Секция «Биоинженерия и биоинформатика»

Проверка выравниваний Pfam по структурным данным

***Нагаев Б.Э.*[1], *Алексеевский Д.А.*[2]**

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

E-mail: bnagaev@gmail.com

Выравнивание последовательностей – один из основных инструментов сравнения белков, установления их гомологичности, восстановления молекулярной филогении, предсказания свойств белков “по сходству”. Вопрос о том, насколько правильными, т.е. биологически обоснованными, они являются, не всегда очевиден. В тех случаях, когда для белков, чьи последовательности входят в выравнивание, известны пространственные структуры, выравнивание может быть верифицировано с помощью совмещения полипептидных цепей белков в пространстве. Известно, что проверка выравнивания, основанная на структурных данных, более надежна, чем проверка, основанная на информации о последовательностях.

Блоком назовем часть выравнивания, составленную из непрерывных фрагментов двух или более последовательностей, лежащих друг под другом. Нами создана библиотека allpy на языке Python для манипуляций с выравниваниями и блоками биологических последовательностей и на ее базе реализована новая версия программы MALAKITE_3D. В выравнивании последовательностей белков, для которых известны пространственные структуры, эта программа находит блоки выравнивания, подтверждаемые совмещением полипептидных цепей. Подчеркнем, что совмещение строится для конкретного блока, а не для полипептидных цепей в целом; таким образом, используется подход гибкого (flexible) выравнивания структур.

Для оценки того, какая часть выравнивания подтверждается структурными данными, мы ввели меру достоверности выравнивания. Рассмотрим выравнивание как множество пар сопоставленных мономеров. Сопоставленными считаются мономеры, стоящие в одной колонке. Достоверностью выравнивания назовём отношение числа таких пар, целиком входящих в какой-либо блок, к общему числу пар сопоставленных мономеров.

Пользуясь разработанным методом, мы оценили выравнивания из базы данных белковых доменов Pfam. Из каждого из 11938 выравниваний были выбраны последовательности только тех белков, для которых известна пространственная структура. Осталось 3022 выравнивания, содержащих более одной последовательности. Из них случайным образом были отобраны 178 выравниваний, содержащих по 8-10 последовательностей. Для каждой из них была рассчитана достоверность. Среднее значение достоверности составило 54% (стандартное отклонение 20%). При этом необходимо обратить внимание на то, что даже в приводимой выборке имелось 16 выравниваний (9% выборки) с достоверностью меньше 20% и ещё 10 с достоверностью от 20% до 30%. Некоторые из них будут проанализированы в докладе. Также в докладе будут приведены данные обработки большей части базы Pfam.

Низкое значение достоверности выравнивания, подтвержденной структурными данными, может быть связано как с действительным положением вещей – отсутствием

биологически осмысленного выравнивания на некоторых участках, так и с ошибками в выравнивании. В обоих случаях этот показатель следует учитывать при анализе семейств белков.

Слова благодарности

Благодарю Алексеевского Андрея Владимировича и Спирина Сергея Александровича