

Поиск распространяющихся микроорганизмов, основанный на анализе представленности k -меров в данных метагеномного секвенирования

Панова В.*, Попов Н., Федоров Д., Манолов А.

НИИ системной биологии и медицины Роспотребнадзора, Москва, Россия

* panova_vv@sysbiomed.ru

Ключевые слова: k -меры; счетчик k -меров; метагеномы; заболевания; пандемия

Мотивация и цель: После пандемии COVID-19 все больше внимания уделяется новым зоонозным инфекциям. Целью данной работы является разработка подхода к идентификации возможных инфекционных агентов на основе увеличения представленности и распространенности коротких подстрок фиксированной длины (k -меров) в данных полногеномного метагеномного секвенирования.

Методы и алгоритмы: Для валидации предложенного метода было проведено моделирование метагеномных прочтений при помощи утилиты InSilicoSeq [1]. Набор и значения представленности видов были получены из анализа секвенирования тотальной РНК выделенной из мазка из носоглотки пациента с COVID-19 (эксперимент SRA SRX8697677, запуск SRR12183113). Для подсчета канонических k -меров был использован Jellyfish [2]. Для оценки таксономического состава образца использовали Kraken2 [3].

Для проверки применимости подхода, нами была смоделирована временная серия из 7 шагов, выбран один организм (*Streptococcus* sp. S5) представленность которого линейно возрастала. Выбор организма основан на его представленности, близкой к медианному значению, рост относительно представленности производился с исходного значения до 61, 71 или 91 перцентиля. Длина k -мера при анализе была выбрана равной 10.

Для подбора оптимального значения длины k -меров был использован тот же набор изучаемых видов. В симуляцию также был добавлен шум: относительная численность каждого вида была распределена в соответствии с отрицательным биномиальным законом и изменялась в течение семи временных шагов. Количество прочтений данного соответствующего таксона принято за неотрицательное среднее распределение (μ). Длины результирующих k -меров были равны 10, 11, 12, 13, 14. Оценка роста производилась с помощью линейной модели (функция `lm` в языке программирования R).

Результаты: В валидационном наборе данных процент растущих k -меров (k -меров, число которых статистически значимо растет во временной серии) из целевого организма от всех растущих k -меров составляет от 84.48 до 85.57 %. В наборе данных, который использовался для подбора оптимального значения длины k -меров, процент растущих k -меров из целевого организма от всех растущих k -меров уменьшается от 99.97 % при $k = 10$ до 52.19 % при $k = 14$. Процент растущих k -меров из всех k -меров целевого организма увеличивается от 1.74 % при $k = 10$ до 82.67 % при $k = 14$. В открытых базах данных (SRA,

HumanMetagenomeDB) собраны метагеномы из России, на которых планируется запустить наш пайплайн.

Выводы: Подход, основанный на анализе динамики k -меров, потенциально позволяет искать возможные агенты (вирусы и бактерии), увеличивающиеся в распространенности. Одной из причин увеличения распространенности может быть появление новых инфекционных агентов. Большая длина k -мера означает рост ложноположительных случаев, меньшая – ложноотрицательных.

Финансирование: Работа выполнена на базе НИИ системной биологии и медицины Роспотребнадзора в рамках госзадания «Разработка алгоритмов для выявления новых, уникальных последовательностей ДНК или РНК в метагеномах и их фенотипическая характеристика *in vitro*», номер 12203090069-4.

Search for spreading microorganisms based on the analysis of the representation of k -mers in metagenomic sequencing data

Panova V.*, Popov N., Fedorov D., Manolov A.

Research Institute for Systems Biology and Medicine (RISBM), Moscow, Russia

* panova_vv@sysbiomed.ru

Key words: k -mers; k -mer counter; metagenomes; diseases; pandemic

Motivation and Aim: After the COVID-19 pandemic, more and more attention is being paid to new zoonotic infections. This work aims to develop an approach to the identification of possible infectious agents based on increasing the representation and prevalence of short fixed-length substrings (k -mers) in the data of genome-wide metagenomic sequencing.

Methods and Algorithms: To validate the proposed method, metagenomic reads were simulated using the InSilicoSeq tool [1]. The set and values of species representation were obtained from the analysis of sequencing of total RNA isolated from a nasopharyngeal swab of a patient with COVID-19 (experiment SRA SRX8697677, launch SRR12183113). Jellyfish was used to calculate canonical k -mers [2]. Kraken2 was used to assess the taxonomic composition of the sample [3].

To test the applicability of the approach, we modeled a time series of 7 steps and selected one organism (*Streptococcus sp. S5*) whose abundance increased linearly. The choice of an organism is based on its abundance close to the median value; an increase in relative representation was made from the initial value to 61, 71, or 91 percentiles. The length of the k -mer was chosen to be 10 in the analysis.

The same set of studied species was used to select the optimal value of the length of k -mers. The noise was also added to the simulation: the relative abundance of each species was distributed according to the negative binomial law and varied over seven-time steps. The number of readings of this corresponding taxon is taken as a non-negative mean distribution (μ). The lengths of the resulting k -mers were equal to 10, 11, 12, 13, 14. The growth was estimated using a linear model (the `lm` function in the R programming language).

Results: In the validation dataset, the percentage of growing k -mers (k -mers, the number of which grows statistically significantly in the time series) from the target organism of all growing k -mers ranges from 84.48 to 85.57 %. In the dataset that was used to select the optimal value for the length of k -mers, the percentage of growing k -mers from the

target organism of all growing k-mers decreases from 99.97 % at $k = 10$ to 52.19 % at $k = 14$. The percentage of growing k-mers from all k-mers of the target organism increases from 1.74 % at $k = 10$ to 82.67 % at $k = 14$.

Metagenomes from Russia were collected from open databases (SRA, HumanMetagenomeDB). Our pipeline is planned to be launched on these metagenomes. *Conclusion:* The approach based on the analysis of the dynamics of k -mers potentially allows us to search for possible agents (viruses and bacteria) that are increasing in prevalence. A longer k -mer length means an increase in false-positive cases, and a smaller one – false-negative ones.

Funding: The research was carried out on the basis of the Research Institute of Systems Biology and Medicine within the framework of the state task “Development of algorithms for identifying new, unique DNA or RNA sequences in metagenomes and their phenotypic characteristics *in vitro*”, number 12203090069-4.

Список литературы/References

1. Gourlé H., Karlsson-Lindsjö O., Hayer J., Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*. 2019;35(3):521-522. doi 10.1093/bioinformatics/bty630
2. Marçais G., Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-770. doi 10.1093/bioinformatics/btr011
3. Wood D.E., Lu J., Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20(1):257. doi 10.1186/s13059-019-1891-0