DOI: 10.25205/978-5-4437-1691-6-36

СРАВНИТЕЛЬНЫЙ АНАЛИЗ БИОИНФОРМАТИЧЕСКИХ АЛГОРИТМОВ ПОИСКА ПРОФАГОВ В БАКТЕРИАЛЬНОМ ГЕНОМЕ *

COMPARATIVE ANALYSIS OF BIOINFORMATIC ALGORITHMS FOR SEARCHING FOR PROFAGES IN THE BACTERIAL GENOME

А. В. Резайкин[1,2], П. В. Микушин[1,3], И. Г. Низовцева[1],
А. А. Чигирева[4], В. И. Дубова[4], А. Е. Глебова[1]

[1] *Уральский федеральный университет им. первого Президента России Б. Н. Ельцина, Екатеринбург*
[2] *Уральский государственный медицинский университет, Екатеринбург*
[3] *Московский физико-технический институт*
[4] *ООО «НПО Биосинтез», Москва*

A. V. Rezaykin[1,2], P. V. Mikushin[1,3], I. G. Nizovtseva[1],
A. A. Chigireva[4], V. I. Dubova[4], A. E. Glebova[1]

[1] *Ural Federal University named after the First President of Russia B. N. Yeltsin, Yekaterinburg*
[2] *Ural State Medical University, Yekaterinburg*
[3] *Moscow Institute of Physics and Technology*
[4] *NPO Biosintez Ltd, Moscow*

✉ pavel.m.sci@gmail.com

Аннотация

В работе проведен сравнительный анализ четырех инструментов для идентификации профагов: PHASTEST, Phigaro, VIBRANT и PhiSpy. С помощью инструментов, основанных на разных алгоритмах (поиске гомологии, машинном обучении или их комбинации), были проанализированы полные геномы трех штаммов *Methylococcus capsulatus*. Результаты выявили существенные расхождения в определении границ, количества генов и функционального состояния предсказанных профагов. Анализ полученных данных показал, что использование нескольких инструментов с различными алгоритмами позволяет получить более точную информацию о профагах и снизить количество ложноположительных и ложноотрицательных результатов.

Abstract

This study presents a comparative analysis of four prophage prediction tools: PHASTEST, Phigaro, VIBRANT, and PhiSpy. The complete genomes of three Methylococcus capsulatus strains were analyzed using these tools, which employ distinct algorithms based on homology searches, machine learning, or a combination of both approaches. The results revealed significant discrepancies in prophage boundary prediction, gene content, and functional assessment. This analysis highlights that utilizing multiple tools with complementary algorithms enhances the accuracy of prophage identification and minimizes both false-positive and false-negative predictions.

Бактериофаги — неотъемлемый компонент микробных экосистем, оказывающий влияние на динамику бактериальных сообществ, метаболизм хозяев и горизонтальный перенос генов [1]. Высокопроизводительное секвенирование позволило изучать фаги без культивирования *in vitro*. Однако отсутствие универсальных генов-маркеров у вирусов затрудняет идентификацию их последовательностей в метагеномных данных. Для решения этой проблемы разработаны инструменты, дифференцирующие вирусные и невирусные последовательности, в том числе профаги в бактериальных геномах [2]. Различия в алгоритмах этих инструментов приводят к неоднозначности результатов, поэтому важна сравнительная оценка их эффективности.

В данном исследовании проведено сравнение четырех инструментов для поиска профагов — PHASTEST (v. 2.3.0) [3], Phigaro (v. 2.4.0) [4], VIBRANT (v. 1.2.0) [5] и PhiSpy (v. 4.2.21) [6], — основанных на различных алгоритмах.

Первоначально все представленные программы определяют протеин-кодирующие гены с помощью prodigal [7]. Далее гены аннотируются как бактериальные, вирусные или неизвестные. PHASTEST идентифицирует фаговые белки путем поиска гомологии с последовательностями в вирусной базе данных RefSeq (NCBI) [3]. В Phigaro реализован аналогичный подход, но с использованием базы данных профилей скрытых марковских

* Исследование выполнено при поддержке РНФ (проект № 24-24-00454).

моделей (HMM) — prokaryotic Virus Orthologous Groups (pVOGs) [4, 8]. PhiSpy использует метод случайного леса. Программа вычисляет набор признаков (длина белка, направленность цепи транскрипции, соотношение AT/GC, частота встречаемости уникальных фаговых слов (определенная последовательность из 12 п. н.), сайты интеграции фагов), на основе которых модель определяет вероятность их фаговой природы [6]. VIBRANT реализует гибридный подход, комбинируя нейронные сети с поиском гомологии [5]. В Phigaro при поиске гомологии используются профили HMM, но из трех различных баз данных: Kyoto Encyclopedia of Genes and Genomes (KEGG) [9], Pfam [10] и Virus Orthologous Groups (VOG) [11].

В итоге потенциальные профаговые гены объединяются в профаговые регионы. При объединении все программы учитывают количество потенциальных генов, расположенных на определенной дистанции, а также дополнительные характеристики. Найденным регионам дается интегральная оценка (score).

Для сравнительного анализа использовали полные геномы трех штаммов *Methylococcus capsulatus (Bath, IO1, Mc7)*, доступные в GenBank (acc. numb. NC_002977, CP079098, CP079095). Результаты включали координаты профаговых регионов, бактериальные и фаговые гены (ORF) внутри них, а также оценку функционального состояния профагов (см. таблицу).

Результаты поиска профаговых последовательностей в геномах штаммов *Methylococcus capsulatus*

Программа	Регион п/п	Размер региона (kb)	Всего ORF в регионе (характерных для профага)	Положение региона в геноме
str. Bath (NC_002977)				
PHASTEST	1	53.3	51 (31)	2819754–2873113
	2	48.9	59 (42)	3093885–3142838
Phigaro	1	50.0	50 (35)	2824960–2874998
	2	45.8	63 (48)	3092748–3138558
VIBRANT	1	61.4	61 (30)	2816780–2878183
	2	50.0	67 (38)	3091049–3141096
PhiSpy	1	31.8	27	2820020–2851790
	2	32.3	40	3096072–3128377
str. IO1 (CP079098)				
PHASTEST	1	48.5	62 (45)	2508110–2556599
Phigaro	1	17.5	15 (9)	104180–121683
	2	50.6	68 (51)	2509623–2560215
VIBRANT	1	8.3	13 (3)	830933–839256
	2	57.5	75 (32)	2512440–2569933
PhiSpy	1	12.3	19	827273–839561
	2	39.5	54	2514897–2554412
str. Mc7 (CP079095)				
PHASTEST	1	25.4	18 (7)	1825683–1851097
Phigaro	*Профагов не обнаружено*			
VIBRANT	*Профагов не обнаружено*			
PhiSpy	1	33.3	40	938889–972163
	2	16.5	24	1444321–1460832
	3	19.9	24	1629311–1649204
	4	17.6	12	2001894–2019505
	5	14.8	20	2542546–2557313

В геноме штамма Bath все инструменты выявили два потенциально жизнеспособных профага. Их расположение совпадало, но наблюдались различия в границах, размере и количестве генов, обусловленные, вероятно, аннотацией пограничных генов с неизвестными функциями. В геноме штамма IO1 все программы обнаружили один общий профаг (2508110–2569933 нт). Phigaro, VIBRANT и PhiSpy идентифицировали дополнительный регион (совпадающий у последних двух). В штамме Mc7 только PhiSpy предсказал пять потенциальных профагов. Регионы 2–5, вероятно, ложноположительные из-за малого размера и числа генов. Первый регион, отличающийся большим размером и количеством ORF, может представлять собой новый, неописанный бактериофаг, что требует дальнейшего исследования.

Для повышения точности идентификации профагов в бактериальных геномах рекомендуется использовать комбинацию инструментов с разными алгоритмами. При оценке результатов важно учитывать размер найденных регионов, количество фаговых генов и особенности алгоритмов. Определение границ профагов следует проводить с учетом аннотаций пограничных генов, полученных разными методами.

Литература

1. Hegarty B., Riddell V. J., Bastien E. et al. Benchmarking informatics approaches for virus discovery: caution is needed when combining in silico identification methods // mSystems. 2024 Vol. 9 (3). P. e0110523.

2. Andrade-Martínez J. S., Camelo V. L. C., Chica C. L. A. et al. Computational Tools for the Analysis of Uncultivated Phage Genomes // Microbiol. Mol. Biol. Rev. 2022. Vol. 86 (2). P. e0000421.

3. Wishart D. S., Han S., Saha S. et al. PHASTEST: faster than PHASTER, better than PHAST // Nucleic Acids Res. 2023. Vol. 51 (W1). P. W443–W450.

4. Starikova E. V., Tikhonova P. O., Prianichnikov N. A. et al. Phigaro: high-throughput prophage sequence annotation // Bioinformatics. 2020. Vol. 36 (12). P. 3882–3884.

5. Kieft K., Zhou Z. Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences // Microbiome. 2020. Vol. 8 (1). P. 90.

6. Akhter S., Aziz R. K., Edwards R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies // Nucleic Acids Res. 2012. Vol. 40 (16). P. e126.

7. Hyatt D., Chen G. L., LoCascio P. F. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification // BMC Bioinformatics. 2010. Vol. 11. P. 119

8. Grazziotin A. L., Koonin E. V., Kristensen D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation // Nucleic Acids Res. 2017. Vol. 45 (D1). P. D491–D498.

9. Kanehisa M., Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000. Vol. 28 (1). P. 27–30.

10. El-Gebali S., Mistry J., Bateman A. et al. The Pfam protein families database in 2019 // Nucleic Acids Res. 2019. Vol. 47 (D1). P. D427–D432.

11. Trgovec-Greif L., Hellinger H.-J., Mainguy J. et al. VOGDB — Database of Virus Orthologous Groups // Viruses. 2024. Vol. 16. P. 1191.