алгоритма нонсенс-трансляции из матрицы *Buchnera aphidicola*. Наличие данного типа нонсенс-соответствий также может быть следствием относительно редкой встречаемости в ДНК этого типа нуклеотидов A и T.

Наличие обычной симметрии и практически полное отсутствие нонсенс-корреляций является типичным не только для *Shigella boydii* Sb227, но и для некоторых других бактерий, например, для *Idiomarina loihiensis* L2TR (NC_006512.1), *Aeropyrum pernix* K1 (NC_000854.1) и др.

С другой стороны, нонсенс-корреляции, характерные для *Buchnera aphidicola* str. ASP, также не являются исключением и, в той или иной степени, могут быть обнаружены у многих микроорганизмов, вирусов, растений. К «носителям» нонсенс-корреляций по типу 2-4 можно отнести бактерию *Myxococcus xanthus* DK 1622 (NC_008095.1), вирус оспы *Vaccinia virus* (NC_006998),

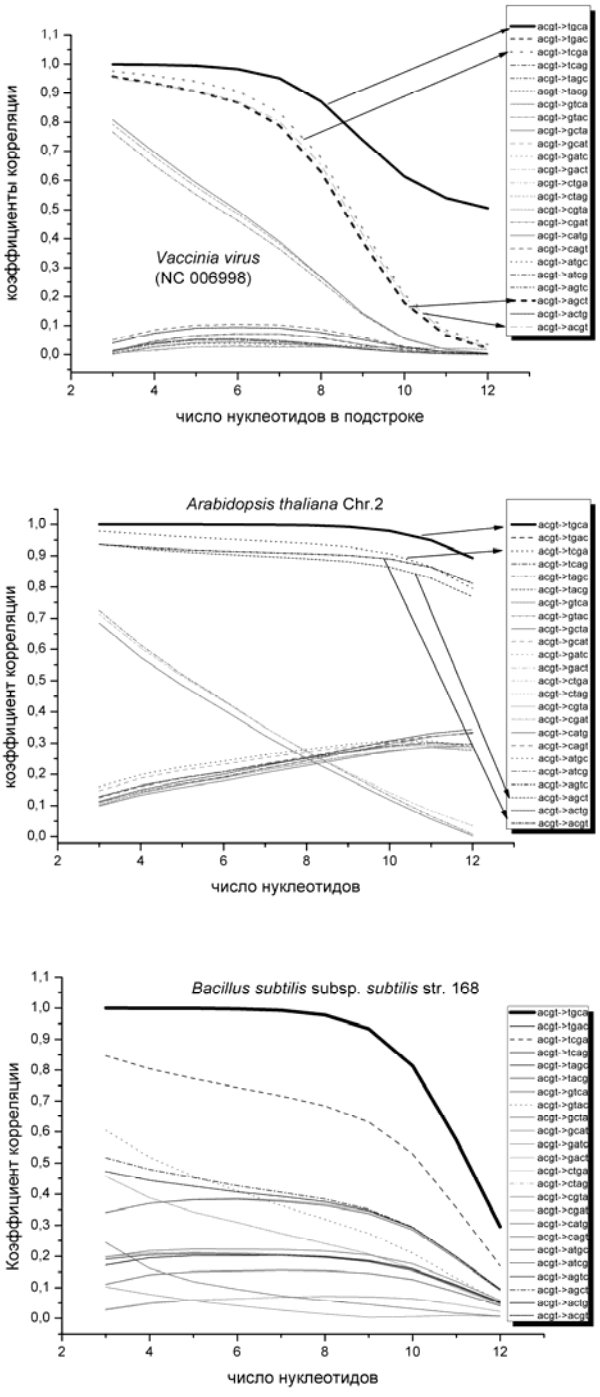

Рис. 11. Нонсенс корреляции в ДНК *Vaccinia virus, Arabidopsis thaliana и Bacillus subtilis*.

хромосомы *Arabidopsis thaliana*. Значимость выявляемых нонсенс-соответствий может располагаться в порядке, ином, чем в случае с *Buchnera aphidicola* str. ASP, но наилучшими соответствиями с матрицей всегда обладают тексты второй цепи ДНК, сгенерированные по природным правилам. Выявление факта наличия и отсутствия нонсенс-корреляций в геномах разных организмов может служить свидетельством доминирования или, наоборот, отсутствия каких-то типов молекулярных событий в процессе эволюции данного организма и его геномной ДНК. О разнообразии таких проявлений, в частности, можно судить, опираясь на зависимости, представленные на рис. 11. Так, в высококонсервативном геноме *Buchnera aphidicola* отчетливо проявляются 3 типа нонсенс корреляций, что также характерно для второй хромосомы растения *Arabidopsis thaliana*. В почвенной бактерии *Bacillus subtilis* присутствует всего один уровень нонсенс соответствия, которое можно назвать значимым, а в патогенном микроорганизме *Shigella boydii* нонсенс корреляции вообще не обнаруживаются. Целый спектр нонсенс корреляций (3 сильных и 3 средних) обнаруживается в геноме вируса оспы *Vaccinia virus*, состоящем всего из 194711 пар нуклеотидов (NC_006998).

6. ВИРТУАЛЬНАЯ «ЭВОЛЮЦИЯ» ПСЕВДОСЛУЧАЙНОГО ДНК ТЕКСТА

Отчетливая лингвистическая симметрия первой и второй цепей геномных ДНК, по всей видимости, является общим свойством организмов любой клеточной организации – как эукариот, так и прокариот. Однако нет никаких оснований полагать, что подобный характер хранения генетических данных функционален и имеет какой-либо молекулярный или биохимический смысл. Наоборот, авторы настоящего исследования склонны считать ее лишь случайным следствием известных молекулярно-биологических процессов[7], происходящих как во время репликации, так и самостоятельно. К числу последних, как известно [14], относятся:

1. Мутации, или точечные замены, включающие удаление или вставку отдельных нуклеотидов.
2. Перемещение и инвазия генетического материала между геномами разных организмов, даже филогенетически удаленных (горизонтальный перенос генов), а также последствия от воздействия фагов.
3. Процессы внутригеномной перестройки ДНК крупными блоками и, в частности:

a) дупликации и мультипликации, приводящие к образованию единичных и многочисленных повторов одного и того же текста;

b) инверсии, следствием которых является возникновение во второй цепи копий фрагмента, функциональный оригинал которых находится в первой цепи;

c) процессы транслокации, следствием которых является перестановка фрагментов текста цепей ДНК из одной координаты цепи в другую.

---

[7] В данном случае имеются основания брать для рассмотрения лишь молекулярно генетические процессы характерные для прокариот, так как лингвистическая симметрия в равной степени проявляется при любой клеточной организации – эукариотической и прокариотической.

4. Процессы крупноблочной делеции ДНК, приводящих к безвозвратной потере каких-то участков генома.

По всей видимости, нет необходимости в деталях обсуждать гипотетическую возможность возникновения в геномной ДНК множественных палиндромов, обычных и комплементарных, в результате точечных мутаций. Мутации, конечно же, являются мощным фактором эволюции первичной структуры ДНК на уровне кода, способствуя через естественный отбор улучшению химической функциональности тех элементов[8], в которые он транслируются. Вместе с тем, предположить существование какого-либо алгоритма мутационного процесса, который бы вел к росту и поддержанию на высоком уровне симметрии цепей ДНК, одновременно улучшая химическую функциональность кода, трудно. Скорее наоборот, представляется резонным предположить, что мутации способны быстро разрушать симметричные структуры в целях улучшения функциональности биохимических компонентов клетки.

Следующий предполагаемый фактор – инвазии чужеродной ДНК, известные в среде прокариот как следствие горизонтального переноса генов, всегда случайны и относительно малозначимы. Очень высока вероятность того, что фрагменты чужеродной ДНК, встраиваемые в геном реципиента, будут по отношению к нему структурно далеки, нежели близки. Таким образом, результирующие изменения не будут систематическими[9], будут либо малозначимыми, либо отрицательными для симметрии. На наш взгляд, их также нельзя всерьез рассматривать в качестве возможной причины существования лингвистической симметрии цепей ДНК.

В итоге для обсуждения и конструирования возможных алгоритмов возникновения рассматриваемых структур ДНК в нашем распоряжении остаются процессы внутригеномной перестройки информации генома крупными блоками, а также безвозвратной потери его участков. Следует отметить, что следы крупноблочной перестройки ДНК обнаруживаются в геномах любых объектов в виде многочисленных повторяющихся фрагментов текста, в том числе инвертированных[10] и имеющих абсолютную гомологию. Их абсолютное количество зависит от чувствительности поиска [18] и, что вполне очевидно, с уменьшением минимальной рамки детектирования возрастает. При рассмотрении очень коротких повторов (10÷15 нуклеотидов) их суммарная протяженность может приблизиться к длине всего генома. Уже сейчас можно говорить о том, что процессы возникновения повторов происходят непрерывно и в масштабах, сравнимых с другими перестройками. Наряду с непрерывным удвоением и мультипликацией информации должна осуществляться непрерывная компактизация генома [14], в про-

[8] Таких элементов, как РНК, не требующих организации на уровне триплетного кода, равно как и белков, транслируемых посредством мРНК и зависимых от триплетного кода.

[9] Не следует забывать и то, что горизонтальный перенос генов, явление свойственное для царства прокариот, а лингвистическая симметрия отчетливо обнаруживается и в текстах эукариотических хромосом.

[10] Термин «инвертированный повтор» подразумевает присутствие в тексте первой цепи ДНК дубликата участка из той же цепи, в том виде, в котором бы он присутствовал во второй. Наличие инвертированного повтора в первой цепи равнозначно наличию прямого повтора во второй цепи.

тивном случае наблюдался бы неограниченный рост длины молекулы геномной ДНК, который практически не имеет места среди прокариот[11].

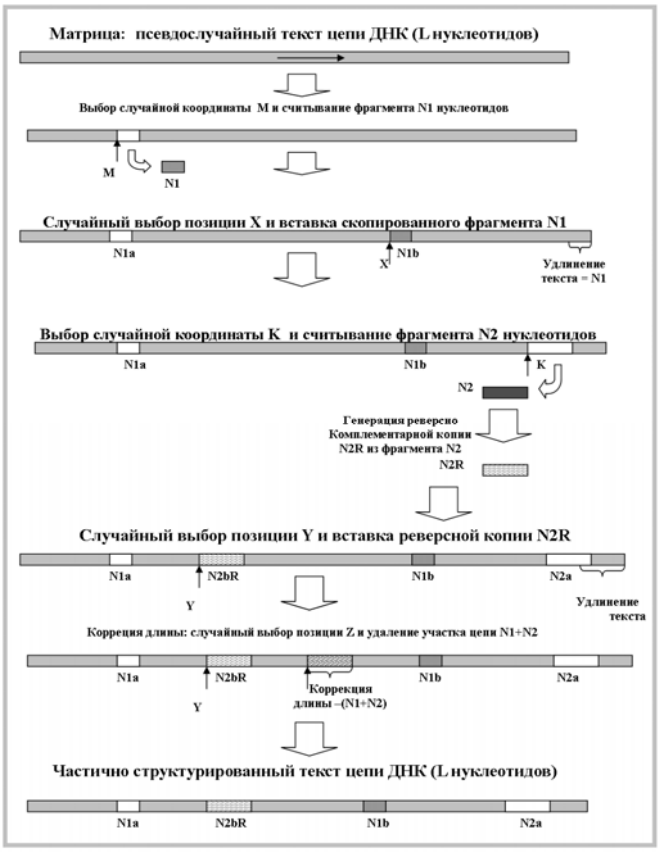

Рис. 12. Модель трансформации текста геномной ДНК, имитирующая образование прямых и инвертированных повторов при условии сохранения постоянства его длины (алгоритм работы программы «Автоморфер»).

Не будем останавливаться на молекулярной сущности этих явлений, пытаясь определить, являются ли они ошибками репликации, или результатами каких-то иных событий, приняв за данность их существование. Построим модель трансформации текста ДНК, в основе которой лежала бы вышеописанная логика возникновения повторов и непрерывной компактизации генома. Остановимся на простейшем случае возникновения одного двойного случайного повтора длиной N1 нуклеотидов в модельном геноме L по случайному адресу M (см. рис. 12) Предположим, что эта копия будет встроена в матрицу по случайному адресу X по механизму вставки, в результате чего в геномном тексте появится один прямой повтор, а его длина увеличится на N1 нуклеотидов. Такая же трансформация, но уже с использованием копии фрагмента N2, считанного со случайной координаты K, переведенного в реверсно-комплементарную форму N2R, и вставленного по случайному адресу Y, приведет к увеличению длины генома на N2 нуклеотидов и одновременно к возникновению одного инвертированного повтора. Восстановление длины ДНК до первоначальной может быть произведено в два этапа: после образования прямого повтора (уменьшением цепи ДНК на N1 нуклеотидов) и после образования одного инвертированного повтора (уменьшением цепи ДНК на N2 нуклеотидов)[12]. В обоих случаях представляется целесообразным использовать процедуру делеции фрагментов N1 и N2 из первой цепи ДНК по случайным координатам B1 и B2. Совокупность операций возникновения прямого и инвертированного повтора вместе с делециями «лишнего» текста ДНК до первоначальной длины генома L обозначим в виде одного цикла «автотрансформации».

[11] Для эукариот (в отличие от прокариот) компактизация хромосомной ДНК не столь характерна [19].

[12] Для упрощения схемы на рис. 13 показана одноэтапный метод делеции лишнего текста ДНК, накопленного в результате вставки прямого и инвертированного повторов.

Компьютерная программа «Автоморфер» реализующая описанный алгоритм трансформации случайного текста ДНК, позволяет сохранять промежуточные («эволюционировавшие») тексты в виде файлов и одновременно рассчитывать для них коэффициенты лингвистической корреляции. Для генерации псевдослучайных последовательностей ДНК можно воспользоваться соответствующей утилитой (random.exe) или же применить в качестве стартовой матрицы готовый текст с известными характеристиками (www.insilico.ru\random_DNA\index.html).

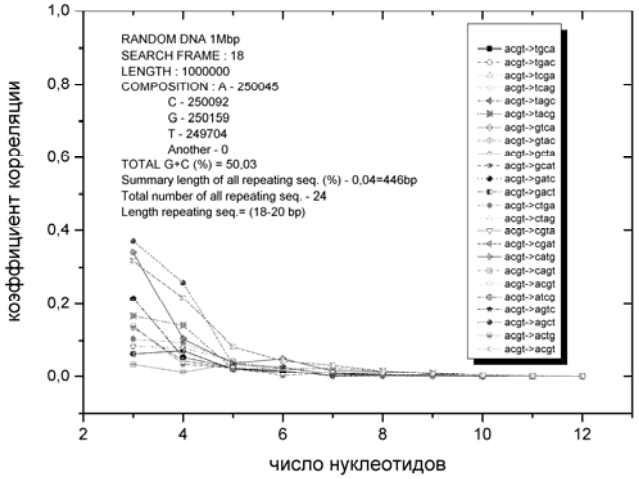

Рис. 13. Исходные характеристики псевдослучайного текста ДНК длиной 1 млн. нуклеотидов, использованного в качестве матрицы.

На рис. 13, в частности, представлены некоторые характеристики текстового массива псевдослучайной ДНК из 1 млн. нуклеотидов[13], который был использован во всех трансформациях, описанных ниже. Для того, чтобы убедиться в отсутствии любых значимых линейных корреляций в этом тексте (в том числе нонсенс-корреляций), также были проведены необходимые вычисления.

Выполним операцию автотрансформации исходной ДНК матрицы 256 тыс. циклами перестановок, пользуясь различными комбинациями вставляемого текста (инвертированный, прямой, прямой + инвертированный), одновременно определяя коэффициенты корреляции. Для модификации воспользуемся моделью генерации повторов фиксированной длины N1 = N2 = 300 нуклеотидов.

Как видно из рис. 14, на котором каждая из кривых представляет собой средние результаты трех измерений, наивысшая скорость установления лингвистической симметрии наблюдаются для варианта трансформации А, использующего только инвертированные повторы, при котором достигаются численные значения коэффициентов корреляции $r_3 \div r_{12}$, лежащие в интервале 0,79÷0,96. Несколько больший разброс $r_3 \div r_{12}$ и меньшие абсолютные значения r = 0,65÷0,94 наблюдаются для случая В, то есть, при использовании смеси равных долей 50:50 прямых и инвертированных повторов. В модели Б, имитирующей возникновение лишь прямых повторов, будет наблюдаться полное отсутствие какой бы то ни было «эволюции» лингвистической симметрии.

Применив алгоритм генерации 100% инвертированных повторов, также нетрудно убедиться в том, что скорость установления упорядоченности возрастает при использовании для автоморфинга (трансформации ДНК по случайному алгоритму, показанному на рис. 12) более длинных фрагментов (см. рис. 15).

[13] Что соответствует реальной длине геномов отдельных бактерий, например, *Mycoplasma gallisepticum* NC_004829, геномная ДНК которой состоит из 996 422 п.н.

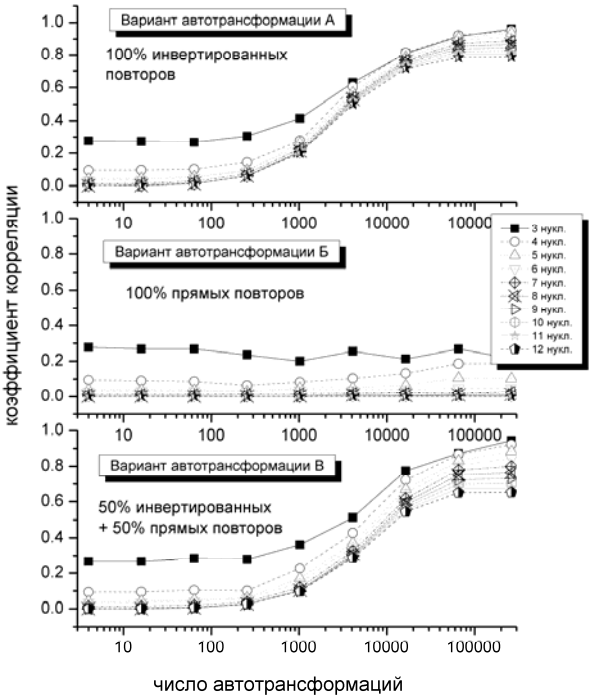

Рис. 14. Динамика роста коэффициентов лингвистической корреляции в тексте ДНК при разных методах генерации повторов.

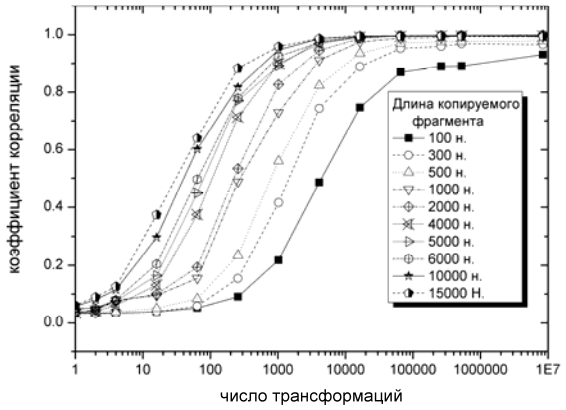

Рис. 15. Динамика возникновения коэффициентов лингвистической корреляции в случайном тексте ДНК в зависимости от длины генерируемых повторов.

Весьма важным и интересным является получение ответа на вопрос, «мешают» ли случайные мутации процессу возникновения и стабильного поддержания лингвистической симметрии, велико ли отношение допустимого числа точечно мутировавших нуклеотидов к числу нуклеотидов, «перенесенных» в процессе генерации повторов?

Воспользуемся алгоритмом генерации повторов (100% инвертированных, $N_1 = 5\,000$ нуклеотидов[14]), со случайной исходной матрицей из 1 млн. нуклеотидов при активизированной функции[15] создания случайных мутаций, по принципу обязательной равновероятной замены. Последнее означает, что после каждого цикла генерации повторов и удаления избыточного текста программа определяет в результирующем массиве случайную координату, считывает в ней нуклеотид и производит его обязательную равновероятную замену на три других возможных.

На рис. 16А представлена серия кривых, характеризующих зависимость коэффициентов лингвистической корреляции на уровне 3-, 4-, 6-, 8- и 10-нуклеотидных сочетаний, от числа мутаций в расчете на один цикл трансформации 1 миллионного текста после 4 194 304 итераций. По всей видимости, в общем случае она носит линейный характер. На рис. 16Б также продемонстрирован характер самоорганизации случайного текста на уровне 6-нуклеотидных комбинаций в процессе его трансформации при фиксированных частотах внесения случайных мутаций. Напомним, что коэффициент r_6 для несбалансированного генома бактерии *Idiomarina loihiensis* (NC_006512.1) составляет значение 0,9632 (см. табл. 3).

[14] В реальных прокариотических системах, наличие повторов длиной более 5 000 нуклеотидов не является редкостью [18].
[15] Встроенная функция программы «Автоморфер».

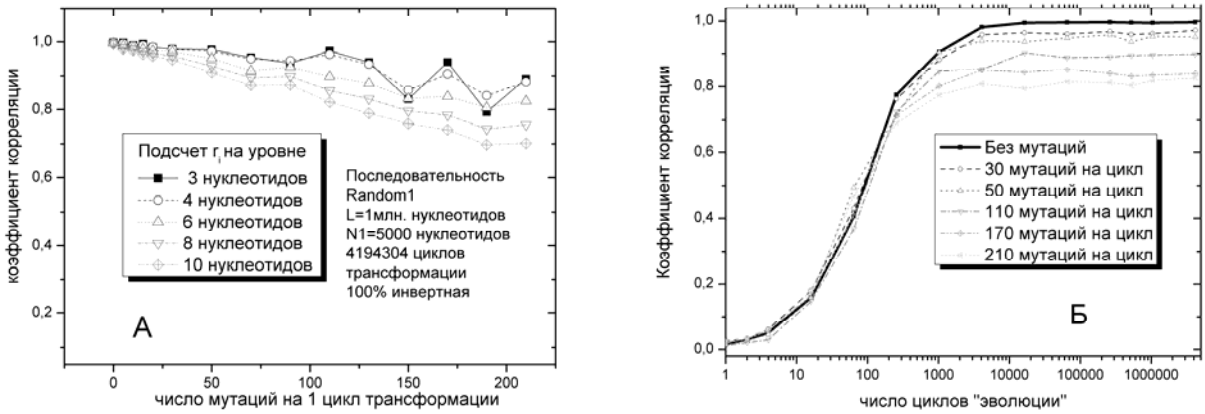

Рис. 16. Влияние «мутаций» на установление лингвистической симметрии псевдослучайного текста ДНК 1 млн. пар оснований при его одновременной трансформации *in silico* инвертированными фрагментами длиной 5000 нуклеотидов.

При этом в модельной системе, при внесении 30 и 50 мутаций на один цикл трансформации с участием вставок по 5 000 нуклеотидов через 256 тысяч итераций возникает и стабилизируется схожий уровень разбалансированности $r_6^{30} = 0{,}97$ и $r_6^{50} = 0{,}95$.

Из полученных соответствий можно сделать вывод о том, что случайные точечные мутации, произведенные методом обязательной равновероятной замены, агрессивно разрушают структуру текста, отвечающую за лингвистическую симметрию. В то же время, молекулярные процессы, ведущие к генерации инвертированных повторов, можно рассматривать в качестве главной причины возникновения и поддержания лингвистической симметрии между 1 и 2 цепями геномных ДНК. При этом обмен информацией между цепями должен протекать непрерывно, в противном случае, достигнутая симметрия будет разрушена относительно небольшим количеством точечных мутаций.

Выводы:

- Геномные ДНК представителей трех известных царств живого, а именно, бактерий, архей и эукариот, обладают высоким уровнем лингвистической симметрии, обнаруживаемой путем подсчета и сопоставления частотных словарей первой и второй цепей молекул хромосомных ДНК.
- Высокий уровень лингвистического сходства обнаруживается между текстами разных хромосом одного и того же организма, которое уменьшается (и даже исчезает) при сопоставлении живых форм, не состоящих в близком родстве.
- Лингвистическая симметрия природных геномов обусловливается присутствием в них реверсных разделенных палиндромов, в сумме занимающих основную часть геномной ДНК. Реверсные разделенные палиндромы могут также рассматриваться в виде совокупности разнесенных совершенных повторов, распределенных по всему тексту генома.

- В компьютерной модели, имитирующей образование инвертированных повторов из линейного массива псевдослучайного текста, продемонстрирована быстрая самоорганизация структуры с образованием лингвистической симметрии, которая агрессивно разрушалась случайными мутациями.
- Лингвистическая симметрия является следствием динамичных процессов изменения геномной ДНК, довлеющих над всеми другими изменениями, и ее главной причиной служит обмен информацией между цепями.

Литература

1. Fleischmann,R.D., Adams,M.D., White,O. et.al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd // Science 269 (5223), 496-512

2. Wheeler D.L., Barrett T., Benson D.A., et.al. (2006) Database resources of the National Center for Biotechnology Information // Nucleic Acids Research, Vol. 34, Database issue D173–D180

3. Вейр Б. Анализ генетических данных. – М. Мир, 1995 400с.

4. Уотермен М.С. (ред). Математические методы для анализа последовательностей ДНК – М. Мир, 1995 349 с.

5. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология. СПб.: Невский Диалект; БВХ-Петербург, 2003.-654 с

6. Александров А.А., Александров В.В., Бородовский Ю.М. и др. Компьютерный анализ генетических текстов. М.: Наука, 1990. 264 с.

7 Abramson G., Cerdeira H.A., Bruschi C., (1999), Fractal properties of DNA walks // Biosystems, 49(1):63-70.

8. Almirantis Y., (1999) A standard deviation based quantification differentiates coding from non-coding DNA sequences and gives insight to their evolutionary history // Journal of Theoretical Biology, 196:297-308.

9. Matthew I. Bellgard, Takashi Gojobori (1999), Significant differences between the G+C content of synonymous codons in orthologous genes and the genomic G+C content // Gene, 238:33-37

10. Wentian, L. (1977) The study of correlation structures of DNA sequences: a critical review // Computers & Chemistry, 21 (4), p.257-271

11 Садовский,М.Г. (2005) Информационно-статистический анализ нуклеотидных последовательностей //Дис. д-ра физ.-мат. наук : 03.00.02 .-И.: РГБ

12. Колесников А. Алгоритмическая загадка молекулярной эволюции (2002). Компьютер и жизнь, №33

13. Марков, А.В. Захаров И.А. (2006) Крупные и мелкие перестройки в эволюции прокариотических геномов // Генетика. № 11. С. 1547-1557.

14. Смирнов Г. Б. (2007) Механизмы приобретения и потери генетической информации бактериальными геномами. // Успехи современной биологии

15 Ратнер В.А. (2002) Что содержит полный геном *Escherichia coli*? //Информационный вестник ВОГиС : ISSN 1814-5558 №2

16. Kunst, F., Ogasawara, N. et.al (2000) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis* .// Nature, vol.390 p.249-256

17. Heidelberg,J.F., Eisen,J.A. et.al. (2000) DNA sequence of both chromosomes of the cholera patogene *Vibrio Cholerae*.// Nature, vol.406 p.477-483

18. Усанов Н.Н., Гильванова Е.А., Усанов Н.Г., Чемерис А.В. Анализ повторов в текстах геномов прокариот свидетельствует о дискретном характере эволюции ДНК// Математическая биология и биоинформатика: 1 Международная конференция. – М.: МАКС Пресс, 2006. С.177-178.

19. Акифьев А.П.(2004) Избыточная ДНК – генетическая квадратура круга? // Природа №10.