Аннотация работы: Программный продукт metaSPAdes для решения задач сборки и анализа метагеномных данных микробиот различной природы

Авторы: младший научный сотрудник лаборатории “Центр алгоритмической биотехнологии” Мелешко Дмитрий Алексеевич и научный сотрудник лаборатории “Центр алгоритмической биотехнологии” Коробейников Антон Иванович

Для участия в конкурсе на соискание премии, присуждаемой Санкт-Петербургским государственным университетом за научные труды в 2019 году (в категории “за вклад в науку молодых исследователей”), выдвинуты следующие публикации:

Nurk S*, **Meleshko D***, **Korobeynikov A**, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017 May 1;27(5):824-34.
*равный вклад

Данная работа посвящена разработке алгоритмов метагеномной сборки. Метагеном представляет собой смесь бактерий в природной среде обитания. Метагеномные данные очень сильно различаются по размеру и сложности: от довольно простых метагеномов на поверхности человеческой кожи, составляющих несколько десятков различных организмов, до сложных почвенных метагеномов, которые могут содержать в себе тысячи различных штаммов. Задача сборки метагеномов состоит в определении геномной последовательности каждого микроорганизма, представленного в большом количестве, в этом образце. Важность сборки метагеномов сложно переоценить. В настоящее время проведено множество исследований, связывающих микробиом человека и заболевания: от аутизма до болезни Крона, исследование здоровой микробиоты позволяет предлагать лечение или реабилитацию с помощью пересадки микробиоты между людьми. Также микробиом может служить источником новых лекарственных препаратов, так как бактерии производят множество активных веществ, которые, в том числе, могут подавлять рост других бактерий. Алгоритмы геномной сборки принимают на вход случайные подстроки геномов, а результатом является так называемая геномная сборка — черновой вариант последовательности геномов организмов.

Задача сборки генома из коротких фрагментов является вычислительно крайне сложной из-за большого объема обрабатываемых данных, ошибок в считанных фрагментах, их относительно короткой длины (обычно 100-1000 нуклеотидов — единичных блоков геномной последовательности). В случае метагеномов к этому прибавляется наличие в данных близкородственных штаммов, горизонтально перенесенных генетических элементов и неравномерное представленность различных видов. Как следствие, проблема метагеномной сборки в общем виде не может быть решена с помощью только коротких фрагментов. Однако, современные метагеномные сборщики используют различные эвристические алгоритмы, позволяющие собирать большую долю высокопредставленых организмов в длинные последовательности - контиги. На данный момент существует несколько десятков метагеномных сборщиков для коротких прочтений.

Представленная на конкурс работа является расширением программы для сборки бактериальных геномов SPAdes (Bankevich et al., Journal of computational biology, 2012), разработка которого была начата в 2011 году в СПбАУ и продолжается в СПбГУ,

начиная с 2015 года. В 2015 году команда, состоящая из Нурка С.Ю. и Мелешко Д.А., начала разработку метагеномного сборщика metaSPAdes. Разработка metaSPAdes по прежнему активна. В течение двух лет с момента старта, была опубликована статья в журнале Genome Research, а также устный доклад был представлен на RECOMB-2016 (Лос-Анджелес, США). В настоящее время metaSPAdes используется в сотнях лабораторий по всему миру, программное обеспечение активно поддерживается, разрабатываются модули для работы с новыми типами данных (Tolstoganov et al., 2019, Bioinformatics) или решения задач, специфических для смежных областей (Meleshko et al., 2019, Genome Research). Вся разработанная научно-техническая продукция находится в открытом доступе на платформе GitHub и сайте лаборатории - http:// cab.spbu.ru/software/spades.

metaSPAdes комплексно решает проблемы сборки метагеномных данных, такие как наличие редких штаммов, большое количество данных, наличие межвидовых повторов, объединяя новые вычислительные идеи и приемы, хорошо зарекомендовавшие себя в более ранних инструментах семейства SPAdes. metaSPAdes осуществляет сборку в широком диапазоне значений покрытия геномов, обеспечивая разумный баланс между точностью и фрагментированностью сборок. При анализе смеси близкородственных штаммов metaSPAdes ставит задачу реконструкции их “консенсусной” последовательности, в том числе за счет потери фрагментов геномов относительно редких штаммов.

В процессе разработки значительное внимание было уделено сравнению metaSPAdes с тремя наиболее популярными на сегодняшний день метагеномными сборщиками: IDBA-UD, Ray-Meta и MEGAHIT, на метагеномных данных различной природы. Проведенный анализ показал, что сборки metaSPAdes менее фрагментированы, чем сборки других программ, при этом количество ошибок внесенных сборщиком редко превышает количество ошибок, которые вносят соперники. Таким образом, metaSPAdes в настоящее время является лучшим выбором для сборки метагеномных данных практически любой природы, что подтверждается большим количеством цитирований исследованиями опубликованными в журналах уровня Nature и Science.

Annotation of work: “metaSPAdes: a programming tool for analysis and assembly of metagenomic data of different origins”
Authors: Meleshko Dmitry Alexeevich, junior research fellow, the laboratory “Center for Algorithmic Biotechnology” and Korobeynikov Anton Ivanovich, research fellow, the laboratory “Center for Algorithmic Biotechnology”

To participate in the competition for the St. Petersburg State University award for scientific works in 2019 (in the category of “contribution to science made by young researchers”), the following publications are nominated:

Nurk S*, **Meleshko D***, **Korobeynikov A**, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome research. 2017 May 1;27(5):824-34.
*equal contribution

This work is dedicated to algorithm development for metagenomic assembly. Metageomic data is a mix of bacteria in their environment. Metagenomic data varies greatly in terms of size and complexity: from the most simple metagenomes on the human skin (starting from 10 species) to complex soil metagenomes (can have more than 1000 strains). Metagenomic assembly aim to determine the nucleotide sequence of each relatively abundant microorganism in the sample. It is hard to overestimate the utility of metagenomic assembly. Currently, there exist multiple studies that revealed connections between microbiome state and diseases: from autism to Crohn's disease. Multiple studies showed that microbiome transplantation from healthy human can cure some diseases or ameliorate symptoms. Finally, microbiome can be a rich source of novel drugs. Bacterias produce multiple biomedically active compounds encoded in their DNA. These compounds can have even antimicrobial effect on other species. Genome assembly algorithms take reads (small random substrings of the genome) as input. Result of genome assembly is draft genome sequences for all species.
Genome assembly from reads is extremely challenging because of large data volume, sequencing errors, relatively short length (100-300 nucleotides). For metagenomes we should also consider potential presence of multiple close strains and different abundance among species, and repetitive elements that are transferred horizontally between species. Thus, metagenomic assembly can’t be solved precisely by using exclusively short-reads. Modern metagenomic assemblers use different heuristics that allow them to assemble most of high covered species into sets of long sequences - contigs. Currently, there are dozens of short-read metagenomic assemblers.
The works submitted for the competition represent extensions of SPAdes -- genome assembly tools for isolate bacterial genomes (Bankevich et al., Journal of computational biology, 2012). SPAdes development started in 2011 and have been held in SPbU from 2015. In 2015 a team of Nurk S.Y. and Meleshko D.A. started devolopment of metagenomic assembler -- metaSPAdes that is still active. In the first two years we published an article about implemented algorithms in Genome Research journal and gave an oral presentation on RECOMB-2016 conference (Los-Angeles, USA). Today metaSPAdes is being used as the main metagenomic assembler in hundred laboratories around the globe, we provide full software support and develop modules for emerging data types (Tolstoganov et al., 2019, Bioinformatics) or tackling specific interdisciplinary problems (Meleshko et al., 2019,

Genome Research). metaSPAdes is freely accessible on GitHub of laboratory website - http://cab.spbu.ru/software/spades.

metaSPAdes comprehensively solves metagenomic assembly problems, such as rare strains in the data, large volumes, interspecies repeats, combining new computational algorithms with approaches that proves their utility in previous SPAdes-family tools. metaSPAdes is able to assemble genomes of completely different coverage, provides a reasonable tradeoff between contiguity and number of assembly errors. When dealing with related strains, metaSPAdes aims to reconstruct the most abundant strain (or “consensus” sequence) utilizing fragments of low-abundant strains.

During development we paid a lot of efforts to comprehensively compare metaSPAdes with the most popular metagenomic assemblers: IDBA-UD, Ray-Meta, MEGAHIT on datasets of different complexity. Analysis revealed that metaSPAdes assemblies are less fragmented that other assemblies. At the same time, the number of assembly errors introduced is comparable of less than in contender’s assemblies. Thus, currently metaSPAdes is the number one choice in metagenomic assembly of any kind, that is proven by the large number of citations and studies published in Nature and Science journals.