Московский Государственный Университет имени М.В.Ломоносова

Факультет биоинженерии и биоинформатики

Использование БД PREFAB для анализа алгоритмов выравнивания аминокислотных последовательностей.

Курсовая работа

студентки 3 курса

Поверенной Ирины Владимировны

Тьютор: Ройтберг Михаил Абрамович,

д.ф.-м.н., зав. лабораторией прикладной математики

Института математических проблем биологии РАН

Москва

2011

ОГЛАВЛЕНИЕ

1. **Введение.**

1.1. Актуальность темы

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 наиболее важными свойствами таких программ являются биологическая точность и вычислительная сложность, т.е. время и требования к памяти (особенно это значимо в случае программ множественного выравнивания). 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 годов) 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 году. 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 за золотой стандарт, для каждого семейства получилась своя база данных 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 пригодных баз данных

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 существуют много различных независимых баз данных эталонных выравниваний аминокислотных последовательностей (test set of reference alignments или benchmark). Однако, несмотря на успехи в развитии баз данных эталонных выравниваний (на английском этот процесс называется benchmarking)[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, касающихся проверки таких баз данных, например, на основе гомологии доменов или соответствия с вторичной структурой белков [8].

1.2. Цель работы.

Цель данной работы – анализ БД эталонных выравниваний PREFAB[9] и определение гомологии выравниваемых последовательностей на основе классификации SCOP[10].

2. **Базы данных эталонных выравниваний.**

2.1. Принципы построения БД эталонных выравниваний. Анализ структурных выравниваний.

Современные БД эталонных выравниваний аминокислотных последовательностей, как правило, построены на основе структурных выравниваний белков, т.е. выравниваний, основанных на совмещении пространственных структур. В тоже время, некоторые БД (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, существует ряд ограничений для такого подхода. Во-первых, надо внимательно следить за тем, чтобы разрешение структур было достаточно высоким (меньше 3 ангстрем), 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, альфа-спирали и бета-тяжа, принято считать заведомо неверным. Разумеется, наличие разных классификаций [13] и нередкое расхождение между методами предсказания вторичной структуры вносят некоторые, порой трудноразрешимые, трудности, однако такая оценка корректности эталонного выравнивания довольно часто используется.

Наиболее популярным методом оценки баз данных эталонных выравниваний является определение выравниваемых структурных доменов и их дальнейшее сравнение, ведь кажется весьма странным выравнивать домены из разных семейств. Самые известные классификации структурных доменов SCOP[10] и CATH[14]. Они отличаются как в плане определения доменов (ручная или автоматическая), так и самой системы классификации. В БД CATH процедура выделения доменов автоматическая; до 2003 года они выделялись тремя алгоритмами: DOMAK[15], DETECTIVE[16] и PUU[17]. В случае расхождения между результатами алгоритмов решение выносили эксперты. С 2003 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. Верхние уровни CATH классификации: класс, архитектура, топология, суперсемейство.

В БД 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, α+β, α/β**), а также несколько специальных (мембранные, маленькие домены и т.д.). Для того чтобы домены принадлежали одной укладке, 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 семейству 30% сходства по последовательности или очень близкие структура и функция.

Вследствие того, что у обеих БД совершенно разная система классификации, семейство в SCOP и семейство в CATH совсем не одно и тоже. На наш взгляд, SCOP

классификация более предпочтительна в том плане, что каждый домен определяется вручную экспертами, а не программой.

2.2.Обзор наиболее популярных БД эталонных выравниваний.

Наиболее популярными на данный момент является такие БД, 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.

BAliBASE

БД BAliBASE является одной из самых первых крупнейших баз данных 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 9 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; последовательностей 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.

HOMSTRAD

Кластеризация БД белковых доменов HOMSTRAD основана на сходстве последовательности и структуры белков. Несмотря на то, что изначально она не была задумана как база данных 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.

OXBench

В БД OXBench содержатся множественные выравнивания белков, построенные с использованием методов как структурного выравнивания, так и выравнивания последовательностей. Вся БД разделена на 3 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.

SABmark

БД SABmark содержит эталонные парные выравнивания последовательностей с известной 3D структурой. SABmark состоит из 2 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].

2.3.БД PREFAB.

2.3.1. Общие сведения и структура.

База данных эталонных выравниваний PREFAB (protein reference alignment benchmark) была сконструирована Р.Эдгаром в 2004 году для тестирования качества работы программ множественного выравнивания.

БД PREFAB содержит:

А. Набор эталонных парных выравниваний.

Б. Выборки последовательностей для тестирования программ множественного выравнивания.

В. Программу оценки качества работы программ множественного выравнивания.

Основное внимание будет уделено только парным выравниваниям.

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

Каждое выравнивание находится в отдельном файле в FASTA формате (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 (4 символа), либо PDB-идентификатор и цепь (5 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. Для оценки качества алгоритмически построенного выравнивания должны учитываться только выровненные позиции. Эталонные выравнивания находятся в директории *./ref*.

2.3.2. Конструирование парных эталонных выравниваний.

Выравнивания, вошедшие в БД PREFAB, были получены следующим образом[9]. Сначала были взяты парные выравнивания из тестовых баз данных SG [32], PP1 и PP2 [33, 34], которые были сконструированы и описаны Садреевым и Гришиным (Sadreyev и Grishin) и Эдгаром и Сьоландером (Edgar и Sjolander), соответственно; выравнивания входящие в указанные базы, были извлечены из БД FSSP [35]. 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.

2.3.3. Другие элементы БД PREFAB.

Помимо эталонных выравниваний БД содержит выборки последовательностей, которые предполагается давать на вход тестируемой программе множественного выравнивания. Это также файлы в FASTA формате, их имена аналогичны именам файлов с соответствующими эталонными выравниваниями. Последовательности в файле аннотированы следующим образом:

>1abcA

>123|1abcA|gi|12345678

В виде PDB идентификатора (просто или с цепью) обозначается последовательность, присутствующая в эталонном выравнивании. Второй случай соответствует хитам – похожим последовательностям, найденным в БД NCBI nr [36] с помощью PSI-BLAST поиска [37, 38].

Поля в данной аннотации означают следующее:

123	номер данного хита в списке[1] всех хитов, выданных PSI-BLAST
1abcA	имя последовательности, для которой производился поиск
gi|12345678	NCBI идентификатор хита

Такая аннотация была выбрана для избегания проблем в таких программах как CLUSTALW[39] и PHYLIP[40], которые в качестве ключа-идентификатора к последовательности используют первое поле аннотации. Множественные выравнивания находятся в директории *./in*.

Также 4 версия БД PREFAB содержит текстовый файл с перечислением выравниваний с протяженными гэпами (>10 а.о.) и программу оценки точности построенных алгоритмических множественных выравниваний.

3. Предобработка PREFAB.

Вся работа проводится только с парными эталонными выравниваниями в PREFAB.

Две основных стадии предобработки 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 считаем последовательность белка из соответствующей PDB записи, такую, что для всех ее аминокислотных остатков известны координаты. Под верификацией эталонных выравниваний мы подразумеваем сравнение SCOP доменов выравниваемых последовательностей.

3.1. Приведение файлов к единому шаблону.

Как уже было сказано в п. 2.2.1., 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

[1] Список, полученный PSI-BLAST до фильтрации по проценту идентичности и случайного выбора 24 последовательностей. Подробнее о конструировании входных выборок см. [9].

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. Т.к. пока не известно, какое из таких выравниваний можно будет считать более правильным, они были оставлены для дальнейшего анализа.

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

3.2. Верификация последовательностей.

3.2.1. Присвоение уникального идентификатора.

К сожалению, в PREFAB одно и то же имя последовательности может означать разные последовательности – 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.

3.2.2. Получение PDB записей.

Для каждой PREFAB последовательности из PDB банка скачивается соответствующую запись, из которой вытягивается PDB последовательность. Надо заметить, что в PDB последовательности все модифицированные остатки были заменены на обычные, например, формилметионин на метионин, моноизопропилфосфорилсерин на серин. Селенометионин, который часто используется в рентгеноструктурном анализе (РСА), был также изменен на метионин.

3.2.3. Построение выравнивания между PREFAB и PDB последовательностями.

Каждая PREFAB последовательность выравнивается с соответствующей PDB последовательностью (строится 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 и PDB последовательностей 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 считаются за опечатку, PREFAB выравнивания, содержащие такую последовательность, редактируются. Если же выравнивание содержит вставки в PDB последовательности (случай Д и Е), т.е. если в PREFAB представлена неполная последовательность, PREFAB последовательность и соответствующее PREFAB выравнивание удаляются. Любые замены в построенном выравнивании считаются за опечатку, PREFAB последовательность в PREFAB выравнивании редактируется согласно ее PDB последовательности.

3.3. Определение SCOP домена.

В качестве источника классификации была взята БД SCOP v1.75. 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 считается 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.

3.4. 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.

3.5. Проверка на повторы.

Последний этап предобработки заключается в отборе файлов-повторов, найденных в п.3.2.1. Из каждой пары файлов выбирается только тот, в котором значения перекрытий выравниваемых последовательностей больше. Если значения совпадают, файл выбирается случайным образом.

4. 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, что у 17,3% PREFAB последовательностей есть пропущенные внутренние фрагменты. Такие неполные белковые последовательности встречаются в 575 (34,2%) PREFAB выравниваниях. Было найдено всего 34 случая наличия вставки в PREFAB последовательности, причем практически всегда это был один дополнительный аминокислотный остаток в начале или в конце последовательности. Вследствие этого, было отредактировано 31 выравнивание. Для 440

(13,07%) 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 неполярный аланин.

Таблица 1. Результаты верификации последовательностей.

Параметры	**Количество PREFAB последовательностей**	**Количество PREFAB выравниваний**
Одиночные замены	440	345
Гэпы в PREFAB п-ти[2]	34	31
Вставка в PREFAB п-ти	580	575

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 2 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.

[2] PREFAB последовательность

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

5. Выводы.

Целью работы было проанализировать БД эталонных выравниваний PREFAB[9] и определить гомологию 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.

Мы планируем продолжить анализ БД PREFAB. Нашим следующим шагом является оценка достоверность PREFAB выравниваний путем построения на их основе структурных выравниваний с последующим вычисления расстояния между каждой парой выровненных остатков.

Благодарим В.В. Яковлева за помощь в реализации методики предобработки и обсуждение, а также М.Ю.Лобанова за консультации и помощь при построении структурных выравниваний.

6. Список литературы.

1. Smith,R.F. and Smith,T.F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng*. **5**, 35-41.

2. Deperieux,E. Baudoux,G., Briffeuil,P. Reginster,I., De Bolle,X, Vinals,C. and Feytmans,E. (1997) MATCH-BOX server: a multiple sequence alignment tool placing emphasis on reliability. *Comput. Appl. BioSci.*, **13**, 249-256.

3. Eddy,S.R. (1995) Multiple alignment using hidden Markov models. *ISMB*, **3**, 114-120.

4. Morgenstern,B., Dress,A. and Werner,T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci, USA.* **93**, 12098-12103.

5. Thompson,J.D., Gibson,T.J, Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence aligment aided by quality analysis tools. *Nucleic Acids Res.*, **24**, 4876-4882.

6. McClure MA, Vasi TK, Fitch WM. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*; **11**: 571-592.

7. Mohamed Radhouene Aniba, Olivier Poch, Julie D. Thompson. (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Res.*;**38(21)**: 7353-7363

8. Edgar R.C. (2010) Quality measures for protein alignment benchmarks. *Nucleic Acids Res.*;**38**:2145-2153.

9. Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*;**32**:1792-1797.

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

11. Hasegawa H, Holm L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*;**19**:341–348.

12. Godzik A.(1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*;**5**:1325–1338.

13. Etchebest C, Benros C, Hazout S, de Brevern AG. (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins.*;**59**:810–827.

14. Orengo,C., Michie,A., Jones,S., Jones,D., Swindells,M. and Thornton,J. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

15. Siddiqui A.S., Barton G.J. (1995) Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci*, **42**, 372-884.

16. Swindells M.B. (1995) A procedure for detecting structural domains in proteins. *Protein Sci*., **4**, 103-112.

17. Holm, L., Sander,C. (1994) Parser for protein folding units. *Proteins*, **19**, 256–268.

18. Harrison,A., Pearl,F., Sillitoe,I., Thornton,J. and Orengo,C. (2002) CATHEDRAL: an effective algorithm to delineate previously seen folds within a multi-domain structure. In preparation.

19. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.,* **31**, 452–455.

20. Thompson JD, Plewniak F, Poch O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics***;15**:87-88.

21. Bahr A, Thompson JD, Thierry JC, Poch O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*;**29**:323-326.

22. Thompson JD, Koehl P, Ripp R, Poch O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins.*;**61**:127-136.

23. Perrodou E, Chica C, Poch O, Gibson TJ, Thompson JD. (2008) A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*;**9**:213.

24. Mizuguchi K, Deane CM, Blundell TL, Overington JP. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*;**7**:2469-2471.

25. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*;**4**:47.

26. Van Walle I, Lasters I, Wyns L. (2005) SABmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics;***21**:1267-1268.

27. Berman HM, Henrick K, Nakamura H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*;**10**:980.

28. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al. (2010) The Pfam protein families database. *Nucleic Acids Res.*;**38**:D211-D222.

29. Boutonnet,N.S., Rooman,M.J., Ochagavia,M.E., Richelle,J. and Wodak,S.J. (1995) Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein Eng.*, **8**, 647–662.

30. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

31. PREFAB v4.0: http://www.drive5.com/muscle/prefab.htm

32. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

33. Edgar,R.C. and Sjolander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, DOI: 10.1093/bioinformatics/bth090.

34. Edgar,R.C. and Sjolander,K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics*, DOI: 10.1093/bioinformatics/bth091.

35. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.

36. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.

37. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

38. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

39. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

40. Felsenstein J.(1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.*, **17**, 368–376.