Математическая биология и биоинформатика. 2011. Т. 6. № 2. С. 211-227.
URL: http://www.matbio.org/2011/Panjukov2011(6_211).pdf

========== **БИОИНФОРМАТИКА** ==========

УДК: (577.214.625+004.93):519.688

Пакет программ aSHAPE для изучения пространственной конформации участков бактериального генома

Панюков В.В.*[1], Назипова Н.Н.[1], Озолинь О.Н.[2]

[1]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия*
[2]*Институт биофизики клетки, Российская академия наук, Пущино, Московская область, 142290, Россия*

Аннотация. Описывается интерактивный пакет aSHAPE для выявления пространственного конформационного различия и сходства в семействах участков ДНК на основе трёхмерной модели двойной спирали. Описано его применение для распознавания промоторных областей в геноме *Escherichia coli*.

Ключевые слова: *конформационная цепочка, деформация, дискриминантный анализ, разделяющая функция, двойная спираль, промотор, ДНК.*

ВВЕДЕНИЕ

Способность геномной ДНК быть матрицей для синтеза РНК в значительной степени определяется структурно-конформационным состоянием её регуляторных участков (промоторов) [1-8]. Первостепенную роль в создании оптимальной пространственной конфигурации промоторной ДНК играют разнообразные белковые факторы, которые в зависимости от потребности способствуют или, наоборот, препятствуют формированию транскрипционного комплекса.

Важно, однако, что способность подвергаться адаптивным конформационным перестройкам заложена в нуклеотидной последовательности промоторов в виде оптимального распределения структурообразующих гомонуклеотидных треков и гибких кинк-образующих динуклеотидов [9-11]. Это значит, что трехмерная структура свободной промоторной ДНК может иметь особую пространственную конфигурацию. О способности типичных для промоторов коротких фрагментов нуклеотидной последовательности формировать стабильные изгибы свидетельствуют данные, полученные методами рентгеноструктурного анализа [12] и ядерного магнитного резонанса (ЯМР) [13]. Однако такого типа анализ пока невозможен для полноразмерных промоторов. Поэтому в данной работе предлагается исследовать архитектурные особенности промоторов, используя трёхмерную структуру протяженных участков ДНК, вычисляемых *in silico* [14]. Интерактивный пакет aSHAPE разрабатывался как инструмент для реализации этого подхода.

Пусть имеется два семейства участков ДНК, где каждый участок определён пространственными координатами входящих в его состав атомов. Требуется найти такие параметры, которые позволяют разделить данные семейства в конформационном пространстве.

* panjukov@itaec.ru

Эта задача является задачей дискриминантного анализа, понимаемого в широком смысле. Следуя этой идеологии, мы отображаем пространство входных данных (если мы изучаем участки ДНК, состоящие из N атомов, то размерность пространства входных данных равна 3N) в одномерное дискриминантное пространство, поскольку размерность последнего должна быть на единицу меньше количества разделяемых семейств. Отображение в дискриминантное пространство осуществляется дискриминантной функцией. В классическом дискриминантном анализе, математический аппарат которого хорошо разработан, отображение в дискриминантное пространство линейно [15,16]. Для разделения семейства участков ДНК мы используем нелинейные дискриминантные функции конформационных параметров.

Категоризация конформационных параметров для протяженных пространственных систем, каковыми являются участки ДНК, является нетривиальной задачей. С методами изучения структурных особенностей ДНК можно познакомиться на сайте [17].

В нашей постановке задачи удобно представлять молекулу ДНК в виде эластичного жгута [14]. В этом случае модели Уотсона-Крика соответствует прямой цилиндрический жгут, а в координатном пространстве - прямой цилиндр. Это идеальная конформация классической B-формы ДНК без деформаций. Любая линия в этом цилиндре, проведенная через одинаковые группы азотистых оснований (назовём её конформационной линией), будет отражать пространственную конфигурацию моделируемого участка ДНК, что дает возможность оценивать произвольную конформацию жгута по изменению формы этой линии. Вследствие дискретности системы конформационная линия, проведенная через цепочку последовательных точек, является ломаной конформационной цепочкой. Поэтому решение задачи конформационного описания молекулы включает в себя два этапа: выбор репрезентативной цепочки последовательных точек и численная оценка её формы. В настоящее время не существует идеальных рецептов по реализации как первого, так и второго этапов.

Известен ряд программ и серверов [18-25], которые, на основе описанного выше подхода, вычисляют глобальную конформацию заданного участка ДНК. Как правило, в качестве реперных точек (вершин) конформационной цепочки используются центры пар оснований молекулы, которые в B-форме ДНК лежат на оси спирали.

Поскольку центры пар оснований можно определять разными способами и с разной степенью точности, то программы, оценивающие конформацию оси двойной спирали, могут давать разные результаты, о чём сообщают Barbic и Crothers [26], сравнив популярные программы 3DNA [20], Curves 5.1 [21], Freehelix98 [24]. Добавим сюда пример программы CURVATURE [23], которая аппроксимирует изогнутую ось спирали непрерывной плоской кривой и меряет деформацию радиусом кривизны этой кривой, в то время как в программе ADN-Viewer [27] в качестве конформационного параметра предлагается вычислять углы, которые образованы соответствующими сторонами треугольника, формируемого тремя последовательными реперными точками.

Методологической особенностью пакета aSHAPE является возможность исследовать конформационные свойства ДНК с использованием нескольких конформационных цепочек.

Каждая вершина базовой конформационной цепочки в пакете aSHAPE ассоциирована с фосфорной парой, образованной атомами фосфора комплементарных пар оснований, находящихся на разных нитях ДНК.

Из базовой цепочки для непосредственных вычислений конформации фрагментов ДНК строятся две рабочих цепочки. В одной цепочке вершинами являются центры

фосфорных пар, а в другой – спиральные центры, ассоциированные с двумя соседними фосфорными парами.

ОПИСАНИЕ ПАКЕТА aSHAPE

Среда отладки пакета

Отладка и разработка инструментария пакета производилась с использований двух семейств (*Prm* и *Cont*) участков ДНК одинаковой длины (300 пар оснований).

В семействе *Prm* собраны последовательности 180 бактериальных промоторов, распознаваемых σ^{70}–субъединицей РНК-полимеразы *E. coli*. Эти промоторы были использованы ранее [11] и характеризуются отсутствием рядом с ними других перекрывающихся промоторов, что допускает позиционное выравнивание их структурных моделей. Последовательности семейства *Prm* выбраны таким образом, что истинная точка старта транскрипции фланкирована участками длиной 150 н.п. слева и 149 н.п. справа.

В качестве контрольного набора *Cont* был использован набор из 51 фрагментов, произвольно взятых из непромоторных областей генома *E. coli*.

Структурное моделирование участков *Prm* и *Cont* было выполнено с помощью интернет-ресурса DNA Tools [14], вычисляющего координаты всех атомов ДНК-дуплекса заданной последовательности и представляющего результат в формате базы данных PDB.

Совокупность полученных при моделировании трёхмерных координат атомов заданной последовательности однозначно определяла геометрию, или конформацию, соответствующего участка ДНК.

Обозначения. Соглашения

Используем обычную иерархию участков геномной последовательности разного относительного размера: геномная последовательность ⊃ участок ⊃ фрагмент ⊃ символ или нуклеотид или основание.

Два семейства из участков ДНК одинакового размера в 300 пар нуклеотидов (п.н.), *Fm*1 и *Fm*2, подаются на вход пакета. Каждый участок позиционирован в –150, то есть считается, что первое основание участка (слева направо) расположено на 150 позиций левее точки старта транскрипции, а последнее на 149 позиций правее. Координаты нуклеотидов участков *Fm*1 и *Fm*2 лежат в отрезке [-150; +149]. Применительно к семейству *Prm* такое позиционирование помещает стартовую точку транскрипции в нулевую позицию. По умолчанию изучаемые участки совпадают с *Fm*1 и *Fm*2, однако, манипулируя позиционированием и размером, можно для изучения выделить любую другую область в участках.

Изучаемый участок покрывается короткими фрагментами молекулы. Пакет aSHAPE изучает конформацию участка посредством вычисления деформации всех входящих сюда фрагментов длины *FragSz*. В участке размера *ArSz* находится $FragNu = ArSz - FragSz + 1$ фрагментов размера *FragSz*, если начало следующего смещено на 1 позицию относительно начала предыдущего участка.

Деформация фрагмента вычисляется функцией деформации. Пусть f некоторая функция деформации фрагментов заданного размера *FragSz*. Применение f к семейству участков даёт ансамбль деформаций, размер ансамбля равен $FragNu \times ArNu$, где *ArNu* количество участков в семействе. Алгоритм вычисления функции f показан на рис. 1.

Для функции деформации вычисляем три гистограммы. Для этого разбиваем размах данной функции на интервалы размера Δ. Обозначим $S_\Delta(x)$ множество

значений функции попавших в интервал разбиения $[x-\Delta/2, x+\Delta/2)$, где x - точка разбиения.

Стандартная гистограмма hv распределения фрагментов в семействе по значению деформации определяется как $hv(x)=|S_\Delta(x)|/(FragNu \times ArNu)$.

Гистограмма плотности hd показывает среднее количество фрагментов с данной деформацией приходящихся на одну зону, $hd(x)=|S_\Delta(x)|/ArNu$.

Гистограмма рейтинга $hr(x)=100\times$ (количество тех зон семейства, в которых имеется хотя бы один фрагмент с деформацией $S_\Delta(x)$)$/ArNu$).

Гистограмма значений деформации hv – рабочая, другие две – вспомогательные.

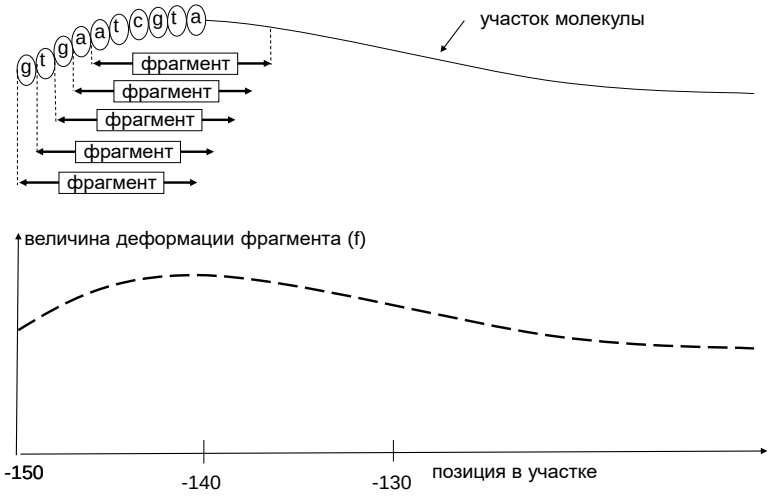


Рис. 1. Вычисление функции деформации f для одного участка, позиционированного в –150.

По заданной функции вычисляются ансамбли деформации $Ens1$ и $Ens2$ изучаемых семейств $Fm1$ и $Fm2$, соответственно, и строятся две гистограммы значений деформации $hv(Fm1)$ и $hv(Fm2)$. Указанные гистограммы по определению расходятся, если они имеют такое взаимное расположение, как показано на рис. 2.

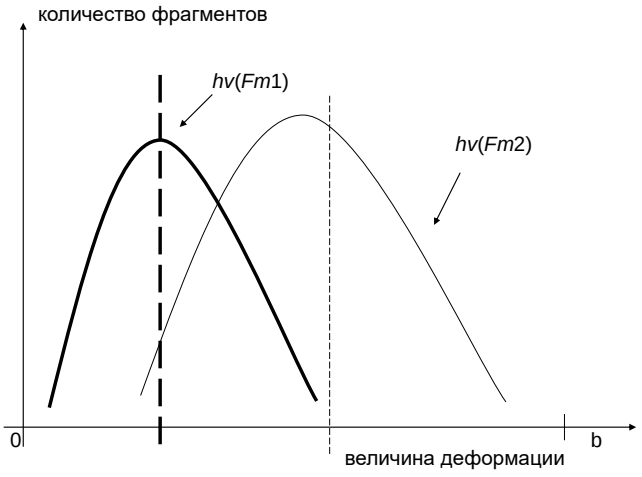


Рис. 2. Формальный пример расходящихся гистограмм для семейств $Fm1$ и $Fm2$. Пунктирные вертикали проведены в точках средних значений величин деформации.

Взаимное расположение гистограмм на рис. 2 отобразим в виде формального неравенства $hv(Fm1) < hv(\mathrm{Fm2})$; такая запись отражает как наличие расхождения гистограмм (знак неравенства), так и порядок их следования (слева направо).

Пусть $\sigma(Ens)$ обозначает величину выборочной дисперсии некоторого ансамбля *Ens*, тогда из рис. 2 вытекает, что $\sigma(Ens1 \cup Ens2) > max\{\sigma(Ens1), \sigma(Ens2)\}$, а это, согласно дискриминантному анализу означает, что ансамбли *Ens*1 и *Ens*2 разделены. Таким образом, расхождение гистограмм характеризует разделение соответствующих ансамблей, а значит, и разделение семейств *Fm*1 и *Fm*2.

Выбор пространственной модели молекулы ДНК

Наиболее адекватно пространственная конфигурация ДНК моделируется с использованием трехмерных структур, полученных для протяженных фрагментов методом рентгеноструктурного анализа или ЯМР, но такие данные пока недоступны для всех участков промоторной ДНК. Поэтому был использован интернет-ресурс DNA tools [14], моделирующий структуру двойной спирали ДНК длиной до 700 нуклеотидных пар с использованием дистанционных и угловых параметров динуклеотидов, определенных экспериментально.

Ресурс DNA tools [14] предоставляет пользователю координаты всех атомов участка молекулы. Это позволяет, кроме традиционной цепочки центров оснований, строить произвольные конформационные цепочки заданного участка, без учета стандартных конформационных угловых параметров *tilt*, *roll*, *twist* [28]. Так как указанные угловые параметры зависимы между собой [29], то их учёт приводит к размытию (уширению) гистограмм фрагментных функций деформации, что крайне нежелательно при использовании метода расходящихся гистограмм.

Конформационные цепочки

В пакете aSHAPE конформация участка молекулы ДНК оценивается величиной деформации цепочки фосфорных пар. В свою очередь деформация цепочки фосфорных пар оценивается посредством двух производных цепочек: цепочки центров фосфорных пар (далее фосфорных центров) и цепочки спиральных центров.

Цепочка фосфорных центров. Если $\boldsymbol{v}_i, \boldsymbol{w}_i$ координатные вектора фосфоров *i*-той пары нуклеотидов, то центр пары $(\boldsymbol{v}_i + \boldsymbol{w}_i)/2$, по определению, является вершиной в цепочке фосфорных центров. Рис. 3 демонстрирует алгоритм построения цепочки фосфорных центров.

Цепочка спиральных центров. Рассмотрим фрагмент *F* некоторой цепочки фосфорных пар, состоящий из двух соседних пар. Пусть $\boldsymbol{v}_i, \boldsymbol{w}_i$ и $\boldsymbol{v}_{i+1}, \boldsymbol{w}_{i+1}$ – координатные вектора двух последовательных пар фосфоров фрагмента *F*. Скажем, что *F* формирует двойную спираль, если его атомы фосфора лежат на некоторой двойной спирали с осью *A* так, что пару $\boldsymbol{v}_i, \boldsymbol{w}_i$ можно совместить с $\boldsymbol{v}_{i+1}, \boldsymbol{w}_{i+1}$ посредством сдвига вдоль *A* и вращения вокруг оси *A*. Легко проверить, что фрагмент *F* формирует двойную спираль, если $|\boldsymbol{v}_i - \boldsymbol{w}_i| = |\boldsymbol{v}_{i+1} - \boldsymbol{w}_{i+1}|$ и $|\boldsymbol{v}_i - \boldsymbol{v}_{i+1}| = |\boldsymbol{w}_i - \boldsymbol{w}_{i+1}|$, где символ |.| означает длину вектора.

Если *F* формирует двойную спираль, то плоскость, проходящая через пару фосфоров $\boldsymbol{v}_i, \boldsymbol{w}_i$ перпендикулярно оси спирали, пересекает ось в точке, которую мы в данной работе называем спиральным центром пары $\boldsymbol{v}_i, \boldsymbol{w}_i$.

Для принятой нами пространственной модели ДНК [14], расчёты для семейства *Prm* и *Cont* показали, что:

- среднее расстояние между атомами фосфора в фосфорной паре ≈ 17.8Å, вариация расстояния ≈ 0.14Å;

– среднее расстояние между атомами фосфора, принадлежащими соседним фосфорным парам нити, равно ≈6.5Å, вариация расстояния ≈ 0.18Å.

То есть условия локальной спирализации в семействах *Prm* и *Cont* выполняются с достаточной точностью, что допускает построение цепочек спиральных центров.

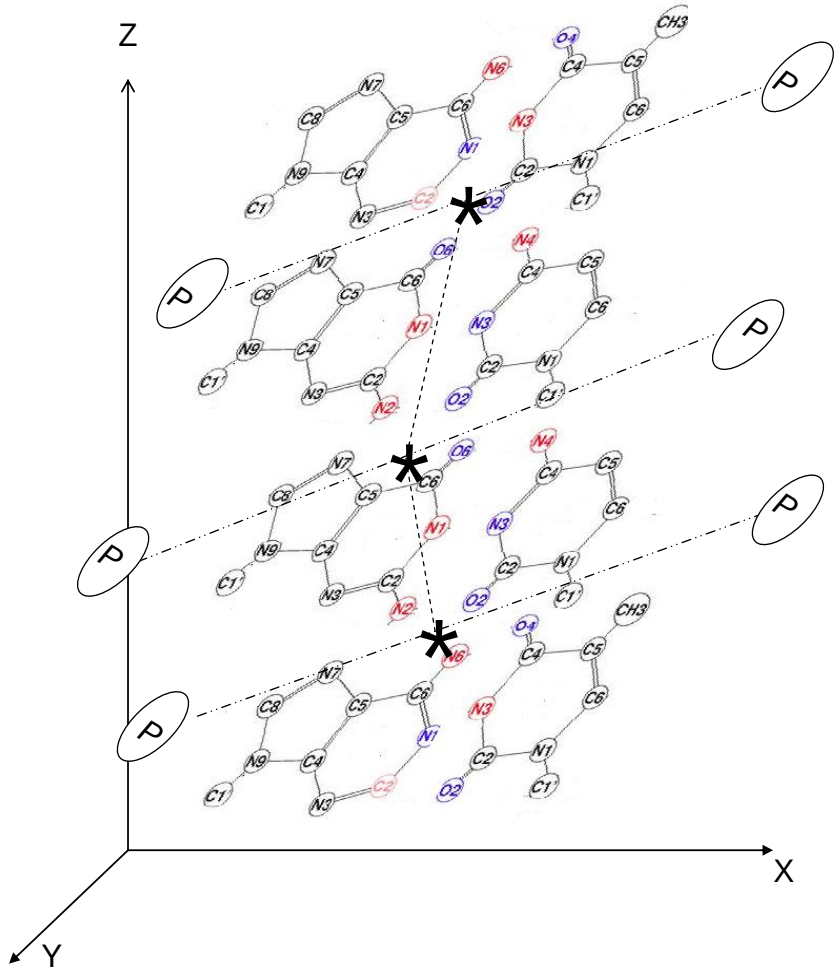


Рис. 3. Схема фрагмента ДНК размером 4 п.н. Атомы фосфора (большие овалы) образуют три фосфорные пары (соединены пунктирными линиями). Центры тяжести пар (звёздочки) дают три вершины цепочки фосфорных центров. Рисунки комплементарных оснований взяты из [30] и адаптированы.

Локальная деформация цепочек

В описываемом пакете деформация фрагментов молекулы исследуется посредством интегральной деформации соответствующих цепочек, которая, в свою очередь, вычисляется на основе локальных деформаций. Дадим определения локальных деформаций произвольной цепочки вершин $S = \ldots \boldsymbol{v}_i, \boldsymbol{v}_{i+1}, \boldsymbol{v}_{i+2}, \boldsymbol{v}_{i+3} \ldots$, где $\boldsymbol{v}_i$.- координатный вектор i-той вершины. Вектор $\boldsymbol{v}_{i+1} - \boldsymbol{v}_i$ будем называть звеном цепочки в вершине $\boldsymbol{v}_i$.

Определим алгоритм вычисления локальных деформаций. Будем считать, что цепочка является аппроксимацией некоторой дифференцируемой кривой. Поместим начало локальной системы координат Эйлера (θ, ϕ, Z) в вершину $\boldsymbol{v}_i$ цепочки S, направив ось Z вдоль звена в этой вершине. Очевидно, что в этой системе координат тензор деформации цепочки диагонален и определяет соответственно локальный изгиб, кручение и деформацию звена цепочки.

Определим изгиб и кручение цепочки как приближение к кривизне и кручению кривой линии (см. рис. 4). Изгиб θ в вершине $\boldsymbol{v}_i$ определяется тремя соседними

вершинами цепи, $\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$, $\boldsymbol{v}_{i+2}$. По определению θ равен углу между звеньями $\boldsymbol{v}_{i+1}-\boldsymbol{v}_i$ и $\boldsymbol{v}_{i+2}-\boldsymbol{v}_{i+1}$.

Кручение ϕ в вершине $\boldsymbol{v}_i$ определяется четырьмя соседними вершинами цепи, $\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$, $\boldsymbol{v}_{i+2}$, $\boldsymbol{v}_{i+3}$. По определению ϕ равно углу между плоскостью, проходящей через $\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$, $\boldsymbol{v}_{i+2}$ и звеном $\boldsymbol{v}_{i+3}-\boldsymbol{v}_{i+2}$.

Деформация звена при вершине $\boldsymbol{v}_i$ определяется тремя соседними вершинами цепи $\boldsymbol{v}_i$, $\boldsymbol{v}_{i+1}$, $\boldsymbol{v}_{i+2}$. По определению деформация звена при вершине $\boldsymbol{v}_i$ равна разности длин соседних звеньев: $\|\boldsymbol{v}_{i+1}-\boldsymbol{v}_i|-|\boldsymbol{v}_{i+2}-\boldsymbol{v}_{i+1}\|$.

Совокупность вершин цепочки, определяющих ту или иную локальную деформацию, будем называть блоком. Таким образом, локальная деформация цепочки есть деформация некоторого её блока.

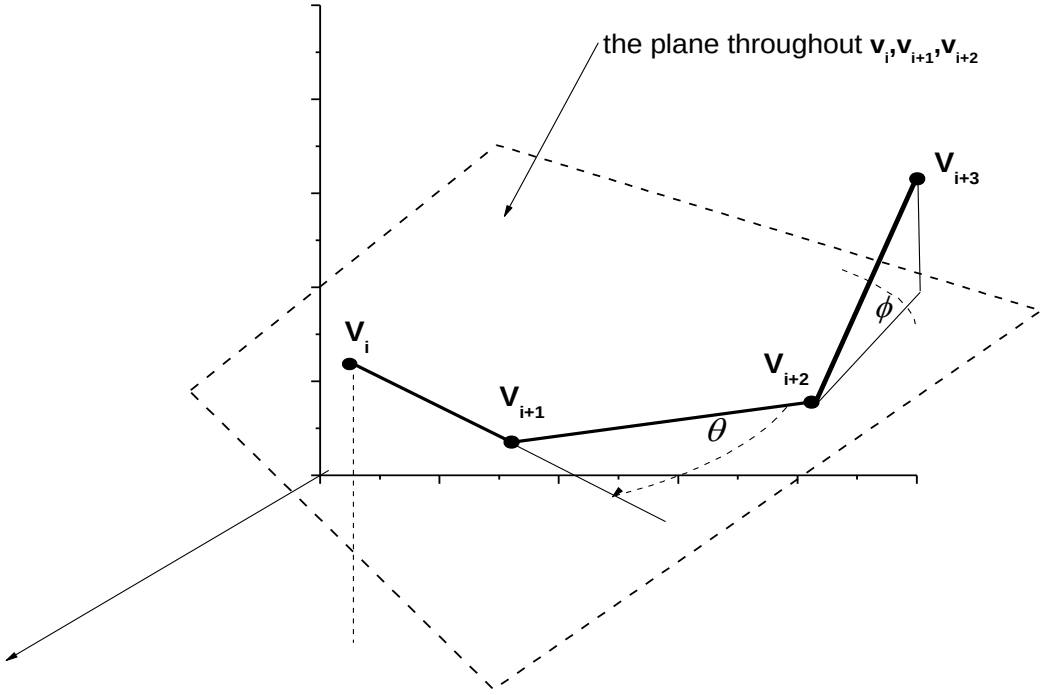


Рис. 4. Схема, отображающая локальные деформации блока цепочки: изгиб θ и кручение ϕ.

Пакет aSHAPE предоставляет возможность визуально оценить наличие или отсутствие зависимости между локальными деформациями. Исследование показало попарную независимость введённых выше локальных деформаций на семействах *Prm* и *Cont* (Рис. 5).

Интегральная деформация и разделяющие функции

Конформация фрагмента в целом оценивается посредством вычисления интегральной деформации соответствующих цепочек. В настоящее время в пакете используется два типа интегральных деформаций, деформация формы и деформация гомогенности.

Дадим принятое нами определение функции, разделяющей семейства. Функция деформации разделяет семейства $Fm1$ и $Fm2$, если гистограммы $hv(Fm1)$ и $hv(Fm2)$, построенные при некотором размере фрагментов для цепочек фосфорных и спиральных центров, удовлетворяют одному из неравенств $hv(Fm2) < hv(Fm1)$ или $hv(Fm1) < hv(Fm2)$.

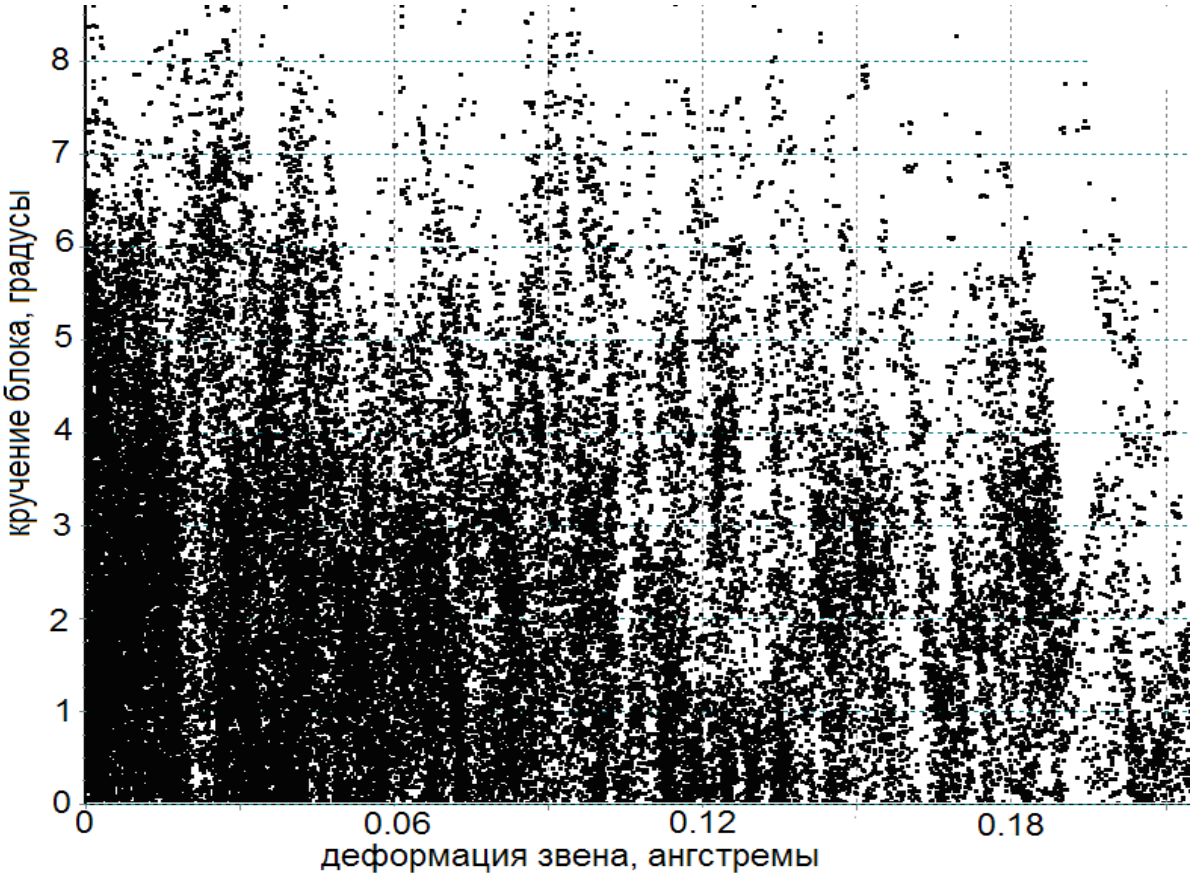


Рис. 5. Диаграмма совместного распределения деформации звена и кручения ϕ блоков семейства *Prm*.

Если функция деформации остаётся разделяющей при варьировании размера фрагмента, то для изучения отношения между *Fm*1 и *Fm*2 мы отбираем гистограммы с возможно меньшей площадью их перекрытия *Sq*(*Fm*1∩*Fm*2).

Деформация гомогенности

Этот вид деформации показывает, насколько велика вариация формы блоков фрагмента. Мы считаем фрагмент гомогенным относительно данной локальной деформации, если деформация всех его блоков приблизительно одинакова.

Приведём четыре примера деформаций гомогенности из пакета, *AvrBendDeform*, *BendDeform*, *ContortDeform* и *LinkDeform*.

Каждому фрагменту ДНК отвечает некоторая цепочка S фосфорных или спиральных центров Пусть $\bar{\theta}$ и $\bar{\varphi}$ обозначают среднее значение локального изгиба и кручения цепочки S, соответственно. Тогда

$$AvrBendDeform(S) = \bar{\theta}, \quad BendDeform(S) = \sum_i \left|\bar{\theta} - \theta_i\right|,$$

$$ContortDeform(S) = \sum_i \left|\bar{\varphi} - \varphi_i\right|, \quad LinkDeform(S) = \sum_i \left|l_i\right|,$$

где θ_i, φ_i, l_i обозначают соответственно локальные деформации изгиба, кручения и звена в i-той вершине цепочки.

Вычисления показали, что деформация *BendDeform* разделяет семейств *Prm* и *Cont* т.к. соотношение $hv(Cont) < hv(Prm)$ сохраняется при вариации вида конформационной цепочки (Рис. 6).

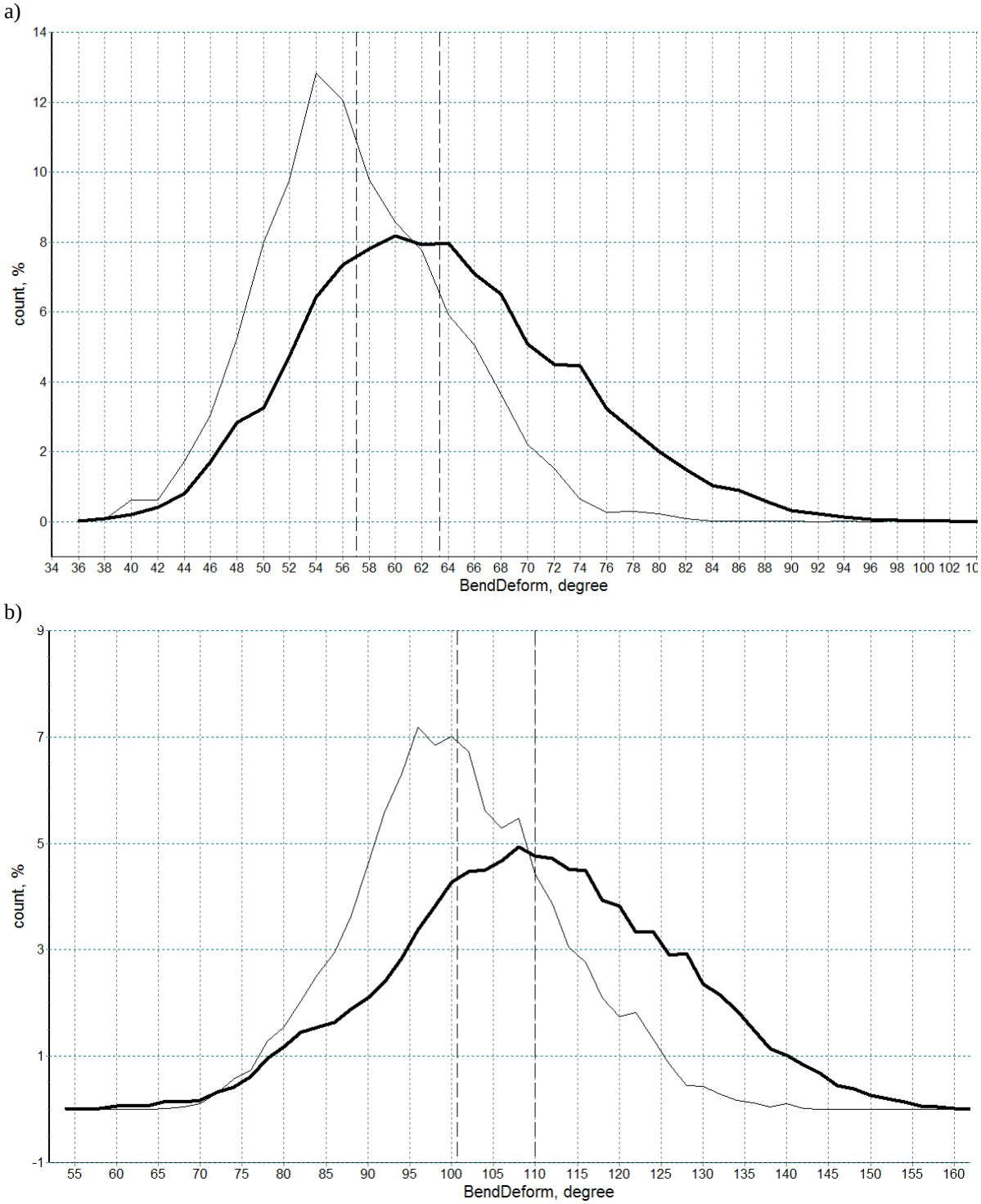
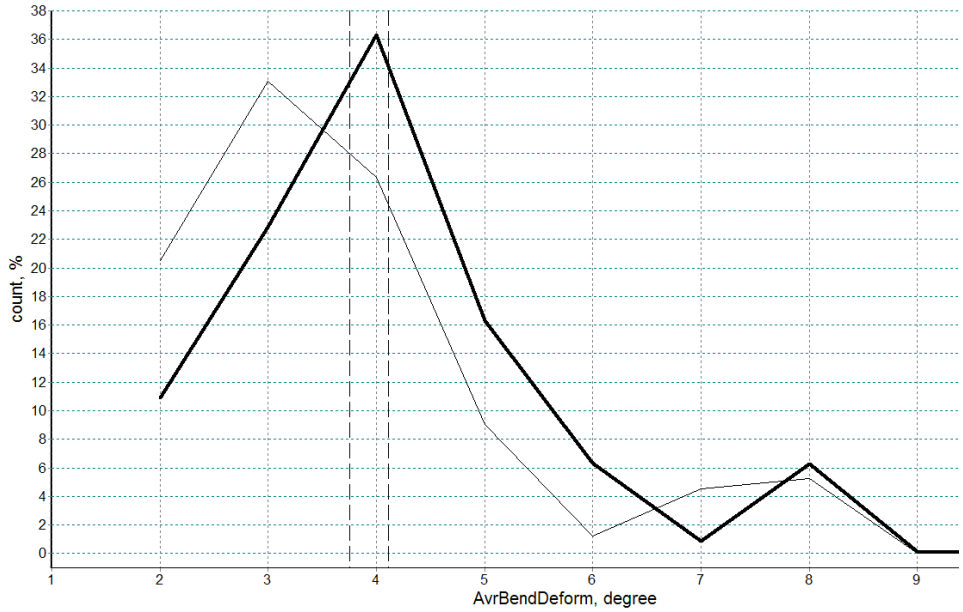


Рис. 6. Гистограммы деформации гомогенности *hv*(*Prm*) (жирная линия) и *hv*(*Cont*) (тонкая линия). Пунктирные вертикали указывают среднюю величину деформации фрагментов. Изучаемая зона совпадает с участком. Размер фрагмента 60 п.н., a) - цепочка фосфорных центров, *Sq*(*Prm*∩*Cont*)=0.69, b) - цепочка спиральных центров, *Sq*(*Prm*∩*Cont*)=0.72.

Аналогичный результат имеет место и для кручения *ContortDeform*, в то время как *LinkDeform* практически не разделяет семейства *Prm* и *Cont*, так как величина перекрытия гистограмм *Sq*(*Prm*∩*Cont*) для данной деформации остаётся близкой к единице при вариации как типа цепочки так и размера фрагмента.

Что касается деформации *AvrBendDeform*, то она, определенно, не разделяет *Prm* и *Cont* (Рис. 7).

a)



b)

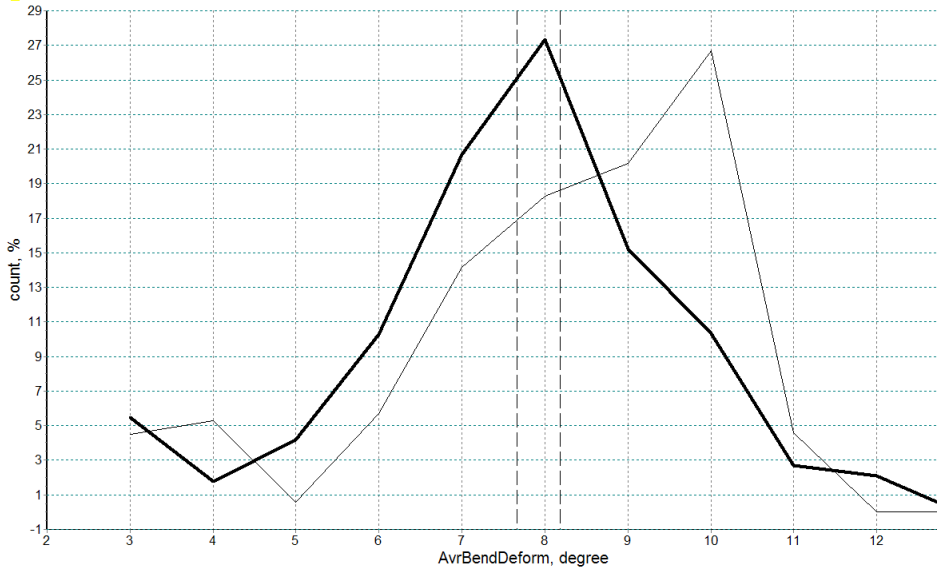


Рис. 7. Гистограммы *hv*(*Prm*) (жирная линия) и *hv*(*Cont*) (тонкая линия) функции *AvrBendDeform*, *FragSz* = 60 п.н.; a) – цепочка фосфорных центров, *hv*(*Cont*) < *hv*(*Prm*), *Sq*(*Prm*∩*Cont*) = 0.65; b) – цепочка спиральных центров, *hv*(*Prm*) < *hv*(*Cont*), *Sq*(*Prm*∩*Cont*) = 0.70.

Деформация формы фрагментов

Этот вид деформации показывает, насколько велико отклонение фрагмента от B-формы. Приведём четыре примера деформаций формы из пакета, *LengthDeform*, *MaxJut* и *RadiusRange*.

Пусть S - цепочка фосфорных или спиральных центров с вершинами $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ и $dist(\boldsymbol{v}_i)$ обозначает расстояние вершины $\boldsymbol{v}_i$ до прямой, тогда

$$LengthDeform(S) = \sum_i \left(\left|\boldsymbol{v}_{i+1} - \boldsymbol{v}_i\right|\right) - \left|\boldsymbol{v}_n - \boldsymbol{v}_1\right| \text{ и } MaxJut(S) = max\{dist(\boldsymbol{v}_i) \mid i=1, \ldots, n\}.$$

В классической B-форме ДНК атомы фосфоров, образуя двойную спираль, лежат на поверхности прямого цилиндра. Аппроксимируем деформированную двойную цепочку фосфоров, обозначаемую *SPP*, прямым цилиндром методом главных компонент [31].

Совокупность координат фосфоров фрагмента даёт трёхмерную корреляционную матрицу *M*, чьи собственные числа являются корнями многочлена третьей степени и поэтому вычисляются по элементарным формулам. Направляющим единичным вектором оси цилиндра, обозначим его $\boldsymbol{h}$, назначается собственный вектор матрицы *M* с наибольшим собственным числом. Пусть $\boldsymbol{w}_1,\boldsymbol{w}_2, \ldots,\boldsymbol{w}_m$ вектора фосфоров двойной цепочки *SPP* в системе координат, начало которой находится в центре тяжести цепочки, а оси направлены по собственным векторам матрицы *M*, в частности, ось *Z* направлена вдоль вектора $\boldsymbol{h}$, то есть, совпадает с осью аппроксимирующего цилиндра. Расстояние *i*-того фосфора до оси *Z* равно $r_i = |\boldsymbol{w}_i - (\boldsymbol{w}_i,\boldsymbol{h})\boldsymbol{h}|$. С цилиндром ассоциируются следующие радиусы:

$$R_{min}(SPP)=min\{r_1,\ldots,r_m\},\ R_{max}(SPP)=max\{r_1,\ldots,r_m\},\ \text{и}\ R_{avr}(SPP)=(r_1+\ldots+r_m)/m,$$

где $m = 2FragSz$. Положим $RadiusRange(SPP) = R_{max}(SPP) - R_{min}(SPP)$.

Очевидно, что для B-формы ДНК выполнено $LengthDeform(S) = MaxJut(S) = 0$ и $RadiusRange(SPP) = 0$.

Отметим, что деформация формы фрагмента есть совместный результат его локальных изгибов, кручений и деформаций звеньев. Использование обоих видов деформаций - гомогенности и формы - увеличивает надёжность результатов по разделению семейств. Так деформация двойной цепочки фосфоров *RadiusRange* (Рис. 8) находится в полном соответствии с деформациями гомогенности *ContortDeform* и *BendDeform* производных цепочек (Рис. 6).

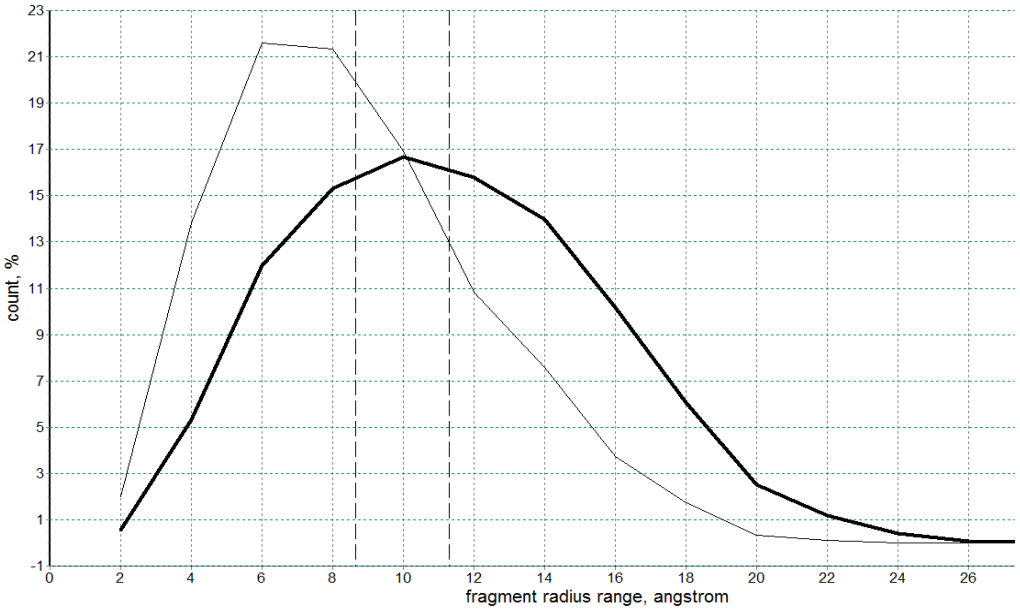


Рис. 8. Деформации формы двойных цепочек фосфоров *RadiusRange* семейств *Prm* (жирная линия) и *Cont* (тонкая линия). Размер фрагмента 60 п.н., $Sq(Prm \cap Cont) = 0.73$.

Дополнительный инструментарий и его применение

После того как обнаружены функции, разделяющие изучаемые семейства, возникает потребность в получении дополнительной информации о семействах. В пакете aSHAPE имеется для этого ряд инструментов.

Пусть f некоторая функция деформации, Fm – семейство участков ДНК и Ens – ансамбль значений функции f на фрагментах заданного размера семейства Fm. Пакет имеет следующие возможности.

- Реализованы команды *CutTop* и *CutBtm*, которые выделяют в ансамбле Ens подансамбль Ens', удовлетворяющий неравенству $CutBtm \le val \le CutTop$, где val – значения деформации.
- Визуально отображается репрезентативность выделяемого подансамбля Ens' в виде графика, показывающего количество фрагментов с деформацией $val \in Ens'$ в каждой зоне семейства Fm.
- Предоставляется информация об окнах в участках, свободных от фрагментов с деформацией $val \in Ens'$. Для данного участка $T \in Fm$ через $FreeWin(T)$ обозначим множество окон размера $FragSz$ свободных от фрагментов с деформацией $val \in Ens'$, то есть, если $w \in FreeWin(T)$, то для любого фрагмента $frag$ с деформацией $val \in Ens'$ выполнено $w \cap frag = \varnothing$. Результат позиционного выравнивания всех свободных окон $FreeWin(Fm) = \bigcup_{T \in Fm} FreeWin(T)$ выводится в таблицу.
- Осуществляется выравнивание фрагментов по значениям деформации f. Точность позиционирования определяется окном Win размером $WinSz$. Пробегая все позиции p_w в окне Win каждого i-того участка семейства, мы получаем совокупность значений деформации, которую удобно представлять в виде матрицы $M = \|f_i(p_w)\|$, где i нумерует строки. Элементы матрицы M, по одному из каждой строки, образуют выборку из матрицы M; понятно, что имеется $WinSz^{ArNu}$ выборок, где $ArNu$ - количество участков в семействе Fm. Среди этих выборок пакет за время пропорциональное $|M|\log_2|M|$ находит все те выборки, которые имеют минимальный размах. Результат выравнивания ансамбля деформаций Ens представлен всеми минимальными выборками, найденными в ходе перебора всех положений окна Win.
- Осуществляется аппроксимация конформационных цепочек прямым цилиндром методом главных компонент [31].
-

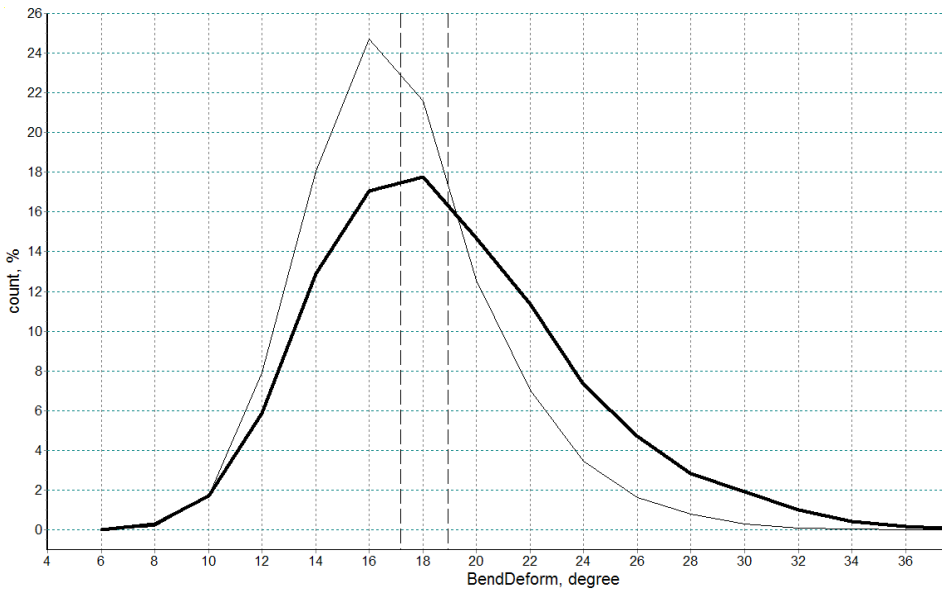


Рис. 9. Гистограммы деформации гомогенности цепочек фосфорных центров семейств *Prm* (жирная линия) и *Cont* (тонкая линия). Размер фрагмента 20 п.н., $Sq(Prm \cap Cont)=0.8$.

Применение дополнительного инструментария

Выше мы показали (Рис. 6), что деформация гомогенности *BendDeform*, разделяет семейства участков *Prm* и *Cont*. Исследуем деформацию фрагментов размера 20 п.н. в семействе *Prm*.

Согласно гистограммам рис. 9, семейство *Prm* в большей степени, нежели *Cont*, использует негомогенные фрагменты. Зададимся порогом негомогенности и выясним частоту использования и местоположение негомогенных фрагментов в семейство *Prm*. Будем считать фрагмент негомогенным, если для соответствующей цепочки фосфорных центров *S* выполнено *BendDeform*(*S*) > 25° (Рис. 9)

Командой *CutBtm* выделим подансамбль *Ens′* негомогенных фрагментов и построим график их использования в *Prm*, который показывает широкий разброс негомогенных фрагментов по участкам, так пять участков семейства *Prm* имеет по два негомогенных фрагмента, один участок содержит 126 таких фрагментов и в 15 участках негомогенные фрагменты отсутствуют (Рис. 10).

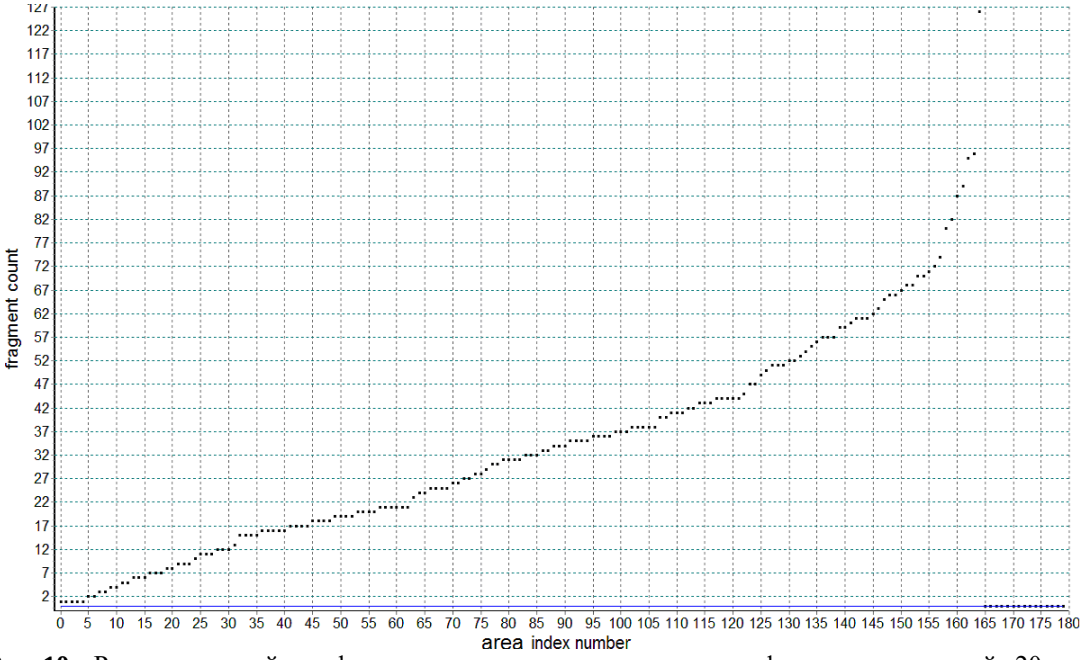


Рис. 10. Ранжированный график использования негомогенных фрагментов длиной 20 п.н. семейством *Prm*.

Командой *FreeWin* получим информацию о расположение негомогенных фрагментов. Вычисления показывают, что в области размером 55 п.н., позиционированной в –67, и в области размером 50, позиционированной в 0, в каждом участке семейства *Prm* имеется окно размера *FragSz* = 20 п.н., свободное от негомогенных фрагментов. Отсутствие негомогенных фрагментов, однако, не означает, что вблизи позиции -67 и вблизи стартовой точки транскрипции имеется какая-то структурная аномалия или особенность.

Для того чтобы узнать, имеются ли в окрестностях промоторов конформационно схожие фрагменты, выполним выравнивание фрагментов размера *FragSz* = 40 п.н. семейства *Prm* по величине деформации формы *MaxJut*. Имеем 180 участков размера 300 и по 261 фрагменту в участке. Ансамбль деформаций, которые достаточно сильно варьируют, представлен на Рис. 11.

Результат выравнивания фрагментов по величине деформации показан на Рис. 12. Размер окна выравнивания *WinSz* определяет допустимый сдвиг фрагментов при переходе от одной нуклеотидной последовательности к другой. Процедура выравнивания выполнялась последовательно для всех позиций окна выравнивания

WinSz в диапазоне [-150; +110]. Опытным путём обнаружено, что для получения хорошего выравнивания размер окна *WinSz* должен быть не меньше размера фрагмента, в данном случае *WinSz* ≥ 40 п.н.

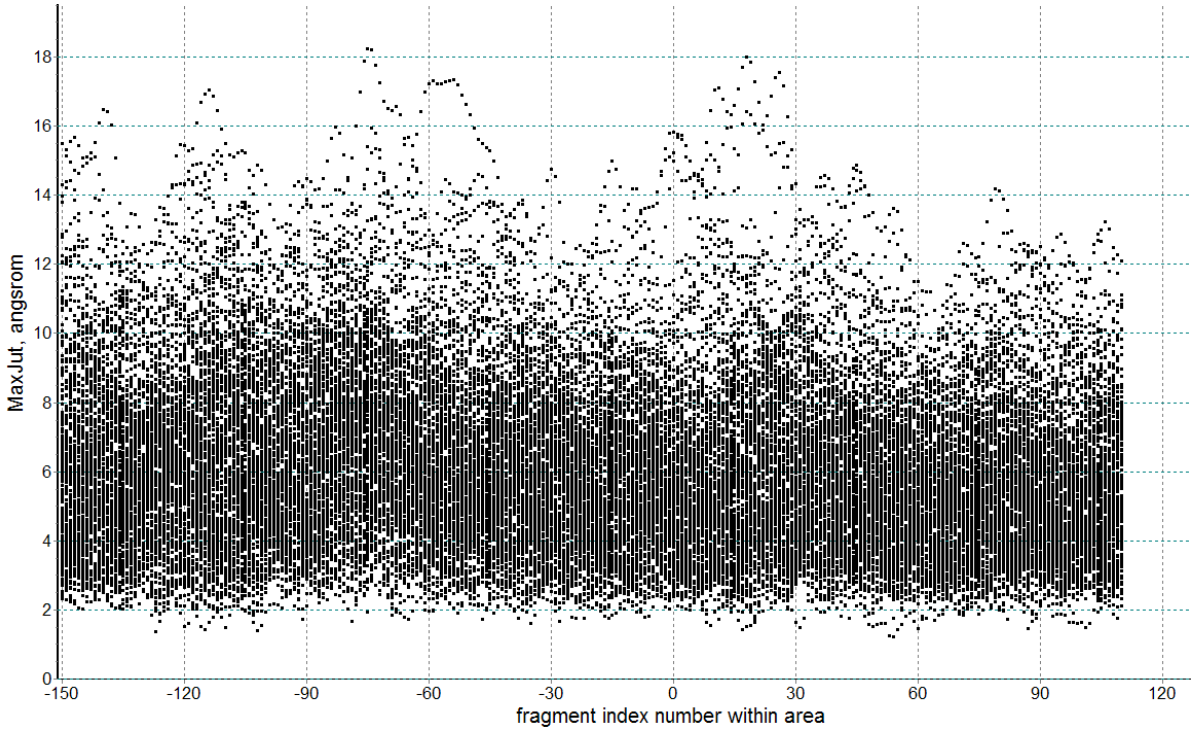


Рис. 11. Значения деформации формы *MaxJut* всех фрагментов 40 п.н. семейства *Prm*.

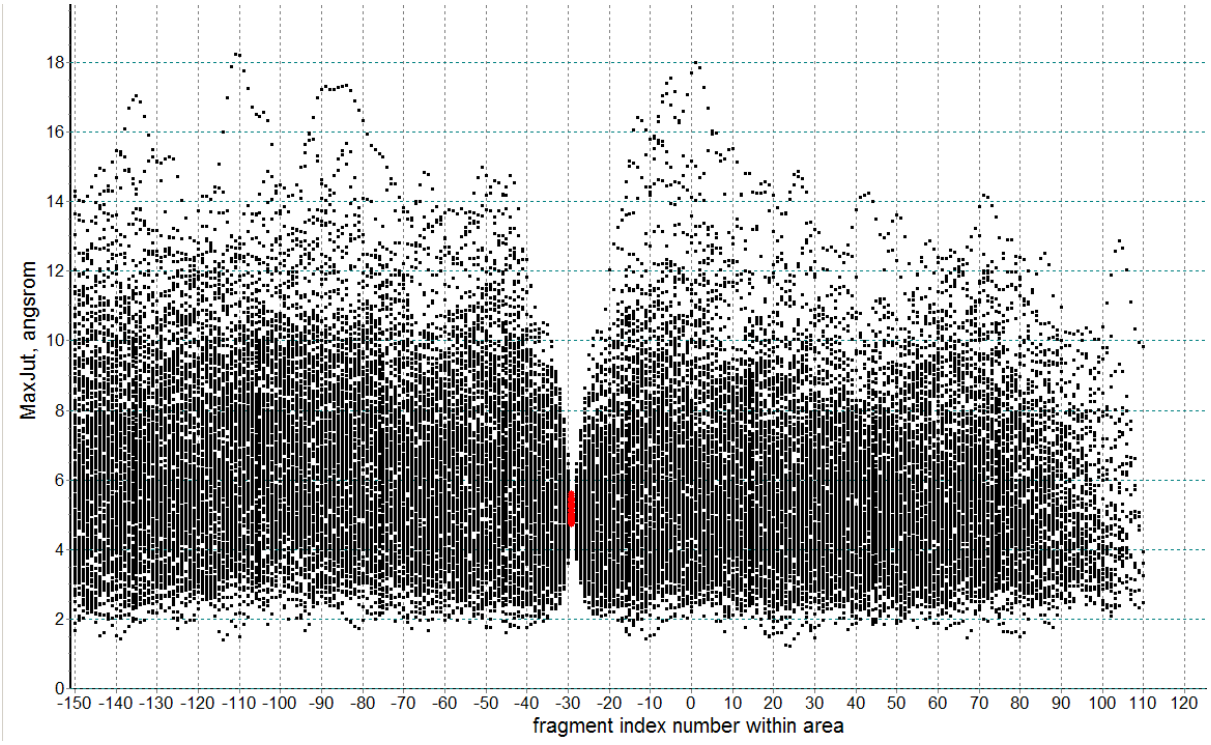


Рис. 12. Результат выравнивания деформации фрагментов 40 п.н. семейства *Prm*.

При *WinSz* = 40 п.н. наилучшее выравнивание получено в позиции –29 (Рис.12), которое говорит, что в области [-29; +11] каждого промотора семейства *Prm* имеется фрагмент с деформацией формы в интервале 4.72Å ≤ *MaxJut* ≤ 5.57Å. Таким образом,

вариация выравненных значений деформации составляет 0.85Å, тогда как размах ансамбля деформаций равен 17.04Å (Рис.11).

ЗАКЛЮЧЕНИЕ

Несмотря на многочисленные попытки построить компьютерную модель промотора, до сих пор эти усилия были направлены на лингвистический анализ нуклеотидных последовательностей (см [11] и ссылки, приведенные в этой публикации), либо на анализ таких физико-химических особенностей промоторов, как термодинамическая нестабильность [32], конформационная подвижность [33] или электростатические свойства [34]. В данной работе предлагается принципиально новый подход, предполагающий полноценное структурное моделирование нуклеотидных последовательностей с помощью сервера DNA tools [14] и исследование пространственных моделей с помощью разработанного пакета программ aSHAPE. Явным преимуществом нового подхода является возможность топологического анализа нуклеотидных последовательностей внутри одного семейства и возможность сравнительного анализа разных семейств.

Проведенный анализ выявил в промоторах два участка, свободных от негомогенных фрагментов. Это может отражать структурные ограничения, накладываемые транскрипционным комплексом на конформацию двойной спирали промоторной ДНК.

Один из этих участков находится в стартовой точке транскрипции, которая находится в непосредственном контакте с РНК-полимеразой, а её окружение играет важную роль при переходе транскрипционного комплекса к продуктивной инициации. Важно, однако, что гомогенность была зарегистрирована для фрагментов длиной 55 п.н., что предполагает участие более протяженной области промоторной ДНК в формировании транскрипционно-компетентных комплексов, чем область непосредственного контакта с ферментом (обычно до позиции +20).

Второй гомогенный участок начинается в позиции -67. Соответствующие фрагменты ДНК (длиной 55 п.н.) перекрываются не только с элементом -35, распознаваемым доменом 2.4 σ-субъединицы РНК-полимеразы, но и с UP-элементом промоторов, взаимодействующим с α-субъединицами фермента.

По-видимому, самым значимым результатом проведенного исследования является обнаруженная возможность выравнивания промоторов по структурной деформации. Наиболее эффективное выравнивание наблюдается вблизи позиции -29, которая находится в спейсере между элементами -35 и -10. Это означает, что в области контакта с РНК-полимеразой у всех промоторов имеется фрагмент длиной 40 п.н. с приблизительно одинаковым отклонением от стандартной B-формы ДНК ($4.72Å \le MaxJut \le 5.57Å$). Не исключено, что именно эта структурная аномалия является первичным сигналом, распознаваемым ферментом и позиционирующим его вблизи специфических модулей.

Методологической особенностью пакета aSHAPE является возможность исследовать конформационные свойства ДНК с использованием разных конформационных цепочек. Сфера его применения не ограничена теми функциями, которые уже имеются в пакете. Важно также, что созданное программное обеспечение может быть использовано для анализа любых функциональных модулей генома.

Работа была частично поддержана грантами РФФИ № 10-04-01218 и № 09-07-00455.

СПИСОК ЛИТЕРАТУРЫ

1. Coulombe B. and Zachary F. B. DNA Bending and Wrapping around RNA Polymerase: a "Revolutionary" Model Describing Transcriptional Mechanisms. *Microbiol. Mol. Biol. Rev*. 1999. V. 63. P. 457–478.
2. Carmona M., Claverie-Martin F., Magasanik B. DNA bending and the initiation of transcription at σ54-dependent bacterial promoters. *PNAS*. 1997. V. 94. P. 9568–9572.
3. Bolshoy A., Nevo E. Ecologic Genomics of DNA: Upstream Bending in Prokaryotic Promoters. *Genome Res*., 2000. V. 10. P. 1185–1193.
4. Hirvonen C.A., Ross W., Wozniak C.E., Marasco E., Anthony J.R., Aiyar S.A., Newburn V.H., Richard L. Gourse Contributions of UP Elements and the Transcription Factor FIS to Expression from the Seven rrn P1 Promoters in *Escherichia coli*. *J. Bacteriol*. 2001. V. 183. P. 6305–6314.
5. Kozobay-Avraham L., Hosid S., Bolshoy A. Involvement of DNA curvature in intergenic regions of prokaryotes. *Nucleic Acids Res.* V. 34. P. 2316–2327.
6. Shultzaberger R.K., Chen Z., Lewis K.A., Schneider T.D. Anatomy of *Escherichia coli* σ^{70} promoters. *Nucleic Acids Res*. 2007. V. 35. P. 771-788.
7. Pul U., Lux B., Wurm R., Wagner R. Effect of upstream curvature and transcription factors H-NS and LRP on the efficiency of *Escherichia coli* rRNA promoters P1 and P2 – a phasing analysis. *Microbiology*. 2008. V. 154. P. 2546-2558.
8. Meysman P., Dang T.H., Laukens K., De Smet R., Wu Y., Marchal K., Engelen K. Use of structural DNA properties for the prediction of transcription-factor binding sites in *Escherichia coli*. *Nucleic Acids Res*. 2011. V. 39. e6.
9. Ozoline O.N., Deev A.A., Trifonov E.N. DNA bendability – a novel feature in *E.coli* promoter recognition. *J. Biomol. Struct. Dynamics*. 1999. V. 16. P. 825–831.
10. Ozoline O.N., Masulis I.S., Buckin V.A. Deformable elements in promoter DNA as a basis for adaptive conformational transitions. *J. Biomol. Struct. Dynamics*. 2001. V. 18. № 6. P. 1002–1003.
11. Shavkunov K.S., Masulis I.S., Tutukina M.N., Deev A.A., Ozoline O.N. Gains and unexpected lessons from genome-scale promoter mapping. *Nucleic Acids Res*. 2009. V. 37. P. 4919–4931.
12. Shatzky-Schwartz M., Shakked Z., Luisi B.F. X-ray and solution studies of DNA oligomers and implications for the structural basis of A-tract-dependent curvature. *J. Mol. Biol*. 1997. V. 267. P. 595–623.
13. McAteer K., Aceves-Gaona A., Michalczyk R., Buchko G.W., Isern N.G., Silks L.A., Miller J.H., Kennedy M.A. Compensating bends in a 16-base-pair DNA oligomer containing a T(3)A(3) segment: A NMR study of global DNA curvature. *Biopolymers*. 2004. V. 75. P. 497–511.
14. Vlahovicek K, Kajan L, Pongor S. DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Res*. 2003. V. 31. № 13. P. 3686–3687. URL: http://hydra.icgeb.trieste.it/dna/index.php (дата обращения:01.08.2011)
15. *Факторный, дискриминантный и кластерный анализ*. Под ред. И.С. Енюкова. М.: Финансы и статистика, 1989.
16. Большаков А.А., Каримов Р.Н. *Методы обработки многомерных данных и временных рядов*. М.: Горячая Линия - Телеком, 2007.
17. Федосеева В.Б. *ДНК: изгибы двойной спирали: структура и функции*. URL: http://medbiol.ru/medbiol/dna_bend/00000a6c.htm#00013419.htm (дата обращения 01.08.2011).
18. Zheng G, Xiang-Jun Lu, Wilma K. Olson W.K. Web 3DNA – a web server for the analysis, reconstruction, and visualization of threedimensional nucleic-acid structures. *Nucleic Acids Research*. 2009. V. 37. W240–W246.

19. Strahs D., Schlick T. Analysis of A-tract bending: Insights into experimental structues by molecular dynamics simulations. 2000. *J. Mol. Biol.* V. 301. P. 643–663. URL: http://www.biomath.nyu.edu/index/software/Madbend/index.html (дата обращения 01.08.2011).
20. Xiang-Jun Lu, Shakked Z., Olson W. K. A-form Conformational Motifs in Ligand-bound DNA Structures. *J. Mol. Biol.* 2000. V. 300. P. 819–840.
21. Dickerson R. E. DNA bending: the prevalence of kinkiness and the virtues of normality *Nucleic Acids Res.* 1998. V. 26. P. 1906–1926.
22. Goodsell D.S. and Dickerson R.E. Bending and curvature calculations in B-DNA. *Nucleic Acids Res.* 1994. V. 22. № 24. P. 5497–5503.
23. Shpigelman E. S., Trifonov E. N. and Bolshoy A. CURVATURE: software for the analysis of curved DNA. *Comput. Appl. Biosci*. 1993. V. 9. P. 435–440.
24. Lavery R., Sklenar H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn*. 1988. V. 6. P. 63–91.
25. Lee S., Park K., Kang C. Z-curve: a computer program calculating DNA helical axis coordinates for three-dimensional graphic presentation of curvature. *Molecules and Cells*. 1999. V. 9. № 4. P. 350-357.
26. Barbic A., Crothers DM. Comparison of analyses of DNA curvature. *J. Biomol. Struct. Dyn*. 2003. V. 21. № 1. P. 89-97.
27. Herisson J., Payen G., Gherbi R. A 3D pattern matching algorithm for DNA sequence. *Bioinformatics.* 2007. V. 23. № 6. P. 680–686.
28. Olson W.K., Bansal M., Burley S.K., Dickerson R.E., Gerstein M., Harvey S.C., Heinemann U., Lu X.-J., Neidle S., Shakked Z., Sklenar H., Suzuki M., Tung C.-S., Westhof E., Wolberger C., Berman H.M. A Standard Reference Frame for the Description of Nucleic Acid Base-pair Geometry. *J. Mol. Biol*. 2001. V. 313. P. 229-237.
29. Suzuki M., Amano N., Kakinuma J., Tateno M. Use of a 3D Structure Data Base for Understanding Sequence-dependent Conformational Aspects of DNA. *J. Mol. Biol*. 1997. V. 274. P. 421–435.
30. Olson W.K. *Nucleic acid structural principles*. URL: http://128.6.69.24/lnotes/BioPhysChem_week5.pdf. (дата обращения 01.08.2011).
31. *Справочник по прикладной статистике*. Под ред. Айвазяна С.А., Тюрина Ю.Н. М.: Финансы и Статистика, 1990. Т. 2.
32. Kanhere A., Bansal M. A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*. 2005. V. 6. № 1.
33. Wang H., Benham C.J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*. 2006. V. 7. № 248.
34. Sorokin A.A., Osypov A.A., Dzhelyadin T.R., Beskaravainy P.M., Kamzolova S.G. Electrostatic properties of promoter recognized by *E. coli* RNA polymerase Esigma70. *J Bioinform Comput Biol*. 2006. V. 4. № 2. P. 455-467.

Материал поступил в редакцию 01.08.2011, опубликован 19.08.2011.