Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 529–536.
URL: http://www.matbio.org/2013/Korotkov_8_529.pdf

========== **КОММЕНТАРИИ** ==========

К вопросу о распознавании скрытой периодичности в последовательностях ДНК

Коротков Е.В[1,2]., Шеленков А.А.[1], Короткова М.А.[2]

[1]*Центр «Биоинженерия», Российская академия наук, Москва, 117312, Россия*
[2]*Национальный исследовательский ядерный университет (МИФИ), Москва, 115409, Россия*

Аннотация. В данной работе мы сравнили метод информационного разложения (ИР) и спектрально-статистический подход (СС). Мы показываем, что СС подход не учитывает влияния малой выборки, а выделение статистически значимого периода в последовательности оснований ДНК СС подход проводит математически некорректно. Обнаружение «профильной периодичности» СС подходом зависит исключительно от соотношения длин скрытых периодов. Обнаруженные недостатки СС подхода показывают, что для поиска районов ДНК со скрытой периодичностью в нуклеотидных последовательностях более корректно использовать статистику *Z* и метод ИР.

Ключевые слова. Скрытая периодичность, информационное разложение, спектрально-статистический подход, гены, триплетная периодичность, профильная периодичность.

1. ВВЕДЕНИЕ

В публикации [1] производится сравнение спектрально-статистического (СС) подхода (1) и метода информационного разложения (ИР) (2). Анализируются также СС подходом последовательности со скрытой периодичностью со вставками и делециями символов, опубликованные в публикации [3]. Как нам кажется, сравнение двух методов, разработанных для поиска скрытой периодичности нуклеиновых последовательностей, было проведено с ошибками, а сам СС подход [1] содержит некоторые серьезные недостатки. В данной работе мы рассматриваем ошибки, допущенные при сравнении СС подхода и метода ИР, а также некоторые недостатки СС ниже по пунктам.

2. ПОИСК СКРЫТОЙ ПЕРИОДИЧНОСТИ ВО ФРАГМЕНТАХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ AF453480 И CO11168X1

На страницах 510–511 публикации [1] авторы попытались сравнить разработанный ими СС подход с методом, использованным нами ранее в работах [2–3]. Смысл публикации [3] состоял в том, чтобы показать существование скрытой периодичности нуклеиновых кислот, которая может обнаруживаться только в присутствии делеций и вставок нуклеотидов. В качестве примера были показаны несколько таких последовательностей, среди них - район последовательности af453480 с 4166 по 4368 нуклеотид и район последовательности co11168x1 с 176412 по 176535 нуклеотид. Скрытая периодичность длиной $\lambda = 2$ нуклеотида была найдена в них только в присутствии некоторых делеций и вставок нуклеотидов. Выравнивания для этих двух последовательностей показаны нами в таблице 4 нашей публикации [3]. Однако авторы публикации [1] применили свой подход к этим двум последовательностям без проделанного нами выравнивания и, вполне естественно, не нашли обнаруженную

нами периодичность. Мы пересчитали результаты рис. 5,а и 5,б из публикации [1] без проведения выравнивания, и результаты этого пересчета показаны на рис. 1,А и рис. 2,А. Видно, что в целом результаты совпадают с рис. 5 из публикации [1]. Однако, если использовать сделанные нами в публикации [3] выравнивания, получается совершенно другая картина, и тогда уже виден период в 2 нуклеотида. Эти результаты показаны ниже на рис. 1,В для последовательности af453480 и рис. 2,В для последовательности co11168x1.

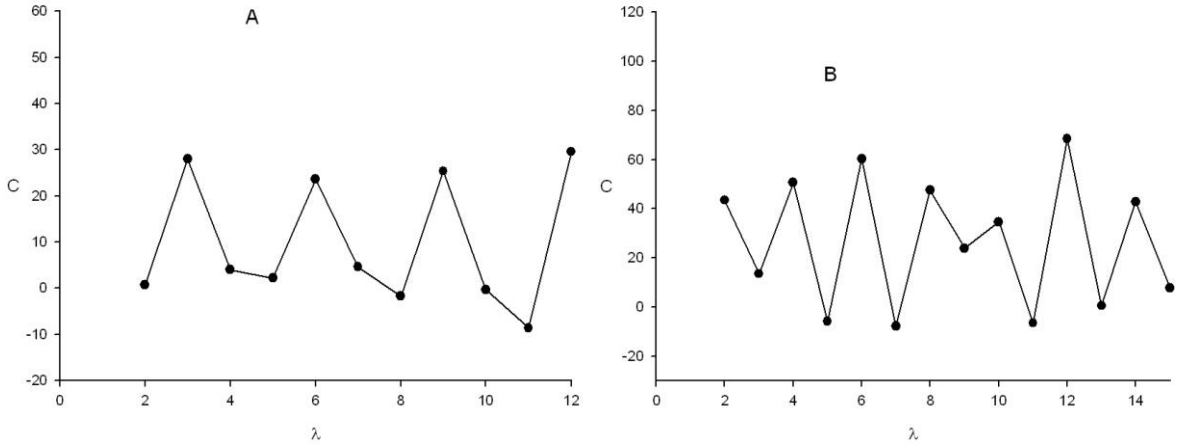

Рис. 1. Характеристические спектры для района последовательности af453480 с 4166 по 4368 нуклеотид без учета выравнивания (А) и с учетом выравнивания (В), сделанного в публикации [3].

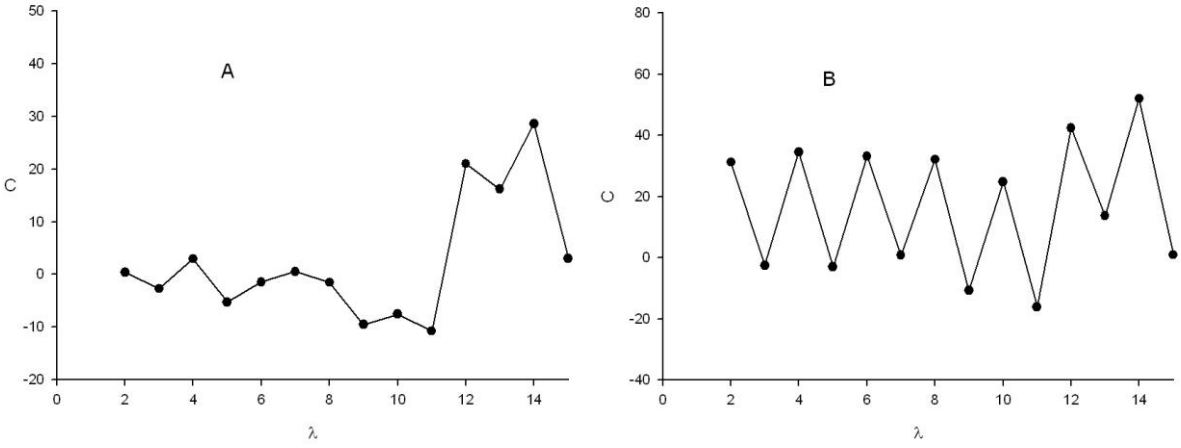

Рис. 2. Характеристические спектры для района последовательности co11168x1 с 176412 по 176535 нуклеотид без учета выравнивания (А) и с учетом выравнивания (В), сделанного в публикации [3].

Построенные нами графики показывают, что в публикации [1] сравнение с нашими результатами сделано с грубой ошибкой, так как оно проведено без учета сделанных нами выравниваний (вставки и делеции нуклеотидов). Поэтому вывод публикации [1] на странице 511 «Таким образом, результаты оценки скрытого периода, полученные в работе [4] без их верификации, могут оказаться некорректными» является неправильным. Видимо, авторы публикации [1] невнимательно прочитали нашу публикацию [3], и поэтому не смогли сделать корректное сравнение результатов.

3. ПРОБЛЕМА МАЛОЙ ВЫБОРКИ В СПЕКТРАЛЬНО-СТАТИСТИЧЕСКОМ ПОДХОДЕ

В публикации [1] используется распределение χ^2 для оценки статистической значимости найденных периодичностей. Как видно из данной работы и из более ранней публикации [4], в них проводился анализ последовательностей короткой длины, где длина района со скрытой периодичностью может содержать всего несколько периодов. В этом случае расчет по формулам (2) и (4) в публикации [1] будет проводиться для N периодов, где N может быть небольшим числом. Длина анализируемой последовательности будет равна $N\lambda$, где λ длина периода. Если число периодов N будет находиться в интервале от 2 до ~20 периодов, то все оценки статистической значимости, выполненные в публикации [1], будут очень неточными, так как этот диапазон N соответствует так называемой «малой выборке». Это явление достаточно широко известно при статистическом анализе данных. Для примера, об этом можно прочитать в книге [5].

Для иллюстрации мы в данной работе оценили влияние малой выборки на вероятность α, которая используется в формуле (4) публикации [1] для длины периода $\lambda = 32$. Вероятность α в работах [1, 4] выбиралась равной 0.05. По значению α для каждой длины периода λ рассчитывалось такое значение $\chi^2_{0.05}$, чтобы вероятность $P(x \geq \chi^2_{0.05})$ была равна α для x, имеющего χ^2-распределение с числом степеней свободы, равным $R(\lambda–1)$. Здесь R – число используемых оснований ДНК, и оно равно 4. На рис. 3 показано значение вероятности α в зависимости от числа периодов N, определенной по методу Монте-Карло. Видно, что значения α для малой выборки могут отличаться в 5 раз от значений α, ожидаемых для большой выборки. Эти отклонения будут тем больше, чем меньше берется вероятность α, и они зависят от частот нуклеотидов в анализируемом районе со скрытой периодичностью. Получается, что оценки статистической значимости, сделанные как в публикации [1], так и в более ранней публикации [4], допускают в некоторых случаях ошибку в несколько сотен процентов. Такие ошибки приводят к искажению спектров «профильной периодичности», построенных по формулам (2) и (4) в публикации [1], и к неправильным выводам, как в рассматриваемой публикации [1], так и в более ранних публикациях этих авторов.

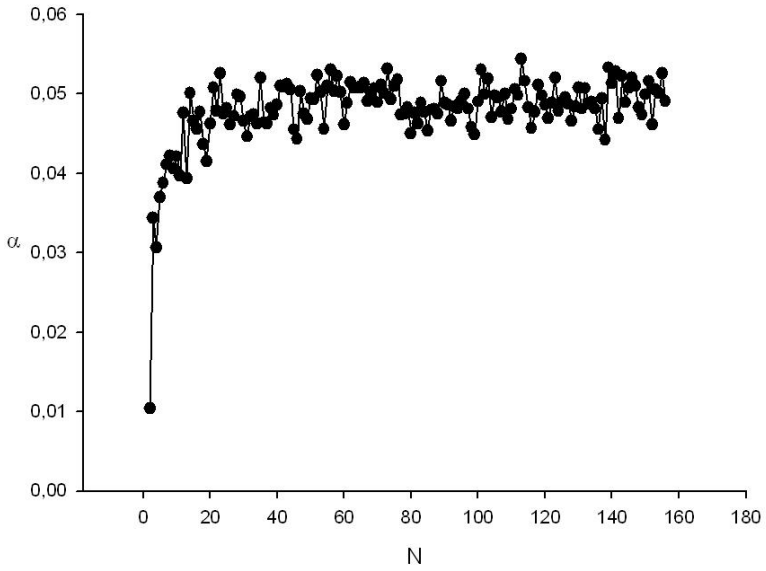

Рис. 3. Показана зависимость вероятности α, которая используется в формулах (2) и (4) публикации [1], от числа периодов N. Расчет сделан для длины периода λ, равной 32 нуклеотидам, и для равномерного содержания букв.

4. ПОИСК НАИБОЛЕЕ СТАТИСТИЧЕСКИ ЗНАЧИМОГО ПЕРИОДА В СПЕКТРАЛЬНО-СТАТИСТИЧЕСКОГО ПОДХОДЕ

В публикации [1] для поиска наиболее значимого скрытого периода и для поиска профильной периодичности используется характеристический спектр $C(\lambda)$ (формула (3) публикации [1]) и спектр $D_L(\lambda)$ (формулы (2) и (4) публикации [1]). Спектр $C(\lambda)$ и спектр $D_1(\lambda)$ полностью аналогичны спектру $2I$ в зависимости от λ, который был получен нами в работах [6, 7]. Только спектр $C(\lambda)$ получен из спектра $2I(\lambda)$ вычитанием для каждого λ математического ожидания $E(2I(\lambda))$, а спектр $D_1(\lambda)$ получен делением значения $2I(\lambda)$ на значения $2I_{0.05}(\lambda)$ такого, что $P(2I > 2I_{0.05}(\lambda)) = 0.05$ для каждого λ. Спектр $C(\lambda)$ используется в работе [1] на странице 505–506 для поиска первого тест-периода с максимальным значением $C(\lambda)$. Авторы публикации [1] внизу страницы 505 пишут «Первый тест-период L с ярко выраженным максимальным значением спектра C служит оценкой скрытого периода в строке *str*.” Далее они также пишут «Эти спектры, согласно сформулированному правилу, позволяют получить правильные оценки скрытых периодов для рассматриваемых тандемных повторов”. Мы считаем эти утверждения в общем случае неправильными. Под общим случаем понимается такой случай, где в анализируемой последовательности оснований ДНК может содержаться как один скрытый период, так и несколько таких периодов. Наше утверждение связано с тем, что мы не можем сравнивать $C(\lambda)$ для различных значений λ, так как при равенстве $C(\lambda_1) = C(\lambda_2) = C_0$ эти значения имеют совершенно разные вероятности $P(C(\lambda_1) \geq C_0)$ и $P(C(\lambda_2) \geq C_0)$ согласно используемому авторами работы [1] распределению χ^2. Мы посчитали для примера вероятности $P1$ того, что $C(\lambda) \geq 30.0$ для значений λ в интервале от 2 до 60, что показано на рис. 4. Из этого рисунка видно, что такие вероятности очень сильно различны. Если учесть вид зависимости, представленной на рис. 4, то видно, что статистически менее значимые длинные периоды всегда будут иметь очень сильное предпочтение перед короткими периодами. Это означает, что истинное значение длины скрытого периода и его статистическая значимость могут определяться неправильно, если использовать спектр $C(\lambda)$.

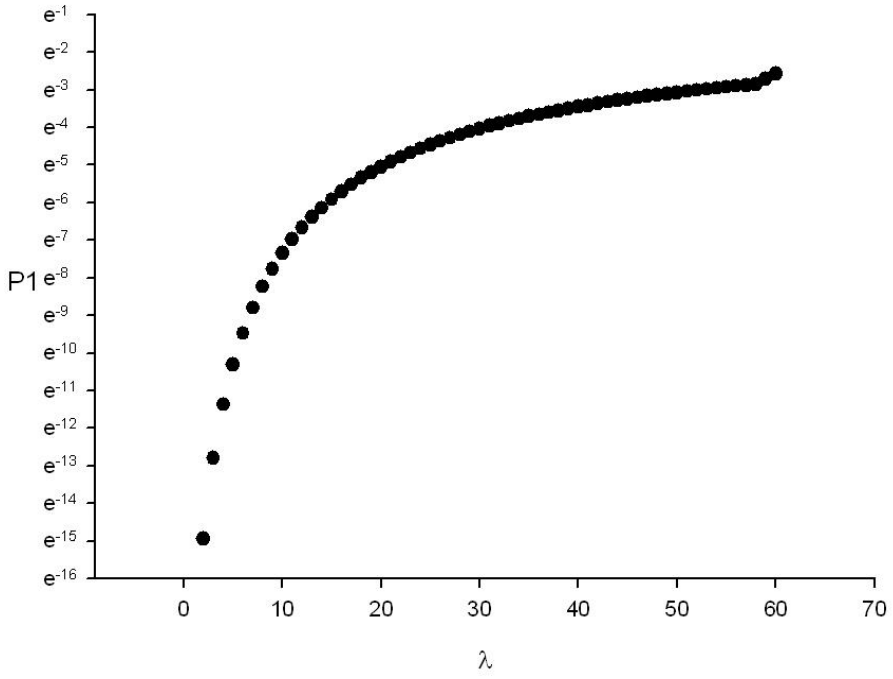

Рис. 4. На рисунке показана вероятность $P1$ того, что $C(\lambda) \geq 30.0$) для значений λ (длина периода) в интервале от 2 до 60.

Также нельзя определить наиболее статистически значимый скрытый период и по спектру $D_1(\lambda)$, так как при равенстве $D_1(\lambda_1) = D_1(\lambda_2) = D_0$ эти значения имеют совершенно разные вероятности $P(D_1(\lambda_1) \geq D_0)$ и $P(D_1(\lambda_2) \geq D_0)$ согласно используемому авторами работы [1] распределению χ^2 (рис. 5).

В работе [1] допущена также ошибка на рис. 2,б. Как видно из рис. 4,б, авторы применяют свое правило, которое мы цитировали выше в этом пункте, и из всех периодов, видимых на рис 4,б, они берут такой, который имеет максимально значение $C(\lambda)$, т. е. пропускают все λ до $\lambda = 84$, где наблюдается максимум $C(\lambda)$. Если же применить это приведённое выше правило выбора наиболее значимого скрытого периода для рис 2,б, то следует признать существование в анализируемой нуклеотидной последовательности периода длиной в 24 нуклеотида, так как из рисунка видно, что $C(12) \approx 48.0$ и $C(24) \approx 60.0$, т. е. в любом случае $C(12) < C(24)$. Почему в данном случае авторы противоречат своему правилу, и почему они считают, что это правило позволяет получить «правильные оценки скрытых периодов», остается загадкой. Авторы публикации [1] не приводят никаких доводов и расчетов в доказательство справедливости своего правила.

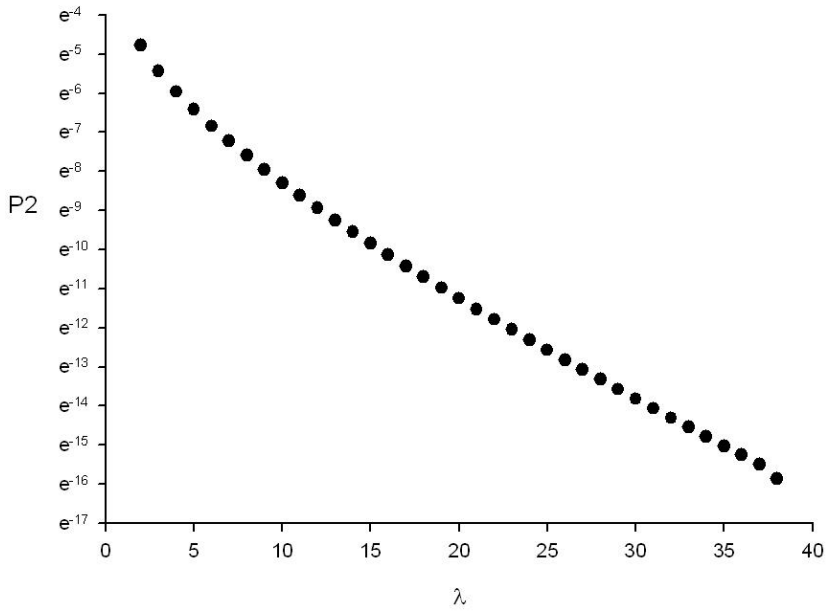

Рис. 5. На рисунке показана вероятность $P2$ того, что $D_1(\lambda) \geq D_0$) для значений λ (длина периода) в интервале от 2 до 38 и для $D_0 = 1.5$.

Следует отметить, что ситуация, когда разработанное авторами правило будет указывать на неверную длину периода, будет очень часто встречаться при анализе нуклеотидных последовательностей. Мы имеем в виду ту ситуацию, когда максимум будет наблюдаться для $C(k\lambda)$, где k меняется от 2 до такого значения k, когда еще $\lambda_{max} \geq k\lambda$. Здесь λ_{max} – максимальная длина периода в спектре $C(\lambda)$ для анализируемой последовательности. Все дело в том, что в этом случае, как мы уже писали ранее в работе [7], значение $\chi^2(k\lambda) \geq \chi^2(\lambda)$ если использовать формулу (1) публикации [1], т. е. $\chi^2(\lambda)$ включен полностью в $\chi^2(k\lambda)$. Это значит, что если из-за случайных факторов значение $C(k\lambda) = C_k$ будет больше, чем значение $C(\lambda) = C_0$, то вероятность $P(C(k\lambda) > C_k)$ может быть больше вероятности $P(C(\lambda) \geq C_0)$. А для самого статистически значимого периода такая вероятность должна быть минимальной. Фактически это будет приводить к тому, что наиболее значимый период, который определяют авторы в публикации [1] и некоторых предыдущих своих публикациях, будет определен неправильно. Как мы считаем, эта ошибка отразилась в работе авторов [8], где авторы на рис. 5 приводят дендрограмму по разделению некоторых кодирующих районов на

группы в соответствии с выявленными свойствами их последовательностей. В публикации [8] произведено разделение кодирующих последовательностей на 4-ом уровне на «3-профильные» и «не 3-профильные». Всего на этом уровне содержалось 12996 последовательностей (73,6% от исходных). Эти последовательности были поделены на 10699 последовательностей, которые обозначены как «3-профильные» (60,6% от исходных), и на 2297 последовательностей, которые обозначены как «не 3-профильные» (13% от исходных последовательностей). В силу вышесказанного, такое разделение, проведенное по методу, описанному в работах [1, 8], будет некорректным в силу того, что длина наиболее статистически значимого периода определяется математически неправильно. Фактически это приводит к сильному занижению числа районов ДНК с периодичностью в 3 нуклеотида и сильному завышению числа периодов с периодичностью другой длины, представленных на рисунках 5–7 в публикации [8].

5. ПОНЯТИЕ «ПРОФИЛЬНОЙ ПЕРИОДИЧНОСТИ»

Мы также считаем, что деление скрытой периодичности на «профильную» и «непрофильную» [1, 8] не имеет большого значения, так как такое деление не определяет свойства скрытой периодичности, а зависит только от соотношения длин скрытых периодов, которые присутствуют в анализируемой последовательности. Это обусловлено тем, что в нуклеотидной последовательности может быть несколько скрытых периодичностей с различной длиной. Согласно данным работы [7], периоды, чьи длины представляют собой взаимно простые числа, не влияют друг на друга. Эти длины периодов могут «наводить» периодичность на всех кратных длинах. Это означает, что $\chi^2(\lambda)$ содержится полностью в $\chi^2(k\lambda)$ если использовать формулу (1) публикации [1], как отмечено выше в пункте 4. Если же длина периода является составным числом, то такая «наводка» осуществляется для всех длин, равных простым числам, составляющих длину этого периода, и для произведений этих чисел. Это явление прекрасно видно на рис. 4,б в публикации [1]. Тогда получается, что если мы имеем в изучаемой последовательности несколько периодичностей различной длины, где длины периодов представляют собой простые числа, то профильным периодом всегда будет период, равный их произведению. Для него все значения D будут всегда меньше 1. Это прекрасно продемонстрировано на рис. 3 в публикации [4]. Тут есть две периодичности, равные 3 основаниям и 11 основаниям. Поэтому профильный период получается равным 33 основаниям. Эта длина меньше λ_{max} (см. пункт 4), поэтому удается выделить «профильную» периодичность. Если бы скрытые периоды имели бы длину, например, 37 оснований и 13 оснований, а $\lambda_{max} = 300$, то в такой последовательности методы, предложенные в работах [1, 4, 5], не выявляют «профильную» периодичность просто из-за того, что произведение $13 \cdot 37 > 300$. Поэтому обнаружение присутствия или отсутствия «профильной периодичности» является не более чем дифференциацией скрытых периодичностей в анализируемой последовательности по произведению длин периодов, которые в ней присутствуют, и никакого иного смысла это понятие не имеет. Поэтому классификация кодирующих последовательностей из базы данных KEGG на профильные и непрофильные (рис.5 в работе [8]) показывает только то, что у 20.8% генов длина «профильного периода» будет больше, чем длина анализируемого гена. Это означает, что многие «непрофильные последовательности» содержат периодичность с длиной периода, равной 3 основаниям, на фоне одного или нескольких более длинных периодов. Поэтому присутствие в последовательности оснований ДНК «непрофильной периодичности» [1, 4] не может служить критерием отсутствия скрытой периодичности в нуклеотидных последовательностях.

6. ДВУХУРОВНЕВАЯ ОРГАНИЗАЦИЯ СКРЫТОЙ ПЕРИОДИЧНОСТИ В ГЕНАХ

Двухуровневая организация кодирования, о которой авторы говорят в статье [8] на странице 150 и в публикации [4], была нами замечена ранее в публикации [7] на странице 204. Однако в работах [4,8] отсутствует ссылка на публикацию [7]. Мы же сравнили этот эффект в публикации [7] с амплитудной модуляцией радиосигналов. В генах скрытая периодичность, кратная трем основаниям, как бы модулирует периодичность, равную трем основаниям. Авторы работ [4,8] поэтому не самостоятельно заметили этот факт, а всего лишь повторяют выводы работы [7], в данном случае, без ссылок на первоисточник.

7. ПРЕИМУЩЕСТВА МЕТОДА ИНФОРМАЦИОННОГО РАЗЛОЖЕНИЯ

После описанных в пунктах 3–5 недостатков СС подхода становится совершенно ясно, что использование статистики Z для поиска скрытой периодичности в нуклеотидных последовательностях (публикации [2, 3, 6, 7]) является более корректным, чем использование спектров $C(\lambda)$, $D(\lambda)$ (формулы (3) и (4) публикации [1]) или использование критерия χ^2 согласно формуле (1) публикации [1]. Во-первых, значительно уменьшается влияние статистки малой выборки. Во-вторых, можно сравнивать по значению Z статистическую значимость скрытой периодичности с различной длиной периода. Если $Z(\lambda_1) = Z(\lambda_2) = Z_0$, то это означает, что вероятности $P(Z(\lambda_1) \geq Z_0)$ и $P(Z(\lambda_2) \geq Z_0)$ будут также равны. Это позволяет по спектру $Z(\lambda)$ найти самый статистически значимый период, который имеет максимальное значение Z. Этого не наблюдается в случае использования статистик $C(\lambda)$, $D(\lambda)$ или прямого использования χ^2 (публикация [1], формулы (3), (4) и (1)).

8. ОТСУТСТВИЕ ДАННЫХ ОБ ОШИБКАХ ПЕРВОГО И ВТОРОГО РОДА

В публикации [1] проводится построение дендрограммы для кодирующих последовательностей при помощи спектрально-статистического подхода. Каждый шаг такой классификации требует определения FDR (false discovery rate) и числа ошибок первого и второго рода. Однако ни FDR, ни число ошибок первого и второго рода в этой работе определены не были. В силу этого разделение кодирующих участков генов на однородные и неоднородные, на профильные и непрофильные, а также на 3-регулярные и 3-нерегулярные не представляется статистически значимым (рис. 5 публикации [1]). Также в этой классификации неявным образом присутствует зависимость от распределения кодирующих участков по длинам и от максимальной длины периода λ_{max} в профильном спектре, чего в математически грамотно выполненной классификации быть не должно. Присутствие такой зависимости делает дендрограмму не имеющей какого-либо биологического смысла. Все это не позволяет рассматривать введенные классы кодирующих областей как математически или биологически обоснованные.

В публикации [9] проводился поиск районов со скрытой периодичностью в последовательностях ДНК из различных хромосом. В этой работе также нет определения FDR (false discovery rate) и числа ошибок первого и второго рода. Кроме того, в созданном банке данных присутствуют все недостатки спектрально-статистического метода (пункты 3–5 выше). В силу этого банк данных содержит некорректную информацию.

ЛИТЕРАТУРА

1. Чалей М.Б., Кутыркин В.А Распознавание скрытой периодичности в последовательностях ДНК. *Математическая биология и биоинформатика*. 2013. Т. 8. №2. С. 502–512.
2. Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method to analyze symbolical sequences. *Physical Letters A*. 2003. V. 312. P. 198–210.
3. Shelenkov A., Skryabin K., Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. *DNA Res*. 2006. V. 13. P. 89–102.
4. Chaley M., Kutyrkin V. Profile-Statistical Periodicity of DNA Coding Regions. *DNA Res.* 2011. V. 18. P. 353–362.
5. Сухорученков Б.И. *Анализ малой выборки. Прикладные статистические методы*. М.: Вузовская книга, 2010.
6. Korotkov E.V., Korotkova M.A. DNA regions with latent periodicity in some human clones. *DNA Sequence.* 1995. V. 5. P. 353–358.
7. Korotkov E.V., Korotkova M.A., Tulko J.S. Latent sequence periodicity of some oncogenes and DNA-binding protein genes. *CABIOS*. 1997. V. 13. P. 37–44.
8. Кутыркин В.А., Чалей М.Б. Структурные различия кодирующих и некодирующих районов последовательностей ДНК генома человека. *Вестник МГТУ им. Н.Э.Баумана. Сер. Естественные науки*. 2012. Спец.выпуск № 3 «Математическое моделирование». С. 146–157.
9. Чалей М.Б., Кутыркин В.А., Тюльбашева Г.Э., Теплухина Е.И., Назипова Н.Н Исследование феномена скрытой периодичности в геномах эукариотических организмов. *Математическая биология и биоинформатика*. 2013. Т. 8. № 2. С. 480–501.

Материал поступил в редакцию 14.10.2013, опубликован 21.11 .2013.