ИМПБ РАН 50 лет. Лаборатория биоинформатики

Назипова Н.Н.

Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша РАН

nnn@impb.ru

The IMPB RAS is 50 years old. Bioinformatics Laboratory

Nazipova Nafisa

[1]*Institute of Mathematical Problems of Biology of RAS – the Branch of Keldysh Institute of Applied Mathematics of RAS*

Исследования в области биоинформатики в нашем институте ведутся непрерывно с начала 80-х годов 20-го века, когда в мире утвердилось новое направление компьютерной биологии, связанное с появлением экспериментальных методов секвенирования и последующей компьютерной обработки нуклеотидных последовательностей. За это время было создано множество оригинальных алгоритмов и прикладных программ для изучения структурно-функционального строения геномов, с их помощью получены интересные результаты [1–15].

Основные направления современных исследований лаборатории биоинформатики

Распознавание неканонических микроРНК

До недавнего времени считалось, что главное содержание генома – это гены, кодирующие белки в процессе трансляции матричной РНК (мРНК). В результате секвенирования генома человека было определено, что в геноме *Homo sapiens*, длина которого приблизительно равна 3 Гб, содержится около 20 тысяч генов, кодирующих белок, что составляет меньше 2 % от общей геномной последовательности. Остальную часть генома назвали мусорной, или эгоистичной. Когда выяснилось, что эукариотический почвенный червь длиной 1мм *Caenorhabditis elegans*, геном которого в 30 раз короче человеческого (1 Мб), имеет такое же количество белковых генов, стало очевидно, что сложность развития и физиологическая сложность биологического организма не может находиться в прямой зависимости от количества белокодирующих генов.

Бурное развитие высокопроизводительных технологий секвенирования и компьютерного исследования геномов вызвало рост числа работ, касающихся возможной функциональной нагрузки на другую часть генома, которая не кодирует белки, и поэтому названа некодирующей. Оказалось, что продуктивной транскрипции подвержены не только области, кодирующие белки, но почти весь геном млекопитающих. Был открыт феномен всепроникающей транскрипции, клеточной активности в результате которой более 90 % генома человека может порождать множество взаимно перекрывающихся молекул РНК, каждая из которых транскрибируется хотя бы в одном типе клеток, хотя бы на одном этапе онтогенеза. До сих пор ведутся споры о том, представляет ли пул жизнеспособных транскриптов лишь транскрипционный «шум», или это функциональные некодирующие РНК (нкРНК), функции, которые просто еще не идентифицированы. В работе [16] сделан обзор нкРНК с известными функциями.

МикроРНК, открытые около 30 лет назад, сейчас представляют собой самый изученный класс молекул РНК, которые играют важную регуляторную роль у различных видов эукариот. Начало изучения этих молекул связано с открытием феномена РНК-интерференции, что произвело революцию в понимании регуляции генов, выявив ряд связанных с этим явлением путей в клетке, в которых небольшие (длиной 20-30 нуклеотидов) нкРНК и ассоциированные с ними белки контролируют экспрессию генетической информации. Хорошо изученный механизм биогенеза микроРНК, обычно сопровождающий описание РНК-интерференции, касается только канонических микроРНК, их первичные транскрипты первоначально расщепляются ядерным ферментом Drosha с образованием шпилек-предшественников микроРНК. После экспорта в цитоплазму шпильки превращаются в дуплексы микроРНК под действием фермента Dicer. Одна дуплексная цепь длиной 22 нуклеотида сохраняется в комплексе белком из семейства Argonaute, она направляет его к комплементарным мишеням на целевой матричной РНК, тогда как ее комплементарная цепь miRNA* разрушается.

Помимо канонического пути, были обнаружены неканонические пути биогенеза микроРНК с участием РНКаз, которые функционируют в других процессах. Главным среди них является путь миртронов, в котором расщепление Drosha заменяется действием сплайсосомы.

Неканонические микроРНК используют специфический путь транскрипции, отличный от пути созревания канонических микроРНК. Они используют машину сплайсинга, которая работает при созревании всех мРНК генов, кодирующих белки у эукариот. Неканонические микроРНК кодируются в интронах белок-кодирующих генов, которые имеют большое число альтернативных транскриптов.

Сравнительное исследование репертуаров миртронов у человека и мыши [17] показало, что в этих геномах кодируются по меньшей мере 478 и 488 миртронов, соответственно, из них 13 общих. Из числа общих только 3 имеют консервативную последовательность, а 10 регулируют трансляцию одинаковых по функции генов у этих двух организмов. В работе были проанализированы библиотеки глубокого секвенирования РНК из клеток разных тканей человека и мыши, в результате были получены достоверные выборки неканонических микроРНК, существование которых экспериментально подтверждено.

С использованием этих достоверных выборок нами ведется работа по определению закономерностей интронов, которые кодируют миртроны, а также особенностей строения и термодинамических параметров вторичных структур миртронов. Это необходимо для создания методов распознавания потенциальных миртронов в интронах белок-кодирующих генов.

Распознавание сильно размытых периодических участков геномов

Изучение длинных периодичностей в геноме важно для того, чтобы найти подтверждения гипотезам о роли периодической ДНК в геноме. Считается, что всего повторяющейся ДНК в геноме человека – около 50 %, в этой фракции находятся все виды периодичности: повторяющиеся участки в кодирующей и в не кодирующей белки частях генома. Последняя занимает более 90 % генома человека.

Нами ранее был разработан метод обнаружения размытых тандемных повторов в геномах с любой длиной паттерна и любой степенью сохранности паттерна [8, 18, 19]. Реализация этого подхода функционирует в виде сервера (http://www.mathcell.ru/model5.php?l=ru) и решает проблему нахождения периодичностей с постоянной длиной копий паттерна. В основу технологии извлечения информации о периодичности в геномах положен разработанный коллективом спектрально-статистический подход [8], применение которого позволило создать комплекс программ, достоверно выявляющих статистически значимые размытые тандемные повторы.

Теоретический предельный уровень дивергенции копий паттерна в тандемных повторах, выявляемых этим подходом, составляет 50 %. Этот метод одинаково хорошо выявляет как микро- и минисателлиты с длиной паттерна от 2 нукл. и более, так и периодичности с длиной паттерна свыше 100 нукл. (макросателлиты).

В процессе анализа геномов мыши и крысы с помощью визуального интерфейса программы спектрального анализа геномов [20–22] нами был обнаружен новый вид скрытой периодичности – мегасателлитные тандемные повторы. Они характерны тем, что имеют длину паттерна порядка более 2 тысяч позиций, а также большую вариабельность в длинах повторов паттерна, которая обусловлена наличием треков переменной длины, представленных простыми повторами, например, ATAT…(это SSR, Simple Sequence Repeats), при достаточно консервативной базовой последовательности (core sequence) паттерна периодичности. Для выявления такого вида периодичностей в геномах мы разработали новый подход, использующий методики выравнивания коротких прочтений в процессе референсной сборки генома. Такие прочтения получаются на выходе секвенаторов.

Для сборки консенсусной последовательности (генома) из набора коротких многократно перекрывающихся между собой фрагментов прочтений разработаны алгоритмы выравнивания, использующие преобразование Барроуза – Уиллера (Burrows-Wheeler Transform, BWT) [22] при обратном обходе префиксных деревьев [23], и позволяющие эффективно выравнивать участки ДНК с тем, чтобы определить степень их сходства. При этом допускаются несовпадения, вставки и выпадения фрагментов. Этот подход быстрее широко используемого алгоритма Needleman-Wunsch [24].

В основе нашего метода поиска мегасателлитных последовательностей лежит применение суффиксных массивов для сжатой последовательности ДНК – сначала все SSR с длинами паттернов от 1 до 3 сжимаются до одной копии. Мы находим все неточные вхождения (задается число допустимых замен на длину затравки) каждой из затравок (seed) и расширяем обнаруженные участки неидеального локального совпадения насколько это возможно в обе стороны путем выравнивания по Needleman-Wunsch. Затравками являются все k-меры (k – параметр алгоритма), которые имеют более 2 вхождений в исследуемую последовательность. Этот метод выявляет разнесенные неточные повторы переменной длины, тандемные повторы являются частным случаем.

Классификация вирусов по особенностям использования кодонов

В сотрудничестве с Омским государственным медицинским университетом Минздрава России лаборатория ведет работу по разработке методов оперативного определения видовой принадлежности вредоносных для здоровья человека вирусов. Для этого используются статистические характеристики распределений кодонов в структурных белках вирусов. При создании метод применялся к классификации видов флавивирусов, тогда использовалась статистика, основанная на величине отклонения распределения кодонов в гене полипротеина вновь секвенированного образца от усредненного распределения [25].

В результате изучения модели организации кодирования структурных белков коронавируса [26] SARS–CoV–2 создан надежный метод, получивший название вариантного подхода. Он был применен к распознаванию рода коронавируса. Вариантный подход использует для распознавания как комбинацию нескольких структурных генов, так и отдельные гены, секвенирование которых займет меньше времени, чем определение последовательности всего генома [27].

Для создания в каждом роде коронавирусов усредненного референсного распределения кодонов использовались соответствующие гены прототипных (образцовых) штаммов рода. В настоящее время род *Alphacoronavirus* (α–CoV) подразделяется на 12 подродов (22 прототипных штамма), род *Betacoronavirus* (β–CoV) – на пять подродов (28 прототипных штаммов), род *Deltacoronavius* (δ–CoV) – на четыре подрода (10 прототипных штаммов) и род *Gammacoronavirus* (γ–CoV) – на два подрода (7 прототипных штаммов).

Программное обеспечение для анализа экспериментов по изучению секретомов бактерий

Разработка методик для исследования данных экспериментов RNA-seq для фракций коротких внутриклеточных и экстраклеточных РНК бактерий и для фракций РНК, секретируемых смешанными популяциями микроорганизмов очень актуальна для изучения еще одного механизма регуляции экспрессии белков.

Массовая секреция ДНК, мРНК и нкРНК до недавнего времени была известна только у эукариот. Фракции эукариотических РНК, секретируемые в составе мембранных везикул, уже подробно исследованы. Было установлено, что они имеют внеклеточные регуляторные функции, связанные с такими важными процессами, как онкогенная трансформация, клеточная пролиферация, окислительный стресс и даже гибель клеток.

Бактериальные внеклеточные РНК были обнаружены только недавно, и пока мало изучены. Требует особого изучения их участие во внутривидовой и межвидовой сигнализации. Однако подходы, разработанные для идентификации и характеристики секретируемых РНК эукариот, используют специфические особенности эукариотических транскриптов и непригодны для картирования бактериальных секретомов. С помощью созданных сотрудниками лаборатории методов в сотрудничестве с лабораторией функциональной геномики и клеточного стресса ИБК РАН (зав. лаб. О.Н. Озолинь) удалось решить ряд задач бактериальной геномики. В частности, стало возможно изучение фракции РНК, секретируемых бактериями нескольких видов в разных условиях роста, включающее исследование закономерностей их биосинтеза и процессинга, а также способности влиять на экспрессию генов в реципиентных клетках.

Созданы и инкорпорированы в общий программный комплекс 64bit консольные вычислители, анализирующих наличие уникальных относительно совокупности мишеней коротких олигонуклеотидов (*k*-меров, $k = 16, 18, 20$ или 22), в геномах *E. coli* и двух модельных бактерий, чтобы определить, какие из них преимущественно секретируются *E. coli* в ответ на присутствие другой бактерии. Это позволило выявить потенциально активные внеклеточные РНК и впервые продемонстрировать способность их синтетических аналогов проникать в клетки *E. coli* и влиять на их рост [28].

Созданные подходы в метагеномных исследованиях, использующие *k*-меры в качестве бар-кодов для идентификации отдельных геномов, создали новую стратегию таксономического анализа и проложили путь к филогении с высоким разрешением. Используя этот подход для филотипирования *Lacticaseibacillus paracasei* ВКМ В-1144 на уровне рода, были выделены четыре филогруппы *L. paracasei* и установлено, что *L. casei* 12A принадлежит к одной из них, а не к кладе *L. casei*, т.е. спецификация этой клады требует модификации. На уровне рода мы обнаружили только одного родственника *L. paracasei* ВКМ В-1144 среди 221 генома, полного или имеющегося в контигах, и показали, что кодирующий потенциал генома этого «редкого» штамма позволяет рассматривать его как потенциальный пробиотический компонент. Четыре набора опубликованных метагеномов использовались для оценки зависимости присутствия *L. paracasei* в микробиоме кишечника человека от хронических заболеваний, диетических изменений и лечения антибиотиками [29]. В данном исследовании наиболее значимым представляется результат, свидетельствующий о благоприятном изменении состава родной микробиоты. Даже после однократного приёма добавок процентное содержание многих потенциально патогенных бактерий (отслеживали 27

таксонов) достоверно уменьшалось, в то время как уровень присутствия большинства доминирующих в резидентной микрофлоре таксонов поддерживался на приблизительно постоянном уровне.

4. Список литературы

1. Лунина Н.Л. *Система обработки нуклеотидных последовательностей HEID*. Пущино: ОНТИ НЦБИ, 1984. (Материалы по математическому обеспечению ЭВМ, выпуск 9).
2. Vernoslov S., Kondrashov A., Roitberg M., Shabalina S., Yureva O., Nazipova N. SAMSON program package for primary structure-analysis of biopolymers. *Molecular Biology*. 1990. Т. 24. № 2. С. 430–435.
3. Nazipova, N.N., Shabalina, S.A., Ogurtsov A.Yu., Kondrashov, A.S., Roytberg M.A., Buryakov, G.V., Vernoslov, S.E. SAMSON: a software package for the biopolymer primary structure analysis. *CABIOS*. 1995. V. 11. No. 4. P. 423–426.
4. Kislyuk O.S., Borovina T.A., Nazipova N.N. Estimation of Redundancy of Genetic Texts by the High Frequency Component of the *l*-Gram Graph. *Biophysics*. 1999. V. 44. No. 4. P. 621–630.
5. Borovina T.A., Nazipova N.N., Overbeek R., Gelfand M.S. Statistical Analysis and Prediction of Prokaryote Ribosomal Binding Sites. *Biophysics*. 1999. V. 44. No. 4. P. 594–601.
6. Чалей М.Б., Назипова Н.Н., Кутыркин В.А. Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях. *Математическая биология и биоинформатика*. 2007. Т 2. № 1. С. 20–35. doi: 10.17537/2007.2.20
7. Назипова Н.Н., Елькин Ю.Е., Панюков В.В., Дроздов-Тихомиров Л.Н. Расчёт скоростей метаболических реакций в живой растущей клетке методом баланса стационарных метаболических потоков (метод БСМП). *Математическая биология и биоинформатика.* 2007. Т. 2. №1. С. 98–119. doi: 10.17537/2007.2.98
8. Chaley M.B., Nazipova N.N., Kutyrkin V.A. Statistical Methods for Detecting Latent Periodicity Patterns in Biological Sequences: The Case of Small-Size Samples. *Pattern Recognition and Image Analysis*. 2009. V. 19. No. 2. P. 358–367.
9. Панюков В.В., Назипова Н.Н., Озолинь О.Н. Пакет программ aSHAPE для изучения пространственной конформации участков бактериального генома. *Математическая биология и биоинформатика.* 2011. Т. 6. № 2. С. 211–227. doi: 10.17537/2011.6.211
10. Matveeva O.V., Nazipova N.N., Ogurtsov A.Y., Shabalina S.A. Optimized models for design of efficient miR30-based shRNAs. *Frontiers in Genetics.* 2012. V. 3. Article No. 163.
11. Matveeva O.V., Nechipurenko Yu.D., Riabenko E., Ragan Ch., Nazipova N.N., Ogurtsov A.Y., Shabalina S.A. Optimization of signal-to-noise ratio for efficient microarray probe design. *Bioinformatics*. 2016. V. 32. No. 17. P. i552–i558. doi: 10.1093/bioinformatics/btw451
12. Nazipova N.N., Isaev E.A., Kornilov V.V., Pervukhin D.V., Morozova A.A., Gorbunov A.A., Ustinin M.N. Big Data in Bioinformatics**.** *Math. Biol. Bioinf*. 2018. V. 13. No. S. P. t1–t16. doi: 10.17537/2018.13.t1
13. Panyukov V.V., Kiselev S.S., Alikina O.V., Nazipova N.N., Ozoline O.N. Short Unique Sequences in Bacterial Genomes as Strain- And Species-Specific Signatures. *Mathematical Biology and Bioinformatics*. 2017. V. 12. № 2. P. 547–558. doi: 10.17537/2017.12.547
14. Matveeva O.V., Ogurtsov A.Y., Nazipova N.N., Shabalina S.A. Sequence characteristics define trade-offs between on-target and genome-wide off-target hybridization of oligoprobes. *PLoS ONE*. 2018. V. 13. No. 6. Article No. e0199162. 10.1371/journal.pone.0199162
15. Nazipova N.N., Shabalina S.A. Understanding off-target effects through hybridization kinetics and thermodynamics. *Cell Biol. Toxicol.* 2020. V. 6. P. 11–15. doi: 10.1007/s10565-019-09505-4
16. Назипова Н.Н. Разнообразие некодирующих РНК в геномах эукариот. *Математическая биология и биоинформатика*. 2021. Т. 16. № 2. С. 256–298. doi: 10.17537/2021.16.256
17. Wen J., Ladewig E., Shenker S., Mohammed J., Lai E.C. Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates. *PLoS Comput. Biol*. 2015. V. 11. № 9. Article No. e1004441. doi: 10.1371/journal.pcbi.1004441
18. Чалей М.Б., Кутыркин В.А., Теплухина Е.И., Тюльбашева Г.Э., Назипова Н.Н. Исследование феномена скрытой периодичности в геномах эукариотических организмов. *Математическая биология и биоинформатика*, 2013. V. 8. № 2. P. 480–501. doi: 10.17537/2013.8.480
19. Chaley M., Kutyrkin V., Tulbasheva G., Teplukhina E., Nazipova N. HeteroGenome: database of genome periodicity. *Database: the journal of biological databases and curation*. 2014. V. 2014. P. 1–18. doi: 10.1093/database/bau040
20. Tetuev R.K., Nazipova N.N. Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern. *Repbase Reports*. 2010. V. 10. № 5. P. 776–776.
21. Tetuev R.K., Dedus F.F., Nazipova N.N. Consensus of repeated region of rat chromosome 4 similar to mouse chromosome 6 repeated region, enclosed in the intergenic region between genes Hrh1 and Atg7. *Repbase Reports*. 2010. V. 10. № 8. P. 1185–1185.

22. Burrows M., Wheeler D.J. *A Blocksorting Lossless Data Compression Algorithm*. Palo Alto, Calif: 1994. (SRC research reports, Vol. 124).
23. Li H., Durbin R. *Bioinformatics*. 2009. V. 25. P. 1754–1760. doi: 10.1093/bioinformatics/btp324
24. Needleman S.B., Wunsch C.D. *J. Mol. Biol*. 1970. V. 48. № 3. P. 443-453. doi: 10.1016/0022-2836(70)90057-4
25. Чалей М.Б., Тюлько Ж.С., Кутыркин В.А. Распознавание видов флавивирусов на основе кодирующих последовательностей полипротеинов. *Математическая биология и биоинформатика*. 2019. Т. 14. No. 2. С. 533-542. doi: 10.17537/2019.14.533
26. Чалей М.Б., Тюлько Ж.С., Кутыркин В.А. Исследование структуры кодирования ORF1ab, S, M и N генов коронавирусов. *Математическая биология и биоинформатика*. 2020. Т. 15. № 2. С. 441–454. doi: 10.17537/2020.15.441
27. Чалей М.Б., Кутыркин В.А. Распознавание рода коронавируса на основе прототипных штаммов. *Математическая биология и биоинформатика*. 2022. Т. 17. № 1. С. 10–27. doi: 10.17537/2022.17.10
28. Markelova N., Glazunova O., Alikina O., Panyukov V., Shavkunov K., Ozoline O. Suppression of Escherichia coli growth dynamics via RNAs secreted by competing bacteria. *Front. Mol. Biosci*. 2021. V. 8. Article № 609979.
29. Frolova M., Yudin S., Makarov V., Glazunova O., Alikina O., Markelova N., Kolzhetsov N., Dzhelyadin T., Shcherbakova V., Trubitsyn V., Panyukov V., Zaitsev A., Kiselev S., Shavkunov K., Ozoline O. Lacticaseibacillus paracasei: occurrence in the human gut microbiota and k-mer-based assessment of intraspecies diversity. *Life*. 2021. V. 11. Article № 1246. doi: 10.3390/life11111246