*БИОХИМИЯ, 2015, том 80, вып. 11, с. 1683 – 1687*

*УДК 577.29*

# StructAlign – ПРОГРАММА ДЛЯ ВЫРАВНИВАНИЯ СТРУКТУР ДНК-БЕЛКОВЫХ КОМПЛЕКСОВ*

**© 2015 Я.В. Попов[1], А.А. Галицына[1], А.В. Алексеевский[1,2,3], А.С. Карягина[2,4,5], С.А. Спирин[1,2,3]****

[1] *Московский государственный университет им. М.В. Ломоносова, факультет биоинженерии и биоинформатики, 119991 Москва; факс: +7(495)939-4195*

[2] *НИИ физико-химической биологии им. А.Н. Белозерского, Московский государственный университет им. М.В. Ломоносова, 119991 Москва; факс: +7(495)939-3181, электронная почта: sas@belozersky.msu.ru*

[3] *Институт системных исследований РАН, 117218 Москва, Нахимовский пр., 36, корп. 1; факс: +7(495)719-7681*

[4] *Центр эпидемиологии и микробиологии им. Н.Ф. Гамалеи Минздрава России, 123098 Москва, ул. Гамалеи, 18; факс: +7(495)193-6183*

[5] *Институт сельскохозяйственной биотехнологии, 127550 Москва, ул. Тимирязевская, 42*

Поступила в редакцию 03.07.15

Сравнительный анализ структур комплексов гомологичных белков с ДНК важен для анализа узнавания ДНК белками. Необходимый этап сравнительного анализа – выравнивание, т.е. установление соответствия аминокислотных остатков и нуклеотидов одного комплекса аминокислотным остаткам и нуклеотидам другого. В настоящий момент доступных программ для выравнивания структур ДНК-белковых комплексов не существует. Мы представляем программу StructAlign, которая призвана заполнить этот пробел. В качестве входных данных программа принимает две структуры комплексов двойной спирали ДНК с белками и выдает выравнивание цепей ДНК, отвечающее наилучшему пространственному совмещению цепей белка.

**КЛЮЧЕВЫЕ СЛОВА:** структурная биоинформатика, ДНК-белковые комплексы, выравнивание, веб-интерфейс.

Пространственные структуры гомологичных белков, как правило, весьма сходны между собой. Сравнительный анализ сходных структур часто помогает выявлять особенности этих структур, важные для функции соответствующих белков. ДНК-связывающие белки образуют семейства, в пределах которых консервативна не только укладка белковой молекулы, но и ее расположение на двойной спирали ДНК. В этом случае имеет смысл проводить сравнительный анализ структур ДНК-белковых комплексов.

Необходимый этап сравнительного анализа структур — их выравнивание. Под выравниванием мы понимаем задание соответствия остатков одной структуры остаткам другой. Выравнивание структур отдельных цепей белка выполняется многими программами (см., например, [1–4]). К сожалению, не существует общедоступных программ, которые выполняют выравнивания макромолекулярных, в т.ч. ДНК-белковых комплексов. Разработанная нами программа с рабочим названием StructAlign призвана заполнить этот пробел. Эта программа в качестве входных данных принимает две структуры комплексов белка с двойной спиралью ДНК и выдает выравнивание нуклеотидных последовательностей, отвечающее лучшему совмещению комплексов, а также численный показатель качества выравнивания. В алгоритме программы используется возможность наложения друг на друга любых двух двойных спиралей ДНК в соответствии с любым выравниванием нуклеотидов, не содержащим вставок и делеций. Тем самым задача выравнивания ДНК-белковых комплексов состоит в выявлении лучшего «сдвига» одной спирали относительно другой; о том же, насколько один сдвиг лучше другого,

* Первоначально английский вариант рукописи был опубликован на сайте «Biochemistry» (Moscow), Papers in Press, BM15-205, 20.09.2015.

** Адресат для корреспонденции.

можно судить по качеству совмещения цепей белка, получающегося при том или ином совмещении ДНК.

## МЕТОДЫ ИССЛЕДОВАНИЯ

Входные данные программы StructAlign состоят из двух структур комплексов белка с двойной спиралью ДНК в формате PDB (см. http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html). Работа программы состоит из следующих этапов:

1) к каждому нуклеотиду обеих структур, подаваемых на вход программе, привязывается своя система координат (алгоритм см. ниже);

2) для каждой пары нуклеотидов, по нуклеотиду из каждой структуры, вычисляется мера *S* сходства расположения белка относительно данных нуклеотидов (алгоритм см. ниже);

3) в каждой из структур находится по отрезку двойной спирали ДНК одинаковой длины, для которых сумма мер *S* по соответствующим нуклеотидам отрезков максимальна среди всех возможных выборов таких пар отрезков; эта пара отрезков задает выравнивание нуклеотидов ДНК двух структур;

4) формируются выходные файлы: входные структуры, совмещенные по паре нуклеотидов с наибольшим значением *S*, в формате PDB, а также выравнивание цепей ДНК в текстовом виде.

**Алгоритм построения системы координат по структуре нуклеотида.** За начало координат принимается центр атома фосфора. За направление оси X берется направление от центра атома фосфора к середине отрезка, соединяющего центры двух атомов кислорода (OP1 и OP2 в обозначениях банка PDB), ковалентно связанных с атомом фосфора и не связанных с остатком рибозы. За направление оси Y берется направление, перпендикулярное X и параллельное плоскости, проходящей через центры атомов P и C1' и параллельной X; из двух таких направлений берется то, которое образует меньший угол с направлением от P к C1'. За направление Z берется направление, перпендикулярное X и Y и такое, что тройка XYZ образует правоориентированную систему (т.е. направление вращения от X к Y – по часовой стрелке, если смотреть в направлении Z).

**Алгоритм вычисления меры сходства *S*.** Пусть имеются два нуклеотида, каждый из своей структуры. Построим по каждому из них систему координат и совместим координатные пространства двух структур по этим системам координат. Рассмотрим для каждого Cα-атома белка из первой структуры ближайший (после совмещения) Cα-атом второй структуры. Точно так же рассмотрим для каждого Cα-атома второй структуры ближайший Cα-атом первой структуры. Если два Cα-атома, один (*a*) из первой структуры, другой (*b*) из второй, таковы, что *a* – самый близкий к *b* из всех Cα-атомов первой структуры, а *b* – самый близкий к *a* из всех Cα-атомов второй структуры, и расстояние $d(a,b)$ между ними менее 4,5 Å, то будем считать эти атомы взаимно ближайшими (относительно данной пары нуклеотидов). Константа 4,5 подобрана как результат тестирования различных значений. Мера сходства пары нуклеотидов получается из суммы Σ значений вида 4,5 Å – $d(a,b)$ по всем парам взаимно ближайших Cα-атомов *a*, *b*. Сумма Σ тем больше, чем более сходно расположение белка относительно двух данных нуклеотидов в их структурах. Значение меры *S* получается из указанной суммы вычитанием константы, равной 31 Å. Последняя константа подобрана по результатам тестирования на нескольких семействах ДНК-белковых комплексов. В нескольких десятках пар родственных комплексов были визуально определены нуклеотиды, одинаково расположенные относительно белка. Для пар соответствующих (т.е. одинаково расположенных относительно белка) и несоответствующих нуклеотидов были найдены взаимно ближайшие Cα-атомы белка, после чего построены два распределения сумм Σ: для соответствующих и несоответствующих пар нуклеотидов. Эти распределения показаны на рис. 1. Константа 31 Å выбрана из тех соображений, чтобы значение $S = \Sigma - 31$ Å для пар несоответствующих друг другу нуклеотидов было, как правило, отрицательным, а для пар соответствующих нуклеотидов – как правило, положительным.

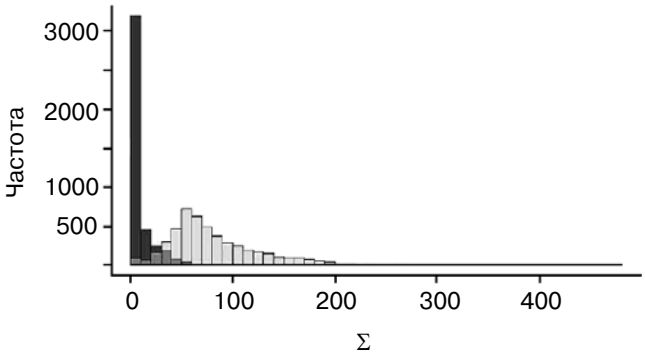


**Рис. 1.** Распределение величины Σ по парам соответствующих (светлые столбики) и несоответствующих (темные столбики) нуклеотидов

Программа StructAlign написана на языке программирования C, веб-интерфейс к программе – на языке Python с использованием CGI.

## РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

**Веб-интерфейс к программе StructAlign.** Веб-интерфейс доступен по адресу http://mouse.genebee.msu.ru/tools/StructAlign.html. Он предлагает пользователю ввести два идентификатора банка PDB и, опционально, идентификаторы цепей белка из соответствующих записей PDB. Если идентификатор цепи не задан, берется первая по расположению в документе цепь белка. Программа выравнивает две структуры, каждая из которых состоит из одной (заданной пользователем) цепи белка и всех имеющихся в данных записях PDB цепей ДНК.

Выходные данные программы представляют собой:

1) вес выравнивания (число, характеризующее качество совмещения структур);

2) файл в формате PDB, содержащий совмещенные структуры;

3) выравнивание последовательностей ДНК в виде преформатированного текста;

4) интерактивное изображение совмещенных структур с использованием Jmol.

**Пример 1. Использование программы StructAlign для сравнения структур комплексов фаговых репрессоров с ДНК.** Были взяты 13 структур комплексов ДНК с белками, отнесенными в базе SCOP [5] к семейству «Phage Repressors», а именно 1LMB (цепи белка 3 и 4, C1-репрессор фага лямбда), 1LLI (цепи A и B, мутантный C1-репрессор фага лямбда), 1RIO (цепи A и B, C1-репрессор фага лямбда), 1PER (цепи L и R, C1-репрессор фага 434), 1RPE (цепи L и R, C1-репрессор фага 434), 3CRO (цепи L и R, белок Cro фага 434), 6CRO (цепь A, белок Cro фага лямбда). Все пары структур были проанализированы программой StructAlign. Выяснилось, что программа хорошо выравнивает между собой структуры 1PER, 1RPE, 3CRO (независимо от цепи белка). Также хорошо выравниваются между собой структуры 1LMB, 1LLI, 1RIO. Веса выравниваний внутри этих групп колеблются между 1300 и 5500 Å, визуально наложение структур очень хорошее (рис. 2, *а*). Между собой структуры из этих двух групп выравниваются хуже, вес выравнивания колеблется между 400 и 1000 Å, наложение хуже (рис. 2, *б*). Совсем плохо с остальными структурами выравнивается структура 6CRO: вес между 110 и 260 Å, при наложении по лучшей паре нуклеотидов совмещаются, по сути, лишь узнающие спирали белка (рис. 2, *в*). Последнее неудивительно, учитывая, что белок Cro из фага лямбда имеет укладку, включающую β-шпильку, в то время как остальные рассмотренные белки состоят из четырех α-спиралей. В данном случае включение этих белков в одно семейство базы SCOP не соответствует малому сходству их третичных структур.

**Пример 2. Использование программы StructAlign для сравнения структур комплексов белков, включающих мотив Zn2/Cys2 и «лейциновые молнии», с ДНК.** Белки из пекарских дрожжей (*Saccharomyces cerevisiae*) GAL4 (PDB-код 3COQ) и HAP1 (PDB-код 1HWT) имеют в своей структуре ДНК-узнающий мотив Zn2/Cys2 и димеризационный мотив «лейциновая молния». Каждый из этих белков связывается с ДНК как гомодимер, при этом в случае GAL4 это симметричный димер, а два

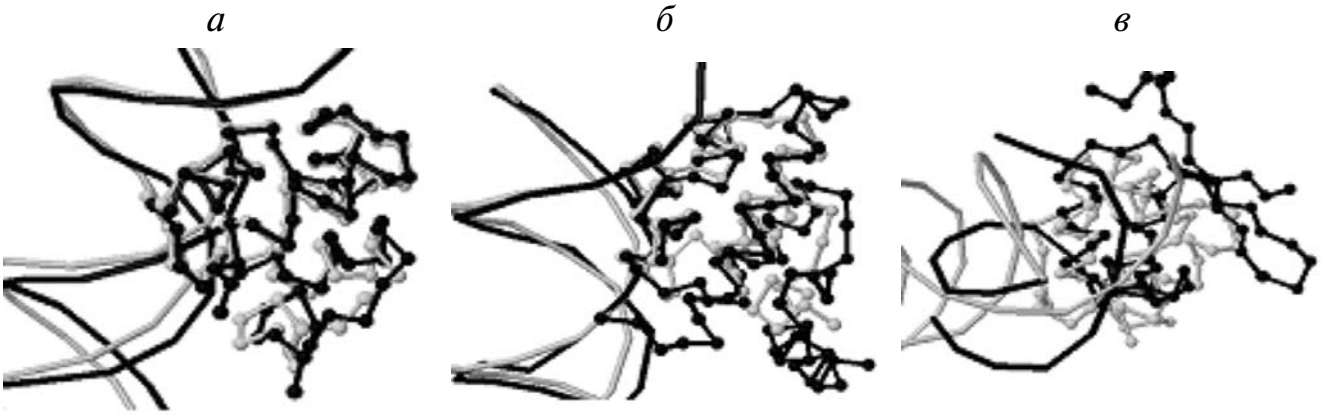


**Рис. 2.** Результаты работы программы на парах структур фаговых репрессоров с ДНК. *а* – 1PER, цепь белка L (серый) и 3CRO, цепь белка L (черный), вес выравнивания 1149 Å; *б* – 1PER, цепь белка L (серый) и 1LMB, цепь белка 3 (черный), вес выравнивания 855 Å; *в* – 1PER, цепь белка L (серый) и 6CRO, цепь белка A (черный), вес выравнивания 116 Å

мономера HAP1 образуют несимметричный димер (рис. 3) [6]. В структуре каждого из мономеров имеется длинная α-спираль (вместе с такой же спиралью другого мономера она образует «лейциновую молнию»). Попарное выравнивание структур комплексов каждого из мономеров

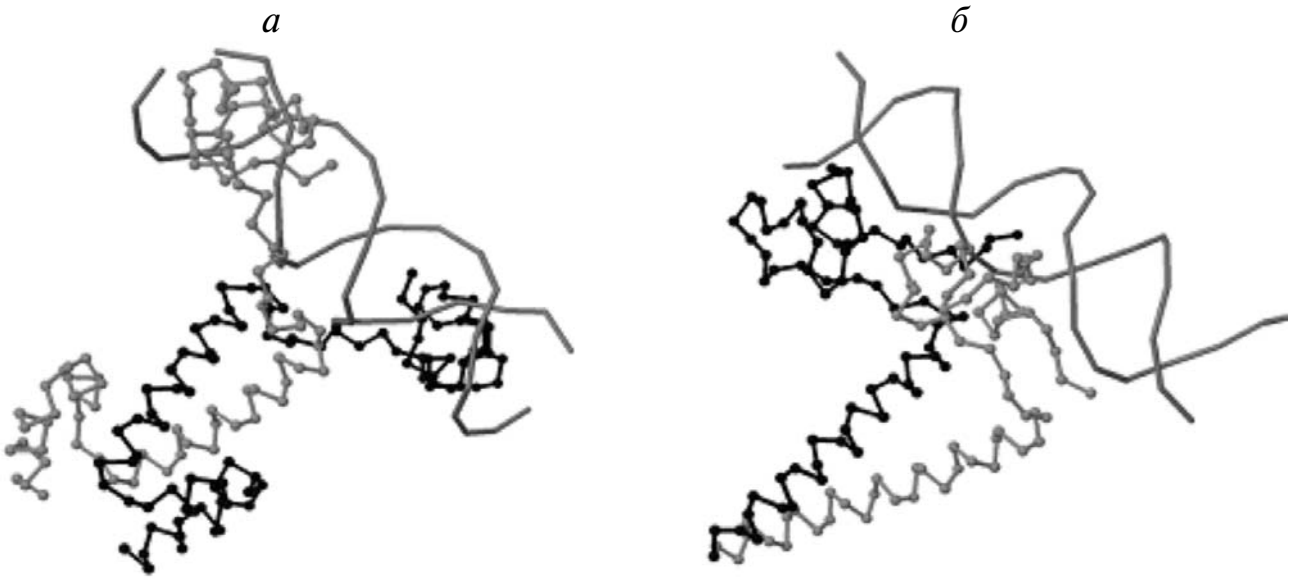


**Рис. 3.** Структура комплексов гомодимеров белков, содержащих ДНК-узнающий мотив Zn2/Cys2 и димеризационный мотив «лейциновая молния», с ДНК. *а* – Симметричный димер белка дрожжей GAL4 (PDB 3COQ), *б* – несимметричный димер белка дрожжей HAP1 (PDB 1HWT). ДНК и одна из цепей каждого белка (цепь A для 3COQ, цепь C для 1HWT) показаны серым, вторая цепь белка (цепь B для 3COQ, цепь D для 1HWT) – черным

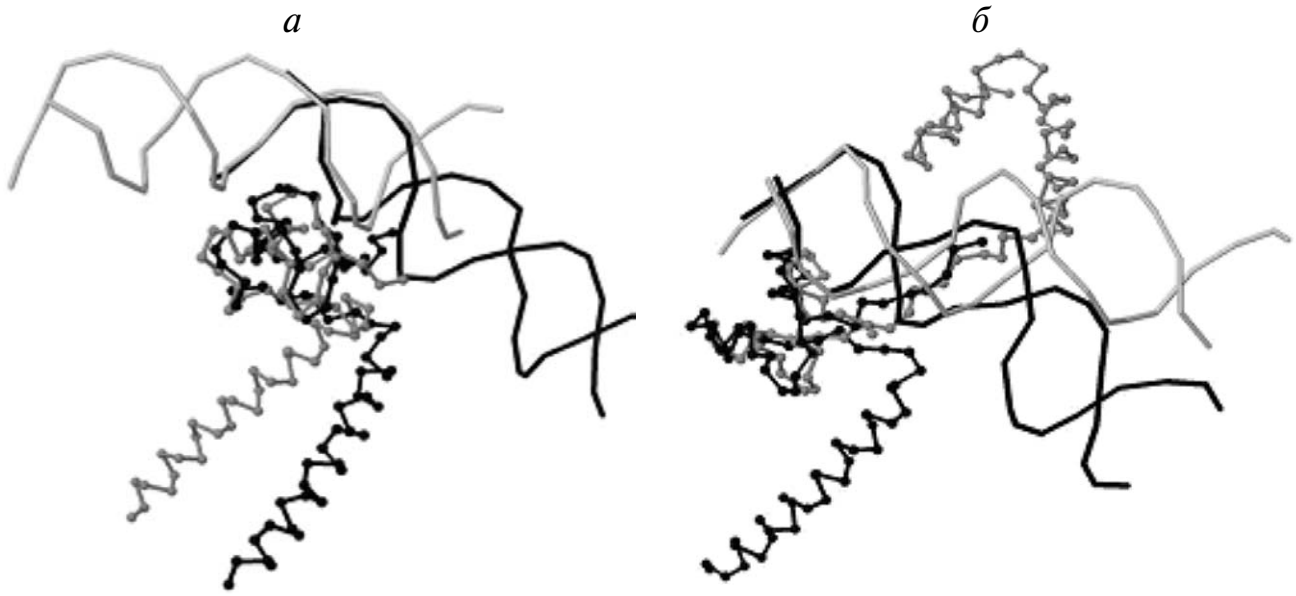


**Рис. 4.** *а* – Результат работы программы на двух мономерах белка HAP1 из структуры 1HWT. Цепь C и одна из копий цепей A и B ДНК изображены серым, цепь D и другая копия ДНК – черным. При совмещении нуклеотидов ДНК, наиболее сходно расположенных относительно цепей белка, ДНК-узнающие домены Zn2/Cys2 совместились, но α-спирали мотива «лейциновая молния» разошлись. Вес выравнивания 246 Å; *б* – результат работы программы на структурах комплексов ДНК с белками GAL4, PDB 3COQ, цепь белка A (серый) и HAP1, PDB 1HWT, цепь белка D (черный). При выравнивании структур по ДНК домены Zn2/Cys2 совместились, α-спирали мотива «лейциновая молния» разошлись. Вес выравнивания 250 Å

обоих белков в отдельности с ДНК позволяет получить хорошее наложение ДНК-узнающих мотивов Zn2/Cys2 (рис. 4). При этом вес выравнивания двух комплексов, включающих разные мономеры GAL4, равен 2789 Å за счет того, что совмещаются оба мотива: Zn2/Cys2 и α-спираль лейциновой молнии. В то же время вес выравнивания двух комплексов, включающих разные мономеры HAP1, равен всего 246 Å, поскольку при выравнивании структур по ДНК совмещаются только мотивы Zn2/Cys2 и часть линкера между мотивами, а α-спирали лейциновой молнии далеко расходятся (рис. 4, *а*). Примерно такой же вес, ~250 Å, имеют выравнивания комплексов, включающих любой из мономеров GAL4, с комплексом, содержащим один из мономеров HAP1 (цепь D согласно обозначениям PDB-записи 1HWT, рис. 4, *б*); выравнивание же с комплексом, включающим другой мономер (цепь C), имеет еще меньший вес – 170 Å.

**Возможные применения программы.** Программа может облегчить сравнительное изучение ДНК-белковых комплексов в нескольких аспектах. Во-первых, она автоматически определяет соответствующие друг другу нуклеотиды ДНК в двух сходных комплексах. Нахождение таких нуклеотидов «вручную» (т.е. с помощью визуализаторов структур) достаточно трудоемко. Во-вторых, она позволяет визуализировать «ДНК-ориентированное» совмещение комплексов, что выявляет сходно расположенные относительно ДНК участки белковой молекулы. В-третьих, определенную ценность может иметь выдаваемый вес выравнивания.

Работа выполнена при финансовой поддержке РНФ (грант 14-50-00029, разработка алгоритмов, тестирование программ) и РФФИ (грант 13-07-00969, программирование).

## СПИСОК ЛИТЕРАТУРЫ

1. Taylor, W.R., and Orengo, C.A. (1989) Protein structure alignment, *J. Mol. Biol.*, **208**, 1–22.
2. Holm, L., and Sander, C. (1993) Protein structure comparison by alignment of distant matrices, *J. Mol. Biol.*, **233**, 123–138.
3. Shindiyalov, I.N., and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng.*, **9**, 739–747.
4. Krissinel, E., and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr.*, **60**, 2256–2268.
5. 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.
6. King, D.A., Zhang, L., Guarente, L., and Marmorstein, R. (1999) Structure of a HAP1-DNA complex reveals dramatically asymmetric DNA binding by a homodimeric protein, *Nat. Struct. Biol.*, **6**, 64–71.

# StructAlign, A PROGRAM FOR ALIGNMENT OF STRUCTURES OF DNA–PROTEIN COMPLEXES

**Ya. V. Popov[1], A. A. Galitsyna[1], A. V. Alexeevski[1,2,3], A. S. Karyagina[2,4,5], S. A. Spirin[1,2,3]***

[1] *M. V. Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow 119991, Russia; fax: +7(495)939-4195*

[2] *A. N. Belozersky Institute of Physico-Chemical Biology, M. V. Lomonosov Moscow State University, Moscow 119991, Russia; fax: +7(495)939-3181, E-mail: sas@belozersky.msu.ru*

[3] *Institute of System Studies, Russian Academy of Sciences, Nakhimovski prosp. 36-1, Moscow 117218, Russia*

[4] *N. F. Gamaleya Center of Epidemiology and Microbiology, ul. Gamalei 18, Moscow 123098, Russia*

[5] *Institute of Agricultural Biotechnology, ul. Timiryazevskaya 42, Moscow 127550, Russia*

Received July 3, 2015

Comparative analysis of structures of complexes of homologous proteins with DNA is important in the analysis of DNA–protein recognition. Alignment is a necessary stage of the analysis. Alignment is matching of amino acid residues and nucleotides of one complex to residues and nucleotides of the other. Currently, there are no available programs for aligning structures of DNA–protein complexes. We present the program StructAlign, which should fill this gap. The program inputs a pair of complexes of DNA double helix with proteins and outputs an alignment of DNA chains corresponding to the best spatial fit of protein chains.

*Key words:* structural bioinformatics, DNA–protein complexes, alignment, web interface