УДК 577.3

АНАЛИЗ ТОНКОЙ CpG СТРУКТУРЫ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Д. Р. Дуплий, Ю. Г. Шкорбатов

Харьковский национальный университет им. М. Н. Каразина
пл. Свободы, 4, Харьков 61077, Украина
E-mails: duplijd@mail.ru, Yury.G.Shckorbatov@univer.kharkov.ua.

Поступила в редакцию 5 марта 2003 г.

Предложены методы описания распределения **CpG** сайтов в последовательностях ДНК с помощью числовых характеристик и графически. Нормализация кривой графика на длину последовательности позволяет сравнивать последовательности по характеру накопления **CpG** и классифицировать их по подобию накопления, а не только по частоте встречаемости. С помощью введенных числовых характеристик и принципа подобия показано существование различий в накоплении CpG сайтов в ДНК генов организмов, отличающихся по эволюционному положению и наличию системы метилирования ДНК. Для видов, у которых метилирование ДНК не имеет регуляторного значения в процессах импринтинга и клеточной памяти, характерно случайное распределение CpG сайтов.

КЛЮЧЕВЫЕ СЛОВА: метилирование ДНК, CpG сайты, метилцитозин, изохоры, гистоновые белки, CG островки, интрон, экзон.

Характерной особенностью генома человека является крайне неравномерное распределение частот нуклеотидов и динуклеотидов **CG** и **AT** как вдоль хромосом, так и между комплементарными цепями [1, 2, 3]. При дифференциальном окрашивании (акрихином или модифицированными методиками Гимза) хромосом млекопитающих выявляется специфическая исчерченность. Различное сродство к красителям связано с нуклеотидным составом ДНК и позволяет различать **AT** и **CG** богатые сегменты [4]. Экзоны характеризуются большей стабильностью по отношению к мутациям, чем межгенные области, что связано с ограничениями, налагаемыми естественным отбором [5]. Таким образом, цепь ДНК находится в компромиссе между мутационным давлением и естественным отбором. На примере бактериального генома было показано, что содержание каждого нуклеотида в геноме пропорционально времени, необходимому для замены половины нуклеотидов данного типа [6].

Для описания областей ДНК с преобладанием **CG** или **AT** пар G. Bernardi ввел понятие изохор, то есть "имеющих одинаковый объем". Изохоры - это области ДНК длиной более 300 тыс. пар нуклеотидов, близкие по уровню содержания гуанина и цитозина [7], а геном млекопитающих представляется мозаикой изохор. Различают **CG**-бедные изохоры, находящиеся в 62% генома и содержащие 34% генов; **CG**-богатые изохоры — 31% генома, 38% генов; и **CG** обогащенные изохоры, составляющие 3% генома с 28% генов [8]. Таким образом, имеется зависимость между составом изохор и количеством генов: плотность генов в **CG** обогащенных изохорах в 10 раз выше, чем в **CG** бедных областях [9].

Следует отметить, что обычно, говоря о **CG** содержании, подразумевают частоту содержания комплементарной пары **CG**, в то время как особый интерес представляет содержание динуклеотидов 5′–**CpG**–3′ в одной цепи. Особое значение **CpG** сайтов связано со способностью цитозина в этом динуклеотиде служить мишенью при метилировании с образованием 5-метилдезоксицитидина [10], который играет важную роль в экспрессии генов. Но 5-метилцитозин - относительно неустойчивое основание, которое в результате спонтанного дезаминирования и последующей замены аминогруппы на оксигруппу превращается в тимин, что приводит к обеднению ДНК парами **CpG**,так как при первом же раунде репликации гуанин в комплементарной цепи заменяется на аденин. Таким образом, вследствие метилирования цитозина в комплементарных цепях ДНК происходит замена **CG** пар на **TA** [11]. В ДНК позвоночных 5′–**CpG**–3′ встречается в пять раз реже, чем 5′–**GpC**–3′ [10]. Частота встречаемости 5′–**CpG**–3′ и 5′–**GpC**–3′ в ДНК прокариот почти одинакова и близка к случайной, как и частоты встречаемости других динуклеотидов, например 5′–**GpA**–3′ и 5′–**ApG**–3′. Известно, что у беспозвоночных, в частности, у бактерий, дрозофилы и

нематоды уровень метилирования цитозина в **CpG** сайтах очень низок, и, по-видимому, метилирование ДНК у этих организмов не является механизмом регуляции генной экспрессии [12]. У дрозофилы имеется фермент метилтрансфераза-3а, который метилирует **CpG** сайты в ДНК эмбриональных клеток, но неактивен в клетках соматических тканей. Этот фермент также метилирует цитозин в сайтах **CpA** и **CpT** [13].

Система метилирования ДНК у бактерий важна при защите клеток от чужеродной ДНК, например, транспозонной или вирусной [14]. Система рестриктаз-метилаз бактерий содержит метилтрансферазы, метилирущие нуклеотиды в специфических сайтах. Функция метилирования ДНК у бактерий заключается в защите собственной ДНК от действия рестриктаз, расщепляющих чужеродную ДНК. У бактерий метилазы, специфически метилирующие цитозин в последовательности **CC(A/T)GG** и аденин в последовательности **GATC**, участвуют в клеточной регуляции и связаны с вирулентностью болезнетворных бактерий [14]. Имеются сведения о метилировании и других нуклеозидов, так, в составе ДНК эукариот встречается N-6 метиладенозин [4], а, например, метилгуанозин в сайте **CpG** является горячей точкой, то есть местом повышенной частоты мутирования в гене p53 при развитии рака легких у курильщиков [15]. Но все же 95% метильных групп в ДНК позвоночных содержится именно в остатках цитозина динуклеотидов 5′–**CG**–3′, и больше половины таких динуклеотидов метилировано по **5'** положению остатков цитозина. У растений можно обнаружить метилирование цитозина в тринуклеотиде 5′–**CNG**–3′, где **N=A, T, C** [10].

При расщеплении ДНК эндонуклеазой HpaII, имеющей сродство к неметилированной последовательности 5′-**CCGG**-3′, образуются низкомолекулярные фрагменты. Использование таких фрагментов в качестве зондов для блот-гибридизации с геномной ДНК выявило наличие "островков" внутри **CG** обогащенных изохор, содержащих неметилированные динуклеотиды **CpG**, названные **CG** или **HTF** (HpaII tiny fragments) островками длиной 1000-2000 п.н.[16, 17, 18]. Существуют более строгие определения, полученные с использованием статистических методов, согласно которым **CpG**-островками считают:

1) последовательности ДНК длиной более 200 п.н., содержащие более 50% **C+G,** при этом отношение наблюдаемой частоты **GC** к ожидаемой (случайной) больше, либо равно 0.6 [19],

2) последовательности от 500 до 2000 п.н., содержащие 60% **C+G**, а также неметилированные **CpG** динуклеотиды и **G/C** боксы - участки, родственные сайту узнавания фактора транскрипции **Sp1-G4CG4C** [7].

CpG-островки располагаются в начале транскрибируемых генов и могут выступать маркерами конкретных структурных генов. Отмечено предпочтительное расположение **CpG** островков в первых кодирующих экзонах [20]. Островки наблюдаются в большинстве генов "домашнего хозяйства" в 5′ фланкирующей последовательности и у 40% тканеспецифичных генов [21]. Экспериментально обнаружено от 30 000 до 45 000 **CpG**-островков в геноме человека [16, 17, 22]. Большое их количество находится также в теломерных участках хромосом *1, 9, 15-17, 19, 20* [21].

Метилированные области ДНК, связываясь с определенными белками, становятся недоступными для действия ряда факторов транскрипции, вследствие чего снижается экспрессия генов [9]. Эффективность генной экспрессии при наличии метилированных групп уменьшается, а активное состояние генов сочетается с гипометилированием в тех же сайтах [20]. Замечено, что метилирование в двух цепях (симметричное) более стабильно и может сохраняться в течение нескольких клеточных поколений, в то время, как метилирование в одной цепи (асимметричное), менее стабильно и подвержено деметилированию [23]. Метилирование **CpG**-островков связано с явлением геномного импринтинга, который установлен в *15,11,14* и *7* хромосомах [24]. Геномным импринтингом называется механизм, обеспечивающий избирательную экспрессию признаков, унаследованных от одного из родителей. Отцовский импринтинг означает, что фенотипической экспрессии аллеля не происходит при передаче его от отца, а материнский импринтинг соответственно — при передаче аллеля от матери. Геномному импринтингу подвержена небольшая доля генов млекопитающих (по различным оценкам, от 100 до 300), эти гены вовлечены в регуляцию эмбрионального развития, причастны к онкогенезу [25]. У человека некоторые импринтируемые гены связаны с наследственной патологией, что прослеживается на примере ряда моно- и полигенных заболеваний. В частности, хорошо изучены такие патологии, передающиеся исключительно или в большей степени от матери, как синдром Беквита-Видемана (ген локализован в хромосоме *11p*), нейрофиброматоз II (*22q*). Преимущественно от отцов передаются такие наследственные патологии, как

HLA-связанный сахарный диабет (*6q*), синдром Прадера-Вилли (кластер импринтированных генов *15q11-q13*) [26]

Механизмы, связанные с метилированием ДНК, обеспечивают инактивацию одной из X хромосом в геноме самок млекопитающих. Инактивации в данном случае подвергается хромосома, полученная либо от отца, либо от матери. Инактивированная хромосома, ДНК которой подвергается дополнительному метилированию, находится в конденсированном, гетерохроматинизированном состоянии. В неактивной хромосоме обнаружены так называемые "центры инактивации", в которых процесс инактивации начинается и от которых затем распространяется вдоль хромосомы [27].

Таким образом, закономерности распределения **CpG** динуклеотидов внутри областей ДНК связаны с регуляцией активности генома, поэтому изучение распределения **CpG** является актуальной проблемой.

МАТЕРИАЛЫ И МЕТОДЫ

Анализируемые последовательности ДНК

Нуклеотидные последовательности были взяты из общедоступной базы данных Американского Национального Центра Биотехнологической Информации (NCBI). Рассматривались последовательности из следующих таксономических групп: бактерий – Thermoplasma acidophilum, Thermotoga maritima, Methanococcus jannaschii, нематод – Caenorhabditis elegans, насекомых – Drosophila melanogaster, костистых рыб подотряда нототениевидных, обитающих в Антарктике – Dissostichus mawsoni, Harpagifer antarcticus, Notothenia coriiceps, Pagothenia borchgrevinki, а также нуклеотидные последовательности человека, кодирующие функционально различные белки, в частности, цитохромы, локус гистосовместимости, семейство гистоновых белков.

Для работы отбирались области ДНК с экспериментально установленными функциями, которые отражены в соответствующих описаниях протоколов GENBANK. Гены, локализованные в комплементарной цепи, а также предположительно некодирующие области не рассматривались. Геномы бактерий анализировались полностью.

Из начального набора последовательностей были выбраны гены метаболических ферментов бактерий, гены антифризовых гликопротеидов костистых рыб, у человека гены цитохромов и семейство генов гистоновых белков (человека и нематоды), а также 24 **CpG**-островка в спейсерных районах, расположенных в **C+G** богатой области хромосомы *6p21.3-22.3*. Причем, распределение **CpG** пар вычислялось отдельно для всего гена, его первичного транскрипта РНК, и кодирующей части (с удаленными интронами, если такие были).

Характеристики содержания CpG сайтов

Для изучения распределения **CpG** пар вдоль одной цепи ДНК, введем следующие количественные характеристики: плотность оснований **C** и **G** равна

$$\frac{N_C + N_G}{L} 100\%,$$

где $N_{\mathbf{C}} + N_G$ — абсолютное количество оснований **C** и **G** в последовательности длиной L, выражающейся числом нуклеотидов. Плотность **CpG** пар есть $\frac{N_{CpG}}{L} 100\%$

Очевидно, что при равномерном распределении всех четырех нуклеотидов плотность $(N_C + N_G)$ будет равна 50 %. При этом в создании **CpG** пар будет участвовать $\frac{1}{4}$ всего цитозина, который также пойдет на образование других динуклеотидов **CT**, **CA**, **CC**. Но в реальности имеет место неравномерное распределение нуклеотидов, особенно на коротких отрезках. Допустим, что ни один цитозин не вовлечен в образование пар **CpG**, тогда число пар на этом участке будет равно 0. Предельно возможному количеству **CpG** пар соответствует случай, когда весь цитозин связывается только с гуанином, после цитозина в одиночной цепи всегда стоит гуанин, тогда количество таких пар будет определяться меньшим из N_C или N_G.

"Вовлеченность **CpG** пар" показывает, какая доля цитозина и гуанина участвует в образовании **CpG** пар

$$\frac{N_{CpG}\text{ наблюдаемое}}{(N_C + N_G)}100\%.$$

"Эффективность образования **CpG** пар"

$$\frac{N_{CpG}\text{ наблюдаемое}}{N\text{ предельно возможное}}100\%,$$

где "предельно возможное" число пар определяется меньшим из чисел N_C или N_G.

Введенные нами величины позволяют количественно характеризовать **CpG**-свойства нуклеотидной последовательности. В силу неравномерного распределения нуклеотидов вдоль одной цепи, вероятно, что эти свойства будут отличаться в функционально и эволюционно различных последовательностях ДНК. Таким образом, плотность **CpG** пар, вовлеченность и эффективность могут быть использованы в качестве характеристик нуклеотидных последовательностей, связанных с их эволюционным происхождением.

Графическое изображение CpG распределения

Ранее описание распределения **CpG** сайтов проводилось либо во всей последовательности в целом [5, 28], либо с помощью "скользящего окна", двигающегося вдоль анализируемой последовательности.

Метод скользящего окна представляет содержание **CpG** сайтов в длинных последовательностях ДНК в виде кривой, пики которой соответствуют наблюдаемым частотам внутри сканирующего окна длиной, например, 500 нуклеотидов и шагом продвижения 10 нуклеотидов [29]. Этот метод не дает информации о характере распределения **CpG** сайтов внутри этого довольно большого окна и о положении **CpG** сайтов относительно границ функционально различных областей.

Нами предложен метод, который позволяет наглядно отобразить характер расположения **CpG** динуклеотидов внутри выбранного отрезка. По оси абсцисс откладывается порядковый номер нуклеотида в последовательности, а по оси ординат - суммарное количество встреченных **CpG** к данному моменту. Такое построение позволяет визуализовать динамику появления пар вдоль последовательности, то есть, тонкую **CpG** структуру. Угол наклона ступенчатой кривой может быть охарактеризован величиной, которую назовем плотностью накопления **CpG** пар. В нашей модели экстремальному случаю, когда последовательность состоит только из **CGCGCGCGCG**... соответствует угол наклона 45^0. Если последовательность не содержит ни одного динуклеотида **CpG**, график представляется горизонтальной прямой, следовательно, угол наклона равен нулю. Все остальные графики располагаются в секторе от 0 до 45^0. Графики были нормализованы на длину последовательности, что позволило их сравнивать по подобию.

РЕЗУЛЬТАТЫ

Проведенный анализ выявил различия в характере накопления **CpG** пар в генах разных белков разных организмов. Среди 997 генов Thermotoga maritima и 767 генов Thermoplasma acidophilum (в основном, это гены метаболических ферментов) содержание (N_C+N_G) больше 50% наблюдалось только у некоторых: в 70 (7,02%) генах Thermotoga maritima и в 51(6,64%) генах Thermoplasma acidophilum. В среднем, (N_C+N_G) состав генов бактерий составлял около 46%, из которых вовлечено в образование **CpG** пар только 10,28% цитозина и гуанина. Содержание **CpG** сайтов у Thermotoga maritima – 4,9% что составляет почти четверть (24%) от максимально возможного числа **CpG** пар при данном наборе нуклеотидов – эффективность образования **CpG** пар (см. табл. 1). В 855 генах Methanococcus jannaschii **G+C** содержание не более 40%, из которых в 27 генах (3,16%) **CpG** пар вообще не было. Среднее содержание (N_C+N_G) составило 31,8%, вовлеченность 22,5%, а эффективность - 64,83%. Примечательно, что у бактерий соотношения нуклеотидов внутри кодирующих частей такие же, как и во всем геноме, чего нельзя сказать о других более высокоорганизованных видах.

При низком содержании **CpG** график накопления пар **CpG** в генах бактерий представляет беспорядочный набор точек. Графики последовательностей, в которых число **CpG** сайтов менее 10 на ген, не классифицировались. Например, область ДНК архебактерии (Methanococcus jannaschii) для гистонов А3 и А2 содержит всего 2-4 **CpG** пары, при длинах областей 250 -300 нуклеотидов, соответственно.

Местонахождение первой пары **CpG** у всех бактерий не показало никакой корреляции относительно начала гена, в то время как в генах человека имелось четкое предпочтение первой парой **5'**- конца гена, независимо от его длины.

Таблица 1. Усредненные величины содержания **CpG** сайтов в проанализированных нуклеотидных последовательностях

Организм	CG%	C+G%	вовл.	эфф.	N генов	C+G% геном	
Thermotoga maritima	4,90	45,93	10,66	24,40	997	46,25	
Thermoplasma acidophilum	4,83	46,41	10,40	23,21	767	45,99	
Methanococcus jannaschii	4,85	31,8	11,86	26,20	855	31,41	
C. elegans II хромосома	3,11	34,81	8,93	18, 28	5	34,82	
кодирущая часть II хр.	4,82	45,55	10,5	23,40	5	-	
гены II хр.	3,5	38,13	9,18	19,41	5	-	
Drosophila melanog. III хр.	4,34	41.85	10,38	20,96	36	41,85	
кодирующая часть	5,37	46,24	11,61	26,20	36	-	
гены Dros.	5,59	48,41	11,54	23,77	36	-	
mRNA Dros.	5,59	48,41	11,39	25,84	36	-	
Dros. митохондр геном	0,68	17,85	3,79	8,92	14	17,84	
тРНК митох. дроз.	0,42	20	2,1	4,12	14	-	
Harpagifer antarcticus	2,43	55,48	4,37	9,57	4		
Pagothenia borchgrevinki	2.25	55.59	4,06	9,13	1		
Dissostichus.mawsoni	2,38	45,13	5,27	10,46	7		
Notothenia corriceps	2,61	55,67	4,69	10,53	2		
Homo sapiens VI хромосома	2,51	43,75	5,74	12,98	-	43.10	
гистоновые гены	8,62	61,52	13,99	29,06	9	-	
CG островки	7,79	62,14	12,54	26,89	24	-	
локус гистосовместимости	5,23	60,61	8,65	22,73	16	-	
цитохром P450	1,00	46,50	2,15	4,35	-	-	

Интересно заметить, что внутри генов транспортной РНК бактерий содержание (N_C+N_G) и **GpC** было зачительно выше, чем в других генах (см. табл. 1), в то время как **CpG** пары в 14 генах транспортной РНК Drosophila melanogaster не наблюдались, кроме одной тРНК для аргинина, содержащей 2 таких пары. Среднее содержание **G+C** было около 20%. Графики распределения **CpG** сайтов в последовательностях генов коллагена и 3 хромосомы нематоды (Caenorhabditis elegans) в общем случае аппроксимировались линейной функцией, хотя на коротких участках можно заметить уплотнение содержания пар **CpG** (Рис. 1). Среди 36 генов L плеча хромосомы *3* Drosophila melanogaster среднее **G+C** отмечалось несколько выше 48,41%, чем во всей хромосоме (41,85%). При этом доля **GpC** сайтов 5,59%, вовлеченность 11,39%, а эффективность 23,77%. В то время, как в митохондриальном геноме дрозофилы наблюдается низкое содержание **G+C**, всего лишь - 17,84%.

В проанализированных нами геномах бактерий нематоды и дрозофилы не отмечалось выраженной неоднородности в распределении **CpG**. Графики накопления **CpG** сайтов в последовательностях ДНК бактерий, дрозофилы и нематоды аппроксимированы линейной функцией, то есть появление **CpG** сайтов в анализируемой последовательности происходит случайно вне связи **CpG** сайта с положением относительно начала и конца гена (Рис. 2).

У позвоночных, в частности у человека, графики располагаются над или под прямой, связывающей максимальное количество **CpG** сайтов в конце анализируемой последовательности с началом координат, и выглядят выпуклыми (Рис. 3) или вогнутыми (Рис. 4) кривыми, что дает возможность систематизировать последовательности по подобию кривых и наглядно видеть распределение **CpG** пар относительно границ гена, кодирующих или сигнальных областей.

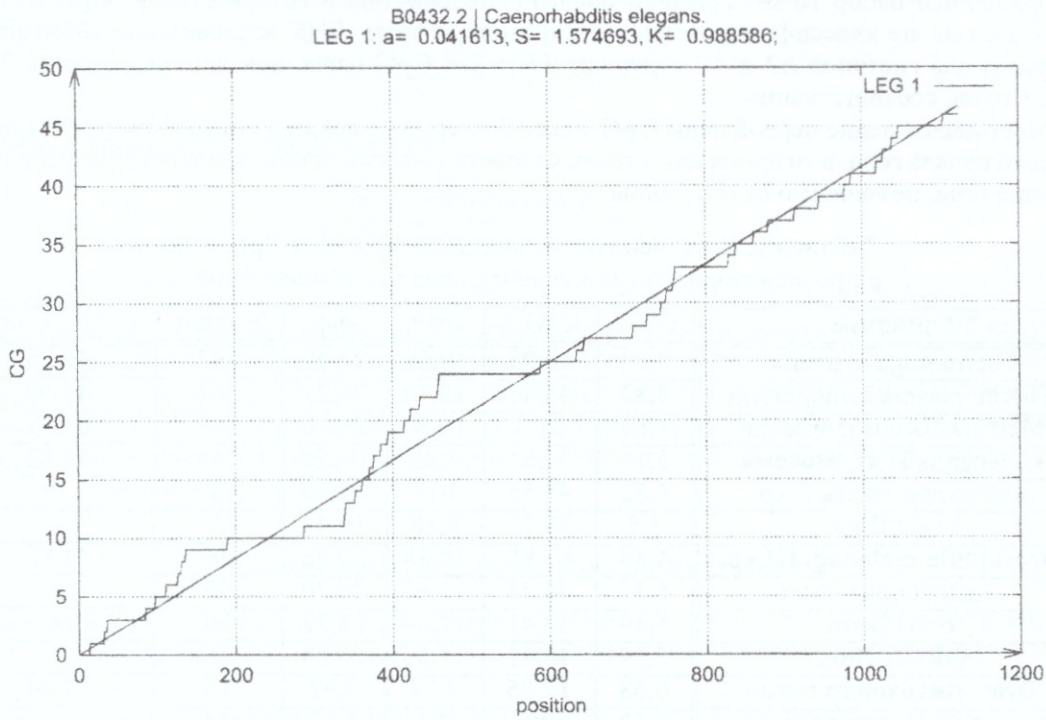

Рис. 1. График накопления CpG сайтов в гене гистонового белка H1 (C. elegans). На всех графиках: ось абсцисс — позиция в последовательности, ось ординат — кумулятивное количество **CpG** пар, первая строка — оригинальное название последовательности из GENBANK, a — угол наклона аппроксимирующей прямой, S — среднеквадратичное отклонение, K — коэффициент корреляции.

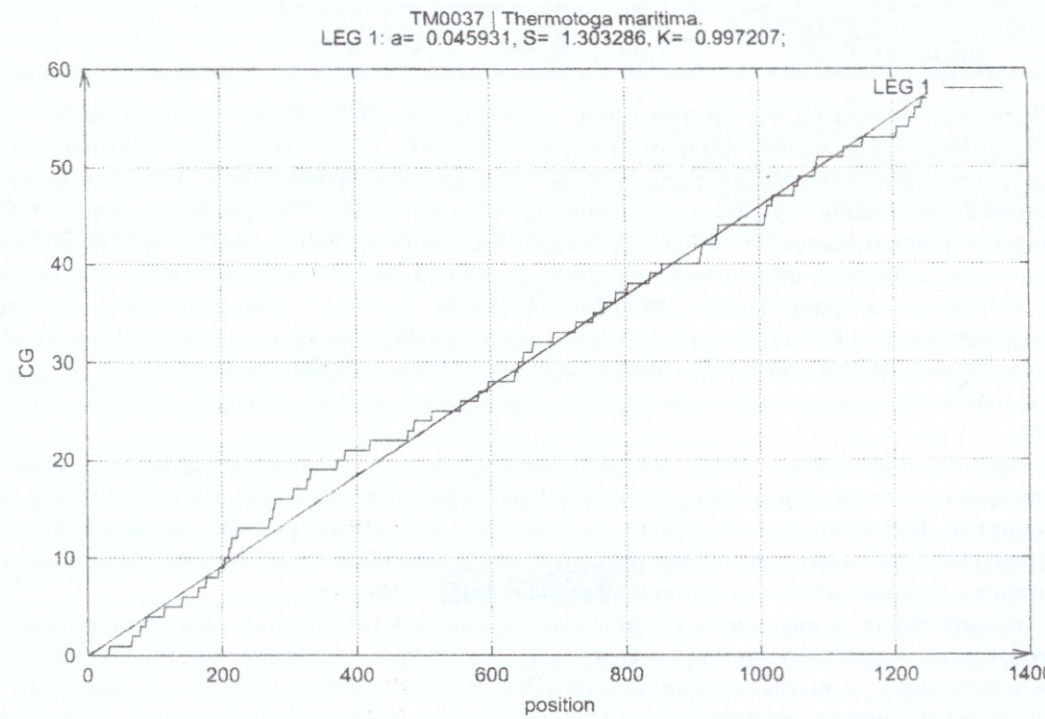

Рис. 2. Типичный график накопления **CpG** сайтов для бактериальных генов Thermotoga maritima.

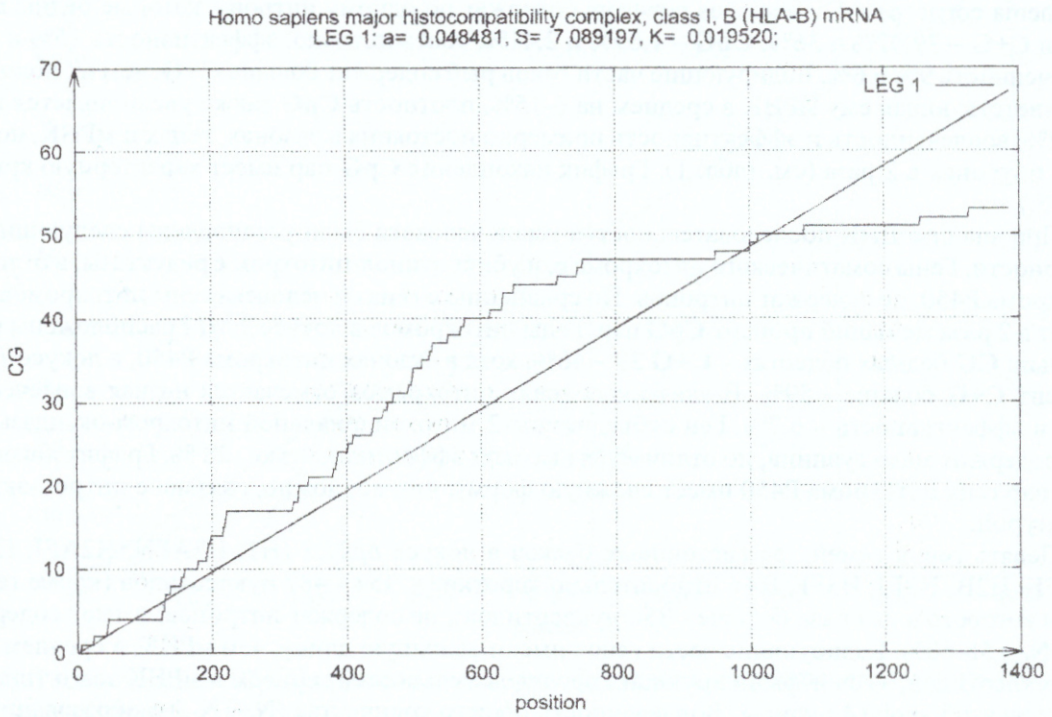

Рис. 3. График накопления CpG сайтов в гене главного комплекса гистосовместимости класса I, B (Homo sapiens). Заметно повышение частоты встречаемости CpG в средней части гена.

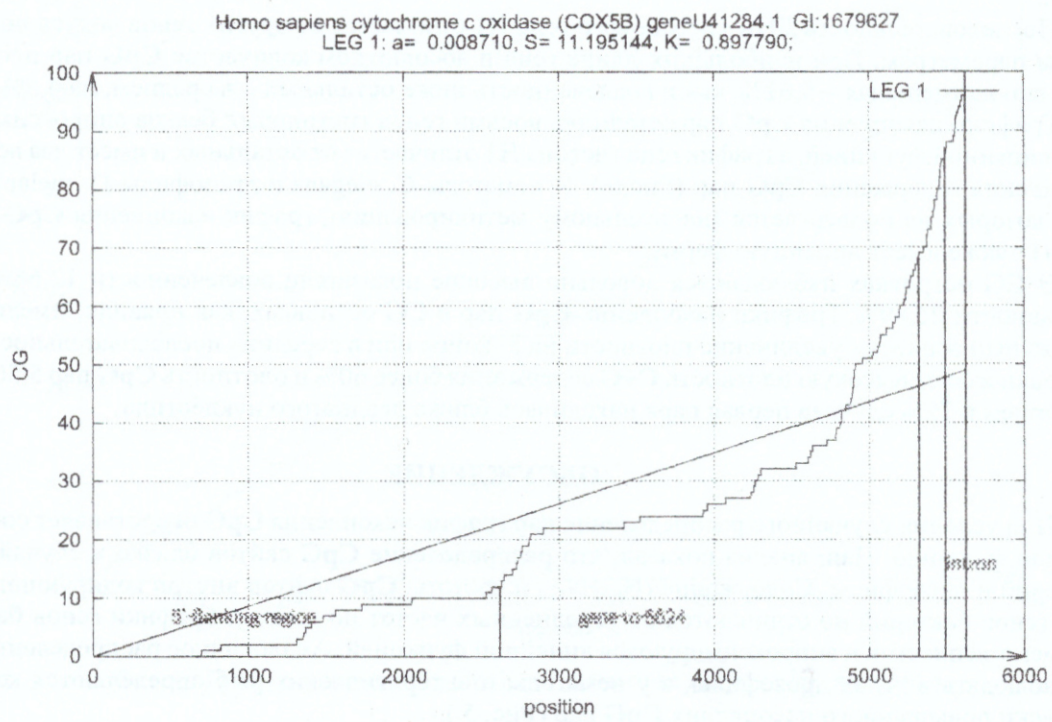

Рис. 4. График распределения CpG сайтов в гене цитохрома с оксидазы человека

Проанализировано 14 генов 4 видов антарктических рыб, в том числе 4 гена антифризовых белков. В генах всех четырех видов отмечается невысокое среднее содержание **CpG** пар 3,55%, низкая вовлеченность – 5,52% и эффективность- 13,24%. Гены антифризовых гликопротеинов Notothenia corriceps и Dissostichus mawsoni содержат по одному интрону, которые бедны в отношении **C+G** – 39,02% и 36%, **CpG** – 1,69% и 2,41%, соответственно, эффективность 15% и 14%, а вовлеченность 9% и 6%. Кодирующие части генов рыб содержат больше **C+G**, чем их полный ген и соответствующая ему мРНК в среднем на 6-15%, плотность **CpG** также увеличивается на 0,08 – 2,42%, вовлеченность и эффективность примерно постоянна в экзонах, генах и мРНК, но ниже, чем в интронах в 2 раза (см. табл. 1). График накопления **CpG** пар имеет характерную кривизну (Рис. 5.).

При анализе ДНК последовательностей генов человека были установлены следующие закономерности. Гены соматического цитохрома **c**, и убиквитинол-цитохром **c** редуктазы, в отличие от цитохрома P450, не содержат интронов. По сравнению с генами человека гены цитохромов мыши имеют в 2 раза меньший процент **CpG** пар. Гены цитохромов в локусе *22q12* расположены в относительно **CG** бедных областях – **C+G** 39 – 46%, хотя в экзонах цитохрома P450, в локусе *15q21.2* процент **C+G** больше - 59%. В указанных генах цитохромов отмечается низкая вовлеченность 2,8% и эффективность – 6,7%. Ген субъединицы 2 митохондриальной цитохром-оксидазы человека содержит мало гуанина, но отличается высокой эффективностью - 28 %. График накопления **CpG** пар гена цитохрома P450 имеет сложную форму, что, возможно, связано с интрон-экзонной структурой.

Девять генов семейства гистоновых белков в локусе *6p21.3* (H1, H2AFN, H2AFI, H2AFD, H2AFE, H2B, H3FJ, H3FF, H4), относительно короткие – 354 - 487 нуклеотидов (кроме гена H1, длина которого в два раза больше - 856 нуклеотидов), не содержат интронов, имеют содержание (N_C+N_G) 58-60%. Кодирующие части генов имеют меньшую длину, чем мРНК, в среднем, на 50-100 нуклеотидов, за счет фланкирующих последовательностей, причем в мРНК этого типа генов отсутствуют 3'-poly(A)-концы. Вовлеченность общего количества (N_C+N_G) в образование пар в генах гистонов выше, чем в генах цитохромов -13,99%, а эффективность образования пар 26%. Абсолютное число **CpG** пар в кодируюшей области примерно такое же, как в мРНК, а плотности (N_C+N_G) содержания и **CpG** пар выше, чем в гене в среднем на 2,4% и на 2% соответственно, вовлеченность примерно одинакова в обоих случаях. Таким образом, пары **CpG** преимущественно находятся не в сигнальных и терминальных областях, а внутри транскрибируемой последовательности.

Последовательность ДНК гистонового гена H1 отличается от других генов локуса по следующим параметрам. При наибольших длине гена и абсолютном количестве **CpG** пар плотность **CpG** пар наименьшая – 5,61%, как и вовлеченность ниже остальных – в среднем, на 4,5%.

Графики накопления **CpG** пар семейства восьми генов гистоновых белков аппроксимировались линейной функцией, а график гена гистона H1 отличается от остальных и имеет два всплеска повышения содержания **CpG** пар (Рис.6). У нематоды C. elegans и дрозофилы D. melanogaster, ДНК которых не подвергается значительному метилированию, график накопления **CpG** сайтов для H1 также имеет линейную форму.

В **CG** островках наблюдаются довольно высокие показатели вовлеченности 12,66% и эффективности 26,89%. Графики накопления **CpG** пар в **CG** островках, как правило, имели слабо выраженый всплеск - увеличение плотности на 5' конце или в середине последовательности. Все островки имеют высокую плотность **C+G** содержания более 60% и плотность **CpG** пар 5,50 -10,40 %, причем в 75% случаев первая пара находилась ближе двадцатого нуклеотида.

ОБСУЖДЕНИЕ

При условии случайного распределение пар график накопления **CpG** представляет собой линейную функцию. Наш анализ показал, что распределение **CpG** сайтов близко к случайному у бактерий и дрозофилы. Содержание (N_C+N_G) и частоты **CpG** сайтов внутри кодирующих областей генов бактерий не отличаются от усредненных частот по геному. Графики генов бактерий не имеют всплесков и аппроксимируются линейной функцией. Аналогичное распределение можно наблюдать в генах дрозофилы, а у нематоды и антарктических рыб определяются короткие всплески повышенного накопления **CpG** пар (Рис. 5).

Гены антифризовых гликопротеидов обладают специфическими свойствами, например, они

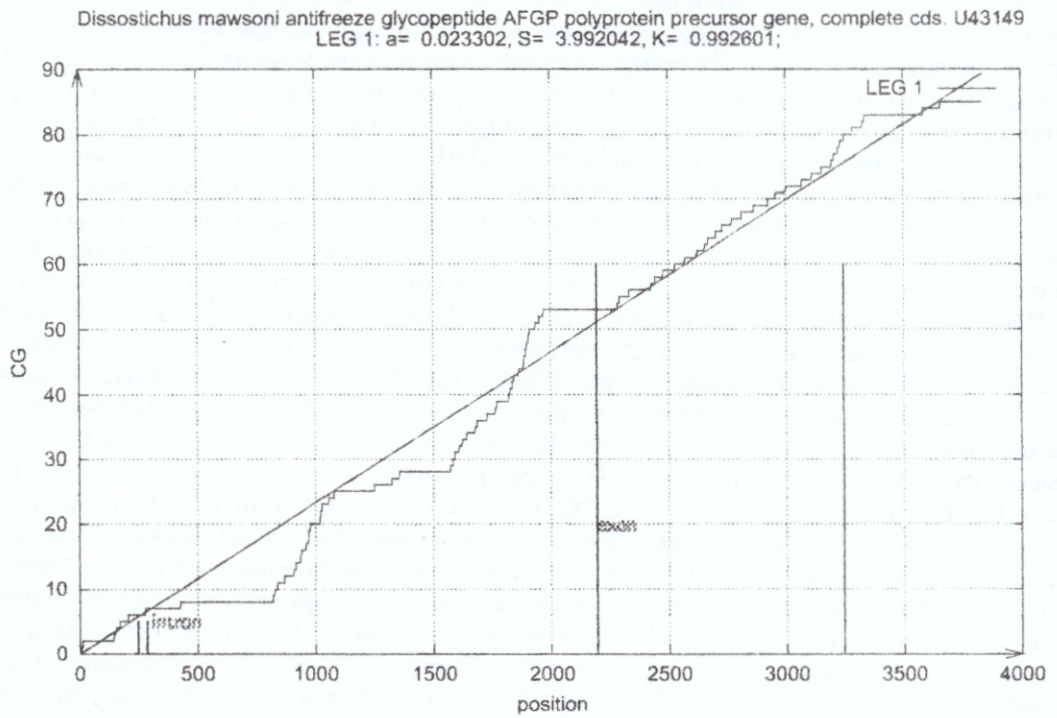

Рис. 5. Распределение CpG сайтов в гене антифризового гликопротеида костистой рыбы Dissostichus mawsoni.

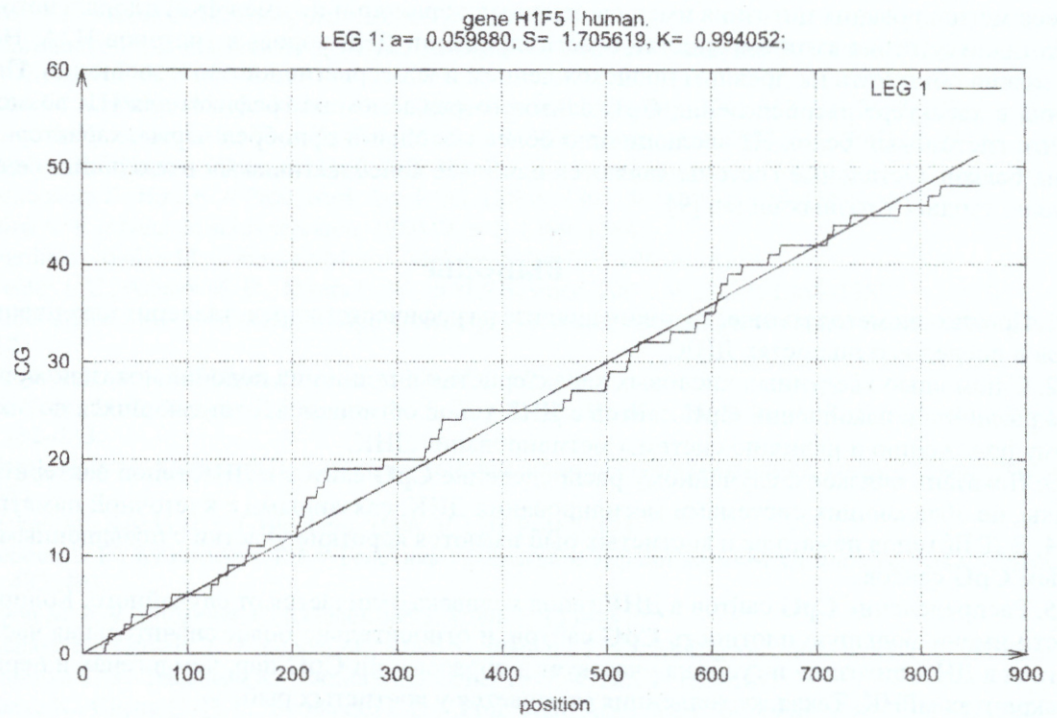

Рис. 6. График накопления CpG сайтов в гене гистона H1 человека.

синтезируют большое число копий пептидов, а сами гликопептиды претерпевают посттрансляционные изменения [30]. В областях, объединяющих интроны и экзоны, наблюдается разный характер накопления **CpG** сайтов. Линейную апроксимацию распределения **CpG** сайтов в ДНК бактерий и дрозофилы можно объяснить отсутствием механизма регуляции экспрессии генов путем метилирования цитозина в положении перед гуанином по 5-му атому и отсутствием (либо малой активностью, в случае дрозофилы) специфической ДНК метилазы в клетках указанных видов.

Для последовательностей генов белков человека показано наличие отклонений от случайного распределения **CpG** сайтов вдоль анализируемой последовательности ДНК. Различия в графиках распределения **CpG** сайтов для генов гистоновых белков можно было бы объяснить кодонной избирательностью (предпочтением **CG** содержащих кодонов) при похожем аминокислотном составе. Например, увеличение плотности **CpG** пар на графике- "всплеск", может обусловлено расположением подряд идущих кодонов для аргинина, который, как известно, кодируется шестью кодонами, четыре из которых содержат **CpG**: **CGA, CGU, CGG, CGC, AGA AGG.** Но аргинином богаты белки H2, H3, H4, а у H1 аргинин встречается всего 3 (1,23%) раза на 242 аминокислоты. Таким образом, "всплеск" **CpG** пар, определяемый на графике, нельзя объяснить скоплением в этом месте кодонов для аргинина. Белок H1 содержит аланин в 59 позициях, (из четырех кодонов которого **CpG** есть только у одного — **GCG, GCA, GCU, GCC**), а лизин в 65 позициях, кодоны которого вообще не содержат **CpG**. Другие аминокислоты в H1 представлены глицином в 16 позициях, серином - 13, валином - 10. У нематоды белок H1 (190 аминокислот) также богат аланином 37 и лизином 46, но, по сравнению с человеческим, богаче аргинином 4 (2,10%) и содержит серин в 11 позициях, а валин в 18. Из этого можно заключить, что всплески содержания **CpG** пар в гене белка H1 являются не следствием избирательности **CpG** богатых кодонов, а возникают в силу каких-то других причин. То же можно сказать о белках локуса гистосовместимости, аминокислоты которого кодируются кодонами без **CpG** [31]).

Последовательности для разных антигенов (класс I -HLA-G, HLA-C, HLA-B, HLA-Cw8.2, класс II антиген HLA-DRB4) имеют повышенную скорость накопления **CpG** сайтов на 5' конце и в середине, отчего графики этих генов имеют подобную друг другу выпуклую форму (Рис. 3.). В отличие от экзонов, интроны локуса HLA имеют повышенное содержание **CpG** сайтов на 3' конце. Если учесть, что гистоновые белки возникли в процессе эволюции еще в то время, когда процесс метилирования цитозина имел спонтанный характер и не имел функционального значения, тогда отсутствие закономерностей а метилировании ДНК коровых гистонов H2A, H2B, H3 и H4 можно объяснить их древним происхождением и консервативностью в эволюции. Причина отличия в характере распределения **CpG** сайтов (отражаемом на графике) для H1, возможно, в том, что гистоновый белок H1 эволюционно более молодой и приобрел черты, характерные для других белков. Остальные гистоны являются наиболее консервативными в эволюции белками и довольно сходны у разных видов [9].

ВЫВОДЫ

1. Предложен метод количественного анализа и графического представления накопления **CpG** сайтов в последовательностях ДНК.

2. С помощью введенных числовых характеристик и принципа подобия показано существование различий в накоплении **CpG** сайтов в ДНК генов организмов, отличающихся по эволюционному положению и наличию системы метилирования ДНК.

3. Показано близкое к случайному распределение **CpG** сайтов в ДНК генов бактерий и дрозофилы, не обладающих системами метилирования ДНК, связанными с клеточной памятью.

4. В ДНК генов нематоды и костистых рыб имеются короткие участки с повышенным содержанием **CpG** сайтов.

5. Распределение **CpG** сайтов в ДНК генов человека отличается от случайного. Кодирующие области имеют большую плотность **CpG** сайтов, и относительно более значительная часть имеющегося в ДНК цитозина и гуанина участвует в образовании **CpG** пар, чем в генах и первичных транскриптах мРНК. Такая же тенденция отмечается у костистых рыб.

6. Интроны антифризовых белков костистых рыб содержат меньше **CpG** сайтов, чем их экзоны. У дрозофилы и нематоды плотности **CpG** сайтов внутри кодирующей части и в первичных транскриптах почти одинаковы. У бактерий плотность **CpG** сайтов внутри генов и в среднем по геному не отличаются. У человека интроны антигенов HLA локуса имеют увеличенную плот-

ность **CpG** сайтов на 3' конце. Имеется предпочтительное расположение **CpG** сайтов на 5' конце в генах локуса гистосовместимости, гена микрофибриллярного белка, а также в **CpG** островках.

7. Гены гистоновых белков имеют высокую плотность **CpG** сайтов – 8-10%, при этом гены гистонов H2AFN, H2AFI, H2AFD, H2AFE, H2B, H3FJ, H3FF, H4 отличаются по характеру накопления **CpG** сайтов от гистона H1.

8. Распределение **CpG** сайтов ДНК у видов разных таксономических групп связано с действием мутационного процесса и естественного отбора. Графически можно выделить различные образцы накопления **CpG** сайтов. Для видов, у которых метилирование ДНК не имеет регуляторного значения в процессах импринтинга и клеточной памяти, характерно случайное распределение **CpG** сайтов.

Благодарности. Авторы выражают благодарность В. Калашникову за создание оригинальных компьютерных программ (Perl, C++) и проявленное терпение, Г. Ч. Куринному за полезные советы по статистическому анализу, Н. А. Чащину за интерес к работе, а также С. А. Дуплию за многочисленные обсуждения и дискуссии в течение всей работы, помощь в проведении и интерпретации числовых расчетов, оформление текста и общую поддержку.

СПИСОК ЛИТЕРАТУРЫ

1. Szczepanik D., Mackiewicz P., Kowakczuk M. // J. Mol. Evol. 2001. V. 52. P. 426–433.
2. Kowalczuk M., Mackiewicz P., Mackiewicz D. // J. Appl. Genet. 2001. V. 42. P. 553–577.
3. Cebrat S., Dudek M. R., Mackiewicz P. // Theory Biosci. 1998. V. 117. P. 78–89.
4. Сингер М., Берг П. Гены и геномы. Т. 1. М.: Мир, 1998. 373 с.
5. Bulmer M. // Mol. Biol. Evol. 1987. V. 4. P. 395–405.
6. Kowalczuk M., Mackiewicz P., Mackiewicz D., Nowicka A., Dudkiewicz M., Dudek M. R., Cebrat S. // BMC evolutionary biology. 2001. V. 17. P. 1–13.
7. Aissani B., Bernardi G. // Gene. 1991. V. 106. P. 173–183.
8. Mouchiroud D., D'Onofrio G., Aissani B., et al // Gene. 1991. V. 100. P. 181–187.
9. Глазко В. И., Шульга Е. В., Дымань Т., Глазко Г. В. ДНК технологии и биоинформатика в решении проблем биотехнологий млекопитающих. Белая Церковь: Белоцерк. гос. аграр. унив., 2001. 487 с.
10. Сингер М., Берг П. Гены и геномы. Т. 2. М.: Мир, 1998. 391 с.
11. Льюин Б. Гены. М.: Мир, 1987. 544 с.
12. Bird A. // Genes and Development. 2002. V. 16. P. 6–21.
13. Ramsahoye B., Biniszkiewicz D., Lyko F., Clark V., Bird A. P., Jaenisch R. // Proc. Natl. Acad. Sci. USA. 2000. V. 97. P. 5237–5242.
14. Low D. A., Weyand N. J., Mahan M. J. // Infection and Immunity. 2001. V. 69. P. 7197–7204.
15. Yoon J.-H., Smith L. E., Feng Z., Tang M.-S., All // Cancer research. 2001. V. 1. P. 7110–7117.
16. Bird A. P. // Nature. 1986. V. 321. P. 209–213.
17. Antequera F., Bird A. // Proc. Natl. Acad. Sci. U.S.A. 1993. V. 90. P. 11995–11999.
18. Bird A. P. // Nucleic acids research. 1980. V. 8. P. 1499–1504.
19. Gardiner-Garden M., Frommer M. // J. Mol. Biol. 1987. V. 196. P. 261.
20. Venter J. C., Adams M. D., Myers E. W., et al // Science. 2001. V. 291. P. 1304–1351.
21. Antonarakis S. E. // Genome linkage scanninig: Systematic or intelligent? Nature Genet. 1994. V. 8. P. 211–212.
22. Cross S. H., et al // Mamm. Genome. 2000. V. 11. P. 373.
23. Vu T. H., Li T., Nguyen D., Nguyen B. T., Yao X.-M., Hu J.-F., Hoffman A. R. // Genomics. 2000. V. 64. P. 132–143.
24. Пузырев В. П., Степанов В. А. Патологическая анатомия генома человека. Новосибирск: Наука, 1997. 265 с.
25. Nakao M., Sasaki H. // J. Biochem. 1996. V. 120. P. 467–473.
26. Зерова Т. Е., Бужієвська Т. І // Генетика і селекція в Україні на межі тисячоліть. Київ. Логос, 2001. С. 482–493.
27. Lyon M. F. // Cytogenet. Cell Genet. 1998. V. 80. P. 133–137.
28. Kanaya S., Yamada Y., Kinouchi M., Kudo Y., Ikemura T. // J. Mol. Evol. 2001. V. 53. P. 290–298.
29. Shimizu T. S., Takahashi K., Tomita M. // Gene. 1997. V. 205. P. 103–107.
30. Hsiao K., Cheng C.-H. C., Fernandes I. E. // Proc. Natl. Acad Sci. USA. 1990. V. 87. P. 9265–9269.
31. Tykosinski M. L., Max E. E. // Nucl. Acids Res. 1984. V. 12. P. 4385–4396.