МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2013, том 47, № 4, с. 689–696

БИОИНФОРМАТИКА

УДК 57.573+577.2;577.113

ПОЛНОГЕНОМНЫЙ ПОИСК СТРУКТУРИРОВАННЫХ НЕКОДИРУЮЩИХ РНК

© 2013 г. С. В. Виноградова[1, 2*], Р. А. Солдатов[2], А. А. Миронов[1, 2]

[1]Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, 119234

[2]Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127994

Поступила в редакцию 13.12.2012 г.
Принята к печати 28.02.2013 г.

Некодирующие РНК (нкРНК) – функциональные транскрипты, не кодирующие белки, – вовлечены в регуляцию множества клеточных процессов. Единственная общая характеристика, объединяющая почти все нкРНК – их способность к образованию вторичной структуры, которая может быть крайне важной для их функций и, следовательно, должна быть консервативной. Этот факт может быть использован для полногеномного поиска нкРНК. Наш подход основан на вычислении весов спаривания нуклеотидов с последующей максимизацией суммарной меры спаривания данного участка последовательности с помощью алгоритма Нуссинов. Мы показали, что предложенный метод позволяет эффективно предсказывать известные классы нкРНК, а также потенциально новые нкРНК.

Ключевые слова: **РНК, нкРНК, микроРНК, компенсаторные замены, сравнительная геномика.**

GENOME-WIDE IDENTIFICATION OF FUNCTIONAL NONCODING RNAS, by *S. V. Vinogradova[1, 2]*, R. A. Soldatov[2], A. A. Mironov[1, 2]* ([1]Moscow State University, Moscow, 119234 Russia, *e-mail: kintany@gmail.com; [2]Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia). Non-coding RNAs (ncRNAs) are functional transcripts that do not encode proteins. They are involved in many regulation pathways. The only general characteristic shared by many (but not all) known RNAs is folding into complex shapes that are crucial to function and thus should be conserved. This fact can be used for genome-wide prediction of ncRNAs. Our approach is based on computing of local base pairing probabilities and further maximizing of probability for a segment with the use of Nussinoff algorithm. We showed that it allows to efficiently predict known ncRNA and possibly some new ncRNAs.

***Keywords*: RNA, ncRNA, miRNA, compensatory substitutions, comparative genomics.**

DOI: 10.7868/S0026898413040162

ВВЕДЕНИЕ

Некодирующие РНК (нкРНК) представляют собой потенциально функциональные транскрипты, не кодирующие белки и участвующие во множестве клеточных процессов. Последние исследования транскриптома показали, что до 90% генома транскрибируется в той или иной степени, с одной или обеих цепей [1], в то время как только 1.2% всего генома кодирует белки [2]. Это предполагает наличие большого количества некодирующих транскриптов. Несмотря на то, что данное наблюдение остается спорным и противоречивым, растущее количество подтверждений функциональности ряда нкРНК предполагает их исключительную важность (по крайней мере части нкРНК) в регуляции многих клеточных процессов. Однако в настоящее время большинство нкРНК остаются недостаточно изученными, а основной интерес представляет собой поиск новых типов РНК, ранее не аннотированных, обладающих потенциально новыми функциями.

Полногеномный поиск *de novo* семейств нкРНК может быть выполнен как биоинформатически, с помощью компьютерного анализа геномных данных, так и экспериментально, с использованием методов высокопроизводительного секвенирования [3]. Основная сложность нахождения новых нкРНК заключается в отсутствии выраженного

Принятые сокращения: нкРНК – некодирующие РНК; мяРНК – малые ядерные РНК; мякРНК – малые ядрышковые РНК; тРНК – транспортные РНК.

*Эл. почта: kintany@gmail.com

сигнала в последовательности, позволяющего идентифицировать участок последовательности как нкРНК. С другой стороны, многие функциональные РНК имеют определенную вторичную структуру, и это можно использовать при их поиске. Самые распространенные подходы к предсказанию вторичной структуры и оценки ее значимости – минимизация энергии структуры [4] и оценка вероятности спаривания двух нуклеотидов с последующим построением структуры из наиболее устойчивых пар (алгоритмы MEA [5]). Время работы алгоритмов в общем случае $O(n^3)$. В случае, если цель не предсказание самой структуры, а поиск структурированных участков, то вместо вероятностей спаривания нуклеотидов можно использовать меру, определяющую склонность двух нуклеотидов взаимодействовать в данном геномном окружении. Оценка суммарной меры спаривания нуклеотидов конкретного участка позволит найти потенциально структурированные участки [6].

Однако показано, что вторичные структуры нкРНК часто не являются значимо более стабильными, чем структуры случайных геномных последовательностей [7]. Именно эволюционная консервативность вторичной структуры РНК позволяет делать выводы о функциональности РНК, и, таким образом, сравнительная геномика – наиболее логичный и многообещающий подход. К настоящему моменту расшифровано большое количество геномов и известно, что большинство из них содержат значительную часть некодирующих элементов, часть из которых может относиться к нкРНК. Но и такой подход обладает своими недостатками и ограничениями.

Во-первых, далеко не все нкРНК обладают консервативной структурой: некоторые нкРНК могут быть специфичными для определенного организма или быть структурированными только на каком-то участке своей последовательности, в то время как остальная их часть может не нести сигнала структурированности. Во втором случае необходима постобработка полученных предсказаний для точного определения границ транскрипта [8].

Во-вторых, для применения сравнительно-геномного подхода к предсказанию вторичных структур РНК необходимо выравнивание. Алгоритмы по предсказанию можно разбить на два класса: основанные на выравнивании последовательностей и основанные на выравнивании структур. Второй класс крайней сложен вычислительно, поэтому плохо подходит при поиске в больших геномах. Алгоритмы первого класса (Evofold [9], qRNA [10], RNAz [11]) требуют установления точного соответствия символов в выравниваниях и поэтому для надежной работы необходимо достаточно сильное сходство последовательностей.

В настоящей работе мы предлагаем сравнительно-геномный подход, который не использует собственно выравнивание (сопоставление символов), но базируется только на факте ортологичности последовательностей. Таким образом, мы сильно снижаем требование к качеству выравниваний. Мы применили наш подход для сравнительно-геномного поиска консервативных вторичных структур в геномах рода *Drosophila*. Проведенный поиск позволил получить общую картину распространения структурированных участков в различных областях геномов а также обнаружить ряд новых нкРНК.

МАТЕРИАЛЫ И МЕТОДЫ

Алгоритм. Веса спаривания нуклеотидов ξ вычисляли с помощью программы RNAplfold [6]. Максимизацию суммарной меры спаривания нуклеотидов $\xi_k[i, j]$ данного участка $[i, j]$ последовательности k проводили с использованием стандартного алгоритма Нуссинов [12] с весами, полученными на первом шаге. Степень структурированности данного участка множественного выравнивания вычисляли как среднюю для всех последовательностей $x[i, j] = \mathrm{avg}_k\{\xi_k[i, j]\}$. Важно отметить, что здесь не использовали выравнивание как таковое (посимвольное соответствие), а только факт, что участки геномов сопоставлены. С учетом полученной степени структурированности применяли алгоритм поиска нкРНК: геномное выравнивание сканировали "скользящим окном" фиксированного размера, в каждом окне вычисляли структурированность множественного выравнивания. Для каждого участка вычисляли z-score:

$$z\text{-score}(i, j) = \frac{x[i, j] - E}{\sigma}, \qquad (1)$$

где $x[i, j]$ – степень структурированности данного окна, E – математическое ожидание степени структурированности, σ – стандартное отклонение степени структурированности. В качестве оценки E и σ применяли значения, полученные в результате полногеномного анализа. Полученное значение z-score использовали в дальнейшем статистическом анализе.

Вычисление FDR. Чтобы вичислить долю ложных предсказаний (FDR), использовали случайную модель, построенную следующим образом. Все рассмотренные выравнивания разбивали на короткие сегменты длиной 30 н. Далее колонки выравниваний в этих сегментах случайным образом перемешивали, и все вычисления проводили

для этого набора выравниваний. Такой подход, с одной стороны, не сохраняет вторичные структуры, а, с другой, – сохраняет локальный уровень консервативности. FDR рассчитывался по стандартной формуле Бенджамини-Хохберга [13]:

$$\mathrm{FDR}(x) = \frac{NF(x)}{n(x)},$$

где N – общее число наблюдений, $n(x)$ – их число со значением веса (z-score), превышающем x, $F(x)$ – ожидаемая доля числа наблюдений со значением z-score, превышающем x, оцененная из случайной модели.

Геномное выравнивание и консервативные элементы. Анализ проводили на следующих геномах рода *Drosophila* (идентификатор сборки по UCSC Genome Browser [14] указан в скобках): *D. pseudoobscura* (dp4), *D. ananassae* (droAna3), *D. erecta* (droEre2), *D. erecta* (droEre1), *D. grimshawi* (droGri2), *D. mojavensis* (droMoj3), *D. persimilis* (droPer1), *D. sechellia* (droSec1), *D. simulans* (droSim1), *D. virilis* (droVir3), *D. willistoni* (droWil1), *D. yakuba* (droYak2). Использование геномов разного уровня сходства позволяет предсказывать как консервативные по последовательности участки, так и разошедшиеся участки со стабильной вторичной структурой.

Поскольку важным критерием наличия функциональной вторичной структуры мы считали ее консервативность, то для анализа использовали только те участки геномов, которые входили во множественные геномные выравнивания MULTIZ [15].

Рассматривали сегменты выравнивания, содержащие фрагменты геномов как минимум из 6 видов (включая *D. melanogaster*), а также обладающие степенью идентичности, не менее 40% в "скользящем окне" 100 н.

Сравнение с аннотацией известных нкРНК. Наши наблюдения мы сравнивали с различными классами нкРНК. Использованы наборы: из 237 микроРНК *D. melanogaster* из miRBase [16]; из 279 тРНК *D. melanogaster* из Genomic tRNA Database [17]; из 286 мякРНК D. melanogaster из snoRNA-LBME-db [18] и из 48 мяРНК *D. melanogaster* из аннотации UCSC Genome Browser.

Сравнение с аннотацией белоккодирующих генов. Чтобы определить геномное расположение предсказанных структурированных участков, использовали аннотацию белок кодирующих генов, полученную из UCSC Genome Browser (Feb. 2009 assembly) [19].

Транскрипционные данные. Мы учли данные: данные проекта modENCODE [20] группы Сюзаны Селникер (http://www.modencode.org/celniker/) из проекта "Comprehensive characterization of the Drosophila transcriptome" по транскрипционному профайлингу эмбриональных клеток *D. melanogaster* и тканеспецифичному транскрипционному профайлингу. Полученные SAM-файлы преобразовали и обработали с помощью программы Samtools [21].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Используя полногеномные множественные выравнивания 12 геномов *Drosophila*, мы провели поиск структурированных сегментов геномов, которые предположительно являются функциональными нкРНК.

Для "скользящих окон" длиной 40, 70, 100 и 150 н. отобраны сегменты суммарной длины ~1.16×10^8 н., что составляет 97% от генома *D. melanogaster*. Мы применили наш алгоритм к этим сегментам, количество найденных структурированных участков при FDR 10% в зависимости от длины окна приведено в табл. 1.

При анализе геномного расположения, пересечения с сайтами сплайсинга и подтверждении транскрипционными данными мы использовали участки, найденные с использованием окна длины 100 н.

Известные классы РНК

Для 4 известных классов РНК, таких как микроРНК, тРНК, мяРНК и мякРНК, наш алгоритм достигает высокого уровня чувствительности, при правильном выборе длины окна (табл. 2). На

Таблица 1. Количество структурированных участков на уровне 10% FDR и покрытие генома, в зависимости от длины окна

Длина окна	Количество участков	Покрытие генома, %
40	42200	1.42
70	21133	1.23
100	15851	1.29
150	20246	2.29

Таблица 2. Чувствительность алгоритма на уровне 10% FDR при предсказании разных типов нкРНК

Тип нкРНК	Чувствительность (наблюдение/общее число)	Длина окна, н.
микроРНК	56% (133/237)	100
тРНК	44% (122/279)	40
мякРНК	21% (59/286)	150
мяРНК	53% (25/48)	40

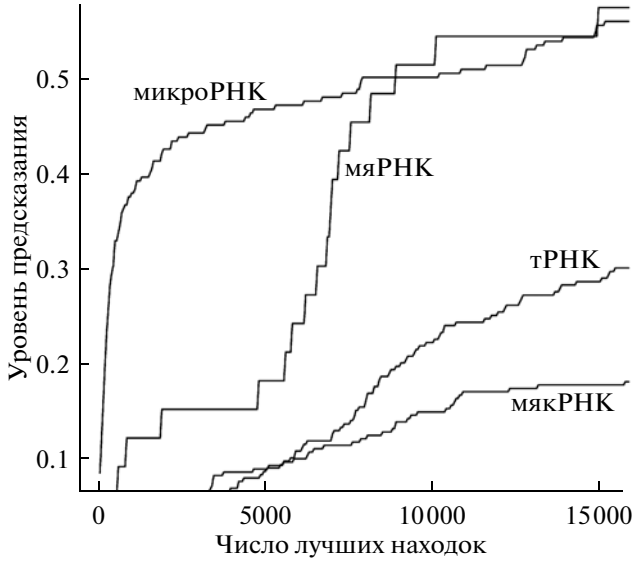

Рис. 1. Уровень предсказания различных классов известных РНК в зависимости от числа выбранных лучших находок.

рис. 1 представлена зависимость доли найденных известных нкРНК от числа найденных структур. Видно, что заметная доля известных нкРНК имеют достаточно низкую склонность к образованию вторичных структур, что соответствует литературным данным [7] (рис. 2). Тем не менее, около половины нкРНК (за исключением мякРНК) предсказана нашим алгоритмом, что сравнимо с эффективностью предсказания других алгоритмов (Evofold [9], RNAz [11]). Это дает основание предполагать, что другие предсказанные нами нкРНК также могут быть функциональными.

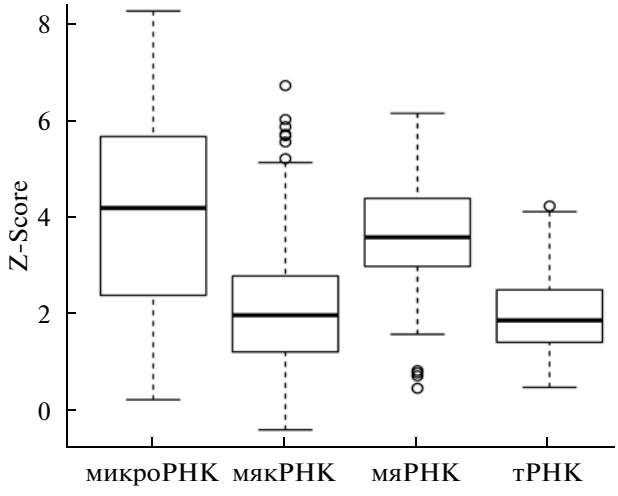

Рис. 2. Распределение *z*-scores известных классов нкРНК.

Геномное расположение

Сначала из числа найденных структурированных участков (длина окна 100 н.) были исключены участки, пересекающиеся с повторяющимися элементами генома, а также участками низкой сложности. После фильтрации осталось 13300 структур. Среди них приблизительно 62% попадают в межгенные или интронные области, 2.5% – в 3'-нетранслируемые области, 2% – в 5'-нетранслируемые области и 32% пересекаются с белок кодирующими генами.

Сайты сплайсинга

Среди найденных структурированных РНК мы провели поиск РНК, пересекающихся с границами интрон-экзонных структур для внутренних интронов. Мы обнаружили 517 пересечений структурированных сегментов с конститутивными сайтами сплайсинга, как донорными, так и акцепторными, и 147 пересечений с альтернативными сайтами сплайсинга, как донорными, так и акцепторными. Количество пересечений с сайтами альтернативного сплайсинга находится на уровне случайного, таким образом, мы не можем сделать вывод о структурированности таких сайтов. С другой стороны, количество наблюдаемых структур в конститутивных сайтах сплайсинга значительно меньше ожидаемого по случайным причинам (p-value ≈ 0), таким образом, конститутивные сайты сплайсинга чаще всего не структурированы.

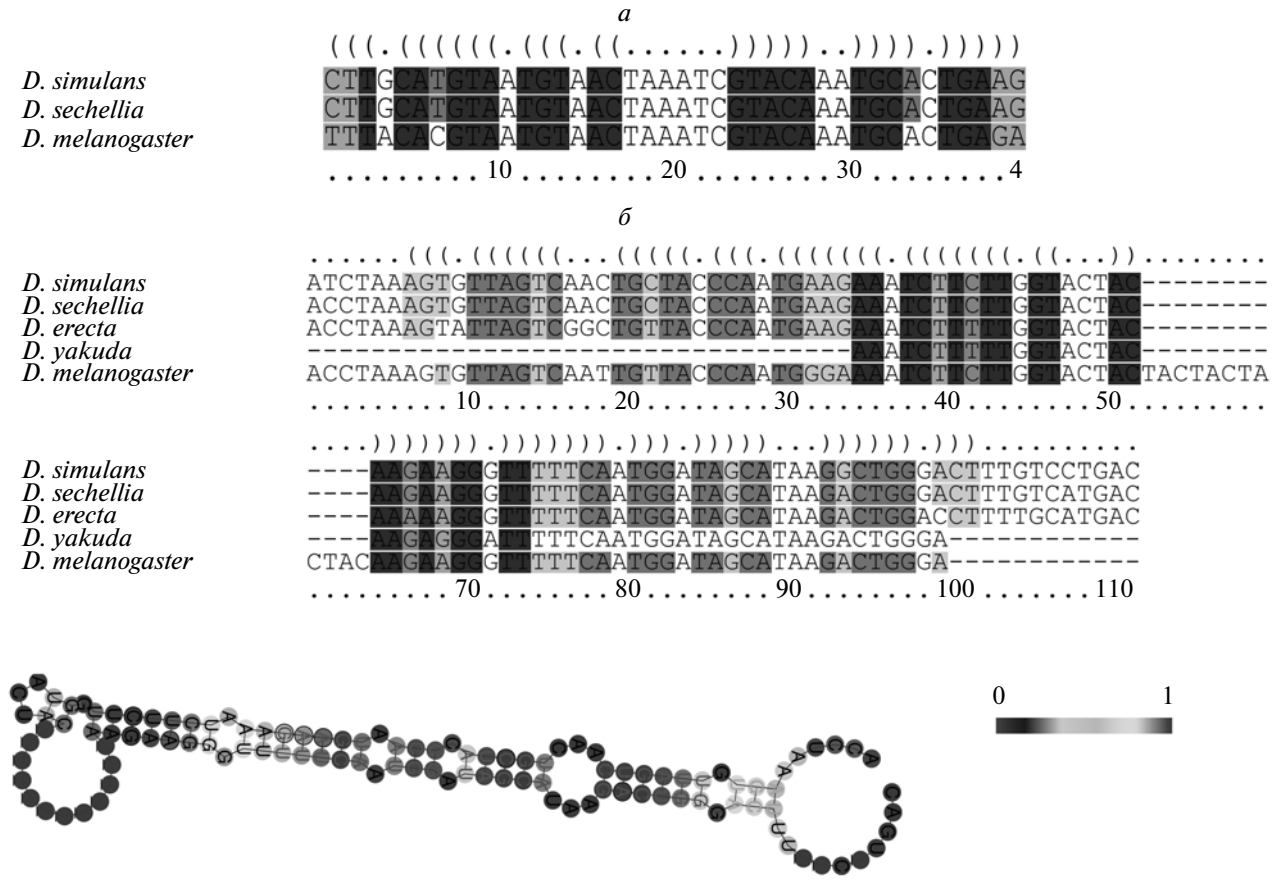

Рис. 3. *а* – Пример компенсаторной замены в 3'-нетранслируемой области гена *r-cap*; *б* – структурированный неаннотированный участок в интроне гена *Mhc*. На выравнивании светло-серым столбцом отмечены компенсаторные замены. Рисунок получен с помощью программы RNAalifold [21].

Компенсаторные замены

В стабильной вторичной структуре РНК изменение одного спаренного нуклеотида может значительно изменить или разрушить структуру. Таким образом, для вторичных структур намного более распространены случаи компенсаторных замен: замены обоих спаренных оснований, сохраняющих структуру. Среди наших кандидатов были найдены примеры компенсаторных замен. Так, 3'-нетранслируемая область гена *r-cap* содержит структурированную область, в которой был обнаружен случай парной компенсаторной замены: пара GC заменилась на пару AT, а пара AT заменилась на пару GT (рис. 3*а*).

Сравнение с Evofold

Программа EvoFold [9] ищет структурированные элементы, основываясь на анализе возможных компенсаторных замен в множественных выравниваниях. Обнаружено 1701 пересечение с предсказаниями EvoFold по геному *D. melanogaster* (при FDR = 10%). При этом 14150 наших предсказаний не пересекаются с предсказаниями EvoFold, а 21043 предсказаний EvoFold – с нашими. Оценка значимости корреляции между расположением структурированных участков, полученных с помощью нашего алгоритма, и участков, предсказанных Evofold, проведена с помощью R пакета GenometriCorr [22]. Результаты статистических тестов показали, что структурированные участки, найденные с использованием нашего алгоритма, расположены не случайно относительно участков EvoFold (p-value < 10^{-10}).

Сопоставление с транскрипционными данными

Найденные структурированные участки, лежащие в межгенных областях, протестированы на пересечение с данными по транскрипции, доступными в открытом доступе. Показано, что 23% структурированных участков, лежащих в межгенных областях, пересекается хотя бы с одним из экспериментально подтвержденных транскриптов.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Мы провели полногеномный поиск структурированных нкРНК в геноме *D. melanogaster*, используя алгоритм, определяющий меру структурированности участка в "скользящем окне" путем подсчета вероятностей нуклеотидов находиться в спаренном состоянии. Известно, что размеры нкРНК могут варьироваться от нескольких десятков до тысяч нуклеотидов. Однако даже в больших структурах взаимодействующие нуклеотиды часто находятся на расстоянии 40–200 н. Чтобы покрыть этот спектр длин, мы осуществляли поиск с окнами 40, 70, 100, 150 н.

В настоящее время в геноме *D. melanogaster* аннотировано менее 1000 функциональных структурированных нкРНК, в то время как последние исследования транскрипционных данных показывают, что вероятнее всего таких РНК в разы больше. Однако оценивать количество нкРНК надо с большой осторожностью, т.к. отсутствие адекватного и общепринятого метода оценки количества ложноположительных наблюдений для большинства алгоритмов предсказания нкРНК сильно усложняет задачу.

Основной подход для таких оценок – создание "случайной", фоновой последовательности, которая с одной стороны не содержит структурированных участков в силу своей случайности, но с другой стороны сохраняет свойства конкретного генома, в частности нуклеотидный и динуклеотидный состав. Кроме того, при сравнительно-геномном анализе необходимо порождать последовательности с заданными свойствами и на заданных эволюционных расстояниях. Для этих целей мы случайным образом отобрали 500000 кусков выравниваний длиной 30 н., перемешали каждый из них и соединили. Полученные "геномы" будут иметь такие же эволюционные расстояния и распределения консервативных участков по геному. При этом не будет структурированных участков, так как локальные перемешивания разрушают структурированные "шпильки". Также мы использовали косвенные подтверждения, такие как пересечения с уже аннотированными функциональными нкРНК, транскрипционные данные, сравнение с предсказаниями других программ, присутствие компенсаторных замен.

Показано, что наш алгоритм имеет высокий уровень чувствительности при обнаружении таких функциональных нкРНК, как микроРНК, тРНК, мяРНК и мякРНК. Таким образом, можно предполагать, что найденные нкРНК содержат в том числе новые, еще не аннотированные нкРНК таких типов. Например, установлено, что интрон гена тяжелой цепи миозина содержит структурированный участок, не аннотированный ранее (рис. 3*б*). Структурированность данного участка подтверждена с помощью программы RNAailfold [23], а также компенсаторными заменами, обозначенными светло-серым фоном на выравнивании.

Большинство найденных нкРНК лежат в межгенных или интронных областях. Мы провели анализ структур, полностью лежащих в интронных областях. Показано, что 719 из 13431 структур пересекаются с аннотацией Evofold (3% от общего числа структур, предсказанных Evofold), 129 с нкРНК, аннотированными в проекте FlyBase [24] (10% от общего числа структур в FlyBase).

Среди оставшихся нкРНК многие структуры могут также быть новыми микроРНК или нкРНК других типов.

Многие структуры, найденные в межгенных областях также могут оказаться функциональными. Например, на хромосоме 2R между генами *Arc1* и *Arc2* предсказана новая структура. При этом наличие данной РНК подтверждено транскрипционными данными.

Кроме того, показано, что найденные структурированные нкРНК часто находятся в нетранслируемых областях мРНК, как 3', так и 5'. Потенциальную структурированность 3'- и 5'-областей мРНК можно использовать для посттранскрипционного контроля транспорта ядро-цитоплазма, эффективности трансляции, внутриклеточной локализации и стабильности мРНК [25]. Эксперименты на дрожжах по определению структурированности мРНК показали, что в среднем 3'- и 5'-нетранслируемые области менее структурированы, чем кодирующие области, а также вокруг старт- и стоп-кодонов отсутствуют локальные структуры [26]. Мы отобрали мРНК, 5'- и 3'-нетранслируемые области которых содержат предсказанные структурированные участки. Получили, что 3% всех мРНК имеют структурированные 5'-нетранслируемые области и 9% – 3'-нетранслируемые области. Мы выровняли для отобранных мРНК 5'-нетранслируемые области относительно старт-кодонов и построили график структурированности (рис. 4). Наш результат согласуется с экспериментальными данными: участок перед старт-кодоном не структурирован, так как потенциальная структурированность может мешать посадке рибосомы, поэтому область перед старт-кодоном не должна содержать локальных вторичных структур [27].

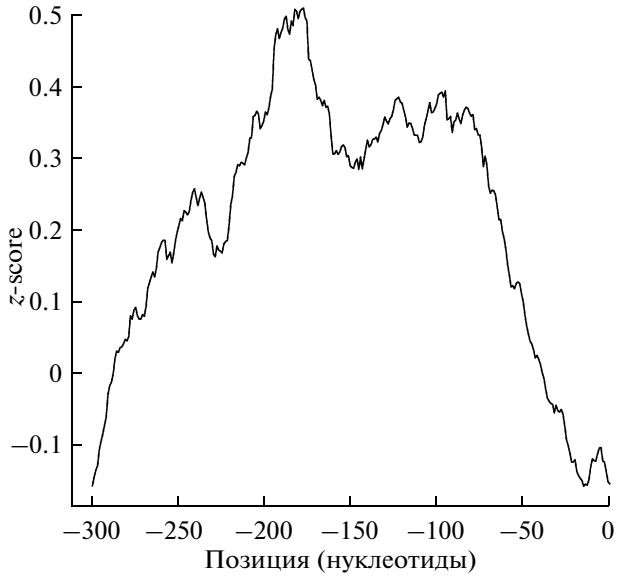

Рис. 4. Структурированность 5'-нетранслируемых областей мРНК, выровненных относительно старт-кодонов.

ЗАКЛЮЧЕНИЕ

Таким образом, наш алгоритм позволяет полногеномно предсказывать структурированные нкРНК, часть из которых может быть функциональна. В числе плюсов данного подхода необходимо отметить возможность поиска структур даже в отсутствие хорошего выравнивания, т.е. практически во всем геноме: при анализе мы рассматривали 97% генома *D. melanogaster*, в то время как алгоритм Evolfold использует лишь 3.7% генома (при анализе геномов позвоночных).

Отметим, что есть подходы, учитывающие динуклеотидный состав последовательности при анализе распределений. Мы не проводим аналогичную процедуру, так как вполне вероятно, что смещение динуклеотидного состава обусловлено структурированностью участка. В таких случаях процедура нормировки будет логически ошибочной. Правда, такой подход может быть оправдан при изучении позвоночных, где логично вычислять параметры распределений для каждой изохоры отдельно.

Среди минусов нашего алгоритма можно выделить использование “скользящего окна”, и таким образом направление дальнейших исследований – снятие этого ограничения и поиск всех локально оптимальных структур.

Исследование выполнено при поддержке Министерства образования и науки Российской Федерации, соглашение 8283.

СПИСОК ЛИТЕРАТУРЫ

1. 2004. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*. **306**, 636–640.
2. Consortium I.H.G.S. 2004. Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931–945.
3. Wang Z., Gerstein M., Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.
4. Zuker M., Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9**, 133–148.
5. Clote P., Lou F., Lorenz W.A. 2012. Maximum expected accuracy structural neighbors of an RNA secondary structure. *BMC Bioinformatics*. **13 Suppl 5**, S6.
6. Bernhart S.H., Hofacker I.L., Stadler P.F. 2006. Local RNA base pairing probabilities in large sequences. *Bioinformatics*. **22**, 614–615.
7. Rivas E., Eddy S.R. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*. **16**, 583–605.
8. Gorodkin J., Hofacker I.L., Torarinsson E., Yao Z., Havgaard J.H., Ruzzo W.L. 2010. De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol*. **2***8*, 9–19.
9. Pedersen J.S., Bejerano G., Siepel A., Rosenbloom K., Lindblad-Toh K., Lander E.S., Kent J., Miller W., Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol*. **2**, e33.
10. Rivas E., Eddy S.R. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*. **2**, 8.

11. Washietl S., Hofacker I.L., Stadler P.F. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*. **102**, 2454–2459.
12. Nussinov R., Pieczenik G., Griggs J.R., Kleitman D.J. 1978. Algorithms for loop matchings. *SIAM J. Appl. Mathematics*. **35**, 68–82.
13. Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy Stat. Soc. B Met*. **57**, 289–300.
14. Dreszer T.R., Karolchik D., Zweig A.S., et al. 2012. The UCSC Genome Browser database: extensions and updates 2011. *Nucl. Acids Res*. **40**, D918–923.
15. Blanchette M., Kent W.J., Riemer C., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. **14**, 708–715.
16. Kozomara A., Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucl. Acids Res*. **39**, D152–157.
17. Chan P.P., Lowe T.M. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucl. Acids Res*. **37**, D93–97.
18. Lestrade L., Weber M.J. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucl. Acids Res*. **34**, D158–162.
19. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006.
20. Celniker S.E., Dillon L.A., Gerstein M.B., et al. 2009. Unlocking the secrets of the genome. *Nature*. **459**, 927–930.
21. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079.
22. Favorov A., Mularoni L., Cope L.M., Medvedeva Y., Mironov A.A., Makeev V.J., Wheelan S.J. 2012. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol*. **8**, e1002529.
23. Bernhart S.H., Hofacker I.L., Will S., Gruber A.R., Stadler P.F. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*. **9**, 474.
24. Crosby M.A., Goodman J.L., Strelets V.B., Zhang P., Gelbart W.M. 2007. FlyBase: genomes by the dozen. *Nucl. Acids Res*. **35**, D486–491.
25. Pesole G., Mignone F., Gissi C., Grillo G., Licciulli F., Liuni S. 2001. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*. **276**, 73–81.
26. Kertesz M., Wan Y., Mazor E., Rinn J.L., Nutter R.C., Chang H.Y., Segal E. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Natur*e. **467**, 103–107.
27. Araujo P.R., Yoon K., Ko D., Smith A.D., Qiao M., Suresh U., Burns S.C., Penalva L.O. 2012. Before it gets started: regulating translation at the 5' UTR. *Comp. Funct. Genomics*. **2012**, 475731.