Секция «Биоинженерия и биоинформатика»

Семейство мер качества выравниваний, основанных на количестве ошибок в колонках в сравнении с эталонным выравниванием

Бурков Б.А.[1]***, Нагаев Б.Э.***[2]

1 - Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, 2 - Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: BurkovBA@gmail.com

Выравнивание последовательностей подразумевает, что любые два мономера, поставленные в одну колонку, родственны друг другу. На практике часто возникает задача сравнения двух выравниваний, построенных различными методами для одного и того же набора последовательностей. Вводится следующая мера различия выравниваний. Одно из выравниваний мы принимаем за эталонное, а другое – за тестируемое. Для каждой колонки эталонного выравнивания считаем, по скольким колонкам тестируемого выравнивания были разбросаны ее остатки. Число этих колонок минус одна признается количеством ошибок первого рода, т.е. недопредсказанием гомологии остатков тестируемым выравниванием. Ошибки второго рода рассчитываются аналогичным способом, но выравнивания меняются ролями. В простейшем случае нормированное число ошибок обоих родов и есть мера.

Проверялось, насколько данная мера отражает относительную эффективность выравниваний в биоинформатических задачах. Для этого были взяты выравнивания из базы данных эталонных выравниваний Balibase [1]. Для тех же последовательностей были построены выравнивания программой muscle [2] с параметрами по умолчанию (default) и с параметрами для грубого, но быстрого выравнивания (fast). Показано, что, в среднем, мера сходства fast muscle с эталонными выравниваниями ниже, чем default muscle. Кроме того, HMM-профилями, построенными для эталонных выравниваний, fast и default muscle находились гомологи в банке данных Swiss-Prot. Было показано, что почти все находки для профилей по muscle нашлись и по эталонным выравниваниям. Fast muscle нашел меньше гомологов, чем default muscle, что коррелирует с поведением меры. Также, филогенетическое дерево, построенное для эталонного выравнивания методом парсимонии с помощью программы TNT [3], лучше воспроизводится выравниваниями с лучшей мерой (default muscle). Таким образом, мера пригодна для использования в задачах сравнения выравниваний.

Литература

1. Thompson JD, Koehl P, Ripp R, Poch O BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins. 2005 Oct 1;61(1):127-36.

2. Edgar RC MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug 19;5:113.

3. Goloboff, P., Farris, J., & Nixon, C. TNT, a free program for phylogenetic analysis. Cladistics. 2008 24:774-786.

Слова благодарности

Авторы работы благодарят А.В. Алексеевского и С.А. Спирина за предложенные ими идеи, помощь и руководство. Работа частично поддержана грантами РФФИ 09-04-92743-ННИОМ_а, 10-07-00685-а , 11-04-91340-ННИО_а и госконтрактом No 07.514.11.4006. Программа TNT свободно распространяется с любезного разрешения Willi Hennig Society, которое НЕ любезно ТРЕБУЕТ ссылаться на себя, иначе нарушишь лицензию.

Иллюстрации

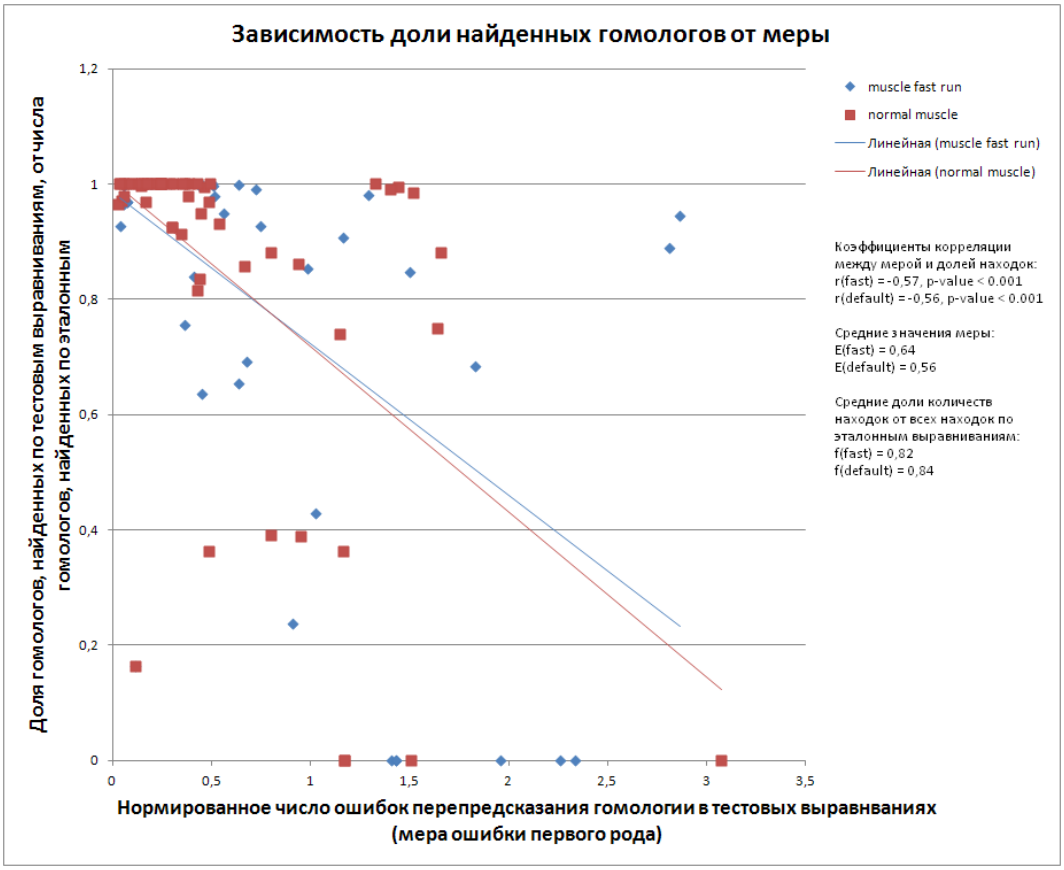

Рис. 1: Зависимость доли найденных по HMM-профилю гомологов от меры. Коэффициенты корреляции для fast и default = -0.57 и -0.56, p-value<0.001, средние значения меры 0,64 и 0,56, средние доли количеств находок от всех находок по профилям эталонных выравниваний 0,82 и 0,84.