БИОФИЗИКА, 2012, том 57, вып. 2, с. 205–211

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.112.8

ВЕРИФИКАЦИЯ БАЗЫ ЭТАЛОННЫХ ВЫРАВНИВАНИЙ PREFAB

© 2012 г. Т.В. Астахова*, М.Н. Лобанов, И.В. Поверенная* ***, М.А. Ройтберг*, В.В. Яковлев****

**Институт математических проблем биологии РАН, 142290, Пущино Московской области;*

***Институт белка РАН, 142290, Пущино Московской области;*

****Факультет биоинформатики и биоинженерии, Московского государственного университета, 119991, Москва, Воробьевы горы*

E-mail: victor@lpm.org.ru

Поступила в редакцию 24.08.11 г.

Проведена верификация базы данных PREFAB, которая выявила значительное число расхождений между последовательностями из базы PREFAB и соответствующими последовательностями из базы PDB. В 575 случаях из PDB-последовательностей были удалены фрагменты, соответствующие выравнивания были исключены из рассмотрения. Для 440 аминокислотных последовательностей в PREFAB были найдены одиночные замены или вставки по сравнению с последовательностью из структуры, взятой из банка данных PDB. В этих случаях последовательности и выравнивания были приведены в соответствие с базой данных PDB. Анализ типов укладки сравниваемых доменов по классификации SCOP показал, что в получившейся выборке только в 502 случаях выравниваются гомологичные домены. После удаления файлов-повторов конечный размер новой базы эталонных выравниваний PREFAB-P, построенной на основе базы данных PREFAB, составил 581 выравнивание.

Ключевые слова: аминокислотные последовательности, эталонные выравнивания, PDB структура, классификация SCOP.

1. БАЗЫ ЭТАЛОННЫХ ВЫРАВНИВАНИЙ

Построение выравниваний аминокислотных последовательностей является одним из ключевых инструментов в биоинформатике, молекулярной биологии и геномном анализе. Выравнивания используются при построении филогенетических деревьев и оценке их качества, нахождении характерных мотивов и консервативных остатков в белковых семействах, построении доменных профилей и решении многих других задач.

Существует много программ как множественного, так и парного выравнивания последовательностей. Для пользователя таких программ, наряду с вычислительной сложностью программ (временем работы и требованиями к памяти), большое значение имеет биологическая адекватность получаемых выравниваний. Таким образом, обычно работа каждой новой программы или алгоритма оценивается через сравнение с работой уже существующих программ по двум параметрам: качество полученных выравниваний и скорость работы. Для такого анализа необходимо иметь так называемые эталонные выравнивания, т.е. выравнивания, которые считаются наиболее биологически корректными.

Изначально (в 1980-х – первой половине 1990-х гг.) в качестве эталонных выравниваний для оценки работы программы авторы выбирали выравнивания сами, исходя из своих собственных критериев (см., например, [1–5]), но, как правило, такие выборки были маленькими. К тому же использование большого количества различных наборов эталонных выравниваний при оценке разных алгоритмов делало их сравнение не слишком удобным. Среди работ этого периода выделим работу McClure et al. [6], опубликованную в 1994 г. Авторы тестировали различные методы множественного выравнивания последовательностей на способность нахождения консервативных мотивов в белковых семействах гемоглобина, киназы, рибонуклеазы H и протеазы, расщепляющей белок по аспарагиновой кислоте. Для всех данных семейств уже были известны и изучены биологически важные мотивы, следовательно, были известны выравнивания последовательностей, принадлежащих каждому из семейств. Такие выравнивания и были взяты в качестве золотого стан-

Сокращения: БД – база данных, БЭВ – база эталонных выравниваний.

дарта (эталона), для каждого семейства получилась своя база эталонных выравниваний. На основе полученных результатов авторы сделали вывод о том, что алгоритмы глобального выравнивания ищут консервативные мотивы лучше алгоритмов локального выравнивания. Однако в то время количество и размеры доступных баз эталонных выравниваний были весьма ограничены, и, следовательно, данный анализ не был достаточно полным, а заключение достаточно обоснованным.

С тех пор количество данных о выравниваниях значительно увеличилось, и на сегодняшний день существует много различных независимых баз эталонных выравниваний аминокислотных последовательностей. Однако, несмотря на успехи в развитии баз эталонных выравниваний [7], по-прежнему остается открытой основная проблема, связанная с тем, насколько можно доверять этим выравниваниям и можно ли считать их золотым стандартом. В последнее время появляются все больше работ, посвященных проверке таких баз данных, например, на основе гомологии доменов или соответствия с вторичной структурой белков [8].

В настоящей работе мы анализируем базу данных (БД) эталонных выравниваний PREFAB [9], в частности, дополняем ее сведениями о гомологичности выравниваемых последовательностей с точки зрения классификации SCOP [10].

1.1. Принципы построения баз эталонных выравниваний. Структурная классификация белков. Современные базы эталонных выравниваний (БЭВ) аминокислотных последовательностей, как правило, построены на основе структурных выравниваний белков, т.е. выравниваний, основанных на совмещении пространственных структур. В то же время некоторые БЭВ (см. ниже) включают выравнивания, полученные только на основе анализа последовательностей. Различные базы отличаются выбором семейств белков, использованными алгоритмами наложения структур, методикой уточнения алгоритмических структурных выравниваний, которое обычно проводится экспертами.

Несмотря на то, что выравнивание последовательностей, построенное с учетом соответствующего структурного выравнивания, считается более верным с биологической точки зрения, для такого подхода существует ряд ограничений. Во-первых, надо внимательно следить за тем, чтобы разрешение структур было достаточно высоким (меньше 3 Å), иначе структурное выравнивание может получиться просто бессмысленным. Во-вторых, различные программы структурных выравниваний могут построить разные структурные выравнивания одних и тех же последовательностей [11,12], и часто трудно определить какое из них более правильное. Поэтому при анализе БД эталонных выравниваний это необходимо учитывать.

Выравнивание разных вторичных структур, например α-спирали и β-тяжа, принято считать заведомо неверным. Разумеется, наличие разных классификаций [13] и нередкое расхождение между методами предсказания вторичной структуры вносят некоторые, порой трудноразрешимые, проблемы, однако такая оценка корректности эталонного выравнивания, основанная на сравнении вторичных структур, используется довольно часто.

Наиболее популярным методом оценки баз эталонных выравниваний является определение типа укладки выравниваемых структурных доменов и дальнейшее сравнение этих типов, так как представляется нецелесообразным выравнивать домены из разных семейств. Самыми известными классификациями структурных доменов являются SCOP [10] и CATH [14]. Они отличаются как способом определения доменов (ручной или автоматический), так и самой системой классификации. В базе данных CATH процедура выделения доменов автоматическая; до 2003 г. они выделялись тремя алгоритмами: DOMAK [15], DETECTIVE [16] и PUU [17]. В случае расхождения между результатами алгоритмов решение выносили эксперты. С 2003 г. основным методом выделения доменов стал CATHEDRAL [18,19]. Его принцип заключается в поиске похожих доменов среди уже выделенных. Если похожий домен не находится, то используется старая процедура.

В базе данных SCOP домены выделяются только экспертами, без участия специальных программ. Верхние уровни классификации: класс (Class), укладка (Fold), Суперсемейство (Superfamily), Семейство (Family). Здесь выделяются четыре основных класса (all α, all β, $\alpha + \beta$, α/β), а также несколько специальных (мембранные, маленькие домены и т.д.). Для того чтобы домены имели общий тип укладки, у них должны быть одинаковые основные элементы вторичной структуры, одинаково расположенные как в пространстве, так и по цепи белка. Принадлежность к суперсемейству означает явные признаки общего происхождения, а к одному семейству принадлежат белки, которые имеют не менее 30% сходства по последовательности или очень близкие структуру и функцию.

На наш взгляд, SCOP-классификация более предпочтительна, поскольку каждый домен анализируется экспертами, а не программой.

1.2. Обзор баз эталонных выравниваний. Наиболее популярными на данный момент являются такие базы данных, как BAliBASE [20–23], PREFAB, HOMSTRAD [24], OXBench [25], SABmark [26]. База PREFAB, последняя по времени создания и являющаяся основным предметом нашего интереса, подробно рассмотрена в разделе 2.3. В этом разделе описаны остальные перечисленные выше БД.

1.2.1. BAliBASE. База данных BAliBASE является одной из самых первых баз множественных эталонных выравниваний. Представленные здесь выравнивания были получены на основе структурных выравниваний (совмещений) с последующей ручной проверкой правильности полученных выравниваний для консервативных аминокислотных остатков. BAliBASE состоит из девяти разделов. Каждый из разделов отражает определенный класс ситуаций, с которыми может столкнуться программа множественного выравнивания. Примеры таких ситуаций: малое число далеких друг от друга последовательностей; последовательности с протяженными негомологичными N/C-концевыми участками или с большими внутренними вставками; выравнивание трансмембранных белков, доменов с повторами и инверсиями и даже линейных мотивов эукариотов. Текущая версия БД содержит 217 выравниваний, в каждом из которых выровнено от 4 до 142 последовательностей.

1.2.2. HOMSTRAD. Кластеризация БД белковых доменов HOMSTRAD основана на сходстве последовательности и структуры белков. Несмотря на то, что изначально она не была задумана как база эталонных выравниваний, многие авторы используют ее в качестве таковой. HOMSTRAD содержит данные не только о последовательности белка, но также о его структуре, предоставляя информацию из различных БД, в том числе из PDB [27], Pfam [28] и SCOP. Последняя версия HOMSTRAD включает в себя 1032 доменных семейства, представленных от 2 до 41 последовательностями, и еще 9602 семейства, в которых есть только один представитель, т.е. одна последовательность.

1.2.3. OXBench. В БД OXBench содержатся множественные выравнивания белков, построенные с использованием методов как структурного выравнивания, так и выравнивания последовательностей. База данных содержит три раздела. Первый – «главный» (master) – раздел состоит из 673 выравниваний доменов белков с известной 3D структурой, от 2 до 122 последовательностей в каждом. Второй раздел (extended – «расширенный») был получен на основе главного раздела с добавлением последовательностей с неизвестной структурой. Третий раздел называется full-length, он также был создан на основе выравниваний из главного раздела, но только здесь выравнивается не один домен, а вся последовательность.

1.2.4. SABmark. БД SABmark содержит эталонные парные выравнивания последовательностей с известной 3D структурой. SABmark состоит из двух разделов: Twilight (последовательности с парным сходством Blast E-value ≥1) и Superfamilies (последовательности с парной идентичностью ≤ 50%). Оба раздела в свою очередь разбиты на группы согласно SCOP классификации: по укладке (для Twilight) и по надсемейству (для Superfamilies). Эталонное выравнивание для каждой пары последовательностей из группы строилось с помощью программ структурного выравнивания: SOFI[29] и CE[30].

2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. База данных PREFAB. *2.1.1. Общие сведения и структура.* База эталонных выравниваний PREFAB (*Protein Reference Alignment Benchmark*) была сконструирована Р. Эдгаром в 2004 г. для тестирования качества работы программ множественного выравнивания.

База данных PREFAB содержит:

1). Набор эталонных парных выравниваний.

2). Выборки последовательностей для тестирования программ множественного выравнивания.

3). Программу оценки качества работы программ множественного выравнивания.

Ниже рассматриваются только парные выравнивания.

Последняя версия PREFAB – PREFAB v.4.0 [31] – была опубликована Р. Эдгаром в марте 2005 г. В ней содержится 1682 эталонных парных выравнивания.

Каждое выравнивание находится в отдельном файле формата FASTA (или, точнее, FSSP FASTA). Имя файла имеет вид NAME1_NAME2, где NAME1, NAME2 – имена выравниваемых последовательностей. Именем каждой последовательности является либо просто ее PDB-идентификатор (четыре символа), либо PDB-идентификатор и цепь (пять символов) в случаях, когда она явно задана. Имена последовательностей в PREFAB соответствуют именам их FSSP структур. Согласно FSSP FASTA формату заглавными буквами в самом выравнивании выделены выровненные позиции, а строчными – невыровненные. При оценке качества алгорит-

мически построенного выравнивания учитываются только выровненные позиции.

2.1.2. Конструирование парных эталонных выравниваний. Выравнивания, вошедшие в БД PREFAB, были получены, как описано в работе [9]. Сначала были взяты парные выравнивания из тестовых баз данных, которые были сконструированы и описаны Садреевым и Гришиным [32] и Эдгаром и Сьоландером [33,34]. Выравнивания, входящие в указанные базы, были извлечены из БД FSSP [35], затем они перевыравнивались с помощью программы структурного выравнивания CE. После этого были отобраны только те выравнивания, для которых FSSP и CE сошлись более чем на 50 позициях. Именно эти выравнивания и составляют выборку эталонных выравниваний БД PREFAB.

2.2. Предобработка PREFAB. Две основных стадии предобработки парных эталонных выравниваний PREFAB – верификация последовательностей и верификация эталонных выравниваний. В первом случае имеется в виду сравнение между собой последовательности из PREFAB-выравнивания и соответствующей ей PDB-последовательности. Под PDB-последовательностью мы понимаем последовательность белка из соответствующей PDB-записи, такую, что для всех ее аминокислотных остатков известны координаты. Под верификацией эталонных выравниваний мы подразумеваем сравнение типов структур сравниваемых доменов по классификации SCOP.

2.3. Приведение файлов к единому шаблону. Как уже было сказано в п. 2.1.1., файлы в PREFAB имеют название NAME1_NAME2, где NAME1 и NAME2 – имена выравниваемых последовательностей. Однако такое название файла вовсе не означает, что последовательности в файле идут в том же порядке, что и в названии файла. А для некоторых программ это может оказаться важным. Поэтому все файлы в PREFAB были приведены к общему шаблону: порядок последовательностей в названии файла соответствует порядку последовательностей в самом файле. В ходе выполнения этой стадии были выявлены так называемые файлы-повторы, т.е. файлы с именами NAME1_NAME2 и NAME2_NAME1. Такие файлы содержат практически идентичные выравнивания. Так как пока неизвестно, какое из таких выравниваний можно будет считать более правильным, они были оставлены для дальнейшего анализа.

На этом же шаге проводилась проверка имен последовательностей на предмет содержания устаревших идентификаторов по сравнению с текущей версией PDB банка. Все устаревшие идентификаторы были заменены на новые.

2.4. Верификация последовательностей. *2.4.1. Присвоение уникального идентификатора.* К сожалению, в PREFAB одно и то же имя последовательности может означать разные последовательности – разные фрагменты одной белковой цепи. К тому же одна и та же последовательность может встречаться в разных выравниваниях. Поэтому, чтобы избежать ошибки, связанные с дальнейшим анализом, каждой последовательности присваивается уникальный идентификатор вида NAME.ALIGN_NAME, где NAME – имя последовательности, а ALIGN_NAME – имя выравнивания, из которого эта последовательность была взята. Последовательности с такими идентификаторами будем называть PREFAB-последовательностями.

2.4.2. Получение PDB-последовательностей. Для каждой PREFAB-последовательности из PDB банка извлекался соответствующий документ, последовательность нужной цепи извлекалась из полей ATOM. При этом все модифицированные остатки были заменены на обычные, например формилметионин на метионин, моноизопропилфосфорилсерин на серин. Селенометионин, который часто используется в рентгеноструктурном анализе, был также изменен на метионин.

2.4.3. Построение выравнивания между PREFAB-последовательностями и PDB-последовательностями. Каждая PREFAB-последовательность выравнивалась с соответствующей PDB-последовательностью (строилось глобальное выравнивание). Возможны следующие варианты:

А. Выравнивания не имеют вставок.

Б. Выравнивания имеют вставки на границах в PREFAB.

В. Выравнивания имеют внутренние вставки в PREFAB.

Г. Выравнивания имеют вставки на границах в PDB.

Д. Выравнивания имеют внутренние вставки только в PDB.

Е. Выравнивания имеют внутренние вставки и в PDB, и в PREFAB.

Если построенное выравнивание PREFAB-последовательности и PDB-последовательности удовлетворяет случаям А и Г, т.е. не содержит внутренние вставки (или гэпы) в PDB и никакие вставки в PREFAB, мы принимаем PREFAB-последовательность для дальнейшего анализа. Случаи Б и В рассматриваются, как опечатки, PREFAB-выравнивания, содержащие такую последовательность, редактируются. Если же выравнивание содержит вставки в PDB-последовательности (случаи Д и Е), т.е. если в PREFAB представлена неполная последовательность,

Результаты верификации последовательностей

Параметры	Количество последовательностей в PREFAB	Количество выравниваний в PREFAB
Одиночные замены	440	345
Вставка в PREFAB-последовательности	34	31
Удаления в PREFAB-последовательности	580	575

PREFAB-последовательность и соответствующее PREFAB-выравнивание удаляются. Любые замены в построенном выравнивании считаются опечатками, PREFAB-последовательность в PREFAB-выравнивании редактируется согласно ее PDB-последовательности.

2.5. Определение домена. В качестве источника классификации доменов была взята база данных SCOP v. 1.75. Для каждой PREFAB-последовательности определяются все возможные SCOP-домены данной белковой цепи. Дальнейшая идентификация SCOP-домена (или доменов) состоит в сравнении соответствующих координат, т.е. координат домена и PREFAB-последовательности согласно последовательности белка. Для каждого домена вычисляется его перекрытие с PREFAB-последовательностью. Перекрытие вычисляется как длина пересечения данного SCOP-домена и PREFAB-последовательности, деленная на длину PREFAB-последовательности. Если перекрытие больше 0,95 (95%), то считается, что PREFAB-последовательность однозначно задается данным SCOP-доменом. Домены, для которых перекрытие равно нулю, из рассмотрения исключаются. Если возможных SCOP-доменов несколько, то каждый домен сначала рассматривается отдельно, и если последовательность однозначно не определяется одним из предполагаемых доменов, то рассматривается суммарное перекрытие оставшихся доменов. Домены принимаются, если оно будет больше 0,95 (95%). Если для PREFAB-последовательности не определен ни один SCOP-домен, то такая последовательность и соответствующее ей выравнивание удаляются.

2.6. Верификация выравнивания. На этом этапе происходит отбор выравниваний путем сравнения SCOP-доменов выравниваемых последовательностей. Выравнивание считается прошедшими верификацию, если SCOP классификация сравниваемых доменов совпадает до семейства. В случае если для одной из PREFAB-последовательностей определено несколько доменов, появляется дополнительное условие отбора: количество доменов у другой последовательности должно быть таким же. Если это соблюдается, то домены сравниваются попарно, согласно их положению по цепи, т.е. первый домен сравнивается с первым, второй со вторым и так далее. Выравнивание принимается, если каждая пара сравниваемых доменов принадлежит одному семейству.

2.7. Проверка на повторы. Последний этап предобработки заключается в отборе файлов-повторов. Из каждой пары файлов выбирается только тот, в котором значения перекрытий выравниваемых последовательностей больше. Если значения совпадают, то принимается первый из файлов в списке.

РЕЗУЛЬТАТЫ

При замене устаревших PDB-идентификаторов выяснилось, что одна из записей (1bef) была признана некачественной и была удалена из банка PDB. Два выравнивания, в состав которых входит эта последовательность, были удалены.

Верификация последовательностей показала (см. таблицу), что у значительной части последовательностей базы данных PREFAB есть пропущенные внутренние фрагменты. Такие неполные белковые последовательности встречаются в 575 PREFAB-выравниваниях. Было найдено 34 случая наличия вставки в PREFAB-последовательности, причем практически всегда это был один дополнительный аминокислотный остаток в начале или в конце последовательности. Эти случаи рассматривались как опечатки, последовательности были отредактированы. Для 440 PREFAB-последовательностей были идентифицированы одиночные несовпадения с соответствующими PDB-последовательностями. Интересно, что из всех аминокислот чаще всего в PREFAB заменялись метионин и цистеин. И если замену метионина в PREFAB на любую аминокислоту, обозначающуюся символом «X», можно легко объяснить тем, что в PDB-записи в этой позиции стоит селенометионин, который мы в процессе предобработки заменили на метионин, то с цистеином (и с любой другой аминокислотой) дело обстоит сложнее. Причем помимо замены на любой аминокислотный остаток бывали случаи, когда полярный незаряженный цистеин заменялся, например, на не-

полярный аланин. При более детальном рассмотрении выяснилось, что чаще всего замены происходили в случае модифицированного аминокислотного остатка в PDB-записи.

Также были обнаружены случаи, когда в PREFAB-последовательности есть участки, для которых в PDB-записи неизвестны координаты, что является весьма странным, ведь эталонное выравнивание с данной последовательностью строилось на основе выравнивания структур.

При определении SCOP-домена выяснилось, что последовательность 1mfa состоит сразу из двух белковых цепей (L и H), каждая из которых содержит свой SCOP-домен. Эта последовательность и соответствующее ей выравнивание (1mfa_1neu) были исключены из рассмотрения.

SCOP-домены были определены для каждой PREFAB-последовательности. Сравнение их классификаций показало, что в 581 выравнивании, т.е. в 31,2% от всей БД PREFAB, выравниваются гомологичные последовательности, чьи домены принадлежат одному семейству. Причем 502 из таких выравниваний содержат последовательности, которые определяются одним SCOP-доменом.

В PREFAB v4.0 была обнаружена 61 пара файлов-повторов. После окончания предобработки таких пар осталось 22, одно выравнивание из каждой пары было выбрано случайным образом.

Полученная в итоге описанной выше работы база данных PREFAB-P доступна по адресу http://server2.lpm.org.ru/static/prefab-p/.

Дополнительные материалы: http://server2.lpm.org.ru/~irina/supplementary.rar.

ВЫВОДЫ

В работе представлен анализ базы эталонных выравниваний PREFAB, включающий в себя определение гомологии выравниваемых последовательностей на основе классификации SCOP.

Мы провели предобработку БД PREFAB и отобрали только те выравнивания, последовательности которых гомологичны друг другу. Было обнаружено, что в некоторых выравниваниях базы данных PREFAB представлены последовательности, для которых SCOP классификация расходится не только на уровне семейства, но и на более высоких уровнях, таких как суперсемейство, укладка и даже класс.

На основании проведенного анализа создана база выравниваний PREFAB-P. Следующим шагом работы будет оценка достоверности отдельных элементов эталонных выравниваний.

Работа выполнена в рамках государственного контракта № 07.514.11.4004, шифр «2011-1.4-514-008-009» при финансовой поддержке Минобрнауки России.

СПИСОК ЛИТЕРАТУРЫ

1. R. F. Smith and T. F. Smith, Protein Eng. **5**, 35 (1992).
2. E. Deperieux, G. Baudoux, P. Briffeuil, et al., Comput. Appl. BioSci. **13**, 249 (1997).
3. S. R. Eddy, ISMB **3**, 114 (1995).
4. B. Morgenstern, A. Dress, and T. Werner, Proc. Natl. Acad. Sci. USA **93**, 12098 (1996).
5. J. D. Thompson, T. J. Gibson, F. Plewniak, et al., Nucl. Acids Res. **24**, 4876 (1997).
6. M. A. McClure, Vasi T. K., Fitch W. M., Mol. Biol. Evol. **11**, 571 (1994).
7. M. R. Aniba, O. Poch, and J. D. Thompson, Nucl. Acids Res. **38** (21), 7353 (2010).
8. R. C. Edgar, Nucl. Acids Res. **38**, 2145 (2010).
9. R. C. Edgar, Nucl. Acids Res. **32**, 1792 (2004).
10. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).
11. H. Hasegawa and L. Holm, Curr. Opin. Struct. Biol. **19**, 341 (2009).
12. A. Godzik, Protein Sci. **5**, 1325 (1996).
13. C. Etchebest, C. Benros, S. Hazout, and A. G. de Brevern, Proteins **59**, 810 (2005).
14. C. Orengo, A. Michie, S. Jones, et al., Structure **5**, 1093 (1997).
15. A. S. Siddiqui and G. J. Barton, Protein Sci. **42**, 372 (1995).
16. M. B. Swindells, Protein Sci. **4**, 103 (1995).
17. L. Holm and C. Sander, Proteins **19**, 256 (1994).
18. A. Harrison, F. Pearl, R. Mott, et al., J. Mol. Biol. **5** (323), 909 (2002).
19. F. M. Pearl, C. F. Bennett, J. E. Bray, et al., Nucl. Acids Res. **31**, 452 (2003).
20. J. D. Thompson, F. Plewniak, and O. Poch, Bioinformatics **15**, 87 (1999).
21. A. Bahr, J. D. Thompson, J. C. Thierry, and O. Poch, Nucl. Acids Res. **29**, 323 (2001).
22. J. D. Thompson, P. Koehl, R. Ripp, and O. Poch, Proteins **61**, 127 (2005).
23. E. Perrodou, C. Chica, O. Poch, et al., BMC Bioinformatics **9**, 213 (2008).
24. K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington, Protein Sci. **7**, 2469 (1998).
25. G. P. Raghava, S. M. Searle, P. C. Audley, et al.,. BMC Bioinformatics **4**, 47 (2003).
26. I. Van Walle, I. Lasters, and L. Wyns, Bioinformatics **21**, 1267 (2005).
27. H. M. Berman, K. Henrick, and H. Nakamura, Nat. Struct. Biol. **10**, 980 (2003).

28. R. D. Finn, J. Mistry, J. Tate, et al., Nucl. Acids Res. **38**, 211 (2010).

29. N. S. Boutonnet, M. J. Rooman, M. E. Ochagavia, et al., Protein Eng. **8**, 647 (1995).

30. I. N. Shindyalov and P. E. Bourne, Protein Eng. **11**, 739 (1998).

31. PREFAB v. 4.0: http://www.drive5.com/muscle/prefab.htm

32. R. Sadreyev and N. Grishin, J. Mol. Biol. **326**, 317 (2003).

33. R. C. Edgar and K. A. Sjolander, Bioinformatics, DOI: 10.1093/bioinformatics/bth090 (2004).

34. R. C. Edgar and K. Sjolander, Bioinformatics, DOI: 10.1093/bioinformatics/bth091 (2004).

35. L. Holm and C. Sander, Nucl. Acids Res. **26**, 316 (1998).

Verification of the PREFAB Alignment Database

T.V. Astakhova*, M.N. Lobanov, I.V. Poverennaya* ***, M.A. Roytberg*, and V.V. Yacovlev***

**Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

***Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

****Faculty of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119991 Russia*

The verification of the PREFAB database containing golden standard protein alignments was performed. It has revealed a significant number of differences between the sequences from PREFAB and PDB databases. It was shown that compared to the sequences given in the PDB database 575 alignments refered to a sequence with a gap; such alignments were excluded. Furthermore, compared to the PDB-sequences a single substitute or the insertions were found for 440 aminoacid sequences from PREFAB database; these sequences were edited. SCOP domain analysis has shown that only 502 alignments in the resulting set contain the sequences from the same family. Finally, eliminating duplicates, we have created a new golden standard alignment database PREFAB-P based on PREFAB; the PREFAB-P database contains 581 alignments.

Key words: aminoacid sequences, golden standards, PDB structure, SCOP classification