

Разработка вычислительного конвейера для поиска семейств мультидоменных белков в геномах растений

Бочарникова М.^{1,2*}, Турнаев И.¹, Афонников Д.^{1,2}

¹ Институт цитологии и генетики СО РАН, Новосибирск, Россия

² Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия

* lachynova@bionet.nsc.ru

Ключевые слова: вычислительный конвейер; мультидоменные белки; филогения

Мотивация и цель: Современные технологии секвенирования привели к огромному потоку данных о последовательностях генома растений. Для определения функций белка эти данные требуют высокопроизводительной компьютерной аннотации. Целью настоящей работы являлось создание вычислительного конвейера OrthoDom, который предназначен для поиска ортологичных белков с учетом их доменного состава.

Методы и алгоритмы: Конвейер реализован на основе контейнерной системы Snakemake [1]. В качестве входных данных конвейер принимает протеомы организмов, у которых пользователь хочет найти белки, наборы референсных белков, для которых известны функции и домены, а также профили НММ [2] для доменов этих белков. Конвейер выводит список ортологичных последовательностей для эталонных белков, а также информацию об их доменном составе и филогенетических деревьях для идентифицированных ортологичных семейств.

Результаты: Мы протестировали точность системы и внедрили ее для анализа растительных белков. Конвейер использовался для идентификации ортологов в геномах злаков.

Выводы: Разработанный нами вычислительный конвейер может успешно применяться для более эффективного поиска семейств мультидоменных белков в геномах растений.

Финансирование: Исследование поддержано Курчатовским геномным центром ИЦиГ СО РАН (проект № 075-15-2019-1662).

Development of a computational pipeline for searching for families of multidomain proteins in plant genomes

Bocharnikova M.^{1,2*}, Turnaev I.¹, Afonnikov D.^{1,2}

¹ Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

² Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* lachynova@bionet.nsc.ru

Key words: computational pipeline; multi-domain proteins; phylogeny

Motivation and Aim: Modern sequencing technologies have led to a huge flow of data on plant genome sequences. These data require high-performance computer annotation

to determine protein functions. The purpose of this work was to create the Orthodox computing pipeline, which is designed to search for orthologous proteins taking into account their domain composition.

Methods and Algorithms: The pipeline is implemented based on the Snakemake container system [1]. As input, the pipeline accepts proteomes of organisms in which the user wants to find proteins, sets of reference proteins for which functions and domains are known, as well as HMM profiles [2] for the domains of these proteins. The pipeline outputs a list of orthologous sequences for reference proteins, as well as information about their domain composition and phylogenetic trees for identified orthologous families.

Results: We tested the accuracy of the system and implemented it for the analysis of plant proteins. The conveyor was used to identify orthologs in the genomes of cereals.

Conclusion: The computational pipeline we have developed can be successfully used to more efficiently search for families of multidomain proteins in plant genomes.

Funding: The study was supported by the Kurchatov Genomic Center of the ICG SB RAS (project No. 075-15-2019-1662).

Список литературы/References

1. Köster J., Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522
2. Eddy S. R. Hidden markov models. *Curr. Opin. Struct. Biol.* 1996;6(3):361-365