## Секция «Биоинженерия и биоинформатика»

### Использование базы данных PREFAB для анализа алгоритмов выравнивания аминокислотных последовательностей.

***Поверенная Ирина Владимировна***
*Студент*
*Московский государственный университет имени М.В. Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия*
*E-mail: ipoverennaya@gmail.com*

При анализе алгоритмов выравнивания аминокислотных последовательностей, нужны эталонные выравнивания, с которыми сравниваются алгоритмически построенные выравнивания. В качестве источника таких эталонов во многих работах (см., например, [2]) использовалась база данных PREFAB v4.0 (см. [4]), база включает 1682 выравнивания. Эти выравнивания построены на основе наложения пространственных структур. К сожалению, для базы PREFAB не описана методика отбора пар выравниваемых последовательностей. При этом для последовательностей указаны только имена PDB-файлов и цепей (в заголовках файлов PREFAB), но не указано, какие именно фрагменты этих цепей взяты.

Целью нашей работы было выяснить, насколько представленные в PREFAB пары последовательностей соответствуют классификации доменов пространственных структур белков SCOP [3]. Говоря более точно, мы считали аминокислотную последовательность допустимой, если она (1) является фрагментом одной из цепей в одном из файлов PDB и (2) этот фрагмент совпадает с одним из доменов, описанных в базе данных SCOP. Выравнивание считалось допустимым только, если домены, соответствующие сравниваемым последовательностям, принадлежат к одному и тому же семейству по классификации SCOP.

Проведенный анализ дал следующие результаты. Выравнивания базы PREFAB v4.0 содержат 1682 последовательности. Оказалось, что в PREFAB есть последовательности, которые образуют домен совместно с фрагментами других цепей того же белка, а другие последовательности содержат более одного домена. Кроме того, выяснилось, что в ряде «однодоменных» последовательностей PREFAB удалены некоторые фрагменты цепей, предположительно, соответствующие неструктурированным участкам белков. Таким образом было отобрано 1294 последовательности, которые удовлетворяют приведенным условиям. Пары, составленные из этих последовательностей составляют 1115 выравниваний PREFAB. Из них 834 выравниваний удовлетворяют условию принадлежности сравниваемых последовательностей одному семейству SCOP. Для этих выравниваний были подсчитаны количество удаленных фрагментов («гэпов») и процент сходства. Среднее количество гэпов в выравнивании примерно равно 8, при этом 13% выравниваний содержат 15 и более гэпов. Построенная выборка предназначена для анализа качества выравниваний, которые строятся методом [1] (адрес сервера: [5]).

### Литература

1. Яковлев В.В., Ройтберг М.А. Увеличение точности глобального выравнивания аминокислотных последовательностей с помощью построения набора выравниваний-кандидатов // Биофизика. - 2010. - Т. 55, N 6. - С. 965-975

2. Edgar, Robert C. (2004), MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.

3. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.

4. БД PREFAB: http://www.drive5.com/muscle/prefab.htm

5. Сервер: http://server2.lpm.org.ru/bio/online/pareto/

**Слова благодарности**

Хочу выразить благодарность моему научному руководителю Ройтбергу М.А., а также Яковлеву В.В., за помощь и поддержку.