Использование программ быстрого поиска гомологии для функциональной аннотации белков

Пронозин Артем[1], Генаев Михаил[1,2], Афонников Дмитрий[1,2]
[1]Институт цитологии и генетики СО РАН, Новосибирск, Россия
[2]Курчатовский геномный центр ИЦиГ СО РАН

АКТУАЛЬНОСТЬ

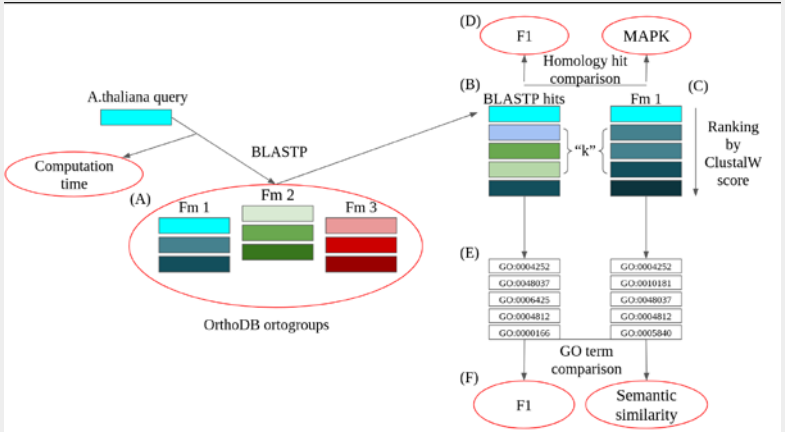
Современные RNA-seq и геномные эксперименты способствуют появлению всё большего количества белок кодирующих последовательностей и требуют их функциональной аннотации.
Основной подход для аннотации белок-кодирующих генов – поиск гомологов с известной функцией. Однако постоянный рост размеров баз данных NCBI's GenBank и UniProt становится существенной трудностью при анализе сходства белков.
Для более быстрого поиска разработаны программы BLASTP-fast, Diamond, Usearch, Mmseq2, которые обеспечивают ускорение поиска в 100-1000 раз по сравнению BLASTP за счет более низкой чувствительности.

ЗАДАЧИ

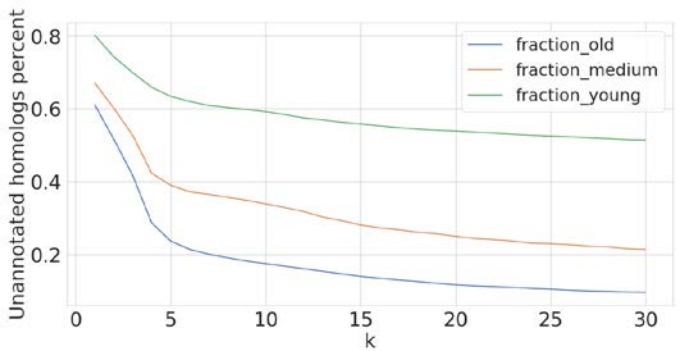
- Оценка точности идентификации ортологов белковых последовательностей на примере *A.thaliana* в базе данных OrthoDB с помощью программ быстрого поиска гомологов.
- Оценка точности идентификации терминов GO для искомой последовательности с помощью программ быстрого поиска гомологов.
- Оценка оптимальных параметров для поиска ортологов и аннотации (количество ближайших гомологичных генов, *k*).
- Оценка зависимости точности поиска ортолгов и аннотации для генов разных возрастов.
- Разработка методов предсказания возраста генов на основе характеристик выравнивания ближайших гомологов.

МАТЕРИАЛЫ И МЕТОДЫ

- Источник данных база: OrthoDB (37 млн. последовательностей, 8.5 млн. групп ортологов, включает аннотацию GO).
- Источник query: 22812 белков *A.thaliana*, входящих в состав 9193 ортогрупп OrthoDB.
- Программы поиска гомологов: BLASTP, BLASTP-fast, Mmseq2, Diamond, UBLAST, Usearch local.
- Стандарт выравнивания и ранжирования последовательностей: программа ClustalW.



СТЕПЕНЬ АННОТАЦИИ ГОМОЛОГИЧНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ



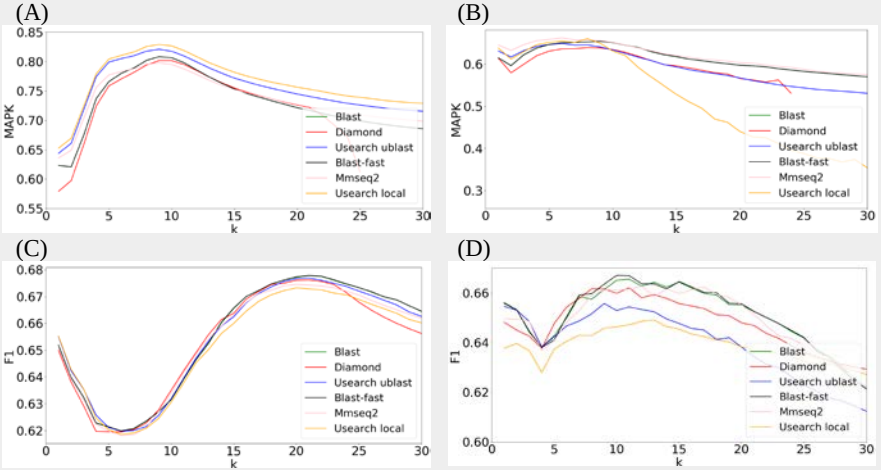
Доля последовательностей query, для которых среди k ближайших хитов, полученных программой BLASTP, не нашлось ни одного с аннотацией терминами GO (ось Y). По оси X отложены значения k.

Больше ~80% генов старой и молодой группы имеют аннотацию на k=30. Эта доля ниже 50% для молодых генов.

ВРЕМЯ ОБРАБОТКИ ДАННЫХ

	Индексирование,мин	Поиск,мин	Размер индекса, ГБ
Blast	22 мин 45 с	91 ч 6 мин	22
BlastFast	22 мин 45 с	9 ч 81 мин	22
Usearch local	53 мин 3 с	10 мин	75
UBLAST	47 мин 8 с	5 ч 23 мин	98
Diamond	3 мин 45 с	11 мин	17
Mmseq2	2 мин 5 с	33 мин	19

РЕЗУЛЬТАТЫ АНАЛИЗА



Ось X – величина *k*. Ось Y – величина MAPK и F1 для GO, соответственно. Панели A, B, представляют среднее значение MAPK для древних и молодых возрастов, соответственно. C, D древние и молодые возраста - F1 для GO, соответственно. Линии различного цвета отвечают за программы.

- Метрика MAPK – для древней группы (A) минимальное значение наблюдается на k =1 (0.65). Значение увеличивается для *k*=10 (0.80-0.85) затем постепенно уменьшается. Для молодой группы (B), значение остается постоянным (0.65) при k = 1-10, затем постепенно уменьшается.
- F1 для GO – древние и молодые группы имеют схожее поведение: высокое значение на k = 1 (0.64-0.66). При k = 3-10, наблюдается минимум, при k = 20 максимум. Различие между древними и молодыми группами заключается в размере минимума для молодых наблюдается при k = 5 для древних при k = 5 -10.

ВЫВОД

- Программы быстрого поиска гомологии показывают схожие с BLASTP результаты идентификации ортологов и GO терминов аннотации на k<30 лучших совпадениях.
- Наблюдаются различия оптимальных параметров k при идентификации ортологов и аннотации GO терминами: при идентификации ортологов лучший результат при k=10, GO аннотация показывают лучший результат при k=20.
- Для лучшей аннотации молодой группы генов рекомендуется брать первый наиболее близкий гомолог.
- Возраст гена может быть предсказан при F1 = 0,89 с использованием 10 лучших совпадений, полученных программой поиска гомологии.

Работа выполнена при поддержке Russian Science Foundation grant 18-14-00293. Вычислительные ресурсы Joint HPC Facility ‘Bioinformatics’ использовались при поддержке бюджетного проекта №0324–2019-0040-C-01.

СПИСОК ЛИТЕРАТУРЫ

[1] A. Conesa,S. Götz,J. García-Gómez,J.Terol,M. Talón,andM.Robles, “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research,” Bioinformatics, 21(18), pp. 3674-3676, 2005.
[2] R. Vaser,, D. Pavlović, and M. Šikić, “SWORD—a highly efficient protein database search,” Bioinformatics, 32.17 , pp. i680-i684, 2016.
[3] Z. Mustafin,et al. “Phylostratigraphic Analysis Shows the Earliest Origination of the Abiotic Stress Associated Genes in A. thaliana,” Genes, 10.12 , pp. 963, 2019.
[4] E. Kriventseva, et al. “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs,” Nucleic acids research, 47.D1, pp. D807-D811, 2019.
[5] S. Altschul,et al. “Basic local alignment search tool,” Journal of molecular biology,215.3, , pp. 403-410, 1990.
[6] B. Buchfink, C. Xie, and D. Huson, “Fast and sensitive protein alignment using DIAMOND,” Nature methods, 12.1,, 59, 2015.
[7] E. Usearch, “Lawrence Berkeley National Laboratory (LBNL),” Berkeley, CA (United States) (2010).
[8] M. Hauser, M. Steinegger, and J. Söding, “MMseqs software suite for fast and deep clustering and searching of large protein sequence sets,”Bioinformatics, 32.9,pp. 1323-1330, 2016.
[9] J. Thompson,, G. Desmondand T.Gibson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” Nucleic acids research, 22.22, pp. 4673-4680, 1994.
[10] N. Pentreath, “Machine learning with spark,” Packt Publishing Ltd, 2015.
[11] V. Rijsbergen CJ, “Information retrieval,” 2nd edn. Butterworths, London, 1979.
[12] C. Pesquita, “Semantic similarity in the gene ontology,” The gene ontology handbook. Humana Press, New York, NY, 2017. 161-173