Секция «Биоинформатика»

Реконструкция филогении двухдоменных белков

Научный руководитель – Спирин Сергей Александрович

Латорцева Дарья Дмитриевна

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: latortsevad@gmail.com

Чаще всего филогенетическое дерево строится на основании множественного выравнивания последовательностей. Ошибки в выравнивании, высоковариабельные участки, гомоплазия могут привести к ошибкам построения дерева. Существует ряд биоинформатических программ, которые фильтруют такие последовательности, например, удаляют ненужные участки. В данной работе исследуется эффективность одной из таких программ - Noisy [1]. Мы хотели проверить, улучшает ли Noisy качество реконструкции филогенетических деревьев.

Исследование проводилось с использованием последовательностей двухдоменных белков. Такой выбор был обусловлен предположением, что скорость мутаций в домене и вне его - разная, а это может привести к ошибкам реконструкции, с которыми должна справиться Noisy. Белки были взяты из восьми групп: Растения, Эукариоты, Грибы, Митохондриальные белки многоклеточных, Актинобактерии, Хордовые и Протеобактерии, отдельно все Многоклеточные животные (из базы Pfam35).

Чтобы выяснить, какая информация из записи о белке важна для построения точного дерева, а какая только ухудшает его, исходя из предположения, что последовательности доменов в белке самые консервативные, мы разделили записи на пять групп по наличию спейсера или домена:

(1) полные записи о двухдоменном белке; (2) записи, содержащие только первый или (3) только второй домен; (4) слитые последовательности двух доменов; (5) записи со спейсером между доменами.

Далее выполнили выравнивание с помощью программы muscle, программой FastMe [2] построили отдельно деревья, прошедшие фильтрацию Noisy, и без нее. Оценили, насколько далеки полученные деревья от референсных подсчетом расстояния Робинзона-Фоулдса [3] для групп с фильтрацией и без.

Анализ важных для построения дерева частей показал, что почти для всех групп организмов лучше всего деревья строились по полным записям белков, чуть хуже по записям доменов со спейсером между ними. Для эукариотических двухдоменных белков, и белков из митохондрий многоклеточных, наоборот, деревья лучше строятся по доменам со спейсером. Корреляции между длиной спейсерной последовательности и качеством построенного дерева не было обнаружено.

Основываясь на результатах нашей работы, точно можно сказать, что для всех групп и типов последовательностей программа Noisy ухудшает качество построенных филогенетических деревьев.

Работа поддержана грантом РНФ номер 21-14-00135.

Источники и литература

1) Dress, A.W., Flamm, C., Fritzsch, G. et al. Noisy: Identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol 3, 7 (2008). https://doi.org/10.1186/1748-7188-3-7

2) Vincent Lefort, Richard Desper, Olivier Gascuel, FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program, Molecular Biology and Evolution, Volume 32, Issue 10, October 2015, Pages 2798–2800, https://doi.org/10.1093/molbev/msv150

3) Briand, S., Dessimoz, C., El-Mabrouk, N. et al. A generalized Robinson-Foulds distance for labeled trees. BMC Genomics 21, 779 (2020). https://doi.org/10.1186/s12864-020-07011-0