9. McMackin E. A. W., Marsden A. E., Yahr T. L. H-NS Family Members MvaT and MvaU Regulate the *Pseudomonas Aeruginosa* Type III Secretion System // J. Bacteriol. 2019. Vol. 201. № 14. P. 1–15.
10. Chambonnier G., Roux L., Redelberger D. et al. The Hybrid Histidine Kinase LadS Forms a Multicomponent Signal Transduction System with the GacS/GacA Two-Component System in *Pseudomonas Aeruginosa* // PLoS Genet. 2016. Vol. 12. № 5. P. 1–30.
11. Cooper C. E., Nicholls P., Freedman J. A. Cytochrome *c* oxidase: structure, function, and membrane topology of the polypeptide subunits // Biochem. Cell Biol. 1991. Vol. 69. № 9. P. 586–607.
12. Cabeen M. T., Leiman S. A., Losick R. Colony-morphology screening uncovers a role for the *Pseudomonas aeruginosa* nitrogen-related phosphotransferase system in biofilm formation // Mol. Microbiol. 2016. Vol. 99. № 3. P. 557–570.

НЕОБХОДИМОСТЬ РУЧНОЙ ВАЛИДАЦИИ ДАННЫХ АВТОМАТИЧЕСКОЙ СБОРКИ И АННОТАЦИИ ГЕНОМОВ НА ПРИМЕРЕ МИТОХОНДРИАЛЬНЫХ ГЕНОМОВ НАСЕКОМЫХ

С. С. Левыкина, Н. В. Воронова

Белорусский государственный университет, Минск, Беларусь
E-mail: s.lewykina@yandex.by

Рассмотрена необходимость ручной валидации полученных данных автоматической сборки и аннотации геномов. Ручная валидация применима при анализе небольших геномов, таких, как митохондриальные геномы. Необходимо совершенствовать методы биоинформатики, чтобы необходимость ручной валидации отсутствовала.

Ключевые слова: методы биоинформатики; митохондриальный геном; тли.

На сегодняшний день для реализации таких процессов, как сборка и аннотация геномов, разработано множество доступных программных пакетов, позволивших автоматизировать выполнение такого рода задач [1–2]. Однако, при всём многообразии информационных технологий для обработки данных секвенирования, а также анализа нуклеотидных последовательностей, полученные результаты часто нуждаются в ручном пересмотре со стороны эксперта в данной конкретной области исследования.

В данной работе необходимость ручной валидации данных рассматривается на примере митогеномов насекомых, в частности тлей (Aphidoidea) [3].

Митохондриальный геном тлей представляет собой двуцепочечную ДНК-молекулу, размером обычно не превышающую 15–20 Kb. Структура генома обычно включает в себя 37 генов, среди которых 22 – гены

транспортных РНК, 2 – гены, кодирующие рибосомальные РНК, 13 – белок-кодирующие участки, а также 2 протяженных некодирующих участка, обозначенных как регион формирования D-петли и участок, богатый тандемными повторами [4]. Установлено, что митохондриальный геном тлей является высококонсервативной структурой, богатой аденин-тиминовыми основаниями (83–86%) [5]. В большинстве митогеномов тлей реализуется предковый для насекомых порядок генов [6]. Наличие любых генных транслокаций или делеций рассматривается как важный эволюционный и филогенетический признак.

База данных GenBank NCBI к моменту написания данной работы содержала 40 аннотированных митогеномов тлей, 5 из которых были собраны, аннотированы и депонированы сотрудниками СНИЛ биоинформатики и молекулярной эволюции животных [7]. Одной из пяти депонированных последовательностей был митохондриальный геном *Aphis craccivora*, сборка и аннотация которого была проведена по аналогии с уже представленным в генетическом банке геномом того же вида, депонированным китайскими исследователями в 2016 году (см. рис. 1.).

Митохондриальный геном данного вида тли представлял интерес для исследования в связи с присутствующей в нем транслокацией гена, кодирущего транспортную РНК аминокислоты тирозина (Y), место локализации которого обычно следует за геном тРНК-цистеина (C), а в работе китайских исследователей ген данной транспортной РНК был обнаружен в составе некодирующего участка, обозначенного как область формирования D-петли (d-loop) [8]. Заинтересовавшись данной перестройкой и её наличием в митохондриальном геноме представителя белорусской популяции *A. craccivora*, мы провели аналогичную работу по сборке и аннотации митогенома данного вида.

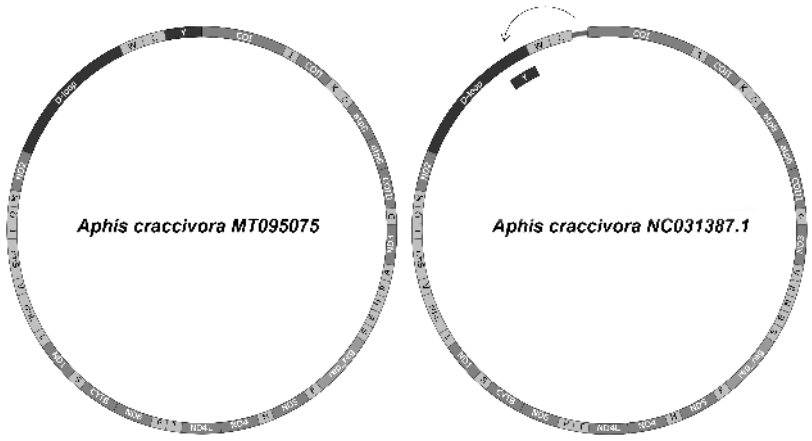

Рис. 1. Два варианта аннотации митохондриального генома *Aphis craccivora*: результаты, полученные авторами статьи [MT095075] и китайскими исследователями [NC031387.1]

В результате более детального исследования, проведенного авторами ранее [7], было установлено, что транслокации как таковой в геноме данного вида нет, и обнаруженная в составе области D-петли последовательность является изоформой соответствующей тРНК. Принятая за транспортную РНК китайскими исследователями изоформа, при выравнивании с другими тРНК-тирозина, извлеченными из остальных митогеномов в GenBank, демонстрировала визуально фиксируемые различия в нуклеотидном составе между ней и другими генами тРНК-тирозина. Также было установлено, что область, обозначенная китайскими исследователями как ген тРНК-тирозина, на самом деле является одним из важных консервативных участков D-петли, который не перекрывается с последующим геном в остальных митогеномах [9].

Причиной данной ошибки было то, что автоматическая аннотация не прошла ручную валидацию. Китайские исследователи обозначили старт-кодон белок-кодирующего гена COI примерно на 27 нуклеотидов раньше, чем это было сделано во всех остальных геномах. При переносе рамки считывания гена COI в типичную для нее локализацию, искомая тРНК обнаружилась между генами тРНК-цистеина и COI геном, а также сформировала типичную для нее вторичную структуру с характерным антикодоном.

Данная ситуация послужила поводом к пересмотру всех аннотаций митогеномов, депонированных в GenBank. В результате ручной валидации 35 аннотаций, депонированных сторонними исследователями, были обнаружены ошибки в 7 геномах, среди которых 4 содержало запись об отсутствующей перестройке, и в 4 не были обозначены границы региона повторов. Отсутствие в аннотациях данных о наличии данной области, при проведении сравнительного анализа с использованием депонированных в GenBank митогеномов, впоследствии привело бы к ошибкам.

Программы, обеспечивающие автоматическую аннотацию геномов, без сомнений, значительно сокращают затрачиваемое на исследование время и ресурсы, однако на практике оказывается недостаточным использование исключительно программных пакетов без проверки полученных данных компетентным в данной области исследователем. Ручная валидация данных, полученных в ходе автоматической аннотации, не представляет труда в случае анализа небольших геномов, таких как митохондриальные геномы, однако коррекция результатов аннотации полногеномных данных весьма затруднительна. В связи с этим, не оставляет сомнений необходимость совершенствования существующих методов биоинформатики, чтобы необходимость ручной валидации данных отпала как таковая, что само по себе открывает множество возможностей и задач для исследователей, специализирующихся в данной области.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Hahn C., Bachmann L., Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads – a baiting and iterative mapping approach // Nucleic Acids Res. 2013. Vol. 41. № 13. P. 1–9.
2. Seemann T. Prokka: rapid procaryotic genome annotation // Bioinformatics. 2014. Vol. 30. № 14. P. 2068–2069.
3. Воронова Н. В., Бондаренко Ю. В., Левыкина С. С., Шулинский Р. С. Митохондриальный геном Aphis Fabae Mordvilkoi Borner & Janisch, 1992 // Молекулярная и прикладная генетика: сб. науч. трудов. 2018. Т. 25. С. 73–83.
4. Lee J., Park J., Lee H. et al. The complete mitochondrial genome of *Paracolopha morrisoni* (Baker, 1919) (Hemiptera: Aphididae) // Mitochondrial DNA Part B. 2019. Vol. 4. № 2. P. 3037–3039.
5. Voronova N. V., Warner D., Shulinski R. S. et al. The largest aphid mitochondrial genome found in invasive species *Therioaphis tenera* (Aizerberg, 1956) // Mitochondrial DNA Part B. 2019. Vol. 4. № 1. P. 730–731.
6. Cameron S. L. Insect mitochondrial genomics: implications for evolution and phylogeny // Annu. Rev. Entomol. 2014. Vol. 59. P. 95–117.
7. Voronova N. V., Levykina S. S., Warner D. et al. Characteristic and variability of five complete aphid mitochondrial genomes: Aphis fabae mordvilkoi, Aphis craccivora, Myzus persicae, Therioaphis tenera and Appendiseta robiniae (Hemiptera; Sternorrhyncha; Aphididae) // International Journal of Biological Macromolecules. 2020. V. 149. P. 187–206.
8. Sun W., Huynh B. L., Ojo J. A. et al. Comparison of complete mitochondrial DNA sequences between old and new world strains of the cowpea aphid, *Aphis craccivora* (Hemiptera: Aphididae) // Agri Gene. 2017. Vol. 4. P. 23–29.
9. Chen L., Chen P.-Y., Xue X.-F. et al. Extensive gene rearrangements in the mitochondrial genomes of two egg parasitoids, *Trichogramma japonicum* and *Trichogramma ostriniae* (Hymenoptera: Chalcidoedea: Trichogrammatidae) // Scientific Reports. 2018. Vol. 8. № 1. P. 1–11.

АВТОЭНКОДЕРНАЯ НЕЙРОННАЯ СЕТЬ ДЛЯ ГЕНЕРАЦИИ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ ВИЧ-1 МЕТОДАМИ ГЛУБОКОГО ОБУЧЕНИЯ

Г. И. Николаев[1], Н. А. Шульдов[2], А. И. Анищенко[2], А. В. Тузиков[1], А. М. Андрианов[3]

[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь
[2]Белорусский государственный университет, Минск, Беларусь
[3]Институт биоорганической химии Национальной академии наук Беларуси, Минск, Беларусь
E-mail: reshaemvsem@gmail.com

Методами глубокого обучения разработан генеративный состязательный автоэнкодер для рационального дизайна потенциальных ингибиторов проникновения ВИЧ-1, способных блокировать участок белка gp120 оболочки вируса, критический для