МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2009, том 43, № 3, с. 521–526

МАТЕМАТИЧЕСКАЯ И СИСТЕМНАЯ БИОЛОГИЯ

УДК 577.150.2

НОВОЕ СТРУКТУРНОЕ ДРЕВО (α + β)-БЕЛКОВ, СОДЕРЖАЩИХ abCd-ЕДИНИЦЫ

© 2009 г. А. Б. Гордеев, А. В. Ефимов*

Институт белка Российской академии наук, Пущино, Московская обл., 142290

Поступила в редакцию 15.07.2008 г.
Принята к печати 30.09.2008 г.

Создана база данных (α + β)-белков, содержащих abCd-единицы, которая включает в себя 926 белков из Банка белковых данных (PDB), в том числе негомологичных – 401 (выборка из 2636 PDB-файлов). Построено обновленное структурное древо для белков этого класса, которое содержит 286 возможных укладок полипептидной цепи. На основе построенного древа разработана современная структурная классификация (α + β)-белков, содержащих abCd-единицы. Вся информация о базе данных, а также структурное древо доступны в Интернете по адресу: http://strees.protres.ru/.

***Ключевые слова*: моделирование, классификация, база данных, сворачивание белков, структурный мотив, структурное сходство.**

NOVEL STRUCTURAL TREE FOR (α + β)-PROTEINS CONTAINING ABCD-UNITS, by *A. B. Gordeev, A. V. Efimov** **(Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia, *e-mail: efimov@protres.ru). A database of 926 (α + β)-proteins and (α + β)-domains containing abCd-units (among them 401 are nonhomologous) has been compiled from the Protein Data Bank (total 2636 PDB entries). A novel structural tree for this structural class of proteins that is composed of 286 possible polypeptide chain folds has been constructed. The structural classification of (α + β)-proteins containing abCd-unit based on the structural tree has been developed. Both the database and the structural tree are accessible at the web-site (http://strees.protres.ru/).**

***Key words*: modeling, classification, database, protein folding, structural motif, structural similarity.**

Структурное древо белков – это совокупность всех разрешенных промежуточных и конечных пространственных структур, которые можно получить из одной корневой (стартовой) структуры путем последовательного добавления к ней других элементов вторичной структуры; возможные пути роста структур показываются линиями, которые в итоге объединяют все структуры в одно древо. Присоединение элементов вторичной структуры к растущим структурам происходит в соответствии с набором правил, которые выведены из известных принципов структурной организации белков. В качестве корневой структуры древа берется соответствующий структурный мотив, имеющий уникальную укладку цепи в пространстве.

Первые варианты структурных деревьев построены более 10 лет назад. Тогда же опубликовано семь структурных деревьев для наиболее крупных белковых суперсемейств [1–4]. В 2008 г. в белках обнаружена новая супервторичная структура – φ-мотив, также построено структурное древо для белков, содержащих этот мотив [5]. За это время существенно выросло количество расшифрованных структур в Банке белковых данных (Protein Data Bank). Поэтому возникла необходимость построения обновленных структурных деревьев, содержащих все имеющиеся в Банке данных белки соответствующих классов. Нами уже построено обновленное структурное древо для класса β-белков, содержащих abcd-единицы, и проведен его анализ. На основе обновленного структурного древа создана иерархически организованная база данных всех β-белков, содержащих abcd-единицы. База данных включает в себя 528 белков и доменов и доступна в Интернете (http://strees.protres.ru/) [6].

Значительно увеличилось количество белковых структур и для других структурных деревьев. Так, например, в состав структурного древа варианта 1997 г. [2] входило порядка пятидесяти известных структур (α + β)-белков, содержащих abCd-единицы, а к настоящему времени мы собрали базу данных из 926 белков и доменов этого класса. Это предопределило необходимость построения обновленного структурного древа для этого класса белков.

* Эл. почта: efimov@protres.ru

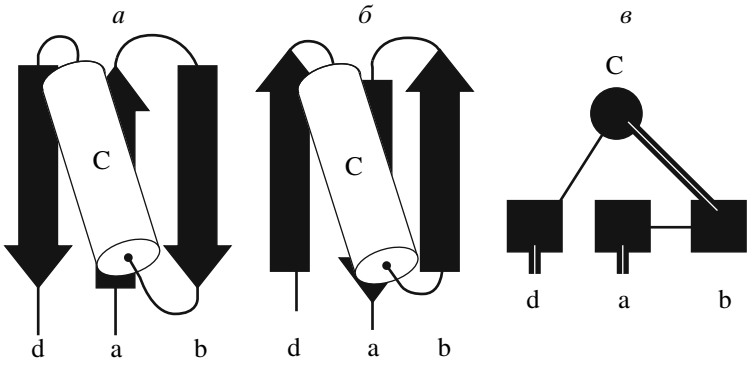

Рис. 1. Схематическое изображение abCd-единицы с прямым (*а*) и обратным (*б*) направлением цепи. β-тяжи показаны в виде стрелок, направленных от N- к С-концам, α-спираль – в виде цилиндра. *в* – Вид abCd-единицы с торца. Здесь β-тяжи изображены квадратиками, α-спираль – окружностью; перетяжки, направленные к наблюдателю, показаны двойными линиями, а удаленные – одинарными линиями. Буквами a, b, C и d обозначены соответствующие β-тяжи и α-спираль.

Построение и анализ обновленного древа позволяет исследовать новые пути роста структур, вести поиск новых способов укладки полипептидной цепи в пространстве, обнаружить новые закономерности в белковых структурах и с учетом всего этого модернизировать правила построения структурных деревьев. Решению этих задач и посвящена настоящая работа.

Структурные деревья – удобный и перспективный инструмент, с помощью которого можно решать целый ряд задач, например, проводить поиск всех возможных (как известных, так и пока неизвестных) укладок полипептидной цепи в компактные пространственные структуры, моделировать пути сворачивания белков, изучать механизм их сворачивания, исследовать структурное сходство белков и др. Одно из важных применений структурных деревьев – это разработка на их основе структурной классификации белков. Такая классификация базируется только на сходстве пространственных структур и на общности смоделированных путей сворачивания. Этим она принципиально отличается от других широко известных систем классификации белков [7–9], в которых учитываются в той или иной степени гомология аминокислотных последовательностей, а также информация о функциях и эволюционном родстве белков (см. также обзоры [10, 11]).

ОБЪЕКТЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

База данных (α + β)-белков, содержащих abCd-единицы, сформирована с помощью системы структурной классификации SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/), версия 1.73. Отбор белков осуществляли вручную. Всего отобрано 926 белков и доменов, содержащих abCd-единицы, из них 401 – негомологичных. С целью выявления возможной гомологии использовали программу BLAST для попарного выравнивания (http://www.ncbi.nih.gov/BLAST/) [12]. Анализ структуры белков проводили визуально с использованием программы молекулярной графики RasMol [13].

Построение структурного древа проводили в соответствии с правилами, сформулированными ранее [1–4]:

1. В качестве корневой структуры древа была взята abCd-единица, которая имеет уникальную пространственную укладку цепи [14]. abCd-единица представляет собой вариант abcd-единицы, характерной для β-белков, в котором вместо с-тяжа находится α-спираль С. Тяжи b и d и спираль С также, как и в случае abcd-единицы, образуют правую суперспираль bCd. Однако структурное сходство (α + β)-белков этого класса и β-белков, содержащих abcd-единицы, этим не ограничивается. Так же как в β-белках, в (α + β)-белках abCd-единицы часто располагаются на краях слоевых структур. Многие белки и домены этих двух классов имеют по существу одинаковую укладку цепей в пространстве, если не принимать во внимание конформации элементов вторичной структуры. abCd-единица может быть в двух вариантах – с прямым и обратным ходом полипептидной цепи (рис. 1*а* и 1*б*), однако, как будет показано ниже, примерно 77% abCd-единиц, обнаруженных нами в негомологичных белках, имеют обратный ход полипептидной цепи. При построении древа abCd-единица, также как и все другие структуры, представлена в упрощенном виде (рис. 1*в*) без указания направления цепи. При этом подразумевается, что разрешен как прямой, так и обратный ход цепи в каждом случае.

2. Пристраивание α-спиралей и/или β-тяжей к растущим корневой и промежуточным структурам проводили последовательно, шаг за шагом; при этом каждую структуру, полученную на предыдущем этапе, сохраняли в составе последующей и т.д. На каждом этапе элемент вторичной структуры,

расположенный ближе других к растущей структуре по цепи, пристраивался первым [2, 14].

3. В соответствии с принципом плотной упаковки все полученные структуры должны быть компактными.

4. Пересечение перетяжек [15] и образование узлов [16] запрещено.

5. Все структурные мотивы (т.е. не только корневые мотивы) должны иметь свойственную им хиральность и пространственную укладку цепи. В (α + β)-белках все β-α-β-единицы должны находиться в форме правых суперспиралей [14, 17].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

На рис. 2 представлен фрагмент построенного структурного древа. Полное структурное древо белков, содержащих abCd-единицы, размещено на сервере Института белка РАН по адресу (http://strees.protres.ru/). Обозначения для этого структурного древа точно такие же, как и для ранее построенного древа β-белков, содержащих abcd-единицы [6]. На структурном древе можно выделить несколько уровней, в каждом из которых находятся возможные укладки полипептидной цепи, состоящие из одинакового количества элементов вторичной структуры (β-тяжей и/или α-спиралей). Элементы вторичной структуры изображены в упрощенном виде: β-тяжи схематично представлены квадратиками, α-спирали – окружностями; перетяжки, направленные к наблюдателю, показаны двойными линиями, а удаленные – одинарными линиями. Все укладки в древе пронумерованы, и каждая укладка цепи имеет свой номер, состоящий из двух чисел, разделенных точкой. Первое число указывает количество добавленных к abCd-единице элементов вторичной структуры. Например, укладки первого уровня (ряда) содержат один добавочный элемент вторичной структуры (β-тяж или α-спираль) и обозначаются цифрой 1, второго ряда – цифрой 2 и т.д. Второе число обозначает порядковый номер укладки, если считать их слева направо в каждом ряду. Например, в третьем ряду находятся всего 57 укладок, и они имеют номера с 3.1 до 3.57.

С другой стороны, структурное древо имеет несколько ветвей. В составе одной ветви укладка цепи, находящаяся на более высоком уровне, содержит в себе укладки, расположенные ниже. Укладки из разных ветвей содержат в себе одну и ту же укладку, находящуюся в месте разветвления. Чем выше в структурном древе находится точка разветвления, тем выше уровень структурного сходства между белками или доменами соответствующих ветвей.

Следует отметить, что некоторые укладки могут быть образованы несколькими различными путями из разных родительских структур. Это вызывает определенные затруднения при определении структурного сходства между белками или доменами, а также делает древо громоздким за счет повторения отдельных ветвей. Чтобы избежать подобных ситуаций, при моделировании структурного древа мы использовали эмпирический подход. Он заключался в том, что сначала проводили моделирование всего древа, учитывая все теоретически возможные укладки и пути их роста. Затем, если получились две одинаковые ветви, произошедшие от разных родительских структур, подсчитывали заселенность этих ветвей белками с известными структурами. В древе оставляли только одну ветвь из нескольких одинаковых, ту, которая имела наибольшую заселенность известными структурами. В большинстве случаев это дало возможность определить предпочтительный путь роста структур.

При моделировании структурного древа (α + β)-белков нами введены некоторые ограничения. Отметим наиболее важные из них. Во-первых, мы не рассматривали структуры, состоящие из более чем трех слоев, поскольку подавляющее большинство белков этого класса (среди негомологичных – 386 из 401) состоят из двух (α-слой + β-слой) или трех слоев (α + β + α-слои). Во-вторых, отношение количества α-спиралей в α-слое к числу β-тяжей в β-слое не должно превышать 2 : 3 (иначе получаются некомпактные структуры – см. правило 3). При моделировании структурного древа мы не рассматривали пути роста, проходящие через такие структуры.

До уровня четырех добавочных элементов вторичной структуры в древо включены все теоретически разрешенные укладки. На более высоких уровнях показаны только те пути роста структур, которые ведут к структурам известных белков. Одна из причин этого состоит в том, что количество теоретически возможных укладок в верхних уровнях резко возрастает, и в целях экономии места на древе показаны только те ветви, которые содержат известные белки. Другая причина заключается в том, что заселенность верхних уровней известными белками быстро снижается, т.е. белки этого класса имеют ограниченный размер. Например, нами найдено всего два белка (из 401), имеющих по 9 добавочных элементов вторичной структуры.

С другой стороны, анализ древа показывает, что чаще всего встречаются белки, состоящие из abCd-единицы и двух-трех добавочных элементов (более 60%), т.е. большинство белков имеют такой оптимальный размер. В качестве примера на рис. 3 представлена укладка, которая состоит из abCd-единицы и двух добавочных элементов – α-спирали и β-тяжа. Одна только эта укладка встречается в 108 негомологичных белках из 401, т.е. непропорционально часто. Всего обновленное древо содержит 286 укладок, из них 80 найдены в 401 негомологичном белке, и 25% этих белков имеют одну и ту же укладку. Отметим, что высокую частоту встречаемости этой

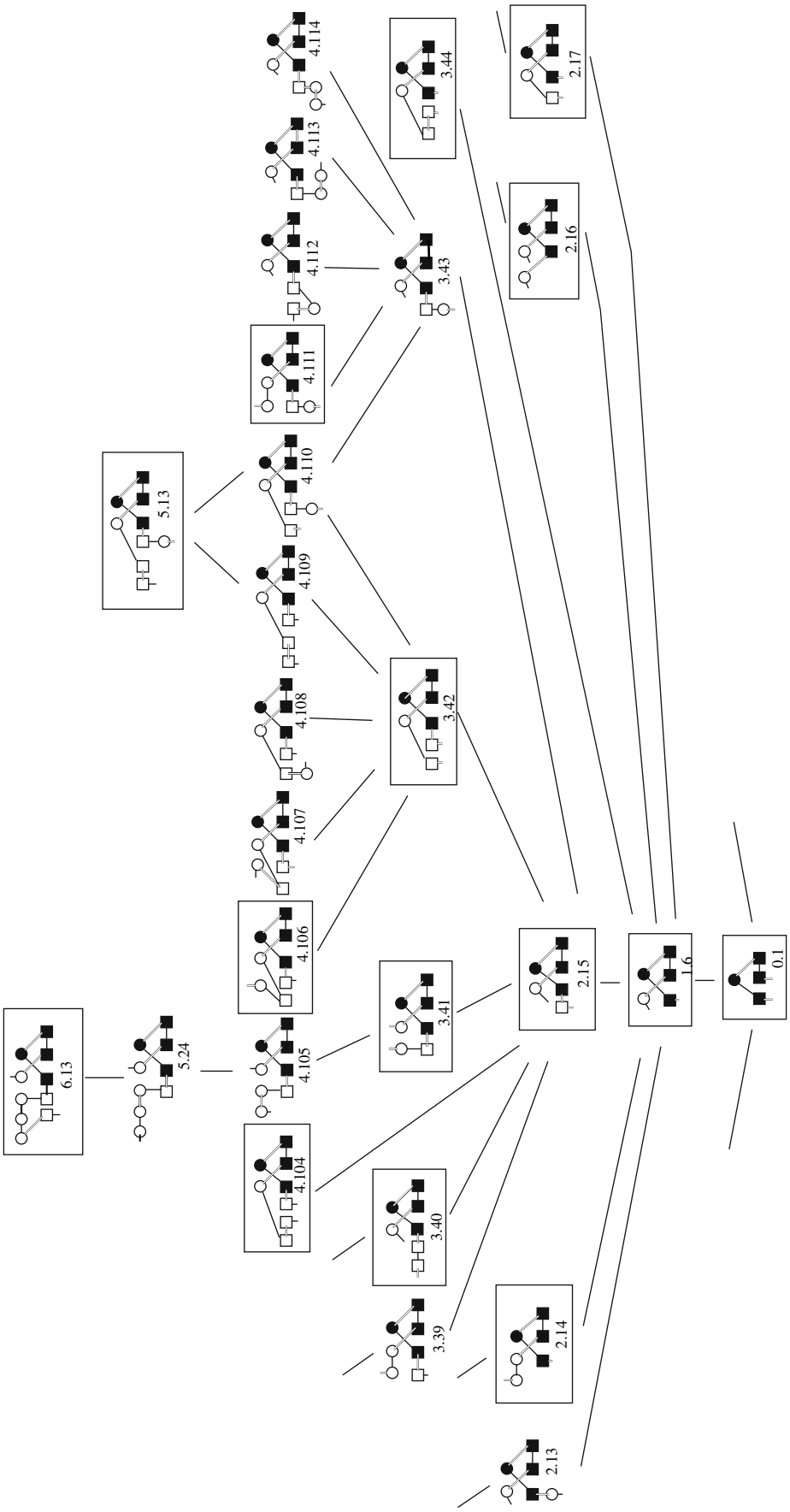

Рис. 2. Фрагмент структурного древа ($\alpha + \beta$)-белков, содержащих abCd-единицы.
Все структуры ориентированы одинаковым образом, и показан их вид с торца, как на рис. 1*в*. Укладки цепи, найденные в белках, обведены в рамочки.

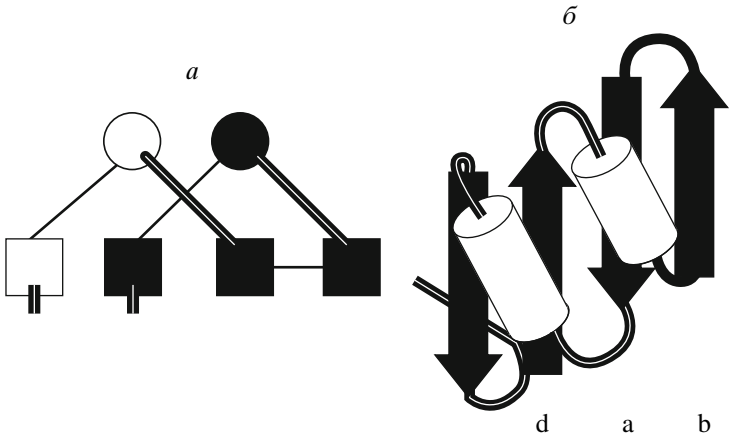

Рис. 3. Часто встречаемая укладка, обнаруженная в 256 белках и доменах. *а* – Вид укладки с торца. *б* – Схематическое изображение укладки на примере белка ферредоксин (PDB-код 1ir0).

укладки в белках наблюдали и ранее [14, 18], хотя статистика была не столь впечатляющая.

Анализ обновленного древа показывает также, что обратный ход полипептидной цепи в abCd-единицах в белках встречается значительно чаще, чем прямой. Из 405 abCd-единиц, найденных в 401 негомологичном белке, 92 abCd-единицы имеют прямой ход цепи и 313 – обратный (77%). Причина такого преимущества пока до конца не выяснена и требует дальнейшего изучения.

Как отмечено выше, одно из наиболее важных применений структурных деревьев – это их использование для структурной классификации белков. Все белки и домены, входящие в состав одного древа, могут быть отнесены к одному структурному классу или суперсемейству. Белки и домены, принадлежащие ветвям структурного древа, образуют подклассы. Как видно, такая классификация базируется только на сходстве пространственных структур и на общности смоделированных путей сворачивания. В ней не учитываются аминокислотные последовательности, а также информация о функциях и эволюционном родстве белков, что в той или иной степени учитывается в других известных классификациях [7–9].

Еще одна область применения структурных деревьев – это изучение с их помощью взаимосвязи между структурой и функцией белков. Имеется ряд ярких примеров, когда белки с одинаковой укладкой цепи имеют одинаковую функцию. Так многие ДНК-связывающие белки содержат одинаковую укладку в виде ТНТ-мотива и группируются в левой ветви структурного древа белков, содержащих α-α-уголки [1, 2], а в правой ветви этого древа находятся Ca^{2+}-связывающие белки, имеющие EF-пары спиралей. Белки, содержащие OB-фолд, имеют строгую тенденцию связывать олигонуклеотиды и олигосахариды [19] (см. также соответствующую ветвь древа белков, содержащих S-образные β-листы [4]). В центральной части древа (α + β)-белков, содержащих abCd-единицы, группируются РНК-связывающие белки (рис. 2, см. также [14, 18]). Однако анализ показывает, что гораздо чаще наблюдается "обратная" картина: белки, имеющие схожие или даже одинаковые укладки цепи, имеют разные функции (см., например, укладку, представленную на рис. 3), а белки, имеющие одинаковые функции, имеют совершенно разные укладки цепи и принадлежат разным структурным классам. По-видимому, это связано с тем, что функция белка определяется не только общей укладкой цепи (что, собственно, принимается во внимание при построении структурных деревьев), но и в значительной степени зависит от тонкой структуры белка и особенно от тонкой структуры активного центра. С другой стороны, анализ структурных деревьев и структурных мотивов позволяет сделать вывод о том, что в основе сходства белковых структур лежит не эволюционное родство белков и не общность функции, а общие физико-химические закономерности, которые “отбирают” наиболее выгодные укладки полипептидной цепи [1–5].

На основе обновленного структурного древа нами создана иерархически организованная база данных всех обнаруженных нами (α + β)-белков, содержащих abCd-единицы, которая доступна в Интернете (http://strees.protres.ru/). WEB-сайт включает в себя: страницы с иерархически организованной базой данных белков данной группы, структурное древо с нанесенными на него укладками, встречающимися в существующих белках, страницу с инструкцией для облегчения работы с базой данных, систему поиска нужного белка по введенному PDB-коду и PDB-файлы белков. Страницы логически связаны друг с другом с помощью контекстных переходов. Можно загрузить PDB-файлы всех содержащихся в базе данных белков (всего 2636 PDB-файлов для 926 белков и их мутантов) и посмотреть их с помощью любой программы молекулярной

графики. На WEB-сайте доступны также структурное древо для класса β-белков, содержащих abcd-единицы и соответствующая иерархически организованная база данных. В настоящее время ведется работа по созданию баз данных и построению обновленных структурных деревьев других структурных классов. Все структурные деревья и базы данных будут доступны в Интернете.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (07-04-00659).

СПИСОК ЛИТЕРАТУРЫ

1. Efimov A.V. 1996. A structural tree for α-helical proteins containing α-α-corners and its application to protein classification. *FEBS Lett*. **391**, 167–170.
2. Efimov A.V. 1997. Structural trees for protein superfamilies. *Proteins*. **28**, 241–260.
3. Efimov A.V. 1997. A structural tree for proteins containing 3β-corners. *FEBS Lett*. **407**, 37–41.
4. Efimov A.V. 1998. A structural tree for proteins containing S-like β-sheets. *FEBS Lett*. **437**, 246–250.
5. Ефимов А.В. 2008. Структурные деревья белков, содержащих φ-мотивы. *Биохимия*. **73**, 29–35.
6. Гордеев А.Б., Кондратова М.С., Ефимов А.В. 2008. Новое структурное древо β-белков, содержащих abcd-единицы. *Молекуляр. биология*. **42**, 323–326.
7. 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.
8. Holm L., Sander C. 1999. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res*. **27**, 244–247.
9. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M. 1997. CATH – a hierarchic classification of protein domain structures. *Structure*. **5**, 1093–1108.
10. Day R., Beck D.A.C., Armen R.C., Daggett V. 2003. A consensus view of fold space: Combining SCOP, CATH and the Dali Domain Dictionary. *Protein Sci*. **12**, 2150–2160.
11. Novotny M., Madsen D., Kleywegt G.J. 2004. Evaluation of protein fold comparison servers. *Proteins*. **54,** 260–270.
12. Tatusova T.A., Madden T.L. 1999. Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett*. **174**, 247–250.
13. Sayle R.A., Milner-White E.J. 1995. RASMOL – biomolecular graphics for all. *Trends Biochem. Sci*. **20**, 374–376.
14. Efimov A.V. 1995. Structural similarity between two-layer α/β and β-proteins. *J. Mol. Biol*. **245**, 402–415.
15. Лим В.И., Мазанов А.Л., Ефимов А.В. 1978. Стереохимическая теория пространственной структуры глобулярных белков. I. Высокоспиральные промежуточные структуры. *Молекуляр. биология*. **12**, 206–213.
16. Richardson J.S. 1977. β-Sheet topology and relatedness of proteins. *Nature*. **268**, 495–500.
17. Rao S.T., Rossman M.G. 1973. Comparison of super-secondary structures in proteins. *J. Mol. Biol*. **76**, 241–256.
18. Janin J. 1993. Shared structural motif in proteins. *Nature*. **365**, 21.
19. Murzin A.G. 1993. OB (oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J*. **12**, 861–867.