

Предсказание функции белков на основе гомологии

Малюгин Е.^{1,3*}, Мустафин З.^{2,3}, Пронозин А.^{2,3}, Генаев М.^{2,3}, Афонников Д.^{2,3}

¹ Новосибирский государственный университет, Новосибирск, Россия

² Институт цитологии и генетики СО РАН, Новосибирск, Россия

³ Курчатowski геномный центр ИЦиГ СО РАН, Новосибирск, Россия

* evgeny.malyugin98@gmail.com

Ключевые слова: предсказание функции; геновая онтология; возраст генов; гомология; ортогруппы

Мотивация и цель: С появлением методов секвенирования нового поколения число секвенированных последовательностей геномов растет огромными темпами. Источниками поступления такого большого количества данных являются различные геномные, метагеномные и транскриптомные проекты. Все это позволяет распознавать структуру генома и предсказывать локализацию белок-кодирующих генов с высокой точностью. Однако для распознанных генов функция, как правило, оказывается неизвестной. Существующие для решения этой задачи методы преимущественно используют для поиска гомологичных последовательностей с известными функциями, описанными в терминах Gene Ontology. Возраст генов имеет прямое отношение к задаче аннотации функций генов: в задачах предсказания функции белков по гомологии важной проблемой является наличие аннотированных гомологов для искомым последовательностей в базах данных. Очевидно, что если гомологичные последовательности для исходной последовательности отсутствуют, то это не позволяет получить информацию для ее аннотации [1, 2]. Такая ситуация характерна для молодых или таксон-специфичных (орфанных) генов, представленных в одном или нескольких ближайших видах и отсутствующих у других организмов. Для них предсказание функции затруднено, и его точность оказывается ниже, чем для древних генов, гомологи для которых представлены более широко. Таким образом, создание метода аннотации функций генов, который бы работал с высокой точностью вне зависимости от возраста генов, является достаточно актуальным.

Методы и алгоритмы: Мы предложили алгоритм аннотации функций генов на основе поиска сходных последовательностей методом k ближайших гомологов, анализа ортологических групп в базе данных OrthoDB v. 10.1 [3] и фильтрации терминов на основе сходства последовательностей методом логистической регрессии. Поиск гомологичных последовательностей проводился в базе данных с использованием алгоритма USEARCH v 11.0.667 [4]. Методы были опробованы на последовательностях белок-кодирующих генов ряда геномов растений, включая *Arabidopsis thaliana* и др. Оценка возрастов генов была использована с помощью программы Orthoscape [5]. Мы также предложили использовать для оценки точности методов предсказания функций генов меру эффективности предсказания (PE), которая позволяет избежать влияния использования различных аннотаций при сравнении разных методов предсказания.

Результаты: Точность предложенного алгоритма при анализе генома *A. thaliana*, *Oryza sativa* и *Zea mays* оказалась сравнимой с методами Blast2GO [6] и eggNOG [7]. Так, эффективность предсказания (PE) на данных *A. thaliana* для нашего

метода составила 85.03, 87.17, 89.24 % для генов молодого, среднего и старого возраста соответственно против 85.32, 84.48, 81.69 % для метода Blast2GO. Показано, что при определенных параметрах алгоритма точность предсказания функций у генов разных возрастов (в том числе молодых) оказывается близкой. Таким образом, предложенный метод позволяет нивелировать различия в точности предсказания функции для генов разных возрастов, при этом в лучшую сторону.

Выводы: Точность предсказания генов молодого возраста гораздо ниже, чем точность предсказания средних и старых генов. Использование ортогрупп и метода логистической регрессии позволяет повысить точность предсказания для генов молодого возраста.

Финансирование: Исследование поддержано Курчатовским геномным центром Института цитологии и генетики СО РАН, соглашение с Министерством образования и науки Российской Федерации № 075-15-2019-1662.

Predicting protein function based on homology

Malyugin E.^{1,3*}, Mustafin Z.^{2,3}, Pronozin A.^{2,3}, Genaev M.^{2,3}, Afonnikov D.^{2,3}

¹ Novosibirsk State University, Novosibirsk, Russia

² Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

³ Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* evgeny.malyugin98@gmail.com

Key words: function prediction; gene ontology; gene age; homology; orthogroups

Motivation and Aim: With the advent of next-generation sequencing methods, the number of sequenced genome sequences is growing at a tremendous rate. Various genomic, metagenomic, and transcriptomic projects are the sources of this large amount of data. All of these make it possible to recognise genome structure and predict the localisation of protein-coding genes with high accuracy. However, for novel genes, the function is usually unknown. The methods to predict function for novel genes predominantly use the search for homologous sequences with known functions described in terms of Gene Ontology. The age of genes is directly related to the task of gene function annotation: in the tasks of protein function prediction by homology, an important problem is the availability of annotated homologues for the searched sequences in databases. Obviously, in the case if homologous sequences for the query sequence are not available, gene function prediction for these methods is difficult [1, 2]. This situation is typical for young or taxon-specific (orphan) genes represented in one or more relative species and absent in other organisms. For them, function prediction is difficult and its accuracy is lower than for ancient genes for which homologues are more widely represented. Thus, the creation of a method of gene function annotation that would work with high accuracy regardless of the age of genes is quite relevant.

Methods and Algorithms: We proposed an algorithm for gene function annotation based on searching similar sequences by *k* nearest homologues method, analysing orthologous groups in OrthoDB database v. 10.1 [3] and filtering terms based on sequence similarity by logistic regression. Homologous sequences were searched in the database using the USEARCH v 11.0.667 algorithm [4]. The methods were tested on sequences of protein-coding genes from a number of plant genomes including *Arabidopsis thaliana* and

several other plant species. Estimation of gene ages was performed by the Orthoscape method [5]. We also proposed a novel prediction performance measure, the prediction efficiency which is independent on the usage of different annotation source for different prediction methods.

Results: The accuracy of the proposed algorithm in analysing the genome of *A. thaliana*, *Oryza sativa* and *Zea mays* was comparable to the Blast2GO and eggNOG methods [6, 7]. Thus, the prediction efficiency (PE) on *A. thaliana* data for our method was 85.03 %, 87.17 %, 89.24 % for young, middle-aged and old genes respectively, compared to 85.32 %, 84.48 %, 81.69 % for the Blast2GO method. It is shown that at certain parameters of the algorithm the accuracy of function prediction for genes of different ages (including young genes) is close. Thus, the proposed method allows levelling the differences in the accuracy of function prediction for genes of different ages, at the same time for the better.

Conclusion: The prediction accuracy of young age genes is much lower than the prediction accuracy of middle-aged and old genes. The use of orthogroups and logistic regression method can improve the prediction accuracy for young age genes.

Funding: The study is supported by the Kurchatov Genomic Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation, No. 075-15-2019-1662.

Список литературы/References

1. Liebeskind B.J., McWhite Claire D., Marcotte E.M. Towards consensus gene ages. *Genome Biol Evol.* 2016;8(6):1812-1823
2. Mustafin Z.S., Zamyatin V.I., Konstantinov D.K., Doroshkov A.V., Lashin S.A., Afonnikov D.A. Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in *A. thaliana*. *Genes (Basel)*. 2019;10(12):963. doi 10.3390/genes10120963
3. Kriventseva E.V., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simão F.A., Zdobnov E.M. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;10(12):963. doi 10.3390/genes10120963
4. Robert C.E. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 2010;26(19):2460-2461
5. Mustafin Z.S., Lashin S.A., Matushkin Yu.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a Cytoscape plugin for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics.* 2017;18(Suppl. 1):1427
6. Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 2005;21(18):3674-3676
7. Huerta-Cepas J., Szklarczyk D., Heller D. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(1):309-314