

Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский государственный университет имени М.В.Ломоносова»

На правах рукописи

Занегина Ольга Николаевна

**Сравнительная характеристика структур ДНК-
белковых комплексов**

03.01.09 – Математическая биология, биоинформатика

Диссертация на соискание учёной степени кандидата биологических наук

Научный руководитель
кандидат физико-математических наук
Сергей Александрович Спирин

Москва – 2015

Содержание

Введение	5
Актуальность работы	5
Цели и задачи исследования	6
Научная новизна и практическая значимость работы	7
Основные результаты и положения, выносимые на защиту	8
Публикации. Степень достоверности и апробация результатов.....	9
Личный вклад автора.....	9
Структура и объём работы	9
Глава 1. Обзор литературы.....	10
1.1 Физические основы взаимодействия белков с нуклеиновыми кислотами	10
1.1.1 Водородные связи.....	10
1.1.1.1 Физические основы.....	10
1.1.1.2 Водородные связи в ДНК-белковых комплексах	10
1.1.1.3 Роль воды в ДНК-белковых взаимодействиях	12
1.1.2 Гидрофобные взаимодействия.....	15
1.1.2.1 Физические основы	15
1.1.2.2 Гидрофобные взаимодействия в ДНК-белковых комплексах.....	15
1.2 Алгоритмы выделения водородных связей и гидрофобных взаимодействий	17
1.2.1 Определение водородных связей.....	17
1.2.2 Определение гидрофобных взаимодействий.....	18
1.3 Механизмы ДНК-белкового узнавания	21
1.3.1 Поиск универсального кода и закономерностей ДНК-белкового узнавания	21
1.3.2 Специфичность узнавания ДНК белком	23
1.3.2.1 Специфичность ДНК-белкового узнавания с точки зрения последовательностей	23
1.3.2.2 Специфичность ДНК-белкового узнавания с точки зрения структур	24
1.3.2.3 Специфичность ДНК-белкового узнавания с точки зрения физико-химических параметров взаимодействия	28
1.4 Классификации ДНК-белковых комплексов	30
1.5 Базы данных макромолекул и их комплексов	33
1.5.1 Базы ДНК-белковых взаимодействий.....	34
Глава 2. Классификация ДНК-белковых взаимодействий	38
2.1 Материалы и методы.....	38

2.1.1	Основные понятия	38
2.1.2	Материалы	43
2.1.3	Методы	45
2.1.3.1	Совмещение структур комплексов и выравнивание	45
2.1.3.2	Процедура определения способа взаимодействия белкового домена с ДНК	46
2.1.3.3	Построение классификации семейств белковых доменов SCOP по классам взаимодействия с ДНК	48
2.2	Результаты	49
2.2.1	Способ взаимодействия для каждой структуры ДНК-белковых комплексов	49
2.2.2	Закономерности взаимодействия, наблюдаемые в разных структурах одного белкового домена	53
2.2.3	Закономерности строения белковых доменов, взаимодействующих с ДНК одним и тем же способом	54
2.2.4	Вклад гидрофобных кластеров в ДНК-белковое узнавание	56
2.2.5	Закономерности ДНК-белкового узнавания внутри семейства SCOP	59
2.2.5.1	Вариации контактов внутри семейства	60
2.2.5.2	Вариации контактов внутри семейства, обусловленные подвижностью боковых цепей белка	61
2.2.5.3	Вариации контактов внутри семейства, обусловленные различиями во вторичной структуре белка	62
2.2.5.4	Вариации контактов внутри семейства, обусловленные аминокислотной заменой в месте контакта	63
2.2.5.5	Вариации контактов внутри семейства, обусловленные разницей во взаимном расположении белка относительно ДНК	64
2.2.6	Классификация семейств ДНК-контактирующих доменов SCOP	66
2.2.6.1	Семейства, для которых появление новых структур может повлиять на класс взаимодействия	69
2.2.6.2	Семейства без общего класса взаимодействия	70
2.2.6.3	Случай разделения семейства на два подсемейства	72
2.2.6.4	Закономерности строения семейств белковых доменов, относящихся к одному и тому же классу взаимодействия	73
Глава 3. Поиск консервативных молекул воды в белковых и ДНК-белковых комплексах		77
3.1	Материалы и методы	78
3.1.1	Программа wLake	78
3.1.1.1	Определения	78

3.1.1.2	Алгоритм	78
3.2	Результаты: применение wLake для анализа макромолекулярных комплексов	80
3.2.1	Поиск консервативных молекул воды в структурах транскетолаз	80
3.2.2	Консервативные молекулы воды в семействах SCOP ДНК-связывающих доменов.....	83
Глава 4.	База данных NPIDB.	86
4.1	Расширение функционала базы данных NPIDB.....	93
4.1.1	Подробное описание семейств ДНК-связывающих доменов SCOP	93
4.1.2	Интеграция классификации ДНК-белковых взаимодействий в базу данных NPIDB	95
Глава 5.	Поиск консервативных особенностей в семействах гомологичных белков....	99
5.1	ТАТА-бокс связывающий домен и белок.....	99
5.1.1	Структуры ТАТА-бокса связывающего домена.....	101
5.1.2	Консервативные особенности комплексов ТРВ с ДНК.....	102
5.1.3	Анализ прямых водородных связей для ТРВ и входящих в его состав доменов.....	103
5.1.4	Гидрофобные контакты для ТРВ и входящих в его состав доменов	105
5.1.5	Симметрия второго порядка в β -листе ТРВ	107
5.1.6	Опосредованные водой водородные связи для ТАТА-бокса связывающего белка и входящих в его состав доменов	108
5.1.7	Консервативность контактов со стороны ДНК	111
5.1.8	Биологическая роль консервативных контактов ТРВ с ДНК.....	112
5.2	Семейство LAGLIDADG homing эндонуклеаз	113
5.2.1	Структуры LAGLIDADG_1 хоминг эндонуклеаз	115
5.2.2	Консервативные контакты LAGLIDADG_1 хоминг эндонуклеаз с ДНК	116
Заключение.....		121
Список публикаций		123
Список литературы		125
Приложение.....		134

Введение

Актуальность работы

Развитие методов рентгеноструктурного анализа, ядерного магнитного резонанса и электронной микроскопии привело к экспоненциальному росту количества пространственных структур макромолекул. За последние 25 лет было расшифровано более 3000 структур комплексов белков с ДНК, и ежегодно их количество увеличивается на ~10 %. Это позволяет изучать не только особенности ДНК-белкового узнавания в конкретных структурах, но и искать закономерности такого рода взаимодействий.

Понимание механизмов узнавания белком нуклеиновой кислоты может помочь в предсказании специфичности ДНК-белкового взаимодействия, а также в направленном мутагенезе, особенно в тех случаях, когда получение пространственной структуры ДНК-белкового комплекса затруднительно. Кроме того, ДНК-белковые взаимодействия могут быть рассмотрены с эволюционной точки зрения. Например, известно [1], что взаимодействующие с ДНК остатки белка являются более консервативными.

К сожалению, к настоящему времени не существует общепринятых подходов к предсказанию НК-белковых взаимодействий. В большой степени это связано с тем, что на узнавание белком нуклеиновой кислоты влияет множество факторов, и единого кода соответствия белку нуклеиновой кислоты не существует не только для пространственно разнообразных структур РНК, но и для двухцепочечных ДНК.

Первым шагом к поиску закономерностей ДНК-белкового узнавания является анализ и систематизация взаимодействий, наблюдаемых в пространственных структурах комплексов белков с нуклеиновыми кислотами. Так, для ДНК был разработан ряд классификаций ДНК-белковых взаимодействий (см. обзор литерату-

ры), рассматривающих все доступные на момент классификации структуры. В большинстве случаев эти классификации были актуальны на момент публикации и в дальнейшем не обновлялись новыми структурами. Поэтому представляется целесообразным создание классификации, открытой к уточнению, дополнению и расширению при появлении новых пространственных структур.

Следующим шагом в изучении ДНК-белковых взаимодействий является создание специализированных баз данных, предоставляющих информацию не только о последовательностях и структурах ДНК-белковых комплексов, но и о характеристиках ДНК-белковых взаимодействий, а также обладающих инструментами для анализа этих взаимодействий. Аналогично классификациям, часть существующих баз данных ДНК-белковых взаимодействий устарело и не обновляется автоматически, а также не содержит инструментов анализа ДНК-белковых комплексов.

Таким образом, создание расширяемой и дополняемой классификации ДНК-белковых взаимодействий и её интеграция в качестве инструмента анализа в автоматически обновляемую базу данных ДНК-белковых комплексов представляется эффективным способом изучения закономерностей ДНК-белкового узнавания.

Цели и задачи исследования

Целью исследования было выяснение закономерностей ДНК-белковых взаимодействий на основе анализа доступных пространственных структур комплексов белков с ДНК. Для достижения этой цели в работе решались пять задач:

1. Разработка новой дополняемой классификации структур комплексов ДНК-связывающих белковых доменов с ДНК.
2. Разработка дополняемой классификации семейств гомологичных ДНК-узнающих доменов, основанной на консервативных взаимодействующих элементах структур домена и ДНК.

3. Интеграция разработанных классификаций в базу данных НК-белковых взаимодействий NPIDB [2].
4. Разработка подхода к описанию консервативно расположенных молекул воды в структурах гомологичных макромолекул, прежде всего в структурах ДНК-белковых комплексов, включающих гомологичные белки.
5. Сравнительный анализ комплексов ДНК с белками нескольких семейств с целью описания консервативных особенностей ДНК-белкового взаимодействия.

Научная новизна и практическая значимость работы

1. Предложенная в данной работе классификация структур ДНК-белковых комплексов имеет ряд преимуществ перед ранее предложенными. Возможность дополнения позволяет классификации оставаться актуальной при появлении новых структур. Интеграция классификации в базу данных НК-белковых взаимодействий облегчает использование этой классификации для анализа ДНК-белковых структур.

2. Предложена первая классификация семейств ДНК-узнающих белковых доменов. Такая классификация позволяет оценивать потенциальные контакты с ДНК тех белков, для которых пока не доступны пространственные структуры в комплексе с ДНК, но которые содержат домены, родственные тем, что рассмотрены в классификации.

3. Выделение консервативных контактов в ДНК-белковом комплексе позволяет выявить наиболее функционально важные контакты и вариации узнавания ДНК внутри одного семейства структурных доменов. Данные о консервативных контактах даже для единичного ДНК-белкового комплекса могут помочь в планировании экспериментов, например, по направленному мутагенезу. В настоящей работе предложен новый подход к описанию консервативности такой важной

составляющей ДНК-белкового интерфейса, как связи, опосредованные молекулами воды. Кроме того, впервые проведён по-дробный анализ всех консервативных элементов ДНК-белкового взаимодействия для семейств TATA-box узнающих белков и белков семейства LAGLIDADG_1.

Основные результаты и положения, выносимые на защиту

1. Разработана классификация структур комплексов белковых доменов с ДНК. Процедура классификации применена к анализу 1942 структур, относящихся к 314 ДНК-контактирующих белковых доменов. Выделено 97 способов взаимодействия структур белковых доменов ДНК. Для семейств структурных белковых доменов, представленных тремя и более различными белковыми доменами, определен один из 17 классов ДНК–белкового взаимодействия. При этом класс определяет особенности взаимодействия, характерные для семейства в целом.

2. Описано распределение способов взаимодействия по доступным структурам комплексов, типичные вариации способов взаимодействия между различными структурами комплексов с ДНК одного белкового домена и между комплексами различных доменов одного семейства, а также распределение классов взаимодействия по достаточно представленным (три и более различных домена) семействам ДНК-связывающих белковых доменов. Кроме того, оценен вклад прямых и опосредованных молекулами воды водородных связей, а также гидрофобных кластеров в формирование ДНК-белковых контактов.

3. Функционал базы данных НК-белковых взаимодействий NPIDB был дополнен подробным описанием семейств ДНК-связывающих доменов SCOP, включающим описание консервативных молекул воды, а также классификацией ДНК-белковых взаимодействий.

4. Основываясь на разработанной классификации и используя функционал базы данных NPIDB, был разработан и применен подход к структурному анализу

консервативных особенностей гомологичных белков. В частности, для семейства транскетолаз была рассмотрена роль консервативных молекул воды. Для семейств TATA-бох связывающих белков и хоминг-эндонуклеаз семейства LAGLIDADG_1 были найдены и проанализированы консервативные контакты с ДНК.

Публикации. Степень достоверности и апробация результатов

По материалам диссертации было опубликовано 10 печатных работах, из них 4 статьи в рецензируемых изданиях, 6 статей в сборниках трудов конференций. Результаты были представлены на научных конференциях Ломоносов'06, BGRS'06, MCCMB'07, BGRS'08, BGRS'14, ECCB'14. Значительная часть результатов доступна на сайте NPIDB (<http://npidb.belozersky.msu.ru/>). Список публикаций приведен в конце работы.

Личный вклад автора.

Постановка изложенных в диссертации задач была сделана научным руководителем соискателя к.ф.-м.н. С.А.Спириным. Изложенные в диссертации результаты получены лично автором. В совместных публикациях {II, III, IV, IX, X} диссертантом выполнена работа по анализу ДНК-белковых взаимодействий. В публикациях {I, V, VI, VII, VIII} — работа по анализу консервативных молекул воды.

Структура и объём работы

Работа состоит из введения, пяти глав, заключения, списка публикаций автора, списка литературы и приложения. Список литературы содержит 137 наименований. Объём работы составляет 138 страниц, включая 6 таблиц и 50 рисунков.

Глава 1. Обзор литературы

1.1 Физические основы взаимодействия белков с нуклеиновыми кислотами

1.1.1 Водородные связи

1.1.1.1 Физические основы

Водородные связи представляют собой донорно-акцепторные взаимодействия атома водорода из группы Н–А (А = О, N, Cl, F, S) с электроотрицательными атомами О, F, N, S, входящими в состав одной молекулы (внутримолекулярные связи) или разных молекул (межмолекулярные связи). Отдельно выделяют водородные связи, опосредованные молекулами воды. Энергия водородной связи принимает значения в интервале 4–29 КДж/моль, что ставит её между слабыми вандер-ваальсовыми взаимодействиями и типичными ковалентными связями.

В молекулах белка и нуклеиновых кислот водородные связи играют важную роль в формировании пространственной структуры. Например, амидные группы полипептидной цепи образуют сеть водородных связей, формируя α -спирали и β -листы. Гидроксильные группы боковых цепей участвуют в образовании водородных связей между собой, а также с ионизованной карбоксильной группой и пептидной связью. Полный список возможных водородных связей между остатками белка приведен в работе Торнтон [3]. Структура нуклеиновых кислот также обусловлена сетью водородных связей между азотистыми основаниями полинуклеотидных цепей.

1.1.1.2 Водородные связи в ДНК-белковых комплексах

Роль водородных связей в формировании ДНК-белковых комплексов была изучена в ряде работ. В работах Диксит и Джаярам [4] было показано, что водо-

родные связи хотя и не являются основной движущей силой формирования конкретного ДНК-белкового комплекса, но служат для стабилизации и оптимизации расстояний и углов в нем.

Ключевую роль во взаимодействии с ДНК вносят боковые цепи белка, а не полипептидный остов. Обладая достаточной подвижностью, они обеспечивают не только химическое, но и пространственное узнавание ДНК [5]. При этом способность образовывать водородные связи с нуклеотидами зависит от природы бокового радикала аминокислоты. Наиболее часто водородные связи образуют полярные и заряженные аминокислоты: аргинин, лизин, серин, треонин, аспарагин и глутамин [6], [7]. В случае, если один и тот же аминокислотный остаток образует несколько, например две, водородные связи с одним гетероциклическим основанием ДНК, такое взаимодействие называется бидентантным, а если с несколькими основаниями — комплексным [7].

Взаимодействие белка с сахаро-фосфатным остовом не является специфическим. Тем не менее, оно вносит значительный вклад в стабилизацию ДНК-белкового комплекса [7]. Наиболее значимыми для узнавания являются контакты аминокислотных остатков с гетероциклическими основаниями ДНК [6]. Водородные связи боковых цепей белка с нуклеотидами обладают значительной, хотя и не однозначной специфичностью. Например, там же [6], [7] было показано, что аргинин и лизин предпочитают образовывать водородные связи с гуанином, а треонин и серин в основном контактируют с сахаро-фосфатным остовом.

Несмотря на склонность аминокислот предпочитать те или иные нуклеотиды, однозначного соответствия аминокислотных остатков нуклеотидам выявить не удалось. Более того авторы [6] подчеркивают, что пространственная доступность атомов гетероциклических оснований не менее важна для образования водородных связей, чем природа ее потенциальных участников.

1.1.1.3 Роль воды в ДНК-белковых взаимодействиях

Молекулы воды вносят большой вклад в стабилизацию структур белка [8], [9] и ДНК [10]. Они являются движущей силой при сворачивании белка и формировании гидрофобного ядра [11], [12], а также опосредуют контакты с лигандами, обеспечивая специфичность связывания (как в случае с триптофановым репрессором [13], [14]), или же наоборот, позволяя связываться большому диапазону лигандов за счет заполнения полостей белка молекулами воды (как в главном комплексе гистосовместимости [15]).

В ряде работ была показана важность консервативных молекул воды для функционирования таких белков как цитохром с [16], тимидилатсинтетаза [17], белок, связывающий жирные кислоты (FABP) [18]. Также была отмечена роль консервативных молекул воды в образовании стабильных димеров аланинарацемазы (AlaR) [19], а также в узнавании и связывании ДНК гомеобелками [20], триптофановым репрессором [14] и доменами семейства ETS [21].

При образовании ДНК-белкового комплекса большая часть гидратной оболочки удаляется из зоны контакта [22]. В некоторых случаях на ДНК-белковом интерфейсе остаются молекулы воды, которые участвуют в образовании связей белка с ДНК (около 6% от всех молекул воды, находящихся на ДНК-белковом интерфейсе [23]). Они являются посредником водородных связей между аминокислотными остатками и участками ДНК, образуя так называемые “водные мостики”. Такие опосредованные молекулами воды водородные связи вносят не меньший вклад в стабильность и специфичность ДНК-белкового взаимодействия, чем прямые водородные связи [24].

Одним из первых был изучен комплекс триптофанового репрессора с ДНК [13]. В нем было найдено 24 прямых и 6 опосредованных молекулами воды водородных связей с сахаро-фосфатным остовом ДНК. Для данного комплекса не были найдены прямые водородные и гидрофобные контакты с основаниями ДНК, кото-

рые бы отвечали за специфичность узнавания ДНК, и было показано, что именно опосредованные молекулами воды контакты белка с основаниями вносят вклад в специфичность взаимодействия. Для взаимодействия узнающей α -спирали гомеодоменов с большой бороздкой ДНК необходима сеть из около 33 водородных связей, 13 из которых опосредованы молекулами воды [25], [26].

Количество связей, опосредованных молекулами воды, больше в случаях неспецифического контакта белка с ДНК. В этих случаях вода играет роль своеобразной смазки, позволяющей белку скользить вдоль ДНК в поисках сайта связывания [27].

Наличие пространственных структур гомологичных белков позволяет находить мостиковые молекулы воды, находящиеся в одном и том же месте в нескольких структурах. Если такие молекулы воды образуют водородные связи с одинаковыми (соответствующими) аминокислотными остатками, нуклеотидами, лигандами или ионами, то такие мостиковые молекулы воды называются *консервативными*. Можно предположить, что консервативные молекулы воды расположены в сайтах гидратации макромолекул. Это подтверждается данными ЯМР [28] и молекулярной динамики [19], которые показывают, что каждый сайт занят одной молекулой воды относительно долгое время.

Поиск консервативных мостиковых молекул воды в соответствующих пространственных структурах был произведен для олигопептидов [29], комплексов белков с субстратами [28], [30], [31] и ДНК-белковых комплексов [20], [21].

В первых работах поиск консервативных молекул воды проводился среди нескольких структур. Например, среди трех структур гомеодоменов методами сравнительного анализа было исследовано положение молекул воды на ДНК-белковом интерфейсе, и были выявлены две полости на интерфейсе, где находятся консервативные молекулы воды. Появление новых 17 структур гомеодоменов позволило выявить дополнительные шесть консервативных молекул воды на ДНК-

белковом интерфейсе [20]. Эти консервативные молекулы воды находятся на поверхности гомеодомена еще до взаимодействия с ДНК, а в дальнейшем образуют водные мостики между белком и сахаро-фосфатным остовом ДНК.

Увеличение количества пространственных структур гомологичных белков потребовало автоматизации процесса поиска консервативных молекул воды в совмещенных пространственных структурах. Для этого был разработан полуавтоматический метод поиска кластеров молекул воды в наборе совмещенных структур с использованием программы wLake, подробно описанной далее [32]. Каждый кластер содержит молекулы воды (не более одной от каждой структуры), которые занимают близкие позиции в пространственном совмещении структур. Молекулы воды, входящие в один кластер, считаются консервативными.

С использованием программы wLake были найдены консервативные молекулы воды для семейства ДНК-связывающих доменов ETS [21]. Совмещение 20 доступных на момент написания работы структур ETS было проанализировано на предмет поиска консервативных контактов — водородных связей, гидрофобных кластеров и мостиковых молекул воды. Всего было найдено 12 водных кластеров, содержащих в среднем по 8 молекул воды из разных структур.

На примере анализа структур ETS доменов с ДНК видна проблема определения границ понятия консервативности молекул воды. Какова должна быть доля структур, содержащих в данном месте пространственного совмещения молекулу воды, чтобы такая молекула воды уже считалась консервативной? На данный момент единого мнения по этому вопросу нет, а границы консервативности в различных работах не совпадают.

1.1.2 Гидрофобные взаимодействия

1.1.2.1 Физические основы

Агломерация гидрофобных элементов системы в присутствии полярного растворителя (например воды) называется гидрофобным взаимодействием. В отличие от водородных связей, основу гидрофобного взаимодействия составляет отталкивание полярного растворителя, а не притяжение элементов системы [33], [34].

Гидрофобное взаимодействие является одной из основных движущих сил при сворачивании белковой глобулы. Неполярные группы атомов белковой цепи группируются, чтобы уменьшить площадь контакта с водой [33]. Так образуются гидрофобные кластеры, формирующие гидрофобное ядро белка.

Степень гидрофобности варьирует у различных аминокислот и зависит от природы бокового радикала аминокислоты. Экспериментально к гидрофобным аминокислотам были отнесены: валин, лейцин, изолейцин, фенилаланин, триптофан, метионин, цистеин, тирозин и аланин [35].

Кроме определения гидрофобности аминокислоты в целом, возможно оценивать гидрофобность отдельных групп атомов. Например алгоритм программы CluD [36], использованной в данной работе, при поиске гидрофобных кластеров рассматривает гидрофобные взаимодействия неполярных групп атомов углерода и серы: $-\text{CH}_3$, $-\text{CH}_2-$, $-\text{CH=}$, $-\text{SH}$, и $-\text{S}-$.

1.1.2.2 Гидрофобные взаимодействия в ДНК-белковых комплексах

Вопросам гидрофобного взаимодействия белков с ДНК посвящено лишь несколько работ.

Термодинамика образования ДНК-белкового комплекса описана в работах Ха [37] и Норберга [38], где показано, что удаление неполярных поверхностей из

контакта с водой является движущей силой в образовании гидрофобных контактов ДНК-белкового комплекса.

Роль конформационных изменений ДНК, происходящих при образовании гидрофобных связей между белком и ДНК рассмотрена в работах Толсторукова [39] и Житниковой [40]. Показано, что пиримидиновые сахара в В-подобной форме и пуриновые в А-подобной способствуют большей гидрофобности большой бороздки и, наоборот, повышают гидрофильность малой бороздки. Для изменения полярного/гидрофобного профиля ДНК необходимы соответствующие конформационные переходы между А и В-подобной формами сахаров.

Анализ консервативных гидрофобных ядер как в белках, так и в ДНК-белковых комплексах был проведен в работе Карягиной [41]. Для этого был разработан пакет CluD, примененный в данной работе. CluD позволяет находить гидрофобные ядра в молекулах и комплексах белков и нуклеиновых кислот.

В рамках поиска универсальных правил ДНК-белкового взаимодействия В работе Ласкомбе [7] были проанализированы прямые и опосредованные водой водородные связи для 129 ДНК-белковых комплексов. Гидрофобные взаимодействия при этом рассмотрены не были, но впервые был рассмотрен вклад ван-дер-ваальсовых взаимодействий.

Стоит отметить, что анализ гидрофобных взаимодействий представляется более оправданным, чем анализ ван-дер-ваальсовых взаимодействий. Это связано с тем, что Ван-дер-ваальсово притяжение очень слабое и не зависит от растворителя, в то время как гидрофобное взаимодействие зависит от природы растворителя, а по силе сравнимо с водородными связями. Кроме того, авторы установили порог на расстояние в $3,9\text{\AA}$, что слишком много для быстро убывающего с расстоянием ван-дер-ваальсова взаимодействия, но не достаточно для гидрофобного взаимодействия.

1.2 Алгоритмы выделения водородных связей и гидрофобных взаимодействий

1.2.1 Определение водородных связей

Поскольку атомы водорода не всегда представлены в структурах белков и нуклеиновых кислот, при определении водородной связи могут рассматриваться как реально присутствующие, так и атомы водорода, положение которых восстановлено по окружающим атомам.

В 1983 году Вольфганг Кабш и Кристиан Сандер разработали критерии определения водородных связей и вторичной структуры белков [42]. Идеальная водородная связь имеет длину 2.9 Å, атомы донора, водорода и акцептора лежат на одной прямой, энергия связи составляет -3, 0 ккал/моль. Максимальное значение длины связи составляет 5.2 Å, угла “водород-донор-акцептор” — 63°, энергии — до -0, 5 ккал/моль. Сервис DSSP [42], основанный на этих критериях и правилах, лежит в основе определения вторичной структуры в PDB, RasMol, Jmol и GROMACS.

В отличие от DSSP, в котором используются фиксированные значения параметров водородных связей, алгоритм STRIDE [43] рассматривает непрерывный спектр параметров водородных связей. Такой подход лучше отражает вариабельность положения цепи белка/ ДНК и их подвижность. Алгоритм STRIDE используется в программе для визуализации структур VMD [44].

Алгоритм программы HBPLUS [45], специально разработанной для определения водородных связей, основан только на геометрических критериях. Поскольку положение водорода в структуре не всегда известно, то вначале определяют наиболее вероятное положение водорода по окружающим атомам. При этом совершенно не обязательно, что водород окажется на прямой, соединяющей донор и акцептор в водородной связи. Поэтому для полученного треугольника “донор-водород-акцептор” были выбраны следующие геометрические критерии водород-

ной связи: расстояние от атома донора до атома акцептора — менее 3.9 Å; расстояние от атома водорода до атома менее 2.5 Å; длина связи между водородом и донором — от 1 Å до 1.33 Å; углы “донор-водород-акцептор”, угол “водород - акцептор - следующий за акцептором атом”, “донор - акцептор - следующий за акцептором атом” должны быть более 90°.

1.2.2 Определение гидрофобных взаимодействий

В настоящее время описан ряд алгоритмов, позволяющих найти гидрофобные кластеры в молекулах белков и нуклеиновых кислот [46]. В основном авторы считают аминокислоты некими неделимыми единицами и относительно них рассматривают окружение. Оценку гидрофобного эффекта производят на основании того, какая часть площади поверхности взаимодействующих молекул скрывается от водного окружения при образовании гидрофобного контакта.

Например, в работе Херинга и Аргоса [47] два аминокислотных остатка считаются сближенными, если расстояние между ними меньше 4.5 Å. Если площадь контакта для обоих остатков превосходит некоторый порог, то такие остатки считаются контактирующими. В группе из трех остатков хотя бы для одного остатка площадь контакта должна превышать порог. Две группы из трех остатков образуют кластер из четырех остатков, если у них два остатка общие. Когда кластер перестает расти, производят проверку остатков. Если остаток имеет больше общих контактов с остатками извне, чем с остатками внутри кластера, то такой остаток удаляется. Если менее 20% площади контакта остатка с другими остатками приходится на остатки кластера, то он тоже удаляется. В результате такой проверки в кластере должно оставаться не менее трех остатков, иначе он аннулируется.

Согласно алгоритму Свинделлс [48] производят группировку гидрофобно взаимодействующих неэкспонированных на поверхности остатков. Для этого отбирают слабо экспонированные остатки, принадлежащие α -спиралям или петлям. Оказывается, что для таких остатков более 75% контактов их атомов с другими

атомами могут быть классифицированы как гидрофобные. Гидрофобным контактом считается сближение двух атомов углерода на расстояние менее чем сумма ван-дер-ваальсовых радиусов плюс $1\text{\AA}$. Два остатка из отобранных считаются взаимодействующими гидрофобно, если число гидрофобных межатомных контактов превосходит число иных межатомных контактов. Для нахождения гидрофобных ядер строится граф, где вершинами являются отобранные остатки, а ребра отражают их гидрофобные взаимодействия. Связные компоненты графа, содержащие 5 или более остатков, называются гидрофобными ядрами.

Алгоритм Цефус [49] основан на поиске компактных групп боковых цепей атомов. Вводится понятие меры компактности как отношение доступной для растворителя поверхности к минимальной возможной поверхности. Группы атомов выращиваются остаток за остатком жадным алгоритмом. Далее по статистике кластеров из данного числа остатков выбираются наиболее компактные группы. Часто они состоят в основном из гидрофобных остатков.

Основываясь на определении компактных групп атомов, Цай и Нуссинов [50] предложили свой метод определения гидрофобных кластеров. Для компактных групп определяется мера изоляции как отношение площади неполярной доступной для растворителя поверхности группы к общей площади поверхности группы. Степень гидрофобности рассматривается как отношение площади поверхности неэкспонированных неполярных остатков к общей неполярной площади поверхности.

Каннан и Вишвешвара [51] предложили определять гидрофобные кластеры в белках, ориентируясь на боковые цепи белков. Взаимодействия боковых цепей аминокислот были представлены помеченным графом. В вершинах находились C^β атомы аминокислот, а ребра отражали расстояния между C^β атомами. Для определения гидрофобных кластеров рассматривали только гидрофобные аминокислоты. Два остатка, удовлетворяющие пороговому критерию перекрывания боковых цепей, соединялись в графе. Согласно авторам, сильный гидрофобный эффект воз-

никает при стекинге гидрофобных остатков или при близком расположении гидрофобных остатков так, что каждый остаток взаимодействует с несколькими другими гидрофобными соседями.

Алгоритм программы CluD [41] в отличие от вышеописанных оперирует не аминокислотными остатками, а отдельными неполярными группами атомов (CH_3 -, $-\text{CH}_2$ -, $-\text{CH}=\text{}$, $-\text{SH}$ и $-\text{S}-$), поскольку такой подход лучше отражает физический смысл гидрофобного взаимодействия. Центром неполярной группы считается центр атома углерода или серы. Атомы водорода при расчетах не учитываются. Строится граф взаимодействия неполярных групп в трехмерной структуре. Каждая неполярная группа определяет одну вершину графа. Две вершины связаны ребром (и считаются взаимодействующими) в том случае, если (а) расстояние между центрами не превышает предельного порога d ; (б) эти группы не экранируются никакой другой группой (полярной или неполярной). Условие экранирования таково: группа C заслоняет взаимодействие между неполярными группами A и B , если шар с центром C и радиусом R_C включает в себя пересечение шаров с центрами A и B и радиусами, соответственно, R_A и R_B . В программной реализации алгоритма все радиусы приняты равными $2.7 \text{ \AA}$, что соответствует минимально возможному расстоянию от центра группы, образованной атомом углерода, до атома кислорода молекулы воды.

В программной реализации d является пользовательским параметром. Максимально предусмотренное значение $d = 5.4 \text{ \AA}$, что соответствует сумме двух радиусов R_A и R_B для неполярных групп с углеродом. Значение d по умолчанию равно $4.5 \text{ \AA}$.

Ребра графа соответствуют двум существенно различным типам отношений между группами: (1) структурно (в том числе ковалентно) связанные группы; (2) собственно гидрофобные взаимодействия. Неполярные группы A и B считаются "структурно связанными", если расстояние между ними фиксировано в силу химических связей, энергия которых существенно превышает свободную энергию

гидрофобного взаимодействия пары атомов. Примерами структурно связанных неполярных групп являются: две группы, центральные атомы которых ковалентно связаны с третьим; любые две группы из ароматического кольца. В описываемом алгоритме и программе типы ребер не различаются.

Из графа удаляются "слабые" рёбра, после чего связные компоненты оставшегося графа объявляются гидрофобными кластерами. В варианте "по умолчанию" ребро слабое, если при его удалении подграф из вершин, непосредственно связанных с концами этого ребра, становится несвязным (в терминологии авторов такое ребро называется $(1,1)$ разрезом). Во общем случае слабое ребро определяется как (k,l) -разрез, где k и l - натуральные числа, являющиеся параметрами программы, но на практике применяется только вариант $k=l=1$.

Для каждого гидрофобного кластера рассчитываются следующие параметры. (1) Число неполярных групп в кластере. (2) Среднее число взаимодействий одной неполярной группы внутри кластера. Этот параметр характеризует форму кластера: чем компактней расположение гидрофобных групп и чем ближе форма кластера в пространстве к идеальному шару, тем больше среднее число взаимодействий неполярной группы внутри кластера. (3) Полуоси эллипсоида инерции, описывающего кластер в пространстве. При расчете гидрофобные группы, составляющие кластер, рассматриваются как материальные точки в пространстве и для их совокупности рассчитывается эллипсоид инерции. Соотношение полуосей отражает форму кластера в пространстве.

1.3 Механизмы ДНК-белкового узнавания

1.3.1 Поиск универсального кода и закономерностей ДНК-белкового узнавания

Одним из главных вопросов, стоявших при изучении механизмов ДНК-белкового узнавания, был поиск универсального и однозначного кода соответст-

вия аминокислотной и нуклеотидной последовательности в сайте связывания. К сожалению, уже в 1988 году было показано [52], что единого кода не существует. Тем не менее, в дальнейших работах был сделан ряд попыток поиска локальных закономерностей для групп белков и белковых мотивов. В работах Судзуки [53], [54] правила узнавания исследовались как на уровне последовательностей, так и на уровне структур. Было выяснено, что если на уровне последовательностей еще можно выделить закономерности узнавания внутри одного ДНК-узнающего мотива, то для структур разнообразие порой слишком велико и не позволяет найти некие универсальные правила для всех близких последовательностей. Одним из удачных примеров поиска кода узнавания является мотив цинкового пальца [55]. Но даже в этом случае каждому нуклеотиду (в зависимости от положения в цепи) были сопоставлены от одной до четырех аминокислот.

Отсутствие однозначного кода соответствия может быть объяснено наличием шести степеней свободы (3 для положения и 3 для ориентации) для ДНК и белка в комплексе. Это делает совпадение взаимного расположения остатков белка и ДНК в неродственных комплексах маловероятным.

Тем не менее, некоторые *закономерности и предпочтения*, наблюдаемые в ДНК-белковых комплексах, были выявлены:

1. Для образования водородных связей с нуклеотидами ДНК чаще всего используются такие заряженные аминокислотные остатки как аргинин, лизин, глутамин, глутаминовая кислота, аспарагин и аспарагиновая кислота.
2. Глицин, аланин и валин склонны образовывать водородные связи посредством атомов остова с сахара-фосфатным остовом ДНК [6].
3. Для сходных мотивов с соответствующими ДНК-белковыми взаимодействиями наблюдается высокая степень пространственной консервативности.

4. Наличие разных связей с основаниями для одного и того же остатка в разных ДНК-белковых комплексах чаще всего объясняется наличием существенных пространственных различий в этих комплексах.
5. Для белков, узнающих большую бороздку посредством α -спирали, можно выделить 2 суперсемейства. В первом аминокислотные остатки с номерами i , $i+4$, $i+8$ находятся на прямой, параллельной большой бороздке. При этом ось самой α -спирали может и не быть параллельной бороздке. Во втором суперсемействе, к которому относится и мотив цинкового пальца, шаг аминокислотных остатков меньше: i , $i+3$, $i+6$ [56].

1.3.2 Специфичность узнавания ДНК белком

Говоря об узнавании белком ДНК, исследователи довольно часто применяют понятие специфичности. Это понятие является довольно широким и однозначно не определено. Специфичность узнавания может рассматриваться в нескольких аспектах [57], указанных ниже.

1.3.2.1 Специфичность ДНК-белкового узнавания с точки зрения последовательностей

При рассмотрении последовательностей традиционно изучаются количество, вид и последовательности взаимодействующих аминокислотных остатков белка и азотистых оснований ДНК. Специфическое взаимодействие трактуется с точки зрения однозначности соответствия узнающей последовательности белка нуклеотидной последовательности ДНК.

В ранних работах изучались закономерности узнавания конкретных нуклеотидов и пар оснований. Было показано [58], что одной водородной связи между аминокислотным остатком и парой азотистых оснований не достаточно для различения пар оснований и бороздки ДНК — для этого необходимы как минимум две водородные связи.

В дальнейших работах (например в [59], связывание σ -репрессора), описывающих структуры ДНК-белкового комплекса, было показано, что для узнавания специфического сайта ДНК необходимо формирование сети водородных связей аминокислотных остатков белка с нуклеотидами ДНК. Однако вскоре стало понятно [60], что наличия консенсусной последовательности еще не достаточно для специфического связывания. Например димер λ -репрессора, состоящий из идентичных субъединиц, связывается с последовательностью ДНК асимметрично — одна субъединица прочнее связывается с консенсусной половиной оператора, чем вторая. Это связано с небольшими различиями в геометрии контактов двух частей димера, которые, тем не менее, оказывают ощутимое влияние на специфичность ДНК-белкового узнавания. Более того, для ряда белков осуществляется неспецифическое с точки зрения последовательностей связывание, основанное на комплементарности пространственных структур (см. следующий раздел).

1.3.2.2 Специфичность ДНК-белкового узнавания с точки зрения структур

При рассмотрении специфического узнавания структур авторы многих работ анализируют пространственную доступность для контакта и комплементарность структур, а также конформационные изменения в районе сайта связывания, происходящие при образовании контакта белка с нуклеиновой кислотой. Считается, что специфическое взаимодействие возможно, если форма поверхностей (вторичная и третичная структуры) белка и нуклеиновой кислоты комплементарны друг другу, а различные типы взаимодействий образуют необходимый паттерн.

В одной из первых подобных работ [61] была изучена стереохимия узнавания ДНК мотивом “цинковый палец”. Авторы выделили 2 вида ДНК-узнающих мотивов, принципиально по-разному взаимодействующих с ДНК и покрывающих разное количество оснований. Также было установлено, что для правильной ори-

ентации α -спирали мотива относительно ДНК необходим перепад длин боковых радикалов между N и C-концом α -спирали.

Анализ 75 структур белковых комплексов с нуклеиновыми кислотами позволил выявить ряд структурных закономерностей в размерах НК-белкового интерфейса, его заряде, плотности водородных связей, роли водных молекул и конформационных перестроек [62]. Авторы выяснили, что площадь контактов в целом варьирует в широких пределах (от 1120 до 5800 Å²). Однако для транскрипционных факторов ДНК можно выделить узнающие участки длиной в 12±3 нуклеотида и 24±6 аминокислот площадью 1600±400 Å². Для ДНК-узнающих ферментов площадь контакта примерно на 600 Å² больше, что связано с необходимостью формировать активный центр. Оценка полярности белковых интерфейсов на основе анализа аминокислотных последовательностей показала, что в среднем белковые интерфейсы в комплексах с ДНК являются более полярными, чем аналогичные интерфейсы в белок-белковых комплексах. Для проанализированных комплексов с высоким разрешением средняя плотность водородных связей (как прямых, так и опосредованных молекулами воды) для комплексов оказалась равной 1 связи на 125 Å². Авторы показали, что плотность упаковки атомов белка на ДНК-белковом интерфейсе аналогична таковой внутри белковой глобулы. При образовании ДНК-белкового комплекса происходит ряд конформационных изменений ДНК и белка, масштаб которых зависит от площади контакта. Для белков характерен широкий ряд перестроек, таких как движения доменов и изменение третичной и четвертичной структуры. При этом на ДНК-белковом интерфейсе остаются молекулы воды. Они играют роль посредников для обеспечения комплементарности поверхностей белка и ДНК.

Появление все большего числа структур ДНК-белковых комплексов позволило проанализировать принципы строения ДНК-связывающих белков, а также структурные вариации связывания и узнавания специфической последовательности ДНК [63]. Большинство ДНК-белковых комплексов содержат ДНК в В-форме

с ограниченным уровнем изгибов и деформаций. Белок обладает большими возможностями к конформационной подстройке, в частности поэтому говорят, что именно белок узнает ДНК, а не наоборот. В основном это происходит посредством α -спиралей и β -листов, которые входят в бороздки ДНК. Архитектура ДНК-узнающих белков отличается большим разнообразием. В связи с этим возник ряд классификаций возможных архитектур.

Тем не менее, можно выделить ряд основных *закономерностей структурного узнавания*:

1. ДНК-узнающие укладки часто содержат глобулярный домен, аминокислотные остатки на поверхности которого взаимодействуют с ДНК. Однако ряд ДНК-узнающих мотивов взаимодействует с ДНК посредством неструктурированного участка, который структурируется при связывании с ДНК и обеспечивает узнавание оснований и остова ДНК. Например, в случае λ -репрессора неструктурированный N-конец узнает большую бороздку ДНК, а в случае гомеодоменов — малую.
2. На размер узнаваемого сайта ДНК влияют конформация и размер белкового домена, число узнающих доменов, а также олигомерность белка.
3. α -спирали наиболее часто используются для узнавания оснований, в основном по большой бороздке. Для узнавания малой бороздки (Lac репрессор) спиралью необходим изгиб ДНК, чтобы α -спираль могла поместиться в бороздку. Спектр размещения α -спирали внутри большой бороздки разнообразен: от перпендикулярной позиции (как в случае Trp репрессора), через наклонную (CAP) до параллельной (например, Zn-finger domain).
4. β -листы значительно реже встречаются в качестве ДНК-узнающих элементов. Это могут быть как β -листы, образованные одной цепью белка, так и димерные антипараллельные β -листы (как в Arg-репрессоре). В основном в узнавании участвуют два тяжа β -листа. Исключением является ТАТА-связывающий домен, 10 β -

тяжей которого взаимодействуют с малой бороздкой ДНК, заставляя ее сильно изгибаться и раскрываться.

5. Неструктурированные элементы (петли, изгибы, концевые участки) в основном являются дополнительными узнающими элементами. В качестве основных они выступают относительно редко. Так, например, для суперсемейства SCOP иммуноглобулин-подобных доменов наблюдается большое разнообразие последовательностей, укладок и взаимного расположения доменов относительно ДНК. Объединяет их то, что основные контакты формируются за счет неструктурированных участков, а не β -листов домена.

6. Специфическое узнавание формируется за счет образования сети оптимальных контактов между белком и ДНК. Конформация ДНК в известной мере определяется последовательностью, что создает на поверхности ДНК специфические паттерны, форму и последовательность которых узнает белок [64]. Большая доступность оснований со стороны большой бороздки упрощает их распознавание. В связи с этим основная масса белков узнает специфические последовательности именно таким образом.

7. Отсутствие однозначного кода для узнавания не мешает формировать аналогичные по форме и физико-химическим характеристикам поверхности даже для разных последовательностей белков. При этом основную роль в специфическом узнавании играет сеть водородных связей. Важно заметить, что не существует однозначного взаимного расположения узнающего аминокислотного остатка и нуклеинового основания. Одна и та же аминокислота в зависимости от окружения может формировать одну или несколько связей с основанием, остовом, молекулой воды или другой аминокислотой.

8. Молекулам воды традиционно отводится роль посредника в образовании водородных связей. Известно, что некоторые молекулы воды могут быть предымобилизованы на поверхности белка и ДНК [65]. В результате взаимодействия гидратированных поверхностей белка и ДНК часть молекул воды удаляется, а часть ос-

тается, заполняя пустоты на интерфейсе. Такие “водные мостики” делают систему энергетически более выгодной.

9. Ван-дер-ваальсовы и гидрофобные взаимодействия в силу своей ненаправленности меньше влияют на специфичность. Тем не менее, большое количество ван-дер-ваальсовых и гидрофобных взаимодействий создают стерические затруднения, косвенно влияя на специфичность.

10. Мультимеризация белка и кооперативное узнавание повышает способность к специфическому узнаванию ДНК. Для белков, способных взаимодействовать как в виде мономера, так и в виде димера (например семейство nuclear hormone receptor), сочетание различных партнеров позволяет узнавать разнообразный набор последовательностей. Кооперативное взаимодействие (например для МАТ α 2 and МАТ α 1 гомеодоменов) приводит к более прочному связыванию и специфичности.

11. Связывание белка может приводить к конформационным изменениям ДНК. Это может быть не только изгиб (как например в структуре 1jj4 папилломавируса человека-18 E2), выпетливание и интеркаляция (как в структуре 1emj урацил-ДНК гликозилазы), но и изменение ширины бороздки. Например, малая бороздка расширяется при связывании ТАТА-связывающего домена и сужается при связывании аргинина (Phage 434 repressor) [66], [64].

1.3.2.3 Специфичность ДНК-белкового узнавания с точки зрения физико-химических параметров взаимодействия

Одним из менее распространенных подходов к оценке специфичности ДНК-белкового узнавания является измерение различных физико-химических параметров. Например в работе Норберга [38] изучали электростатические потенциалы поверхностей контактирующих молекул.

Специфичность связывания была охарактеризована величиной изменения теплоемкости при образовании ДНК-белкового комплекса [67]. Для неспецифиче-

ского связывания характерно положительное значение изменения энтропии, меньшее отрицательное значение изменения энтальпии, высокая зависимость энтальпии от ионной силы раствора и небольшое изменение теплоемкости [68].

Необходимость учитывать как узнающие последовательности, так и конформационные особенности белка и ДНК для определения специфичности была подтверждена статистически и термодинамическими расчетами [69].

Неравноценность узнавания большой и малой бороздок была изучена в работе Привалова [70]. Анализ энергетических профилей значительного количества ДНК-белковых комплексов показал, что при сравнимой свободной энергии Гиббса взаимодействие с большой бороздкой в основном определяется энтальпийным фактором, в то время как взаимодействие с малой бороздкой невыгодно с точки зрения энтальпии, но покрывается энтропийным выигрышем энергии. Это связано с различиями в гидратации большой и малой бороздок. А-Т богатые участки ДНК со стороны малой бороздки содержат структурированную воду, удаление которой дает выигрыш в энтропии связывания белка.

Кроме термодинамического подхода также применяется кинетический подход к изучению образования НК-белкового комплекса. В экспериментальной практике принято считать взаимодействие специфичным, если константа связывания превышает $K_a \approx 10^6 \cdot \text{M}^{-1}$. Тем не менее, встречаются комплексы с меньшей афинностью и специфическим связыванием [71].

Разница между специфичным и неспецифичным взаимодействием отражается в степени диссоциации [72]. Степень специфичности повышается при повышении константы связывания и при увеличении количества центров связывания [73].

Таким образом, оценка специфичности взаимодействия белка с нуклеиновой кислотой является комплексной задачей и, строго говоря, должна учитывать все перечисленные выше аспекты.

В данной работе понятие специфического взаимодействия не используется, поскольку на основании только структурных данных можно дать объективную оценку специфичности лишь в тех немногих случаях, когда есть структуры одного и того же белка с различными последовательностями ДНК.

1.4 Классификации ДНК-белковых комплексов

Первые классификации ДНК-белковых комплексов возникали при попытках найти однозначный код ДНК-белкового узнавания. Довольно быстро авторы стали кивались с необходимостью описать все разнообразие наблюдаемых взаимодействий, что приводило к возникновению классификаций.

Условно классификации ДНК-белковых комплексов можно разделить на 3 категории: белок-центрированные, ДНК-центрированные и комплекс-центрированные — в зависимости от того, которому из участников комплекса уделялось большее внимание.

Многие классификации опираются на другие общеизвестные классификации. Например структурные домены белка выделяются согласно классификации SCOP [74] или CATH [75].

Структурная таксономия ДНК-связывающих белковых доменов была предложена в работе Харрисона [76]. Авторы выделили 4 группы ДНК-связывающих доменов: спираль-поворот-спираль (НТН), цинк-связывающие домены, лейциновые молнии и домены, содержащие β -листы. Ряд дальнейших работ расширил эту классификацию.

В работе Пабо [77] были выделены структурные семейства транскрипционных факторов: спираль-поворот-спираль (helix-turn-helix, НТН), гомеодомены, цинковые пальцы, стероидные рецепторы, лейциновые молнии и мотив спираль-петля-спираль (helix-turn-helix, НЛН), а также β -листы. В качестве основных контактов выделялись контакты с основаниями и, отдельно, с сахаро-фосфатным ос-

товом посредстве водородных связей (в том числе и опосредованных молекулами воды) и гидрофобных взаимодействий.

Винтьенс с соавторами разработали процедуру автоматической классификации коротких белковых фрагментов [78] и применили ее в классификации ДНК-связывающих доменов спираль-поворот-спираль (helix-turn-helix, НТН) [79].

Джонс с соавторами рассматривали форму “отпечатков” ДНК на белке, то есть форму поверхностей контакта. На основе анализа 26 ДНК-белковых комплексов они выделили 3 класса белков: образующие один кластер контакта с ДНК (single-headed), образующие 2 кластера контакта с ДНК (double-headed) и класс белков, окутывающих ДНК подобно кольцу [80].

В работе Ласкомбе с соавторами [81] 240 ДНК-белковых комплексов были разделены на 8 структурно-функциональных групп, а именно: группа спираль-поворот-спираль (helix-turn-helix (НТН)), группа цинк-связывающих белков, группа белков-аналогов лейциновых молний, группа α -спиральных белков, группа β -листов, группа β -шпилек (β -hairpin/ribbon), группа ферментов и группа гистонов/гистоноподобных белков. В девятую группу попали исключения, которые не получилось классифицировать. В дальнейшем группы были разделены на 54 структурных семейства на основании структурного сходства ДНК-узнающих мотивов белков. Позже эта классификация была дополнена работой Малхотра и Совдхамини [82], в которой существующие семейства были расширены новыми представителями, а также была добавлена группа β -пропеллер.

В работе Пономаренко с соавторами [83] классификация была построена полностью автоматически. На основе структурных сходств ДНК-связывающих белковых паттернов было выделено 338 белковых доменов SCOP из 269 ДНК-связывающих белков.

Прабакаран [84] с соавторами разделили 62 ДНК-белковых комплекса на 7 кластеров, автоматически классифицируя их на основе 11 дескрипторов. Среди них: число водородных связей между белком и ДНК с учётом бороздки (большой

или малой) и остова/не остова, глубина и ширина бороздок ДНК, изгиб ДНК, GC-состав ДНК и площадь ДНК-белкового контакта. В процессе классификации было показано, что сходство строения ДНК-узнающих мотивов не всегда обеспечивает сходство способа узнавания ДНК белком. В связи с этим авторы подчеркивают необходимость использования в построении классификаций ДНК-белковых взаимодействий параметров, характеризующих взаимодействие в целом, а не только свойства ДНК-узнающего белка. Стоит отметить, что используемые авторами дескрипторы характеризуют контакт практически только со стороны ДНК, таким образом, данную классификацию можно считать ДНК-центрированной.

Другая ДНК-центрированная классификация предложена в статье Сатияприя с соавторами [85]. В ней построены графы ДНК-белкового взаимодействия для 118 ДНК-белковых комплексов, после чего ДНК-белковые комплексы разделены на 7 классов в зависимости от того, с какой частью ДНК (фосфатом, сахаром или основанием) контактируют аминокислотные остатки белка. Как и в предыдущей работе, в рамках предложенной классификации одни и те же белковые мотивы (например НТН-мотив) попадали в разные классы, а каждый класс включал разные мотивы.

Бисвас с соавторами [86] рассматривали и классифицировали ДНК-белковые контакты на основе того, какой элемент вторичной структуры белка находится на ДНК-белковом интерфейсе. В ряде случаев класс белка (a, b, a/b, a+b и др.) согласно классификации SCOP отличался от того, к какому интерфейсному классу (a, b, ab, nonregular) авторы отнесли данный белок. Кроме того, структурно-функциональные группы, предложенные в работе Ласкомбе [81], были охарактеризованы с точки зрения интерфейсных классов.

Шнайдер с соавторами [87] анализировали ДНК-белковые комплексы на основе классификации взаимодействующих белковых блоков из пяти последовательных остатков с одной стороны, и структур динуклеотидных конформеров с другой стороны. Результатом работы стало статистическое описание частоты кон-

тактов различных структурных элементов для различных классов белков, таких как транскрипционные факторы, нуклеазы и прочие.

1.5 Базы данных макромолекул и их комплексов

Основными широко используемыми базами данных являются PDB и NDB. PDB (Protein Data Bank) [88], белковый банк данных, — это единый открытый архив пространственных структур больших биологических молекул, таких как белки, нуклеиновые кислоты и их комплексы. Информацию о пространственных структурах получают методами рентгеновской кристаллографии, ЯМР-спектроскопии, а в последнее время всё чаще, методом крио-электронной микроскопии и вносят в PDB. PDB содержит первичные данные о структуре биологических молекул. Четырехсимвольные идентификаторы PDB используются чаще других для обозначения пространственной структуры молекул. Данные PDB активно используются при построении других банков данных, категоризирующих и выявляющих закономерности между структурами.

NDB (Nucleic Acid Database) [89,90] — база нуклеиновых кислот, которая, в частности использует данные PDB, позволяет искать ДНК, РНК и их комплексы с белками. Возможность поиска по типу нуклеиновой кислоты, белка и по их структурным и функциональным особенностям делает NDB удобным инструментом при работе с нуклеиновыми кислотами. Кроме первичных структурных данных, NDB содержит атлас пространственных мотивов РНК, неперепредставленный список РНК-содержащих структур, атлас триплетов РНК, инструмент нуклеотидного выравнивания для РНК, а также сервер для поиска, выравнивания и анализа рекуррентных пространственных мотивов РНК.

Примером базы данных, отражающей классификацию, является SCOP [74] — структурная классификация белковых доменов. Хотя в данной классификации рассматриваются только белковые домены, без ДНК, она является основопола-

гающей для многих последующих работ и баз данных. Согласно SCOP можно выделить четыре уровня гомологии белковых доменов по структуре и последовательности: (1) классы, (2) укладки, (3) суперсемейства, (4) семейства. Отнесение белкового домена к тому или иному классу было произведено автоматически с последующей экспертной оценкой.

Аналогом SCOP является классификация и созданная на ее основе база данных CATH [91]. Свое название она получила по четырем уровням иерархии: класс (Class), архитектура (Architecture), топология (Topology) и гомологичные суперсемейства (Homologous superfamily). Домены выделяются в белках автоматически на основе структурного сходства с ранее выделенными доменами. Новые домены выделяются при совпадении результатов трех алгоритмов DETECTIVE [92], PUU [93] и ДОМАК [94]. При несовпадении результатов решение о классификации домена принимает эксперт.

1.5.1 Базы ДНК-белковых взаимодействий

Все базы ДНК-белковых взаимодействий основаны на структурных данных из PDB. Кроме PDB, база данных может быть связана с другими базами, несущими информацию о последовательностях белка и ДНК. Стандартно в базах ДНК-белковых взаимодействий предлагается сервис по визуализации структур, а также различный набор инструментов анализа ДНК-белковых комплексов.

Одной из первых баз данных стала японская база 3DinSight [95], в настоящее время неактивная. Авторы использовали данные базы белковых мотивов и функциональных сайтов PROSITE [96] в 3DinSight, что позволило автоматически отражать на структуре белковые мотивы и значимые сайты. Для отображения мутаций была использована информация из базы данных белковых мутаций PMD [97], а для анализа аминокислотного состава были разработаны собственные инструменты.

Вопросу взаимодействия белка и ДНК была посвящена база AANT (the Amino Acid-Nucleotide Interaction Database) [98]. В ней были охарактеризованы все взаимодействия аминокислот с нуклеотидами в известных на тот момент 930 структурах. Для каждой пары “аминокислота – нуклеотид” было сделано пространственное совмещение этого взаимодействия, взятое из всех доступных структур. Сходные взаимодействия были объединены в кластеры. Эта информация была представлена в виде схем и сервиса для визуализации структур. В настоящее время база данных неактивна.

Термодинамическим аспектам ДНК-белковых комплексов посвящена база ProNIT [99]. Кроме стандартной информации о структуре и последовательности белка и ДНК, в базе есть ряд термодинамических параметров, а также экспериментальные условия получения структур и ссылки на литературу. Данные по термодинамике были взяты на основе анализа литературы для каждого комплекса. Предсказанные (неэкспериментальные) термодинамические данные в базу включены не были. С некоторого момента база данных также неактивна.

BIPA (Biological Interaction database for Protein-nucleic Acid) [100] — это английская база данных ДНК-белковых взаимодействий. В ней представлена информация о таких параметрах ДНК-белкового интерфейса, как размер, форма, склонность остатков к образованию связей, вторичная структура участников и межмолекулярные взаимодействия. Также там содержатся множественные пространственные совмещения ДНК-белковых комплексов с описанием локального окружения контактирующих аминокислот. После 2009 года веб-интерфейс базы данных позволяет только скачать базу целиком, а не воспользоваться онлайн-формами.

3D-footprint [101] — это курируемая еженедельно обновляемая база данных, позволяющая оценить специфичность ДНК-белковых взаимодействий для доступных в PDB комплексов. Веб-интерфейс позволяет пользователю находить ДНК-белковый комплекс по ключевым словам; находить белки, узнающие сходный мотив ДНК; находить сервисом BLAST [102] сходные белки и на итоговом выравни-

вании отмечать аминокислотные остатки, находящиеся на интерфейсе. Каждый комплекс в 3D-footprint охарактеризован графом контактов на интерфейсе, частотой встречаемости контактирующих нуклеотидов для аналогичных интерфейсов, дендрограммой сходных ДНК интерфейсов. Ключевые эксперименты (включая сравнения с курируемыми базами данных RegulonDB [103] и TRANSFAC [104]) по определению специфичности конкретного ДНК-белкового взаимодействия подкрепляют основанные на структурном анализе оценки специфичности, данные в 3D-footprint. Данные базы доступны как для загрузки, так и через веб-интерфейс.

Данные о конформационных параметрах нуклеиновых кислот были объединены с информацией о доступной для контакта поверхностью нуклеиновой кислоты в базе данных ProtNA-ASA [105]. На октябрь 2008 года в базе содержалось 214 ДНК-белковых и 24 РНК-белковых негомологичных комплекса. База данных предоставляет информацию о структурных параметрах, описывающих локальную геометрию нуклеотидных пар и торсионных углов. Также в базе приведены результаты расчетов доступных для контакта с белком областей как НК-белковых комплексов в целом, так и отдельно для большой и малой бороздок ДНК. ProtNA-ASA может быть использована для изучения связи конформации нуклеиновой кислоты с доступностью ее атомов для контакта с белком.

PDIdb (Protein-DNA Interface database) [106] — это специализированная база ДНК-белковых интерфейсов. Текущая версия базы содержит 922 записи непохожих между собой структур с разрешением более 2.5 Å. Имеющиеся в базе белки были классифицированы по трем уровням: классы (ферменты, транскрипционные факторы, структурные/ДНК-связывающие белки), типы (фермента, укладки транскрипционного фактора или функция структурного/ДНК-связывающего белка) и подтипы (конкретная функция белка). Эта классификация была получена в результате экспертного анализа данных PDB, PubMed, CATH, SCOP и CoPS [107]. Для ДНК были выявлены такие структурные особенности как наличие выщепления оснований, выпетливания, липких и расплавленных концов, а также модифи-

цированных нуклеотидов и гетеродимерных дуплексов. Для каждого комплекса было определено взаимное расположение белка и ДНК: перпендикулярно, параллельно или комплексно.

Основная цель базы PDIdb — внести вклад в изучение закономерностей ДНК-белкового узнавания. Для этого авторы сконцентрировались на изучении конкретных атомных интерфейсов, а не на паттернах связывания. Каждый комплекс охарактеризован числом эффективных контактов, принадлежностью к кластеру подобных последовательностей и интерфейсов, а также частотами контактов. Таким образом для каждого комплекса выделен его уникальный ДНК-белковый интерфейс.

NPIDB (Nucleic acids – Protein Interaction DataBase) [2] — база данных НК-белковых взаимодействий — это автоматически обновляемая база данных, содержащая актуальный список всех доступных в PDB ДНК- и РНК-белковых комплексов. Также NPIDB имеет сервисы для анализа НК-белковых комплексов. Каждый комплекс имеет собственную страницу, содержащую основную информацию из PDB, а также данные об асимметрических и биологических единицах, списки доменов Pfam [108] и SCOP, термины GO [109], а также данные о последовательностях белка и ДНК и контактах на них. С помощью программы CluD можно найти все гидрофобные кластеры на интерфейсе, а с помощью встроенных сервисов — отобразить потенциальные водородные связи, в том числе и опосредованные молекулами воды. Кроме того, в базу данных интегрирована классификация, представленная в данной работе.

Глава 2. Классификация ДНК-белковых взаимодействий

2.1 Материалы и методы

2.1.1 Основные понятия

В рамках данной работы рассматривались только те записи PDB, на которые ссылается SCOP 1.75. Среди них рассматривались только те, в которых структурные белковые домены, описанные в SCOP, контактируют с двойной спиралью ДНК длиной не менее 10 пар оснований (допускались однонуклеотидные выпетливания или выщепления азотистых оснований). Белковый домен описан в SCOP как фрагмент одной белковой цепи одной записи PDB, то есть характеризуется идентификатором записи (например, 1ic8), идентификатором цепи (например, A), начальным и конечным номерами остатков в цепи (например, 87-180). Если номера остатков отсутствуют, считается, что домен образован всей указанной цепью.

В настоящей работе один домен, описанный в SCOP указанным образом в одной записи PDB, называется *структурой* белкового домена. *Доменом* мы называем участок реального (природного) белка, которому в одной или нескольких записях PDB могут отвечать несколько структур (Например, в структуре ядерного фактора гепатоцитов (1ic8, семейство POU-specific domain, a.35.1.1¹) участки цепи A:87-180 и цепи B:85-179 являются одним и тем же белковым доменом (рисунок 1).



Рисунок 1. Выравнивание представителей белкового домена “POU-specific domain” в цепях A и B структуры 1ic8.

Заметим, что в SCOP семейство доменов делится на собственно "домены", которые далее делятся на "виды" (например, домен POU-specific domain делится

¹ Здесь и далее указываются идентификаторы структур PDB и идентификаторы семейств SCOP

на "Pit-1" крысы и "Hepatocyte nuclear factor 1a", а также "Oct-1" человека). В большинстве случаев "домен+вид" из SCOP соответствует нашему понятию домена, однако это не так в тех случаях, когда домены одного названия встречаются в одном белке одного вида два раза (например, на N-конце и на C-конце) или большее количество раз. В таких случаях мы считаем их разными доменами, поскольку они представляют разные биологические сущности, хотя на сайте SCOP они и попадают в одну "ячейку" вида "домен+вид". Иными словами, для нас "домен" — это участок природного белка независимо от того, сколько раз и как он представлен в PDB, а представление в PDB называется "структурой домена".

Если в одной и той же белковой цепи встречаются несколько одинаковых доменов SCOP, идущих подряд, то их идентичность определяется на основании совпадения их последовательностей.

Например, в структуре димерного ДНК-связывающего домена Tc3 транспозазы (1u78, семейство Recombinase DNA-binding domain, a.4.1.2) есть 2 участка в цепи A (A:2-54 и A:55-104), относящихся к одному и тому же домену SCOP (домен "Transposase tc3a1-65"). Тем не менее в данной работе они отнесены к разным белковым доменам в соответствии с разницей в последовательностях (рисунок 2).

1u78_A(2-54) - P R G S A L S D T E R A Q L D V M K L - L N V S L H E M S R K I S R S R H C I R V Y L K D P V S Y C T S K R
1u78_A(55-104) A P R R K A L S V R D E R N V I R A A S N S C K T A R D I R N E L Q L S A S K R T I L N V I K R S G -----

Рисунок 2. Выравнивание представителей белкового домена Transposase tc3a1-65 в цепях A и B структуры 1u78.

В семействе Interferon regulatory factor (a.4.5.23) есть три белковых домена, представленных одной или несколькими структурами в комплексе с ДНК: первый домен — Interferon regulatory factor 1, IRF-1 мыши (1if1); второй домен — Interferon regulatory factor 3, IRF-3 человека (1t2k, 2o6g, 2pi0); третий домен — Interferon regulatory factor-2, IRF-2 мыши (2irf).

Основным объектом исследования в данной работе служат структуры комплексов доменов с ДНК. Такая структура — это часть записи PDB, описывающая

структуру одного белкового домена и структуру двойной спирали ДНК длиной не менее 10 пар нуклеотидов, которая контактирует с данным доменом.

Считалось, что белок контактирует с ДНК, если между ними есть водородные связи и/или атомы ДНК и белка формируют общие гидрофобные кластеры. Ионные взаимодействия рассматривались как частный случай водородных связей. Контакты через водные мостики в данной работе не учитывались ввиду того, что в структурах с разрешением хуже 2.5 Å молекулы воды отсутствуют.

Взаимодействующий элемент со стороны ДНК — это либо сахарофосфатный остов, либо атомы азотистых оснований, обращенные в большую бороздку, либо атомы азотистых оснований, обращенные в малую бороздку. Атомами остова считались атомы дезоксирибозы и остатка фосфата. Атомами большой бороздки считались атомы N6, N7 аденина, O4 тимина, O6, N7 гуанина и N4 цитозина. Атомами малой бороздки считались атомы N1, N3 аденина, N3 тимина, N2, N3 гуанина и O2 цитозина (рисунок 3).

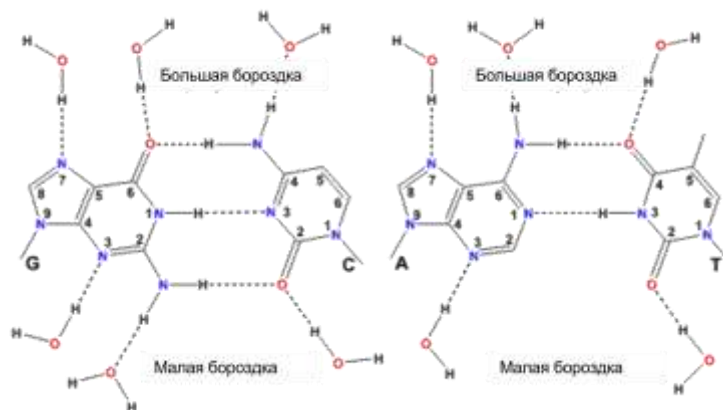


Рисунок 3. Отнесение атомов азотистых оснований к большой и малой бороздкам ДНК.

Взаимодействующий элемент со стороны белка определялся по вторичной структуре (α -спираль, β -лист или петля) контактирующего аминокислотного остатка.

Как для водородных связей, так и для гидрофобного взаимодействия были введены понятия “единицы контакта”.

За единицу контакта для водородных связей было принято взаимодействие одного аминокислотного остатка белка посредством одной или нескольких водородных связей с одним элементом ДНК.

За единицу контакта для гидрофобных взаимодействий было принято взаимодействие одного аминокислотного остатка белка с одним элементом ДНК посредством хотя бы одной пары гидрофобно взаимодействующих атомов.

Гидрофобно взаимодействующими атомами называются два атома, которые: (1) образуют гидрофобный контакт; (2) входят в один и тот же гидрофобный кластер. Гидрофобный контакт образует пара неполярных атомов, если расстояние между их центрами менее 5.4 Å, а отрезок, соединяющий их центры, не пересекается ван-дер-ваальсовой сферой никакого третьего атома. Гидрофобные кластеры определялись программой CluD.

Типы контакта описываются парой взаимодействующих элементов со стороны белка и ДНК. Всего можно выделить 9 типов контакта (в скобках приведены принятые в тексте обозначения):

- Контакты белкового домена с сахаро-фосфатным остовом ДНК посредством:
 - Спирали (в основном α -спирали, в отдельных случаях 3.10 спирали) (**H – Bb**)
 - β -Листа (**S – Bb**)
 - Неструктурированного участка или петли (**L – Bb**)
- Контакты белкового домена с ДНК со стороны большой бороздки посредством:
 - Спирали (в основном α -спирали, в отдельных случаях 3.10 спирали) (**H – Mj**)
 - β -Листа (**S – Mj**)
 - Неструктурированного участка или петли (**L – Mj**)
- Контакты белкового домена с ДНК со стороны малой бороздки посредством:

- Спирали (в основном α -спирали, в отдельных случаях 3.10 спирали) (**H – Mn**)
- β -Листа (**S – Mn**)
- Неструктурированного участка или петли (**L – Mn**)

Следует учитывать, что в конкретной структуре ДНК-белкового комплекса один и тот же *тип контакта* может быть представлен одним или несколькими *контактами* (вне зависимости от природы контакта).

Например, контакт спирали по большой бороздке ДНК может быть реализован:

- одной водородной связью от одного аминокислотного остатка
- двумя водородными связями от одного и того же аминокислотного остатка
- двумя водородными связями от двух аминокислотных остатков
- комплексом водородных связей
- одним или несколькими гидрофобными кластерами
- сочетанием гидрофобных кластеров и водородных связей

Каждая структура белкового домена в комплексе с ДНК характеризуется в данной работе своим способом взаимодействия с ДНК. Под способом взаимодействия понимается список типов контактов (H-Bb, L-Mn и т.п.), наблюдаемых для данного комплекса.

Например, в принятых сокращениях запись “|H-Bb|H-Mj|S-Bb|S-Mn|L-Bb|L-Mn|” обозначает, что белковый домен контактирует с большой бороздкой ДНК посредством остатков α -спирали, с малой бороздкой — посредством остатков β -листа и петли, а также α -спиралью, β -листом и петлей с сахаро-фосфатным остовом.

Способы взаимодействия белковых доменов с ДНК могут сильно варьировать даже для разных структур комплексов одного и того же домена. Поэтому в работе принято, что если хотя бы в одной структуре данного домена обнаружен

некоторый тип контакта, то такой тип контакта считается возможным для домена. Тем самым для каждого ДНК-связывающего белкового домена имеется список возможных для него типов контакта, собранный из всех структур, доступных в PDB.

Тип контакта считается характерным для семейства белковых доменов SCOP, если он встречается хотя бы в одной структуре для каждого домена данного семейства.

Класс взаимодействия с ДНК для семейства SCOP — это список всех характерных для семейства типов контакта белкового домена с ДНК.

2.1.2 Материалы

В качестве источника информации о структурах ДНК-белковых комплексов использовалась база данных NPIDB, которая содержит структуры комплексов белков с нуклеиновыми кислотами. Среди имеющихся в базе данных ДНК-белковых комплексов для дальнейшего анализа было выделено 748 записей PDB (включая комплексы разных субъединиц из одной записи PDB), которые содержат двухцепочечную ДНК, возможно с дополнительными одноцепочечными участками, и при этом ДНК-узнающий домен белка присутствовал в БД SCOP версии 1.75.

Анализировались структуры, полученные как рентгеноструктурным анализом (РСА), так и ядерным магнитным резонансом (ЯМР). Использовались РСА структуры любого разрешения; в случае ЯМР структур рассматривалась только первая модель. В случае, если представленная в PDB асимметрическая единица структуры комплекса не совпадала с биологической единицей, использовалась биологическая единица. При наличии двух и более биологических единиц они все рассматривались как разные структуры.

Из каждого PDB файла были извлечены белковые домены (согласно SCOP), контактирующие с ДНК. Всего выделено 314 различных доменов. Гистограмма числа структур, представляющих один домен, приведена на рисунке 4.

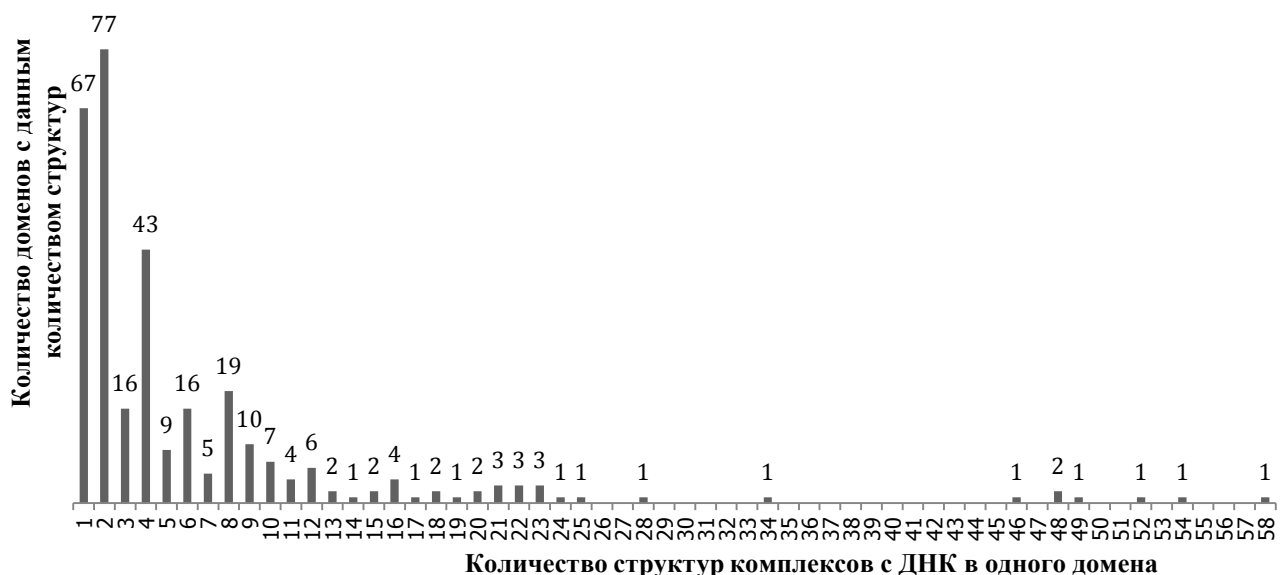


Рисунок 4. Гистограмма распределения количества структур комплексов с ДНК одного домена.

Довольно часто (в ~35% случаев) из одного PDB файла выделяется несколько ДНК-контактирующих доменов. Например в записи 2f8n встречается восемь структур пяти различных доменов семейства Nucleosome core histones (a.22.1.1) в комплексе с ДНК. Это связано с наличием нескольких субъединиц белка внутри одной записи PDB. Подобная же ситуация возникает при наличии в одной белковой цепи двух или нескольких доменов одного или разных семейств. Гистограмма числа доменов в одной записи PDB показана на рисунке 5, а числа структур – на рисунке 6.

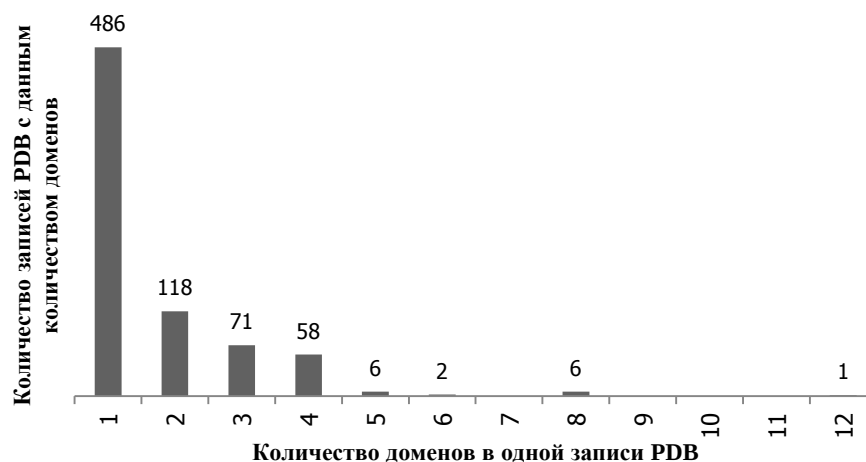


Рисунок 5. Гистограмма распределения количества доменов в одной записи PDB

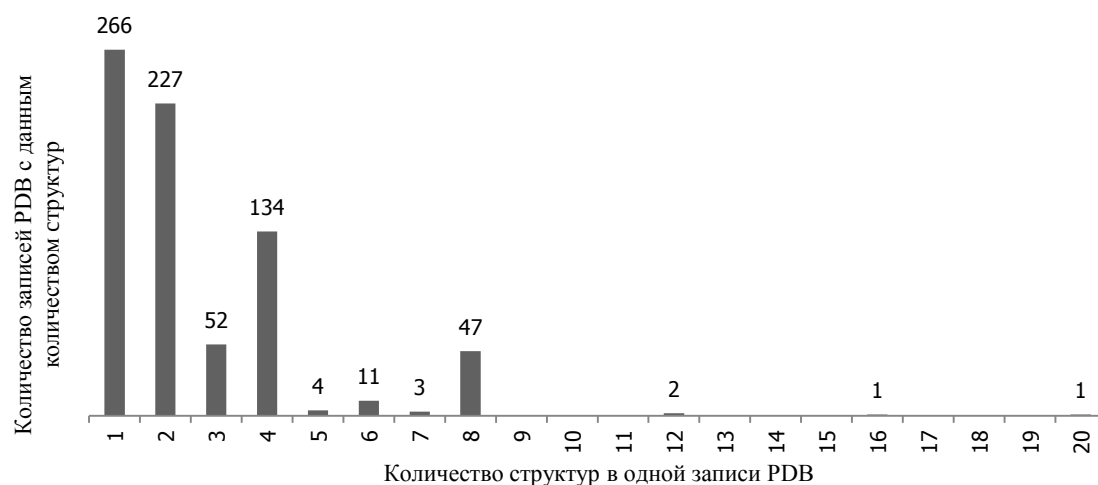


Рисунок 6. Гистограмма распределения количества структур в записи PDB.

В итоге было отобрано 1942 структуры комплексов доменов с ДНК. Эти структуры описывают 314 различных доменов, являющихся представителями 115 семейств по классификации SCOP.

2.1.3 Методы

2.1.3.1 Совмещение структур комплексов и выравнивание

Внутри каждого из 115 отобранных семейств SCOP было сделано пространственное совмещение всех белковых и ДНК-белковых структур доменов, представляющих данное семейство. Для построения пространственного совмещения использовался сервис SSM [110] и, в тех случаях, когда результат совмещения был оценен как неудовлетворительный — сервис CE [111]. Выравнивания белковых последовательностей семейств были автоматически получены как результат работы сервиса SSM. Если SSM не использовался, последовательности были выровнены с помощью Muscle [112].

Для определения вторичной структуры белков применялась программа Stride [113]. Поскольку определение вторичной структуры производится автоматически, возможны ситуации ошибочного определения (например, часть деформированной α -спирали распознается как неструктурированный участок, а корот-

кий участок петли — как полвитка α -спирали). Это может повлиять на дальнейшее определение способа взаимодействия конкретных структур и класса взаимодействия для семейств. Поэтому для семейств в целом была проведена экспертная проверка полученных данных о вторичной структуре и классе взаимодействия.

Для работы с множественными выравниваниями использовалась программа Jalview [114]. Для визуализации и анализа пространственных структур и их со-
вмещений применяли программы Jmol [115] и Rymol [116].

Поиск гидрофобных кластеров (пространственно сближенных гидрофобных групп атомов в одной структуре) на ДНК-белковом интерфейсе проводили с помощью программы CluD (параметры взяты по умолчанию), интегрированной в базу данных NPIDB.

2.1.3.2 Процедура определения способа взаимодействия белкового домена с ДНК в данной структуре комплекса

Процедура позволяет для структуры комплекса белкового домена с ДНК получить характеристику его способа взаимодействия домена с ДНК в данной структуре (рисунок 7). Вначале, используя сервисы NPIDB, происходит определение водородных связей и гидрофобных кластеров на ДНК-белковом интерфейсе. Далее программой Stride определяется вторичная структура белка, а программой 3DNA [117] — структура ДНК, то есть выделяются пары комплементарных оснований. После этого создается список из типов контакта, представленных хотя бы одной единицей контакта любой природы. В результате получается способ взаимодействия белкового домена с ДНК в виде сокращенной формулы, например |H-Bb|H-Mj|S-Bb|S-Mn|L-Bb|L-Mn|.

Дополнительно для каждой структуры проверяется наличие интеркаляции аминокислотных остатков между основаниями ДНК в месте контакта. Случаи интеркаляции возникают при взаимодействии белка с ДНК в зоне изгиба, выпетливания или выплепления азотистого основания. На факт интеркаляции указывает

взаимодействие одного и того же аминокислотного остатка белка с атомами ДНК, относящимися как к большой, так и к малой бороздке. При обнаружении интеркаляции комплекс помечается соответствующей меткой. Далее проверяется, не нужно ли сократить список обнаруженных способов взаимодействия белкового домена с ДНК, например, если взаимодействие α -спирали белка с атомами малой бороздки является следствием интеркаляции, а не реального взаимодействия со стороны малой бороздки.

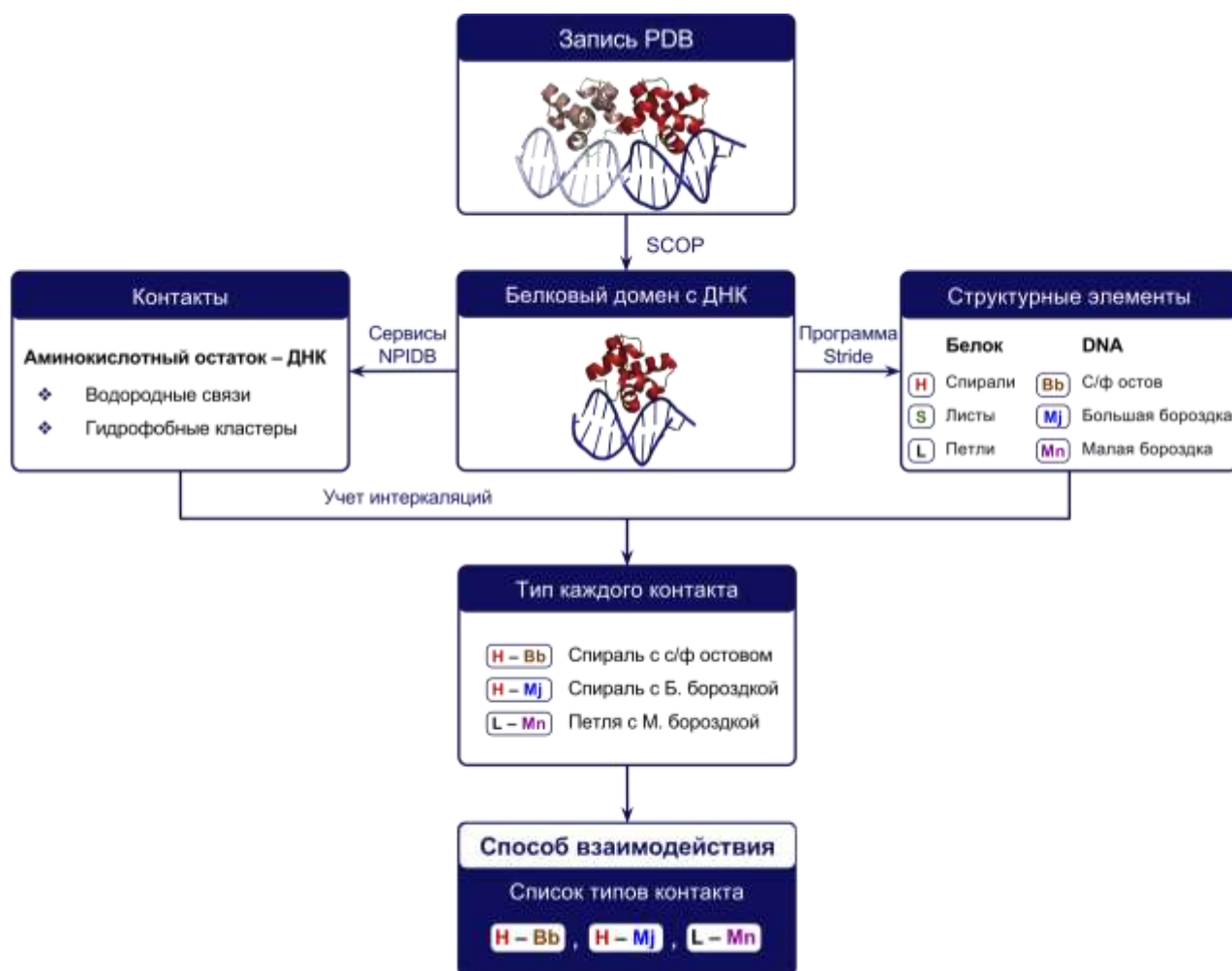


Рисунок 7. Процедура определения способа взаимодействия для структуры белкового домена с ДНК.

2.1.3.3 Построение классификации семейств белковых доменов SCOP по классам взаимодействия с ДНК

Классификации подвергались семейства, представленные структурами, относящимися к трём или более разным белковым доменам.

Для отнесения семейства белковых доменов SCOP к тому или иному классу взаимодействия использовалась следующая процедура (рисунок 8):

1. Для каждого белкового домена выявлялись все типы контакта, которые встретились хотя бы в одной структуре, относящейся к этому белковому домену.
2. Семейству присваивался *класс взаимодействия с ДНК*, представляющий собой список типов контакта, общих для белковых доменов данного семейства.

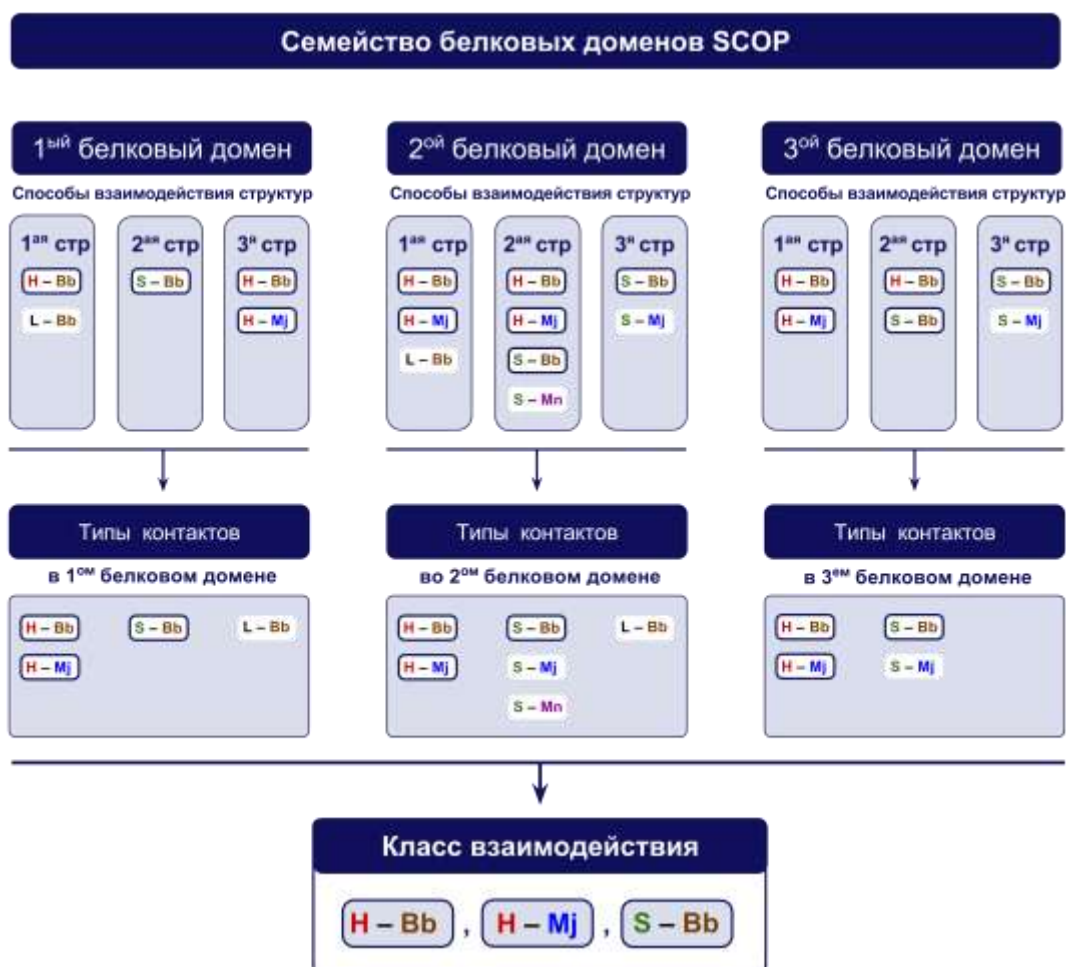


Рисунок 8. Процедура классификации семейств ДНК-белковых комплексов.

2.2 Результаты

2.2.1 Способ взаимодействия для каждой структуры ДНК-белковых комплексов

В рамках данной работы было проанализировано 1942 структуры из 748 записей PDB, описывающих 314 белковых доменов из 115 семейств белковых доменов SCOP. Для каждой структуры был сформирован список типов контактов, реализуемых в данном ДНК-белковом комплексе.

В результате было выделено 97 различных способов взаимодействия белкового домена с ДНК из 511 теоретически возможных (см. таблицу I в приложении и рисунок 41).

При определении способа взаимодействия были учтены возможные интеркаляции белка в структуру ДНК. Аминокислоты белкового домена могут интеркалировать в случае изгиба ДНК, выпетливания или выщепления азотистого основания ДНК, а также если торец ДНК контактирует с белковым доменом. Например, внутри семейства "Histone-like" proteins from archaea (b.34.13.1) белковый домен контактирует с малой бороздкой ДНК посредством β -листа и петель. За счет изгиба ДНК в ряде случаев (например в структуре 1azp) происходит интеркаляция, и аминокислотные остатки аланина, валина и метионина, встраиваясь между нуклеотидами, образуют гидрофобный контакт с атомами азотистого основания, относящимися к большой бороздке. Поскольку фактически контакта со стороны большой бороздки не происходит, подобные контакты не учитывались при определении способа взаимодействия (рисунок 9).

Во всех случаях интеркаляции, возникающей из-за изгиба ДНК, дополнительные контакты с противоположной бороздкой возникали за счет гидрофобных кластеров, а не водородных связей.

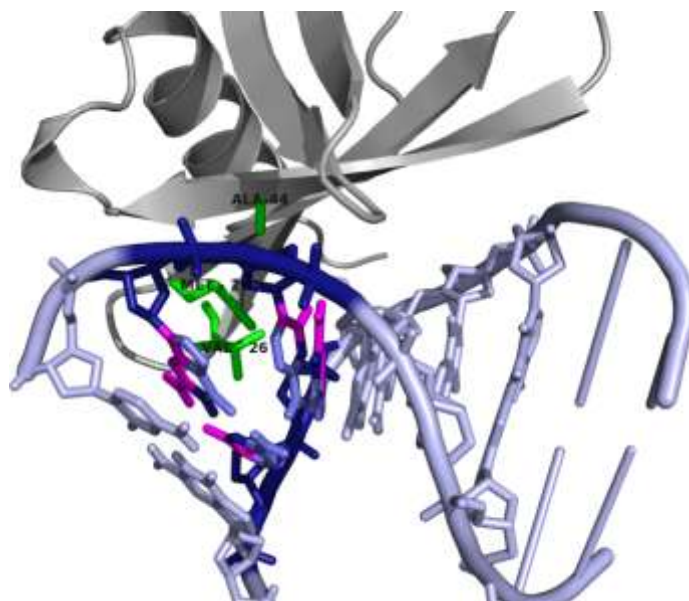


Рисунок 9. Интеркаляция остатков валина (Val26), метионина (Met29) и аланина (Ala44) между азотистыми основаниями ДНК в структуре гипертермофильного хромосомального белка SAC7D (1azp).

В таблице I (см. приложение) представлен список наблюдаемых способов взаимодействия, а также количество SCOP семейств, записей PDB, структур комплексов и различных доменов, для которых такие способы взаимодействия характерны. Отметим, что ни количество доменов, выделенных из полных структур, ни количество самих записей PDB не может быть использовано для анализа распространенности того или иного способа взаимодействия. Это связано с тем, что внутри одной пространственной структуры комплекса может быть несколько субъединиц и/или несколько одинаковых доменов, а наличие самой структуры в банке данных PDB в большой степени зависит от технических особенностей расшифровки структур белков. Поэтому в дальнейшем в данной работе анализируются в первую очередь количества семейств SCOP, внутри которых наблюдаются различные способы взаимодействия с ДНК.

Из таблицы I видна неравномерная распространенность различных способов взаимодействия белкового домена с ДНК. Все способы взаимодействия отсортированы по количеству семейств SCOP, внутри которых они встретились. В верх-

ней части таблицы превалируют контакты α -спиралей и петель с большой бороздкой ДНК, подкрепленные разнообразными контактами с сахаро-фосфатным скелетом.

В нижней части таблицы находятся способы узнавания ДНК, которые наблюдались лишь в единичных случаях (например контакты с участием малой бороздки и β -листов). Это может быть связано как с техническими трудностями разрешения структур таких ДНК-белковых комплексов, так и с действительно редкой встречаемостью этих способов узнавания ДНК в природе. Заметим, что наименьшее количество структур наблюдается в случае сложно организованного контакта белкового домена с ДНК. Сочетание α -спиралей, β -листов и неструктурированных участков белка, одновременно взаимодействующих с обеими бороздками ДНК, в рамках одного и того же домена встречается относительно редко — все подобные способы представлены не более чем двумя структурами, относящимися к одному или двум семействам.

Заметим, что один и тот же способ взаимодействия с ДНК может наблюдаться в семействах, относящихся к разным классам укладок SCOP.

Например точечный контакт α -спирали и петли с сахаро-фосфатным остовом ДНК ($|\text{H-Bb}| \text{L-Bb}|$) есть в семействах Nucleosome core histones (a.22.1.1), p53 DNA-binding domain-like (b.2.5.2), Restriction endonuclease EcoRV (c.52.1.2), Eukaryotic DNA topoisomerase I, catalytic core (d.163.1.2), Erythroid transcription factor GATA-1 (g.39.1.1), Leucine zipper domain (h.1.3.1), относящихся к классам α , β , α/β , $\alpha+\beta$, малых белков и суперспиральных белков соответственно.

Также была проанализирована частота встречаемости каждого из типов контакта в проанализированных семействах (рисунок 10). Полученные данные хорошо согласуются с представлениями о пространственной доступности взаимодействующих элементов ДНК, а также с размерами и подвижностью взаимодействующих элементов со стороны белка.

Со стороны ДНК наиболее распространены контакты по сахаро-фосфатному остову, что связано не только с тем, что сахаро-фосфатный остов больше открыт к контактам чем бороздки ДНК, но и с тем, что практически все контакты белковых доменов с бороздками ДНК неизбежно сопровождаются контактами с сахаро-фосфатным скелетом. Контакты с большой бороздкой ДНК более распространены, чем контакты с малой бороздкой вследствие различия их ширины и доступности.

Со стороны белка наиболее распространенными являются контакты посредством петель и неструктурированных участков. Это объясняется несколькими факторами: (а) неструктурированная белковая цепь обладает достаточной подвижностью, чтобы, изгибаясь, образовывать контакты с ДНК; (б) часто контактирующий участок белка представляет собой α -спираль / β -лист, окаймленный петлями/ неструктурированными участками белка.

Контакты посредством α -спиралей более распространены, чем контакты β -листов, в силу меньшего размера спиралей и большей распространенности α -спиралей в белках. α -спирали более склонны образовывать контакты с большой бороздкой ДНК, чем β -листы и наоборот — в малой бороздке β -листы встречаются чаще, чем α -спирали. При этом связывание β -листов с малой бороздкой приводит к её расширению [118].



Рисунок 10. Частота встречаемости типов контакта в семействах. Указано количество SCOP семейств, в которых есть структуры с данным типом контакта. (H – α -спираль, S – β -лист, L – петля/неструктурированный участок белка, Mj – большая бороздка ДНК, Mn – малая бороздка ДНК).

2.2.2 Закономерности взаимодействия, наблюдаемые в разных структурах одного белкового домена

Напомним, что один домен может быть представлен несколькими структурами. Количество структур зависит от числа записей PDB и содержимого этих записей (многие записи содержат более одной структуры одного и того же домена) и варьирует от одной до более чем десяти (рисунок 6).

В рамках белкового домена может реализовываться несколько способов взаимодействия (рисунок 11). Примерно в половине случаев способ взаимодействия сохраняется для разных структур одного домена. В остальных случаях разница в способе взаимодействия может быть обусловлена подвижностью боковых цепей, различиями во определении вторичной структуры, аминокислотной заменой в месте контакта, а также разницей во взаимном расположении белка относительно ДНК.

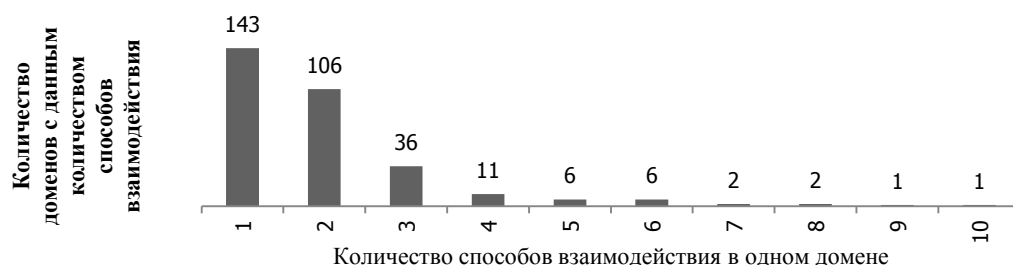


Рисунок 11. Гистограмма распределения количества структур в записи PDB.

Например в структурах гистона H4 *X. Laevis* (семейство a.22.1.1), наблюдается четыре способа взаимодействия, обусловленные разницей во взаимном расположении белка относительно ДНК:

- |S-Bb|L-Bb|
- |H-Bb|S-Bb|L-Bb|
- |H-Bb|S-Bb|L-Bb|L-Mj|
- |H-Bb|H-Mj|S-Bb|L-Bb|

Тем не менее в данном случае можно выделить общий тип контакта (L-Bb), а также способ взаимодействия, включающий все остальные: |H-Bb|H-Mj|S-Bb|L-Bb|.

2.2.3 Закономерности строения белковых доменов, взаимодействующих с ДНК одним и тем же способом

Среди доменов, взаимодействующих одним способом, в некоторых случаях можно проследить наличие общих закономерностей пространственного строения белковых доменов, узнающих ДНК.

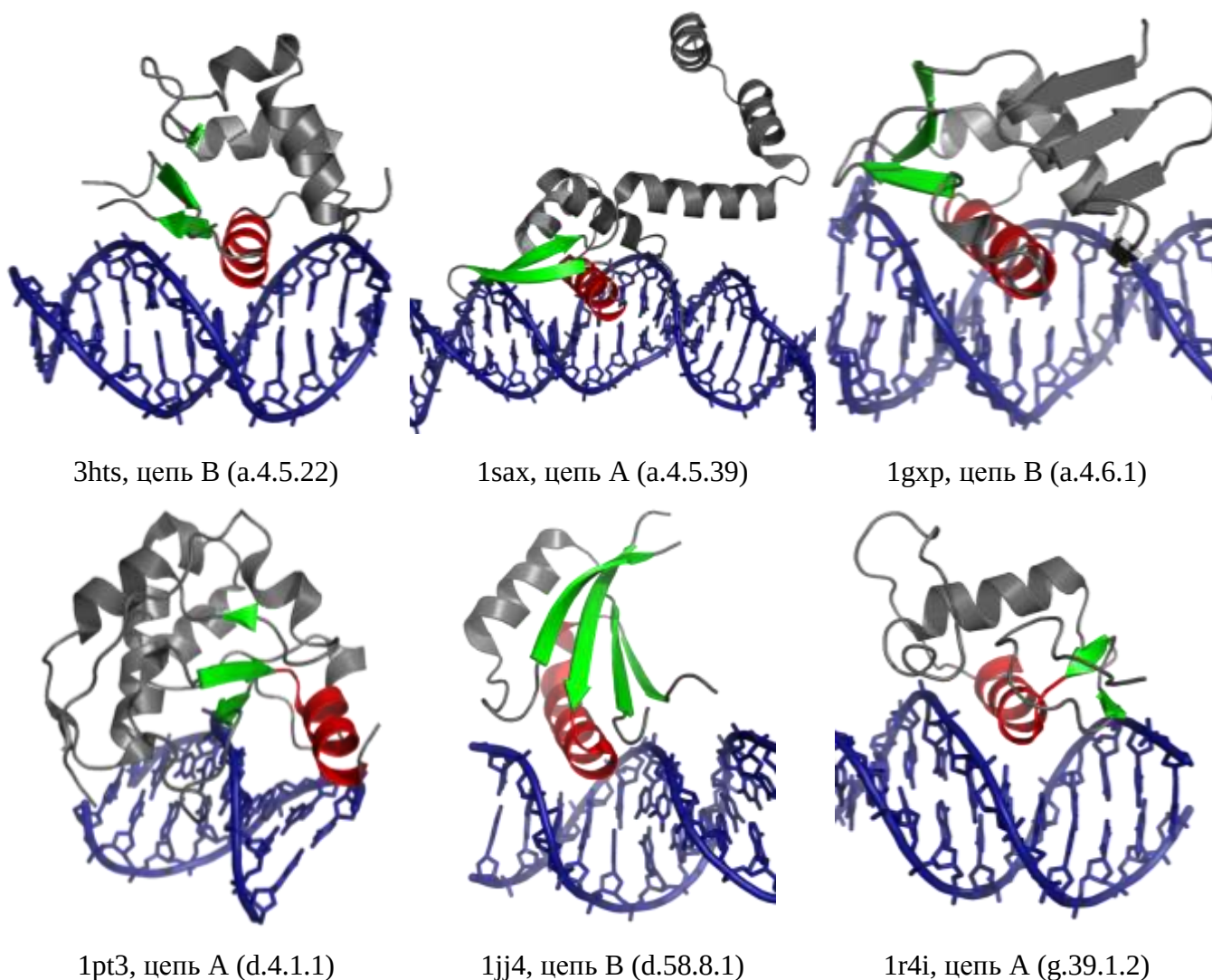


Рисунок 12. Примеры структур ДНК-белковых комплексов, относящихся к разным семействам SCOP, но узнающих ДНК одним и тем же способом ([H-Bb|H-Mj|S-Bb|L-Bb]). Контактирующие α -спираль и расположенный рядом β -лист выделены красным и зеленым цветами соответственно.

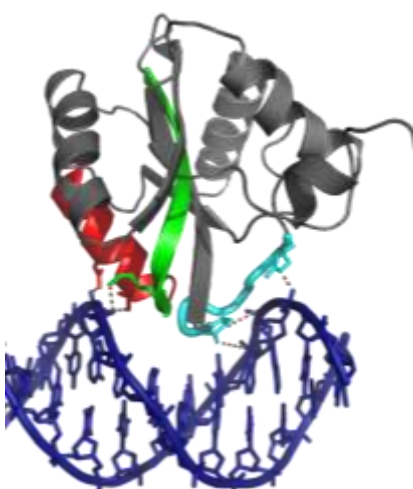
Например, при узнавании большой бороздки ДНК α -спиралью, а сахаро-фосфатного остова α -спиралью, β -листом и петлями ([H-Bb|H-Mj|S-Bb|L-Bb]) в

большинстве случаев можно выделить лежащую внутри большой бороздки α -спираль и расположенный рядом (иногда непосредственно соединенный со α -спиралью) β -лист, контактирующий с сахаро-фосфатным остовом. На рисунке 12 приведены примеры структур комплексов, для которых характерен данный способ взаимодействия. Отметим, что все комплексы принадлежат к разным семействам SCOP и заметно отличаются по укладке вне зоны контакта с ДНК.

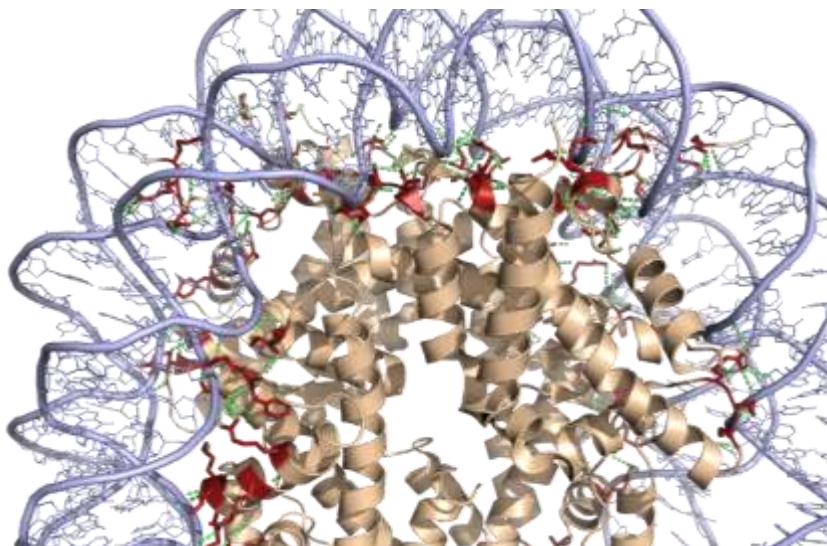
Еще одна закономерность относится к числу и характеру связей с сахаро-фосфатным остовом. Несмотря на то, что эти контакты традиционно относят к неспецифическим, это далеко не всегда означает, что количество подобных контактов мало (рисунок 13).

Максимальное число контактов с остовом формируется при взаимодействии α -спиралей и петель с остовом ДНК, β -листы же играют второстепенную роль. В силу заряженности сахаро-фосфатного остова, контакты в большинстве случаев образуются посредством водородных связей. Например, в структуре комплекса инициации репликации папилломавируса (1ksy, цепь В; семейство Replication initiation protein E1, d.89.1.2) контакт с ДНК образован двумя водородными связями α -спиралей, тремя водородными связями петель и только одной водородной связью β -листа.

Множественные разнородные контакты α -спиралей, β -листов и петель с остовом ДНК составляют основу ДНК-белкового взаимодействия для семейства Nucleosome core histones (a.22.1.1). Число контактов каждого типа варьирует внутри семейства от одного до шести, создавая обширную сеть водородных связей (3b6g).



1ksy, цепь В



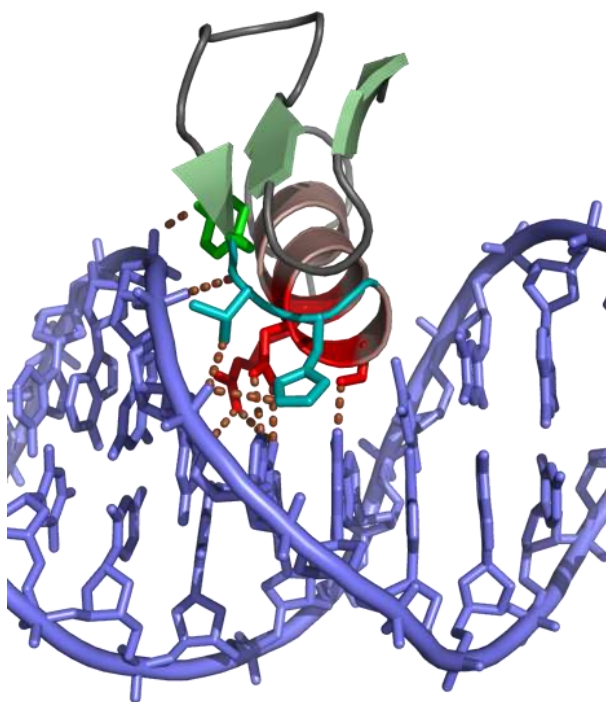
3b6g

Рисунок 13. Контакты белковых доменов с сахаро-фосфатным остовом ДНК в структурах комплекса инициации репликации папилломавируса (1ksy) и гистоновых белков (3b6g).

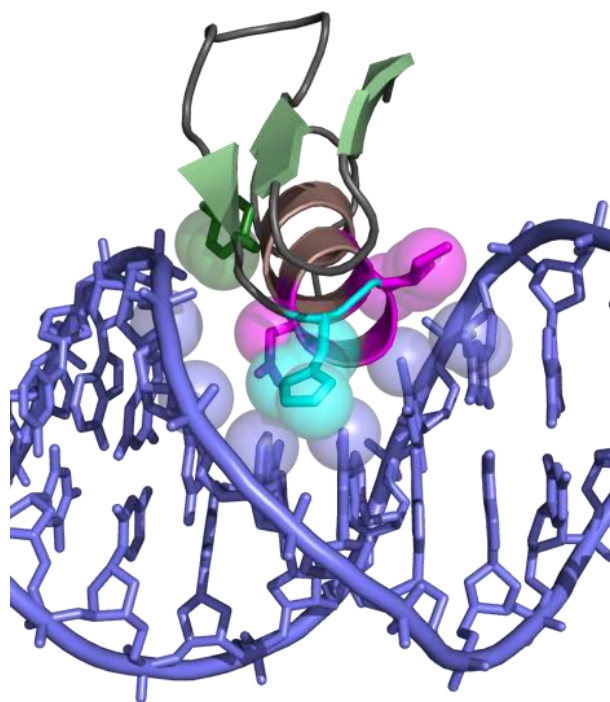
2.2.4 Вклад гидрофобных кластеров в ДНК-белковое узнавание

Традиционно при изучении межмолекулярных взаимодействий в первую очередь рассматривают водородные связи. Гидрофобные взаимодействия анализируются значительно реже, что оправдано далеко не во всех случаях. На основе анализа гидрофобных взаимодействий во всех рассмотренных в данной работе комплексов белковых доменов с ДНК был оценен вклад гидрофобных кластеров на формирование ДНК-белковых контактов.

В большинстве случаев гидрофобные кластеры повторяют контакт всех или части водородных связей. Например в структуре пептидов мотива «цинковый палец» (2drp, семейство Classic zinc finger, C2H2, g.37.1.1) и водородные связи и гидрофобные контакты формируют одни и те же типы контактов, давая в результате следующий способ взаимодействия: |Н-Bb|Н-Mj|S-Bb|L-Bb|L-Mj| (рисунок 14).



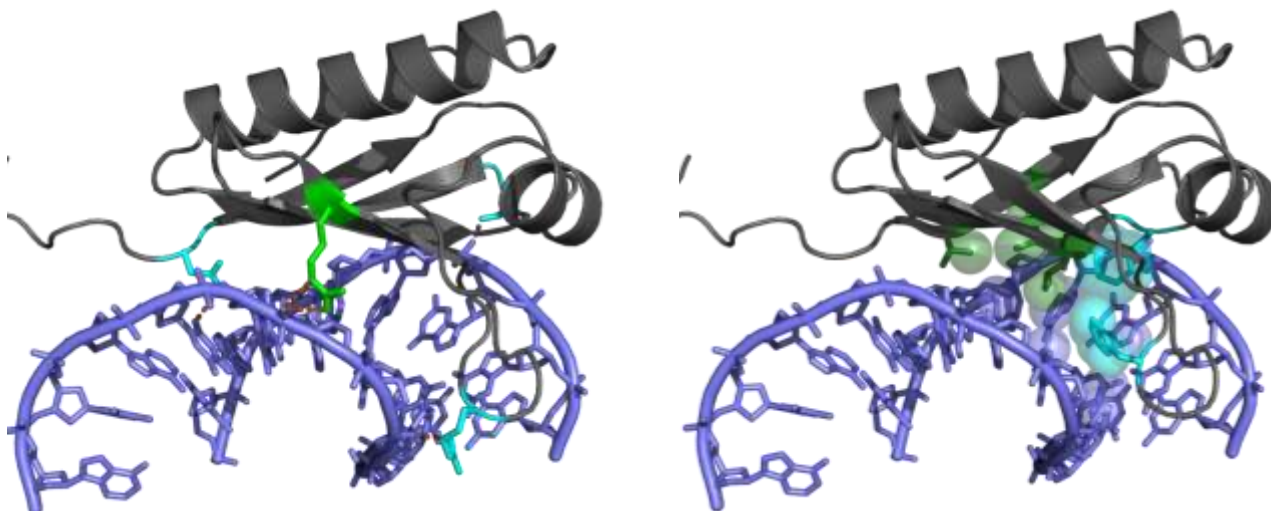
Водородные связи



Гидрофобные кластеры

Рисунок 14. Контакт пептида мотива «цинковый палец» с ДНК (структура 2drp, цепь A(103-139)). Как водородные связи, так и гидрофобные кластеры формируют одни и те же типы контактов: H-Bb, H-Mj, S-Bb, L-Bb, L-Mj.

В отдельных случаях гидрофобные кластеры дополняют водородные связи. Например в структуре ТАТА-box связывающего белка *A. Thaliana* (1qn5, цепь A(115-198)) гидрофобные контакты не только повторяют водородные связи (S-Bb, S-Mn, L-Bb), но и дополняют их контактом петли по малой бороздке (L-Mn). При этом в образовании водородных связей участвуют 4 аминокислотных остатка, а в образовании гидрофобных кластеров — 6 (рисунок 15).



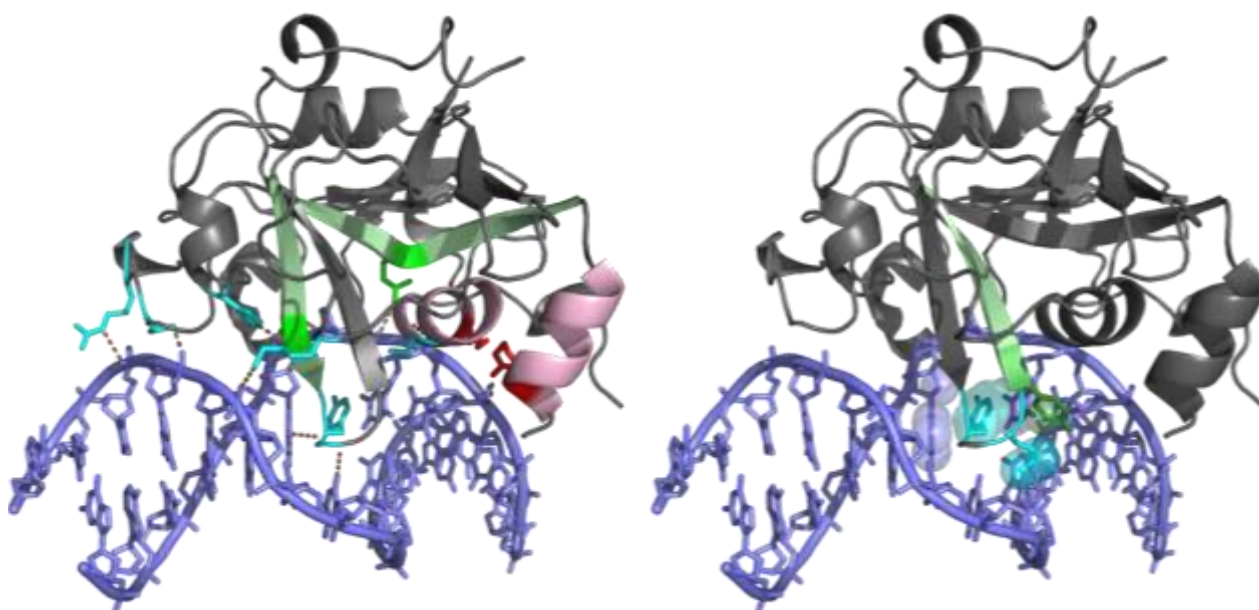
Водородные связи: S-Bb, S-Mn, L-Bb

Гидрофобные кластеры: S-Bb, S-Mn, L-Bb и L-Mn

Рисунок 15. Контакт ТАТА-бокс связывающего белка с ДНК (структура 1qn5, цепь A(115-198)). Гидрофобные взаимодействия не только повторяют типы контактов, образуемых водородными связями (S-Bb, S-Mn, L-Bb), но и дополняют их контактом петли с малой бороздкой (L-Mn).

Особенно важную роль играют гидрофобные кластеры, когда на место выщепленного основания встает ароматическая аминокислота, стабилизируя стекинг. Дополнительных водородных связей при этом не образуется, а вот гидрофобные контакты стабилизируют комплекс. Например, в структуре гликозилазы человека (1bnk, семейство 3-methyladenine DNA glycosylase, b.46.1.2) водородные связи образуются только с сахарофосфатным остовом (H-Bb, S-Bb, L-Bb), в то время как контакты с малой бороздкой (S-Mn, L-Mn) обеспечиваются посредством остатка тирозина (Tyr162), интеркалирующего в месте выщепленного азотистого основания (рисунок 16).

Несмотря на свой вклад в узнавание, гидрофобные кластеры никогда не являются основой контакта, а всегда лишь дополняют водородные связи. Структур, где количество аминокислотных остатков, образующих кластеры, значительно превышало бы количество аминокислотных остатков, образующих водородные связи, найдено не было.



Водородные связи: H-Bb, S-Bb, L-Bb

Гидрофобные кластеры: S-Mn, L-Mn и L-Bb

Рисунок 16. Контакт гликозилазы человека с ДНК (структура 1bnk, цепь A). Водородные связи образуются только с сахаро-фосфатным остовом ДНК, а контакт с малой бороздкой ДНК обеспечивают гидрофобные взаимодействия.

2.2.5 Закономерности ДНК-белкового узнавания внутри семейства SCOP

Для оценки консервативности и разнообразия контактов внутри семейства единицей рассмотрения должен быть один биологический домен белка, а не одна структура из конкретной записи PDB. Напомним, что для одного домена — участ-ка аминокислотной цепи белка с данным названием из данного организма — может быть расшифровано несколько структур.

Рассмотрим для примера семейство Restriction endonuclease HincII (с.52.1.19). Для него доступно шесть записей PDB относящихся к одному и тому же белковому домену Restriction endonuclease HincII из *H. influenzae*. В каждой из записей содержится от одной до четырёх белковых цепей. В итоге для этого се-мейства проанализировано 24 структуры. Все эти структуры высокоомологичны по выравниванию последовательностей, хорошо пространственно совмещаются и очень консервативны в образовании контактов с ДНК. В 21 случае наблюдается

сложный контакт вида $|H-Bb|S-Bb|S-Mj|L-Bb|L-Mj|L-Mn|$, в одном случае отсутствует контакт с малой бороздкой ($|H-Bb|S-Bb|S-Mj|L-Bb|L-Mj|$), в остальных двух, наоборот, есть дополнительный контакт α -спирали по малой бороздке ($|H-Bb|H-Mn|S-Bb|S-Mj|L-Bb|L-Mj|L-Mn|$).

Приведенный пример иллюстрирует общую закономерность: в большинстве случаев внутри семейств наблюдается высокий уровень сходства структур, а взаимное расположение белкового домена и ДНК в районе контакта сохраняется. Способ взаимодействия, тем не менее, может варьировать внутри семейства (рисунок 17).

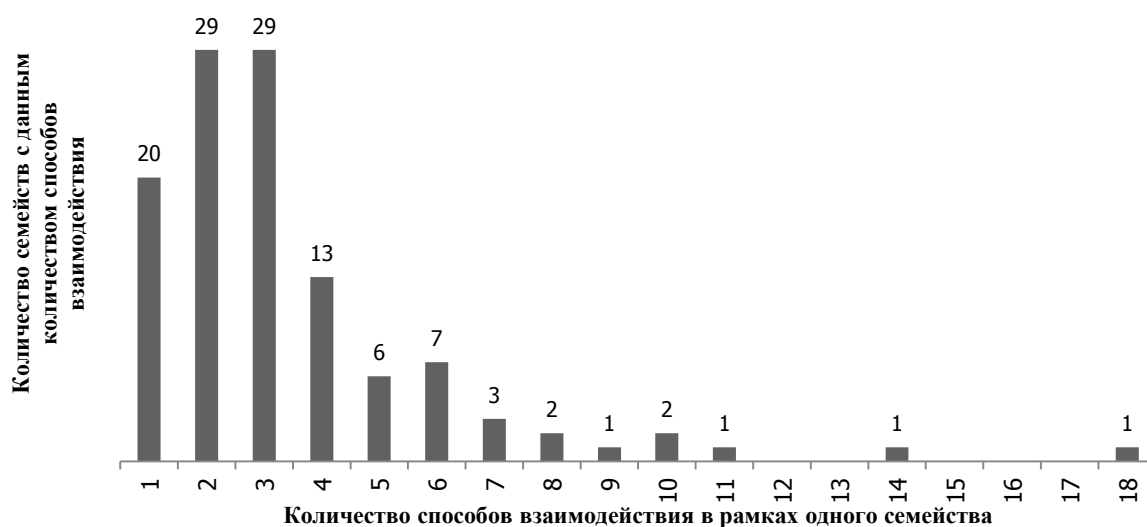


Рисунок 17. Гистограмма распределения количества способов взаимодействия в рамках одного семейства.

2.2.5.1 Вариации контактов внутри семейства

Наличие хорошего выравнивания последовательностей и/или пространственного совмещения структур ДНК-белковых комплексов далеко не всегда обеспечивает идентичность ДНК-белковых контактов. Это может быть обусловлено рядом причин:

- Подвижностью аминокислотных остатков белка
- Аминокислотной заменой в месте контакта
- разницей во взаимном расположении белка относительно ДНК

- различием в определении вторичной структуры белка

2.2.5.2 Вариации контактов внутри семейства, обусловленные подвижностью боковых цепей белка

В семействе эндонуклеаз рестрикции EcoRV (Restriction endonuclease EcoRV, с.52.1.2) в некоторых структурах (например 1bsu) кроме контакта β -листа и петли по большой бороздке ДНК встречается контакт петли по малой бороздке ДНК. Он возникает при небольшом изменении ориентации остатка аспарагина (Asn70, 1bsu, цепи А и В, соответственно, рисунок 18).

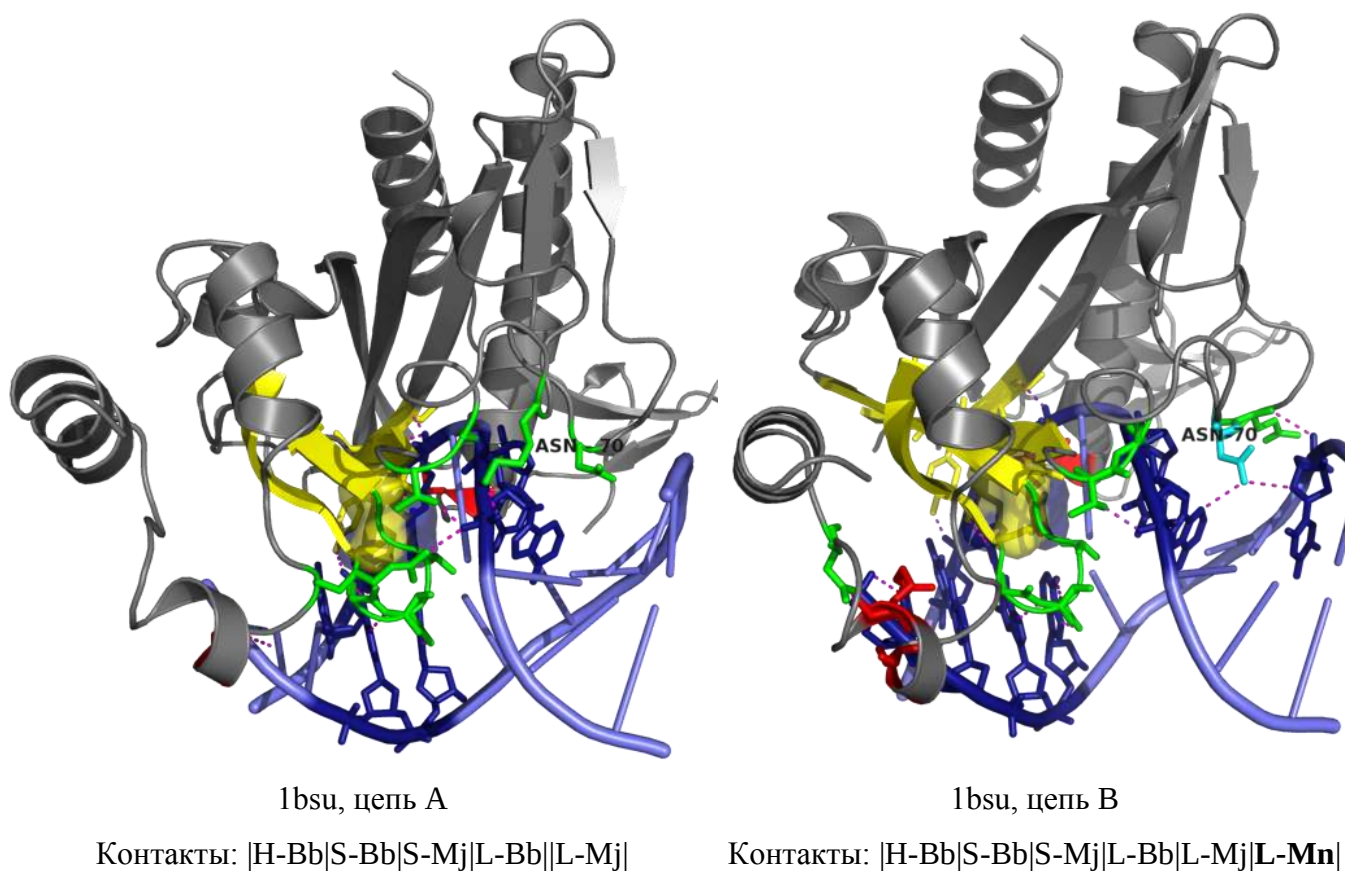


Рисунок 18. Контакт эндонуклеаз рестрикции EcoRV с ДНК (структура 1bsu, цепь А).

В семействе MarR-like transcriptional regulators (a.4.5.28), в димере OhrR-ohrA оператора сенной палочки (1z9c) контакты варьируют от одной цепи к другой не только по типу, но и по числу (рисунок 19). Так, в цепи D есть контакт петли с большой бороздкой и остовом (|H-Bb|H-Mj|S-Bb|L-Bb|L-Mj|L-Mn|), а в цепи С

его нет (H-Bb|H-Mj|S-Bb|L-Bb)). Наличие контакта петли с большой бороздкой (L-Mj) в цепи D объясняется разницей в определении вторичной структуры для серина (Ser68), входящего в состав α -спирали в цепи С. Вариация между контактом петли с сахаро-фосфатным остовом (L-Bb) в цепи С и с малой бороздкой (L-Mn) в цепи D объясняется подвижностью аргинина (Arg94).

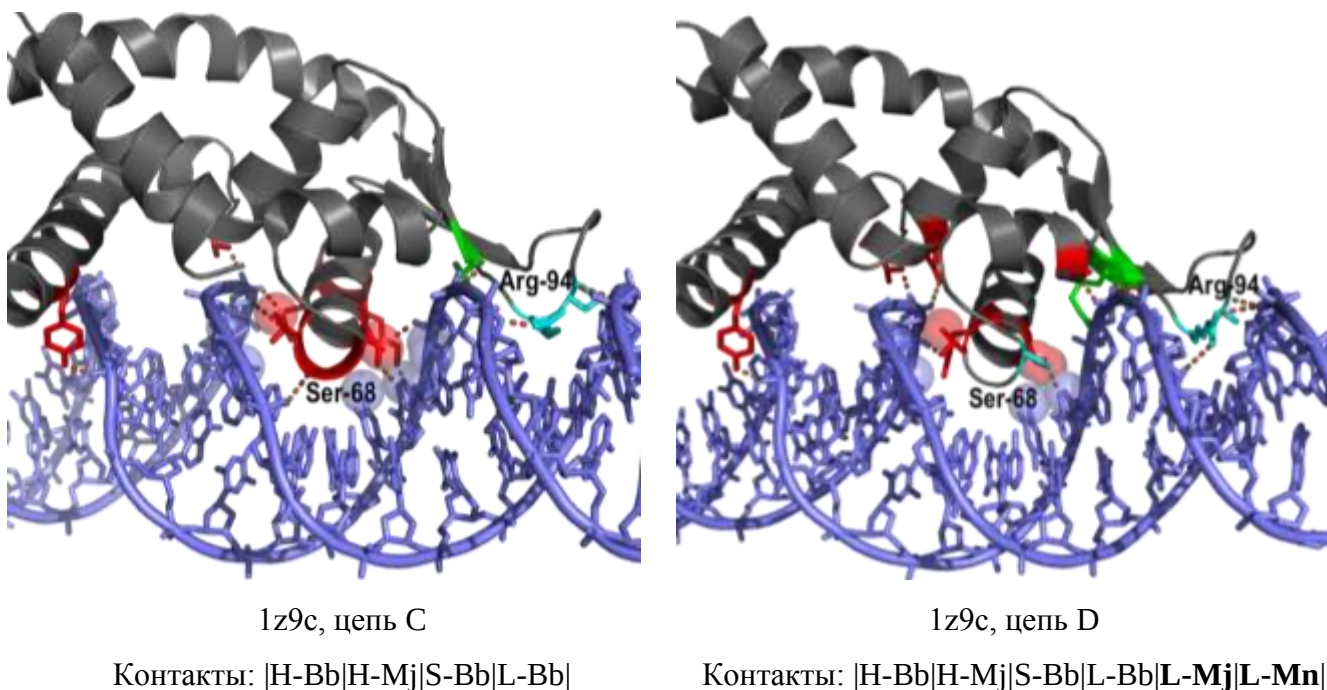


Рисунок 19. Контакт MagR-подобных транскрипционных регуляторов с ДНК (структура 1z9c, цепь С и D). В цепи D по сравнению с цепью С есть дополнительные контакты петель с большой и малой бороздками посредством Ser68 и Arg94 соответственно.

2.2.5.3 Вариации контактов внутри семейства, обусловленные различиями во вторичной структуре белка

Короткие фрагменты α -спиралей и β -листов могут быть распознаны как неструктурированные участки даже при небольших отклонениях в геометрии укладки. Такие вариации могут встречаться как в разных структурах, так и внутри одной и той же записи PDB. Например в семействе Arc/Mnt-like phage repressors (a.43.1.1) встречаются структуры (1bdt, 1par) в которых укладка N-концевого участка (остатки 4-6) белка, контактирующего с большой бороздкой ДНК, определя-

ется программой Stride то как часть 3-10-спирали (1bdt), то как неструктурированный конец (1par). При этом взаимодействие β -листа с большой бороздкой ДНК сохраняется для всех структур комплексов (рисунок 20).

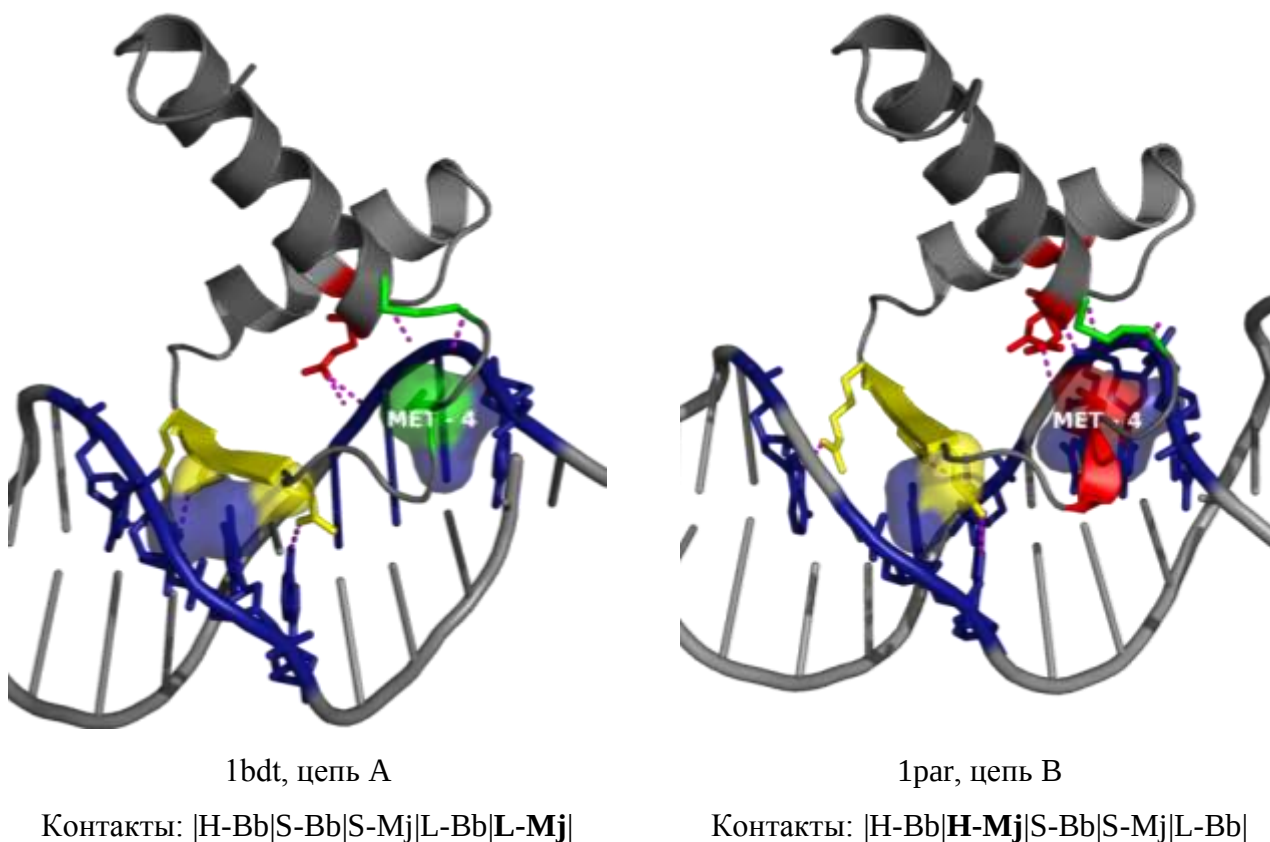


Рисунок 20. Контакт Arc/Mnt-подобных фаговых репрессоров с ДНК. Разница в определении вторичной структуры на участке 4-6 обуславливает разницу в способе взаимодействия белкового домена с ДНК. В структуре 1bdt, цепи A участок 4-6 считается неструктурированным, в то время как в структуре 1par, цепь B этот же изгиб определяется как фрагмент спирали 3-10.

2.2.5.4 Вариации контактов внутри семейства, обусловленные аминокислотной заменой в месте контакта.

В семействе регуляторного фактора интерферона (Interferon regulatory factor, а.4.5.23) в некоторых структурах (например, 2pi0, цепь B) кроме контакта α -спирали по большой бороздке ДНК встречается контакт петли по малой бороздке ДНК. Он возникает при замене аланина (Ala42) на лейцин (Leu42) и достаточном для образования гидрофобного кластера приближении β -поворота белка к атомам

малой бороздки ДНК. Наличие или отсутствие контакта β -листа с сахаро-фосфатным остовом ДНК в данном семействе зависит от вариаций определения вторичной структуры на участке 104-109 (рисунок 21).

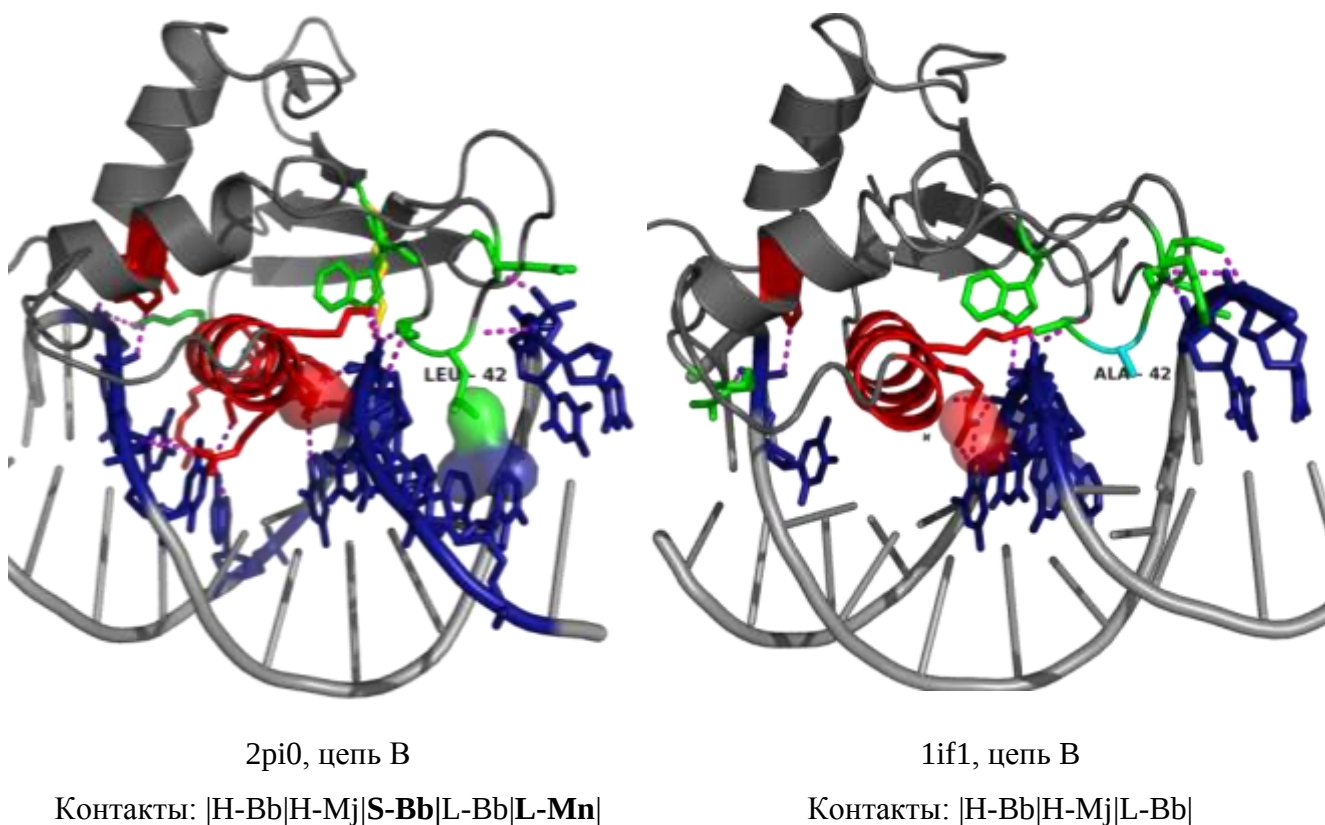


Рисунок 21. Контакт регуляторного фактора интерферона с ДНК посредством α -спирали, β -листа и петель (структура 2pi0) и α -спирали и петель (структура 1if1). Замена лейцина на аланин (Leu42 $\rightarrow$ Ala42) и разница в определении вторичной структуры участка β -листа/петли приводит к исчезновению контактов петли с малой бороздкой и β -листа с сахаро-фосфатным остовом в структуре 1if1(B).

2.2.5.5 Вариации контактов внутри семейства, обусловленные разницей во взаимном расположении белка относительно ДНК.

Одной из причин различия контактов белкового домена с ДНК внутри семейства может стать изменение во взаимном расположении белкового домена относительно ДНК. Такая разница в расположении может быть обусловлена соседними доменами белка, а также находящимися рядом белками.

Например, в семействе Replication initiation protein (a.4.5.10) в структуре инициатора репликации 1rep два домена в цепи С следуют один за другим (остатки 15-143 и 144-246). Положение первого домена влияет на положение второго на ДНК, приводя к сдвигу второго относительно ДНК (рисунок 22). В результате, в отличие от второго, первый домен контактирует β -листом с остовом и петлей с малой бороздкой (|H-Bb|H-Mj|S-Bb|L-Bb|L-Mn|). Второй же домен вместо этого имеет контакт петли с большой бороздкой (|H-Bb|H-Mj|L-Bb|L-Mj|) соответственно.



1rep, цепь С (15-143)

Контакты: |H-Bb|H-Mj|S-Bb|L-Bb|L-Mn|

1rep, цепь С (144-246)

Контакты: |H-Bb|H-Mj|L-Bb|L-Mj|

Рисунок 22. Контакты N- и С-концевого доменов инициатора репликации 1rep с ДНК. Положение первого домена на ДНК влияет на положение второго домена относительно ДНК. В результате способы взаимодействия у двух доменов различаются.

Наличие нескольких способов взаимодействия внутри одного и того же семейства является довольно типичным. Среди 115 проанализированных семейств только 20 представлены единственным способом взаимодействия (рисунок 17), для них способ взаимодействия для каждой структуры и класс взаимодействия, характерный для семейства в целом, совпадают. Для остальных семейств была разработана процедура определения класса взаимодействия.

2.2.6 Классификация семейств ДНК-контактирующих доменов SCOP

Для построения классификации из 115 семейств структурных доменов SCOP были отобраны наиболее полно представленные семейства. Семейство может быть классифицировано, если внутри него расшифрованы пространственные структуры трех и более различных белковых доменов. Достаточное количество структурных данных присутствовало в 34 семействах структурных доменов SCOP. Появление новых структурных данных позволит классифицировать и другие семейства, а также дополнить существующую классификацию.

Напомним, что в процессе классификации из наблюдаемых внутри семейства типов контактов с ДНК выделялись те, которые встречались хотя бы один раз для всех белковых доменов данного семейства. Эти универсальные для всех представителей типы контакта с ДНК определяли класс взаимодействия семейства в целом. В результате все классифицируемые семейства были разделены на 17 классов взаимодействия (таблица 1 и рисунок 43). В данной работе классы взаимодействия выделены курсивом, при записи класса сначала указывается контакт с сахаро-фосфатным остовом, после, через вертикальную черту — контакт с большой и малой бороздками ДНК.

Из таблицы 1 видно, что по сравнению с количеством способов взаимодействия отдельных комплексов (97 шт) классов взаимодействия семейств наблюдается значительно меньше — всего 17. Это связано с тем, что класс взаимодействия представляет собой список не всех встретившихся в доменах семейства типов контакта, а только тех, которые характерны для всех без исключения белковых доменов семейства. Как отмечалось ранее, внутри одного семейства может наблюдаться несколько способов узнавания ДНК. Эти способы узнавания могут быть составлены из всех или из части типов контактов, представленных в классе взаимодействия, а также из других редких типов контакта, не вошедших в класс взаимодействия.

Например, внутри семейства GalR/LacI-подобный бактериальный регулятор (а.35.1.5), представленного тремя белками и 35 записями PDB, наблюдаются 7 способов узнавания ДНК:

1. |H – Bb | L – Bb | H – Mj|
2. |H – Bb | L – Bb | H – Mj | H – Mn|
3. |H – Bb | L – Bb | H – Mj | L – Mj|
4. |H – Bb | L – Bb | L – Mj|
5. |H – Bb | L – Bb | H – Mj | L – Mj | H – Mn|
6. |L – Bb | H – Mn |
7. |L – Bb|

Такое разнообразие обусловлено как отсутствием у части структур С-концевой α -спирали, взаимодействующей с малой бороздкой ДНК, так и артефактами кристаллизации ДНК-белковых комплексов. Несмотря на это, для каждого из трех белков семейства есть хотя бы одна структура, содержащая в себе все наблюдаемые в семействе виды взаимодействия. В результате семейство в целом имеет класс взаимодействия $(H – Bb), (L – Bb) | (H – Mj), (L – Mj) | (H – Mn)$.

Класс взаимодействия	Названия семейств (идентификатор SCOP)
<i>Разнородные контакты</i>	Middle domain of MutM-like DNA repair proteins (a.156.1.2) NF-kappa-B/REL/DORSAL transcription factors, C-terminal domain (b.1.18.1) Classic zinc finger, C2H2 (g.37.1.1) C-terminal, Zn-finger domain of MutM-like DNA repair proteins (g.39.1.8)
(H – Bb)	AraC type transcriptional activator (a.4.1.8) Transcription factor IIB (TFIIB), core domain (a.74.1.2) Leucine zipper domain (h.1.3.1) №2
(H – Bb), (L – Bb)	Restriction endonuclease FokI, N-terminal (recognition) domain (a.4.5.12)
(L – Bb)	DnaQ-like 3'-5' exonuclease (c.55.3.5)
(S – Bb), (L – Bb)	N-terminal domain of MutM-like DNA repair proteins (b.113.1.1)
(H – Bb), (S – Bb), (L – Bb)	Nucleosome core histones (a.22.1.1)
(H – Bb) (H – Mj)	HLH, helix-loop-helix DNA-binding domain (a.38.1.1) Leucine zipper domain (h.1.3.1) №1
(H – Bb), (L – Bb) (H – Mj)	POU-specific domain (a.35.1.1) Phage repressors (a.35.1.2)* Homeodomain* (a.4.1.1) Myb/SANT domain (a.4.1.3) Replication initiation protein (a.4.5.10) ets domain (a.4.5.21) Interferon regulatory factor (a.4.5.23) Nuclear receptor (g.39.1.2)
(H – Bb), (S – Bb) (H – Mj)	Viral DNA-binding domain (d.58.8.1)
(H – Bb), (L – Bb) (L – Mj)	Rel/Dorsal transcription factors, DNA-binding domain (b.2.5.3)*
(H – Bb), (L – Bb) (H – Mj), (L – Mj)	Zn2/Cys6 DNA-binding domain (g.38.1.1) Zinc finger design (k.12.1.1)
(H – Bb), (S – Bb), (L – Bb) (S – Mj), (L – Mj)	Group I mobile intron endonuclease (d.95.2.1)
(H – Bb), (L – Bb) (H – Mn)	HMG-box* (a.21.1.1)
(H – Bb), (S – Bb), (L – Bb) (L – Mn)	Prokaryotic DNA-bending protein (a.55.1.1)
(H – Bb), (L – Bb) (H – Mj) (L – Mn)	Recombinase DNA-binding domain (a.4.1.2) Paired domain (a.4.1.5) SRF-like (d.88.1.1) Lambda integrase-like, catalytic core (d.163.1.1)
(S – Bb), (L – Bb) (S – Mn), (L – Mn)	TATA-box binding protein (TBP), C-terminal domain (d.129.1.1)
(H – Bb), (L – Bb) (H – Mj), (L – Mj) (H – Mn)	GalR/LacI-like bacterial regulator (a.35.1.5)

Таблица 1. Классы взаимодействия семейств белковых доменов SCOP с ДНК. Звездочками помечены семейства, в которых класс взаимодействия может измениться при появлении новых структур. **(H** - α -спираль, **S** - β -лист, **L** - петля/неструктурированный участок белка, **Mj** - большая бороздка ДНК, **Mn** - малая бороздка ДНК, **Bb** – сахаро-фосфатный остов ДНК).

2.2.6.1 Семейства, для которых появление новых структур может повлиять на класс взаимодействия

Формальность процедуры классификации накладывает ограничения на результаты классификации. Например, отмеченные звездочкой (*) в таблице 1 классы взаимодействия могут быть расширены дополнительными типами контакта при появлении новых пространственных структур.

Рассмотрим семейство Z-DNA binding domain (a.4.5.19), которое содержит три белковых домена, образующих контакты трёх типов: H-Bb, H-Mj, и иногда L-Bb. Первый домен контактирует посредством α -спирали: |H-Bb|H-Mj|. Третий домен может контактировать всеми тремя типами контактов в различных вариациях. Вторым белковым доменом, представленным единственной структурой 1j75 (цепь A), контактирует только посредством α -спирали по сахаро-фосфатному остову (H-Bb). Таким образом, общим контактом для семейства оказывается взаимодействие α -спирали с сахаро-фосфатным остовом (H-Bb). Возможно, при появлении новых структур для второго домена, список типов контакта, характерных для семейства, может расшириться взаимодействием α -спирали с большой бороздкой до |H-Bb|H-Mj|.

Семейство Homeodomain (a.4.1.1) представлено 19 белковыми доменами, 17 из которых содержат структуры со следующим способом взаимодействия: |H-Bb|H-Mj|L-Bb|L-Mn|. Поскольку у двух белковых доменов ни в одной структуре нет петли, контактирующей с малой бороздкой, то все семейство получило класс $(H - Bb), (L - Bb) | (H - Mj)$. В данном случае появление новых структур сможет повлиять на класс взаимодействия для семейства, только если в этих структурах будет петля или концевой участок достаточной для взаимодействия с малой бороздкой длины.

2.2.6.2 Семейства без общего класса взаимодействия

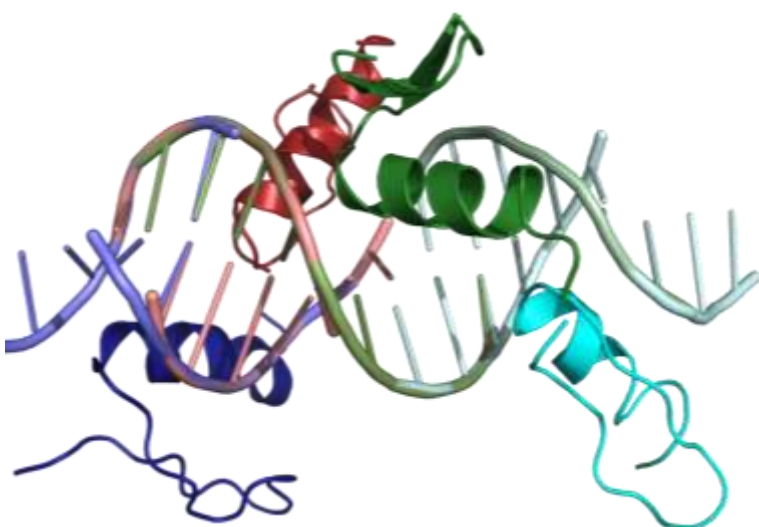
Четыре семейства не имеют однозначного класса взаимодействия и объединены в класс “*Разнородные контакты*”. Это связано либо с тем, что взаимное расположение белка и ДНК немного отличается внутри семейства, либо с небольшими вариациями контактов при сходном взаимном расположении белка и ДНК.

Первый случай можно наглядно продемонстрировать на примере семейства Classic zinc finger, C2H2 (g.37.1.1). В семействе наблюдаются следующие типы контактов: (H – Bb), (H – Mj), (S – Bb), (L – Bb), (L – Mj), но ни один из них не является общим для всех белковых доменов. Это связано с тем, что последовательно идущие в цепи домены цинковых пальцев расположены на ДНК не одинаково. Сдвиг положения белкового домена относительно ДНК обуславливает изменение характера контактов.

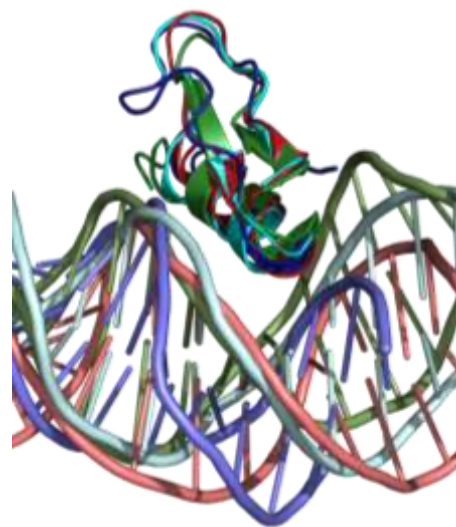
Например в структуре 2gli 4 цинковых пальца идут друг за другом, немного смещаясь относительно ДНК. Совмещение по белковому домену показывает разницу в положении белка на ДНК (рисунок 23). Такая разница в положении доменов относительно ДНК влияет на способ взаимодействия:

- Для первого домена (A:135-167): |H-Bb|H-Mj|L-Bb|L-Mj|
- Для второго домена (A:168-197): |L-Bb|
- Для третьего домена (A:198-228): |H-Bb|H-Mj|S-Bb|
- Для четвертого домена (A:229-257): |H-Bb|H-Mj|S-Bb|L-Mj|

Общего типа контактов у всех четырех доменов не наблюдается не только в данной структуре, но и во всех других структурах данного семейства. Поэтому данное семейство отнесено к классу “разнородные контакты”.



Четыре идущих подряд домена Classic zinc finger, C2H2



Совмещение этих доменов

Рисунок 23. Четыре идущих подряд домена Classic zinc finger, C2H2 (структура 2gli) расположены со сдвигом относительно ДНК и не имеют консервативного типа контакта.

Второй случай возникает в семействе g.39.1.8 — C-terminal, Zn-finger domain of MutM-like DNA repair proteins. Хорошее пространственное совмещение остовов ДНК-белковых комплексов в данном случае оказывается не достаточным для идентичности узнавания ДНК всеми представителями данного семейства (рисунок 24).

С ДНК контактирует β -шпилька, наполовину расположенная в большой бороздке ДНК, а наполовину и находящаяся рядом с остовом ДНК. В зависимости от структуры, контакты могут находиться как на участке шпильки внутри бороздки, так и на участке возле сахаро-фосфатного остова. Шпилька состоит из участка β -листа, соединенного петлей. Каждый из этих структурных элементов шпильки может контактировать с ДНК. В данном семействе встречается четырех различных способа взаимодействия:

- |S-Bb|L-Bb| (структура 1l2c, цепь A:235-274)
- |S-Bb| (структура 1kfv, цепь A:227-271)
- |L-Bb|L-Mj| (структура 1k82, цепь A:225-268)
- |S-Mj|L-Bb| (структура 2oq4, цепь A:224-262)

Как видно из данного списка, общего типа контакта для всех четырех доменов данного семейства нет, поэтому семейство отнесено к классу “разнородные контакты”.

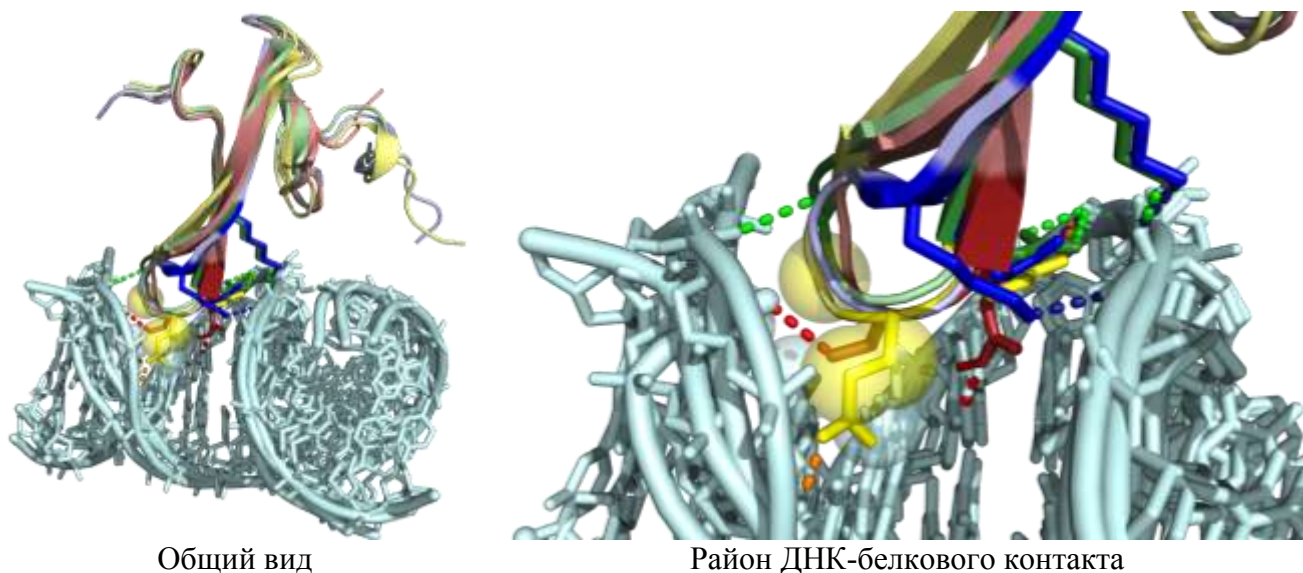
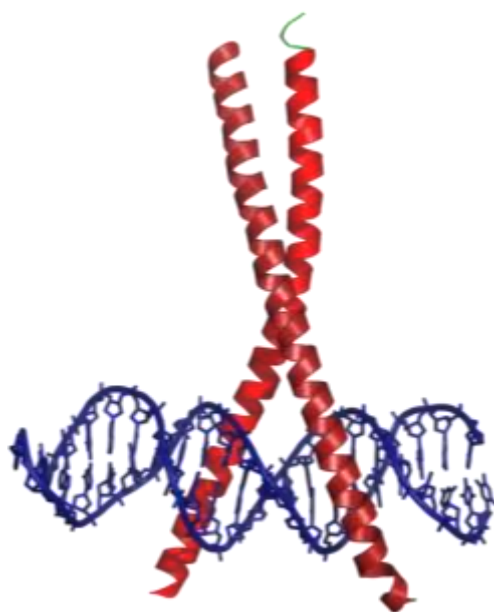


Рисунок 24. Пространственное совмещение комплексов белковых доменов семейства “C-terminal, Zn-finger domain of MutM-like DNA repair proteins” с ДНК (структура 1l2c, цепь A:235-274 – зеленая; структура 1kfv, цепь A:227-271 – синяя; структура 1k82, цепь A:225-268 – красная; структура 2oq4, цепь A:224-262 – желтая).

2.2.6.3 Случай разделения семейства на два подсемейства

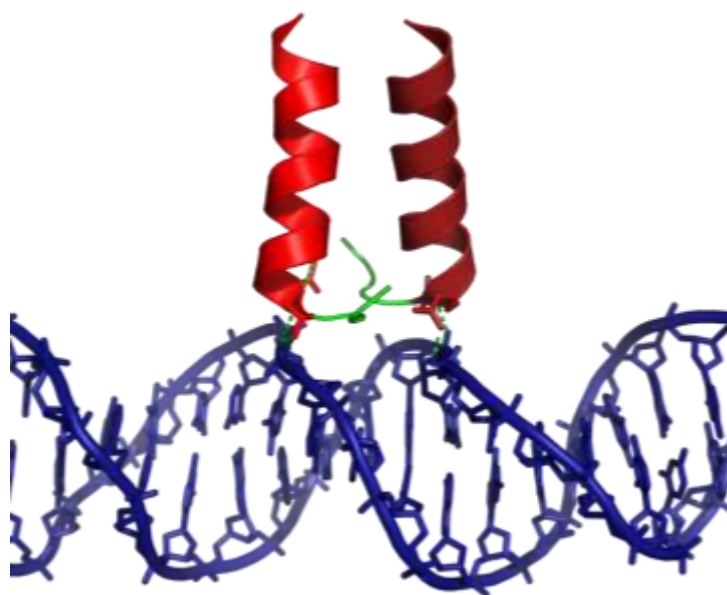
Отдельного внимания заслуживает семейство Leucine zipper domain (h.1.2.1), которое было разделено на два подсемейства, поскольку можно выделить 2 принципиально разных способа узнавания этими белковыми доменами ДНК.

Первый способ узнавания характерен для большинства белковых доменов семейства и представляет собой классическую лейциновую молнию, контактирующую α -спиралью с сахаро-фосфатным остовом и большой бороздкой (класс взаимодействия $(H - Vb) | (H - Mj)$) (рисунок 25a).



(a) 1nwg

Контакты: |H-Bb|H-Mj|



(b) 1d66

Контакты: |H-Bb|L-Bb|

Рисунок 25. Семейство Leucine zipper domain (h.1.2.1). (a) Контакт с остовом и большой бороздкой ДНК посредством α -спирали (1nwg), (b) контакт с остовом посредством α -спирали (1d66).

Второй способ взаимодействия характерен только для трех белковых доменов дрожжей (Baker's yeast), а именно: CD2-Gal4, PUT3 и HAP1. Классической лейциновой молнии тут не наблюдается, но поскольку геометрия белковых доменов аналогична, они были отнесены в то же семейство SCOP (рисунок 25b). Контакта α -спирали с большой бороздкой в этих случаях нет, но возможны четыре точечных контакта с остовом и малой бороздкой типа: H-Bb, L-Bb, S-Bb, S-Mn. Поскольку контакт спирали с остовом встречается во всех трех доменах, данное подсемейство отнесено к классу “H-Bb”.

2.2.6.4 Закономерности строения семейств белковых доменов, относящихся к одному и тому же классу взаимодействия

Большинство классов взаимодействия представлено одним-двумя семействами. Только три класса, а именно $(H - Bb), (L - Bb) | (H - Mj); (H - Bb); (L - Bb) | (H - Mj) | (L - Mn)$ и класс с разнородными контактами представлены большим количеством семейств. Рассмотрим наличие общих черт узнавания ДНК для всех этих

классов, кроме класса разнородных контактов. Последний, очевидно, общих мотивов не содержит.

Наиболее широко представленный класс взаимодействия — это узнавание большой бороздки α -спиралью, подкрепленное контактами α -спирали и петель с сахаро-фосфатным остовом ((*H – Bb*), (*L – Bb*) | (*H – Mj*)) (рисунок 26).

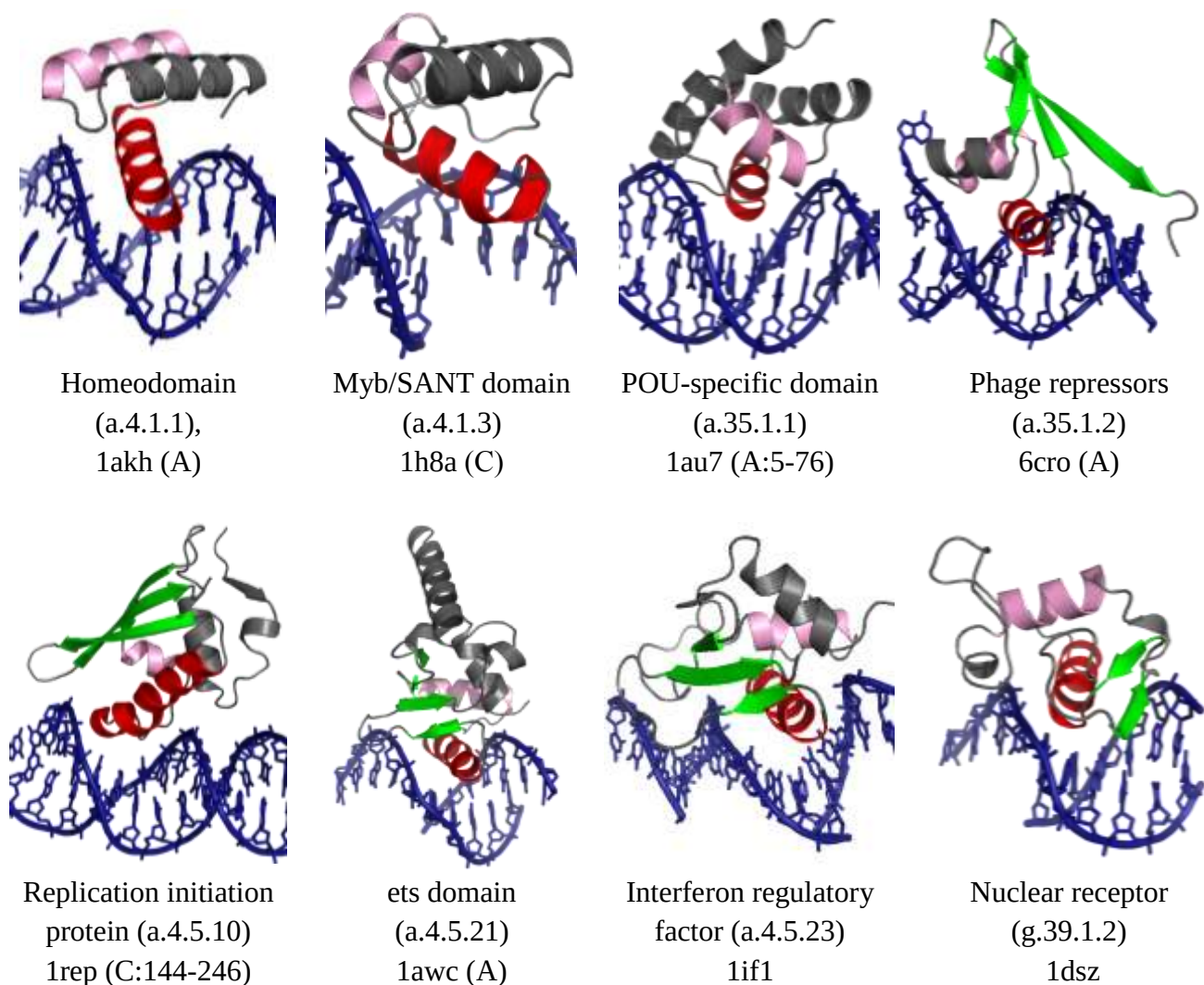


Рисунок 26. Семейства, относящиеся к классу взаимодействия (*H – Bb*), (*L – Bb*) | (*H – Mj*). Контактующая с большой бороздкой ДНК α -спираль показана красным, расположенная с ней рядом α -спираль – розовым, а β -лист – зеленым цветом.

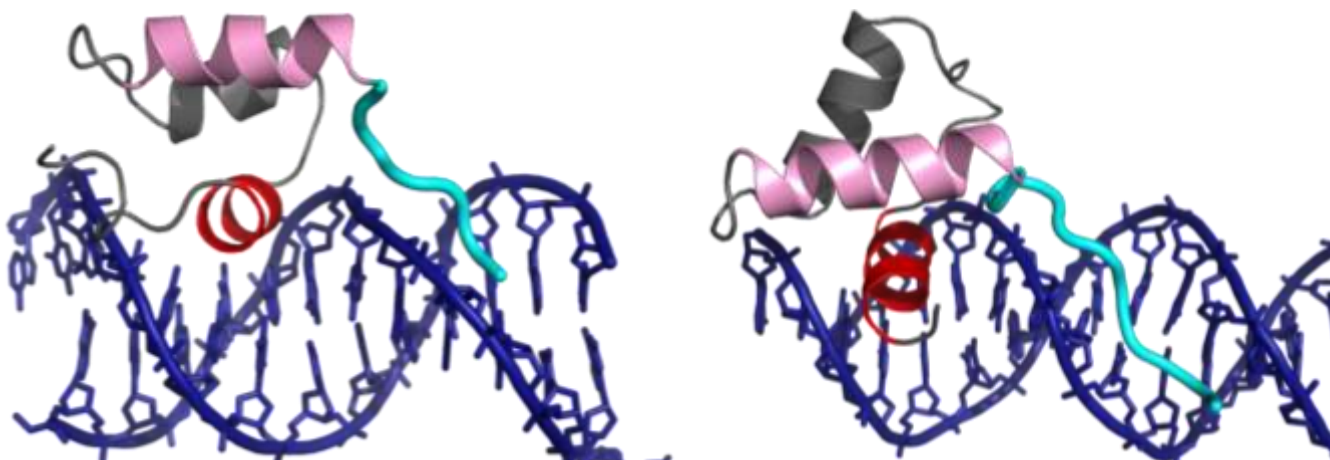
Этот класс взаимодействия характерен для восьми семейств, а именно: POU-specific domain (a.35.1.1), Phage repressors (a.35.1.2), Homeodomain (a.4.1.1), Myb/SANT domain (a.4.1.3), Replication initiation protein (a.4.5.10), ets domain

(a.4.5.21), Interferon regulatory factor (a.4.5.23), Nuclear receptor (g.39.1.2). Это в основном α -спиральные белки различной архитектуры, относящиеся к укладке lambda repressor-like DNA-binding domains, DNA/RNA-binding 3-helical bundle и Glucocorticoid receptor-like (DNA-binding domain) соответственно. В этом классе взаимодействия α -спираль расположена в большой бороздке либо параллельно либо под углом, так что только конец α -спирали взаимодействует с ДНК (красного цвета на рисунке 26). Рядом с ней расположена еще одна α -спираль (розового цвета на рисунке 26).

В семействах, относящихся к укладке DNA/RNA-binding 3-helical bundle (a.4. *. *) а также в семействе Nuclear receptor (g.39.1.2) дополнительно можно выделить β -лист (зеленого цвета на рисунке 26) расположенный рядом с узнающей большую бороздку α -спиралью.

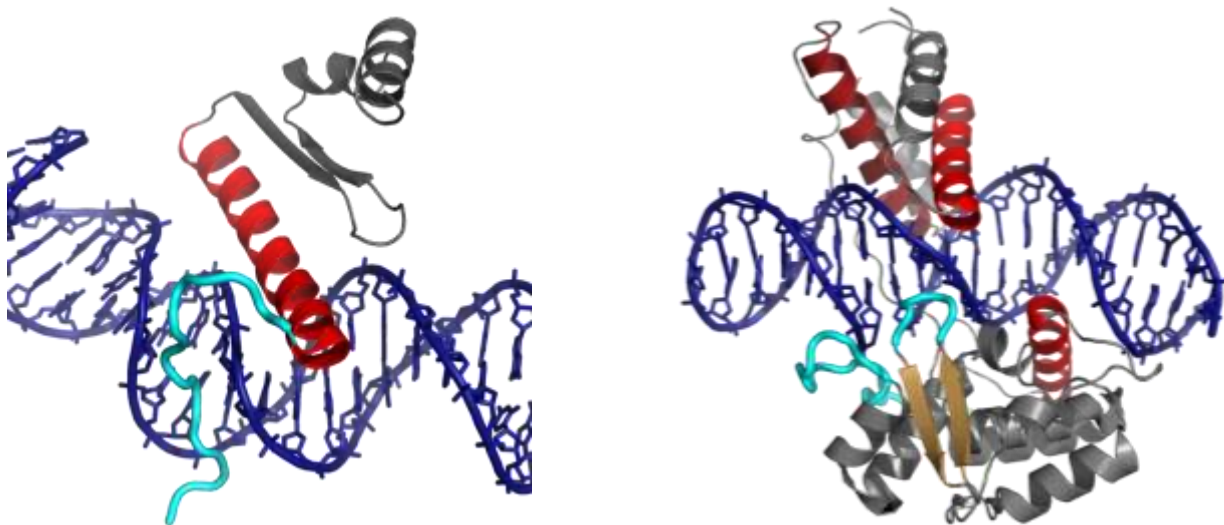
Выделить общий узнающий мотив получается далеко не для каждого класса взаимодействия. Например в классе взаимодействия ($H - Bb$), ($L - Bb$) | ($H - Mj$) | ($L - Mn$), представленном четырьмя семействами, только два семейства имеют сходство узнавания. Это Recombinase DNA-binding domain (a.4.1.2) (например структура 1jj8, цепь C) и Paired domain (a.4.1.5) (структура 1k78, цепь A:82-142), относящиеся к одному суперсемейству Homeodomain-like. Для них характерна узнающая α -спираль (красная) и практически перпендикулярная, лежащая выше α -спираль (светло розовая), соединенная с концевым участком (голубой), узнающим малую бороздку (рисунок 27).

Для семейств SRF-like (d.88.1.1) (структура 1k6o, цепь B) и Lambda integrase-like, catalytic core (d.163.1.1) (структура 1p7d, цепь A) сходных узнающих мотивов выявить не удалось (рисунок 28).



Recombinase DNA-binding domain (a.4.1.2) 1jj8 (C) Paired domain (a.4.1.5) 1k78 (A:82-142)

Рисунок 27. Представители суперсемейства Homeodomain-like, относящиеся к классу взаимодействия $(H - Bb)$, $(L - Bb) | (H - Mj) | (L - Mn)$. Контактирующая с большой бороздкой ДНК α -спираль показана красным, расположенная с ней рядом α -спираль – розовым, а контактирующий с малой бороздкой N-конец – голубым цветом.



SRF-like (d.88.1.1) 1k6o (B) Lambda integrase-like, catalytic core (d.163.1.1) 1p7d (A)

Рисунок 28. Семейства, относящиеся к классу взаимодействия $(H - Bb)$, $(L - Bb) | (H - Mj) | (L - Mn)$ и не имеющие сходства в узнавании ДНК. В семействе SRF-like (d.88.1.1) взаимодействующая с большой бороздкой α -спираль (красная) следует сразу за концевым участком (голубой), узнающим малую бороздку. В семействе Lambda integrase-like, catalytic core (d.163.1.1) в узнавании большой бороздки участвует три α -спирали (красные), а малую бороздку узнает петля из β -листа (голубая).

Глава 3. Поиск консервативных молекул воды в белковых и ДНК-белковых комплексах.

Пространственные структуры записей PDB с разрешением лучше 2.5 Å обычно содержат координаты молекул воды. Молекулы воды, иммобилизованные на поверхности белков или на интерфейсе белков и/или нуклеиновых кислот играют важную роль в обеспечении взаимодействия и стабилизации макромолекулярных комплексов. Наличие пространственных структур гомологичных белков позволяет производить поиск молекул воды, расположенных в одних и тех же местах и образующих водородные связи с соответствующими аминокислотными остатками, нуклеотидами, лигандами и ионами. Такие молекулы воды называются *консервативными*. Можно предположить, что консервативные молекулы воды соответствуют сайтам гидратации макромолекул в растворе. Такие сайты содержат молекулы воды относительно продолжительное время. Это согласуется с данными ЯМР [28] и компьютерными симуляциями методом молекулярной динамики [19].

Возможность поиска консервативных молекул воды на поверхностях и интерфейсах макромолекул имеет ряд практических приложений. Например, если на поверхности макромолекулы имеется сайт связывания лиганда с консервативной молекулой воды в нём, то можно улучшить дизайн нового лиганда, вводя в него полярную группу, способную связываться с консервативной молекулой воды или даже заменять ее в полости сайта связывания.

Для изучения роли консервативных молекул воды может быть проведен направленный мутагенез остатков, контактирующих с такими молекулами воды.

Для поиска консервативных молекул воды в данной работе был использован полуавтоматический метод поиска кластеров молекул воды в пространственно совмещенных гомологичных структурах, основанный на использовании программы wLake [32].

3.1 Материалы и методы

3.1.1 Программа wLake

3.1.1.1 Определения

Базовым понятием, используемым в алгоритме wLake является *кластер молекул воды*. Каждый кластер состоит из молекул воды, принадлежащих разным структурам (не более одной молекулы воды от каждой структуры), которые занимают практически одно и то же положение в пространственно совмещенных структурах. Молекулы воды из одного кластера называются *консервативными молекулами воды*.

Две молекулы воды из разных совмещенных структур являются *близкими*, если расстояния между центрами их атомов кислорода меньше заданного порога (1.5 Å по умолчанию). Таким образом, в кластер попадают попарно близкие молекулы воды из разных структур.

3.1.1.2 Алгоритм

В качестве входящих данных для алгоритма wLake используется файл с координатами пространственно совмещенных структур соответствующих белков или макромолекулярных комплексов.

Цель алгоритма — поиск максимальных (не расширяемых) кластеров, размером не менее заданного числа (*Min*) молекул воды. На первой стадии строится граф G , представляющий набор молекул воды. Каждая вершина графа соответствует единичной молекуле воды из некоторой структуры, и наоборот, каждой молекуле воды из некоторой структуры соответствует одна вершина графа G . Две вершины соединяются ребром, если соответствующие молекулы воды являются близкими и принадлежат разным структурам. Таким образом, максимальные (не расширяемые) кластеры молекул воды соответствуют максимальным кликам в

графе (кликкой называют подмножество вершин графа, таких, что любые две вершины подмножества соединены ребром).

Нижеописанный алгоритм используется для поиска всех максимальных клик из не менее чем Min вершин. Алгоритм рекурсивно проходит по графам из стека, причем изначально в стеке находится граф G . На каждом шаге (1) выбирается и убирается граф G' из стека. Если G' является кликой, то возвращается G' . Если число вершин в G' меньше Min , или число ребер в G' меньше, чем число ребер в клике размера Min , то тогда G' отбрасывается. Иначе (2) выбирается вершина из G' с минимальной валентностью и (3) в стек помещаются два графа: первый, G'' является копией подграфа, включая выбранную вершину и всех её соседей; второй, G''' , получается из G' путем исключения из него выбранной вершины. Перебор прекращается, когда стек оказывается пустым. На выходе алгоритм выдает все максимальные клики размером не менее Min . На шаге (3) каждая максимальная клика из G' представлена или в G'' или в G''' или в них обоих.

Несмотря на экспоненциальную сложность алгоритма, он является эффективным для графов G , представляющих набор молекул воды из нескольких сотен совмещенных структур. Данный алгоритм был реализован в программе wLake, доступной в виде веб-сервиса, а так же под Linux и Windows по адресу: <http://mouse.belozersky.msu.ru/~evgeniy/cgi-bin/wLake/wLake.php>

На вход wLake принимает файл в формате PDB, содержащий пространственное совмещение гомологичных структур. В качестве параметров указывается (1) порог расстояния между парой близких молекул воды (1.5 Å по умолчанию) и (2) минимальное (Min) число молекул воды в кластере (по умолчанию равно трем). В качестве результата программа выдает список кластеров и скрипт для их визуализации в программе RasMol [95].

3.2 Результаты: применение wLake для анализа макромолекулярных комплексов

В структурах многих макромолекулярных комплексов присутствуют молекулы воды. В связи с техническими особенностями метода рентгеноструктурного анализа (РСА) надежность координат молекул воды ниже, чем у атомов белка. В большинстве рентгеновских структур с разрешением $1.5^\circ\text{Å} - 2.5^\circ\text{Å}$ присутствуют молекулы воды, не связанные водородными связями ни с какими другими молекулами [119]. Эти молекулы могут как реально присутствовать в структуре, так и являться артефактами метода.

Мы предполагаем, что найденные на основе анализа гомологичных структур консервативные молекулы воды, связанные водородными связями с другими молекулами, являются более надежными и функционально значимыми, чем молекулы воды, описанные в единичных структурах. В приведенных ниже примерах проведен поиск консервативных молекул воды в гомологичных структурах и анализ их функциональной значимости.

3.2.1 Поиск консервативных молекул воды в структурах транскетолаз

Транскетолазы — это тимидин дифосфатзависимые ферменты, катализирующие перенос гидроксиацетильной группы с кетозы на альдозу. Для каталитической активности транскетолазам необходим тиаминдифосфат (ThDP) и ион магния (Mg^{2+}) или другого двухвалентного металла с октаэдрической координацией (Ca^{2+} , Mn^{2+} , Co^{2+}). В основном эти ферменты являются гомодимерами с молекулярной массой субъединиц $70 \div 74 \text{ kDa}$, состоят из 3 доменов α/β типа и имеют 2 активных центра [120]. Первый, N-концевой домен образован центральным параллельным β -листом, состоящим из 5 тяжей, и окружающими его α -спиралями. Средний (пиримидиновый) домен по структуре очень похож на N-концевой и состоит из 6 тяжелого β -листа, окруженного α -спиралями. Оба эти домена участву-

ют в связывании кофактора, причем каждая молекула тимидиндифосфата взаимодействует с N-концевым доменом одной субъединицы и средним доменом второй субъединицы. С-концевой домен представлен смешанным β -листом, состоящим из пяти тяжей, и α -спиралями, находящимися вокруг него. Этот домен не участвует в каталитической активности белка и, вероятно, несет регуляторную функцию [121] (рисунок 29).

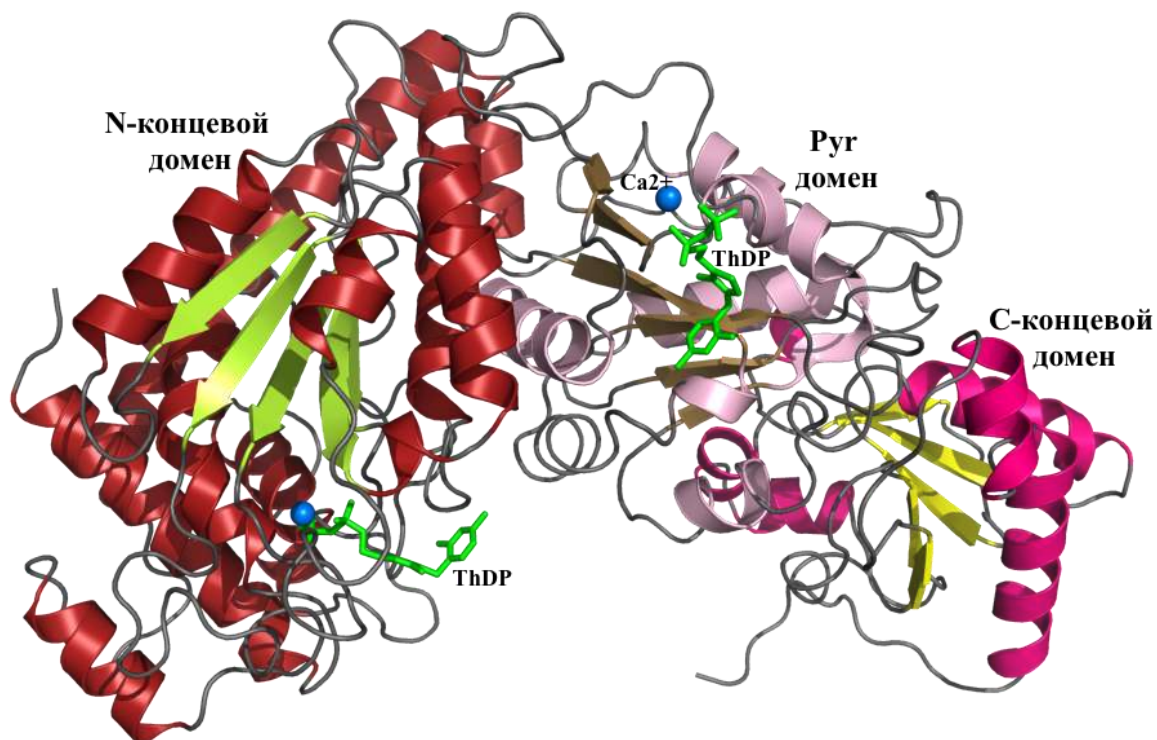


Рисунок 29. Структура мономера транскетолазы дрожжей (1trk, цепь А).

Образование димера происходит за счет взаимодействия между N-концевыми и средними доменами двух субъединиц. Образованная общая поверхность имеет впадины и туннели вследствие неплотного характера контактов между субъединицами. Часть этих впадин заполнено кластерами молекул воды.

Во всех организмах, где проводили поиск транскетолаз, они были обнаружены. На данный момент расшифрованы 34 структуры транскетолаз, принадлежащих 14 организмам: *H. sapiens*, *S. cerevisiae*, *C. jejuni*, *B. anthracis*, *E. coli*, *L. salivarius* UCC118, *P. aeruginosa*, *M. tuberculosis*, *B. pseudomallei*, *B. thailandensis*, *F. tularensis*, *T. thermophilus* HB8 TT0505, *L. mexicana* и *Z. mays*. Каждая структура

содержит в себе от одной до четырех субъединиц (всего 61 структура субъединиц из различных организмов). Поскольку транскетолазы являются гомодимерами, то все структуры были разбиты на субъединицы и пространственно совмещены с помощью сервиса PDBFold (SSM). Несмотря на идентичность только порядка 10% аминокислотных остатков, пространственное совмещение доменов между собой весьма хорошее: RMSD составляет 1.46 Å, при этом совмещаются 500 C-альфа атомов из 613-694.

Используя wLake, был найден 221 кластер консервативных молекул воды во всех пространственно совмещенных субъединицах (порог на расстояние между молекулами воды был выбран равным 1.0 Å, минимальное число молекул воды в кластере — 20). Основная часть кластеров (более 60%) сосредоточены в N-концевом домене, заполняя полости на поверхности и в районе активного центра фермента.

Изначально описанные [121] в активном центре транскетолазы восемь молекул воды были проверены на консервативность, а схема взаимодействий в активном сайте фермента была модифицирована с учетом дополнительных молекул воды (рисунок 30).

Всего в районе активного центра были выделены 11 консервативных молекул воды (рисунок 30). Все они образуют водородные связи либо с белком, либо с кофактором, либо с ионом кальция, формируя водные мостки между ними и поддерживая конформацию активного центра белка. Семь из этих консервативных молекул воды (#1-7, показаны в овалах на рисунке 30) ранее были выделены как структурно значимые для транскетолазы дрожжей [121]. Только одна из описанных ранее молекул воды (#8) оказалась неконсервативной, что может быть связано с наличием только одной водородной связи с ней. Дополнительные четыре консервативные молекулы воды (#9-12), описанные в данной работе, находятся не только на периферии, но и в центре сайта связывания лиганда. Таким образом, на

данном примере видно, что поиск консервативных молекул воды в родственных структурах помогает более точно определять функционально значимые молекулы в гидратной оболочке белка.

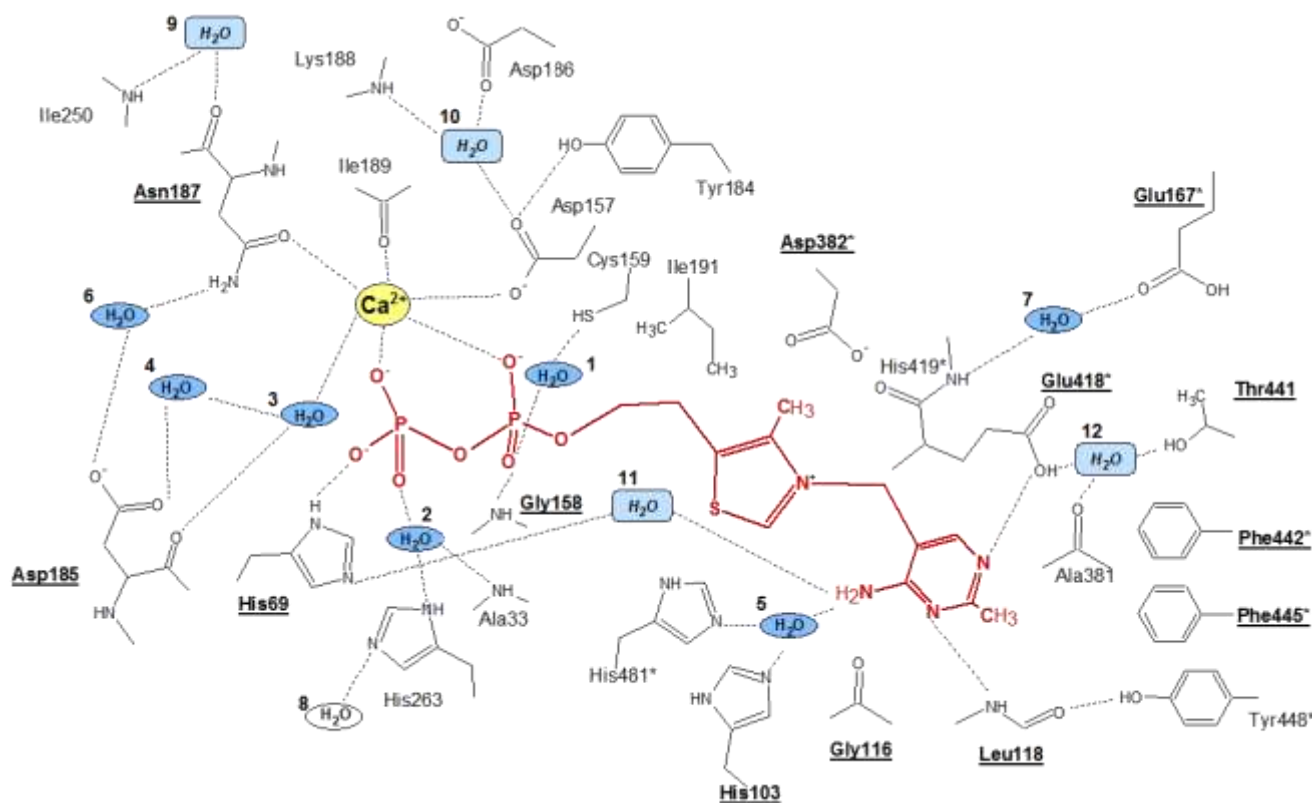


Рисунок 30. Схема взаимодействия тиаминдифосфата, иона кальция и молекул воды в активном центре транскетолазы дрожжей (1trk) (адаптированный рисунок на основе рис. 16 из работы Никкола со соавт. [121]). Консервативные остатки транскетолазы подчеркнуты, остатки из второй субъединицы отмечены звездочкой. Консервативные молекулы воды отмечены голубым. Молекулы воды, отраженные в работе Никкола показаны в овалах, описанные в данной работе — в прямоугольниках.

3.2.2 Консервативные молекулы воды в семействах SCOP ДНК-связывающих доменов

На страницах базы данных NPIDB подробно охарактеризовано 76 семейств ДНК-связывающих доменов SCOP. В частности, для 72 из них предоставлены данные о кластерах консервативных молекул воды на ДНК-белковом интерфейсе. Их можно загрузить в виде файлов с расширениями .flat.txt и .pivot.txt, а также ис-

пользовать визуализацию консервативных молекул воды, указывая необходимый кластер под окном структур Jmol.

Количество кластеров для исследованных семейств колеблется от 1 до 365, и, в первую очередь, зависит от количества молекул воды в каждой из совмещенных структур семейства и не коррелирует с количеством совмещенных структур. В случае большого количества кластеров молекул воды, как например в семействе Lambda integrase-like, catalytic core (d.163.1.1), где найдено 365 кластеров, можно видеть, что многие кластеры находятся рядом, формируя обширные зоны гидратации. В этих зонах гидратации в каждой отдельной структуре могут находиться несколько консервативных молекул воды (рисунок 31).

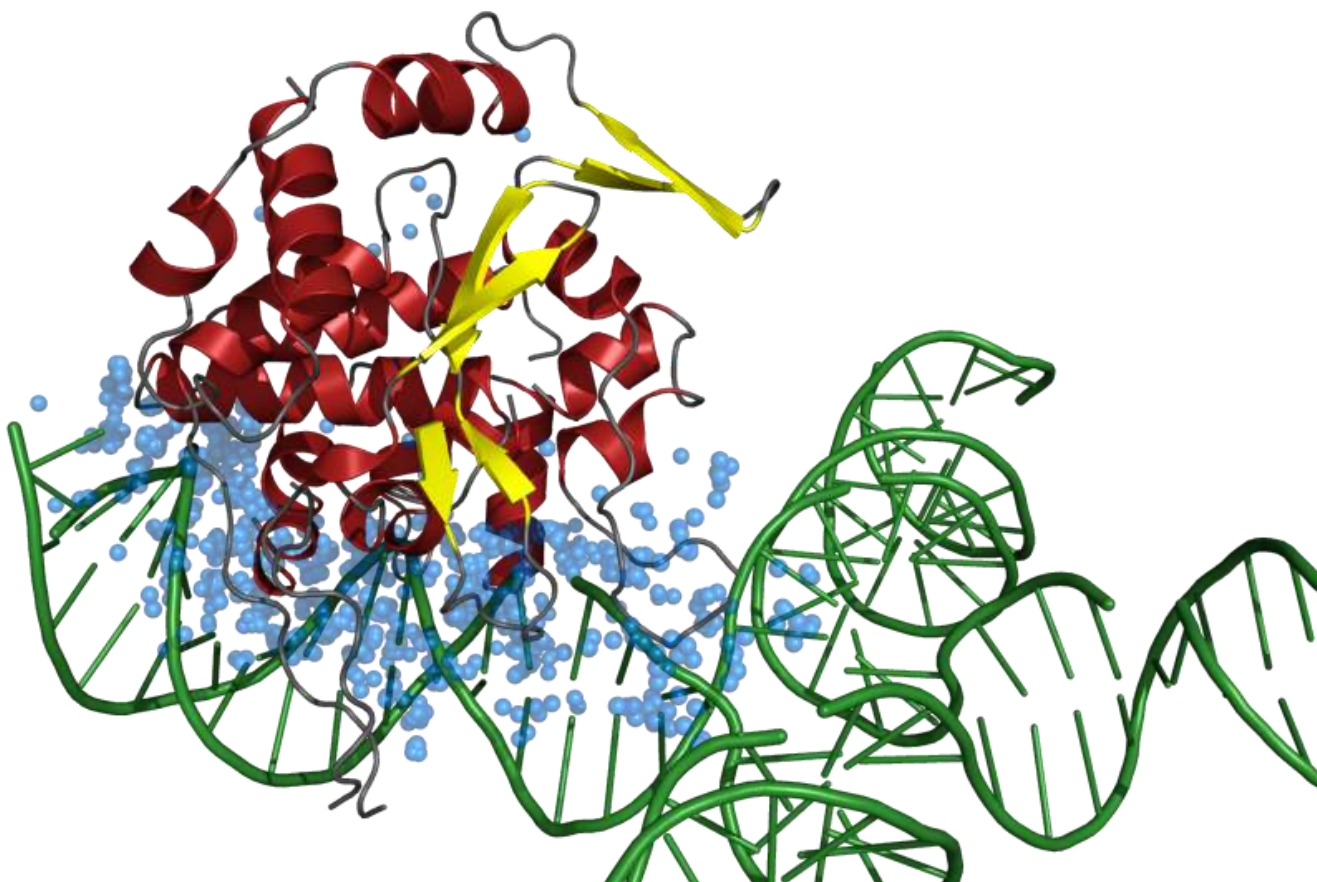


Рисунок 31. Кластеры консервативных молекул воды в семействе Lambda integrase-like, catalytic core (d.163.1.1) (структура 1flo(A)).

На достоверность найденных кластеров консервативных молекул воды влияют следующие факторы:

- количество молекул воды в каждой из наложенных структур,
- количество наложенных структур (не менее трех),
- артефакты определения положения молекул воды в пространственной структуре.

Таким образом, увеличение количества и качества исходных совмещаемых структур позволит уточнить текущие данные о положении консервативных молекул воды в рассмотренных семействах SCOP.

Глава 4. База данных NPIDB.

База данных взаимодействий нуклеиновых кислот с белками NPIDB (Nucleic acids — Protein Interaction DataBase) содержит информацию о всех доступных структурах ДНК-белковых и РНК-белковых комплексах. На данный момент в банке данных PDB доступно порядка 5500 структур комплексов нуклеиновых кислот с белками. Они автоматически извлекаются в NPIDB, после чего расширенная информация о комплексах появляется на соответствующих страницах базы. В NPIDB включает в себя ряд сервисов для анализа взаимодействий белка с нуклеиновыми кислотами. Среди них: поиск водородных связей, потенциальных водных мостиков, гидрофобных кластеров, определяемых программой CluD, а также визуализация комплексов и НК-белковых взаимодействий с помощью Jmol. Для анализа доступны как исходные PDB структуры, так и биологические единицы.

Поиск по базе данных доступен с использованием идентификаторов PDB, Pfam, SCOP, Fuzznuc [122], BLAST, а также терминов GO (рисунок 32).

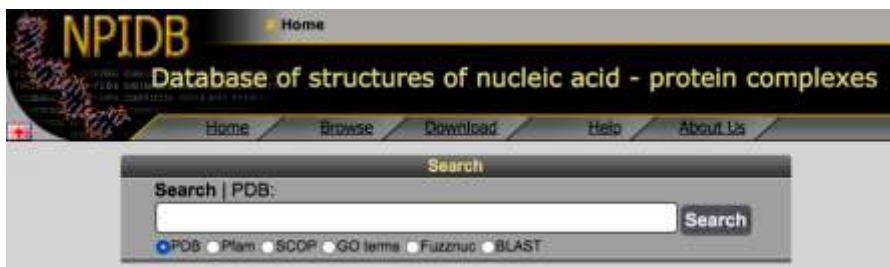


Рисунок 32. Строка поиска по базе данных NPIDB.

Для каждой *цепи белка* в НК-белковых комплексах определена вторичная структура, используя программу Stride, и домены согласно классификациям Pfam и SCOP. Каждый из доменов Pfam/SCOP может быть загружен на локальный компьютер пользователя, а также визуализирован средствами Jmol. Домены SCOP из ДНК-белковых комплексов также охарактеризованы с точки зрения *способа их взаимодействия с ДНК*.

Для каждой *записи* в NPIDB доступна следующая информация и сервисы: файл структуры (загрузка и визуализация), ссылки на на другие базы данных, субъединицы и входящие в состав белка домены SCOP и Pfam (загрузка и визуализация), термины GO для белка, загрузка и визуализация последовательностей белка и нуклеиновой кислоты с отмеченными на них контактами (рисунок 33).

NPIDB Home List of complexes 1TGH structure

Page of complex: 1TGH

Home Browse Download Help About Us

General information

PDB ID: 1TGH
 Title: TATA BINDING PROTEIN (TBP)/RNA COMPLEX
 PDB header: TRANSCRIPTION DNA
 Date: 1996-02-13
 Experimental method: X-RAY DIFFRACTION
 Resolution (Å): 2.0
 Release date:
 Organism: HOMO SAPIENS

Download complex View in Jmol N/A-protein interaction

Quick Links

PDB PDB-e SRS MMDb Jena
 OCA PDBWiki CATH PDBsum HSSP
 PQS DALI Proteopedia BIPA WebPDA
 EDS

Composition of PDB entry and biounits

File	Protein chains	DNA chains	RNA chains	No. of models	Action
pdb1tgh.pdb	A	BC		1	
1tgh.pdb	A	BC		1	

Pfam domains

Name	Type	E-value	Significance	Clan	Prot. chain	Start res.	End res.	Nuc. chain	Action
TBP	Domain	1.5e-32		1 C03407	A	248	332	BC	
TBP	Domain	2e-32		1 C03407	A	157	241	BC	

SCOP domains

Classification	Protein	Prot. chain	Start res.	End res.	Nuc. chain	Int. mode	Action
Class: Alpha and beta proteins (A-B)							
Fold: TBP-like							
Superfamily: TATA-box binding protein-like							
Family: TATA-box binding protein (TBP), C-terminal domain							
Class: Alpha and beta proteins (A-B)							
Fold: TBP-like							
Superfamily: TATA-box binding protein-like							
Family: TATA-box binding protein (TBP), C-terminal domain							

GO terms

Protein chains	GO type	GO description
A=155-339	C	female pronucleus
A=155-339	C	transcription factor TFIID complex
A=155-339	C	nuclear euchromatin
A=155-339	C	male pronucleus
A=155-339	C	cytoplasm
A=155-339	C	transcription factor TFIID complex
A=155-339	P	transcription regulatory region DNA binding
A=155-339	P	sequence-specific DNA binding transcription factor activity
A=155-339	P	spermatogenesis
A=155-339	P	transcription from RNA polymerase II promoter
A=155-339	P	cell death
A=155-339	P	viral reproduction
A=155-339	P	transcription elongation from RNA polymerase II promoter
A=155-339	P	positive regulation of transcription, DNA-dependent
A=155-339	P	transcription initiation from RNA polymerase II promoter
A=155-339	P	viral-host interaction

Sequences

Download file with secondary structure created by Stride

1tgh.seq1.acb: [Download sequences in FASTA format]

B (chr1) Show Click "Show" for more sequences
 C (chr1) Show Click "Show" for more sequences
 A (constant) Show Click "Show" for more sequences

© NPIDB team 2003 - 2016

Рисунок 33. Страница комплекса 1tgh базе данных NPIDB.

[illegible]

88

Для семейств доменов Pfam создана отдельная страница со списком всех имеющихся в NPIDB доменов Pfam (рисунок 35). Для каждого домена приведен его идентификатор (ID), название, число записей в NPIDB, содержащих данный домен, количество данных доменов в NPIDB и число связанных с нуклеиновой кислотой доменов. Из 1037 представленных в базе семейств Pfam, 191 не контактируют с НК. Тем не менее они встречаются в составе НК-контактирующих белков.

NPIDB Home Pfam families

List of available Pfam families

Home Browse Download Help About Us

Download

• List of Pfam domains:

Pfam release : 27.0
Pfam-A families : 14830
Date : 2011-11
Based on UniProtKB : 2012_06

Best resolution complexes of Pfam family representatives, bound with nucleic acid

List of Pfam families detected in NPIDB entries:

Pfam ID ↓	Domain	# of entries	# of domains	# of bound domains
2OG-Fell_Oxy_2	2OG-Fe(II) oxygenase superfamily	43	45	44
4HBT	Thioesterase superfamily	1	4	0
5_3_exonuc	5'-3' exonuclease, C-terminal SAM fold	3	3	2
5_3_exonuc_N	5'-3' exonuclease, N-terminal resolvase-like domain	4	4	1
7kD_DNA_binding	7kD DNA-binding domain	19	24	22
AAA	ATPase family associated with various cellular activities (AAA)	3	16	16
AAA_11	AAA domain	3	4	4
AAA_12	AAA domain	3	4	4
AAA_16	AAA ATPase domain	1	1	1
AAA_19	Part of AAA domain	1	2	2
AAA_21	AAA domain	1	2	0
AAA_22	AAA domain	1	2	2
AAA_23	AAA domain	1	4	2
AAA_30	AAA domain	5	10	8
AAA_33	AAA domain	12	26	20
ABC_tran	ABC transporter	3	6	5
ABC_tran_2	ABC transporter	1	1	1
Aconitase	Aconitase family (aconitate hydratase)	2	3	3
Aconitase_C	Aconitase C-terminal domain	2	3	3
Activator_LAG-3	Transcriptional activator LAG-3	1	1	0

Рисунок 35. Страница со списком семейств белковых доменов Pfam представленных в базе данных NPIDB.

Для каждого домена PFAM создана отдельная страница, включающая список структур, содержащих данный домен, а также интерактивное выравнивание всех структур семейства, созданное сервисом Blocks-3D (рисунок 36).

NPIDB Home Pfam families Pfam family PF01881.11

Cas_Cas6 (CRISPR associated protein Cas6)

Home Browse Download Help About Us

Complexes with domains of Pfam family Cas_Cas6 (PF01881.11)

PDB ID	PDB Header	# of domains ↑	# of bound domains
3qjl	IMMUNE SYSTEM/RNA	2	2
3pkm	HYDROLASE/RNA	2	2
3qjl	IMMUNE SYSTEM/RNA	2	2
3qjp	IMMUNE SYSTEM/RNA	1	1
4	<-- Total -->	7	7

Blocks-3D output

Alignment of the family with reliable blocks

Reliable blocks are found by the program [Blocks-3D](#) with distance threshold 1.5 Å

Download

- Blocks JSON: [\[PF01881.11.blocks.json\]](#)
- Sequences JSON: [\[PF01881.11.dict.json\]](#)
- Blocks TXT: [\[PF01881.11.fasta.bb\]](#)
- Alignment FASTA: [\[PF01881.11.fasta.a\]](#)
- Blocks-3D output HTML: [\[PF01881.11.fasta.bb.html\]](#)

	10	20	30	40	50	60	70	80
3pkm.pdb1.pdb:A:0:101:244	VNGLLMNPEVRLWDERFYLHEIKVLRPEPKKFNSTFVTLSPIAVTVV	-----	YDVPPMEKEFYYSIIKDDLQDKYVMAYGDKPPE					
3pkm.pdb2.pdb:X:0:101:244	VNGLLMNPEVRLWDERFYLHEIKVLRPEPKKFNSTFVTLSPIAVTVVR	-----	KGKSYDVPPMEKEFYYSIIKDDLQDKYVMAYGDKPPE					
3qjl.pdb1.pdb:A:0:91:238	AEAVEKSPEPKIGELSFVLGDIKAIKVKELGKENVFSTLSPIVVRTVKFEGNKLRRHNDLYPHDELPMRLRKVMILRYSEVMGETPKI							
3qjj.pdb2.pdb:A:0:91:238	AEAVEKSPEPKIGELSFVLGDIKAIKVKELGKENVFSTLSPIVVRTVKFEGNKLRRHNDLYPHDELPMRLRKVMILRYSEVMGETPKI							
3qjp.pdb1.pdb:A:0:91:238	AEAVEKSPEPKIGELSFVLGDIKAIKVKELGKENVFSTLSPIVVRTVKFEGNKLRRHNDLYPHDELPMRLRKVMILRYSEVMGETPKI							
3qjl.pdb1.pdb:B:0:91:238	AEAVEKSPEPKIGELSFVLGDIKAIKVKELGKENVFSTLSPIVVRTVKFEGNKLRRHNDLYPHDELPMRLRKVMILRYSEVMGETPKI							
3qjj.pdb1.pdb:B:0:91:238	AEAVEKSPEPKIGELSFVLGDIKAIKVKELGKENVFSTLSPIVVRTVKFEGNKLRRHNDLYPHDELPMRLRKVMILRYSEVMGETPKI							

© NPIDB team 2003 - 2016

Рисунок 36. Страница семейства белковых доменов Pfam Cas_Cas6 в базе данных NPIDB.

Для семейства гомеодоменов (ID PF00046.24) дополнительно представлено детальное описание всего семейства. Для каждой структуры домена имеется подробная информация по качеству структуры, наличию ДНК и молекул воды, происхождению и принадлежности тому или иному белку. Дополнительно представлено множественное выравнивание белковых последовательностей, а также нуклеотидных последовательностей в районе сайта связывания.

На отдельной странице представлен список *семейств структурных доменов SCOP* в виде иерархии, принятой в классификации SCOP (всего 428 семейств). Подробно описанные семейства помечены звездочкой (★) (рисунок 37).

NPIDB Home SCOP families

List of available SCOP families

Home Browse Download Help About Us

Download

List of Scop domains:

Best resolution complexes of SCOP family representatives, bound with nucleic acid

List of SCOP families detected in NPIDB entries:

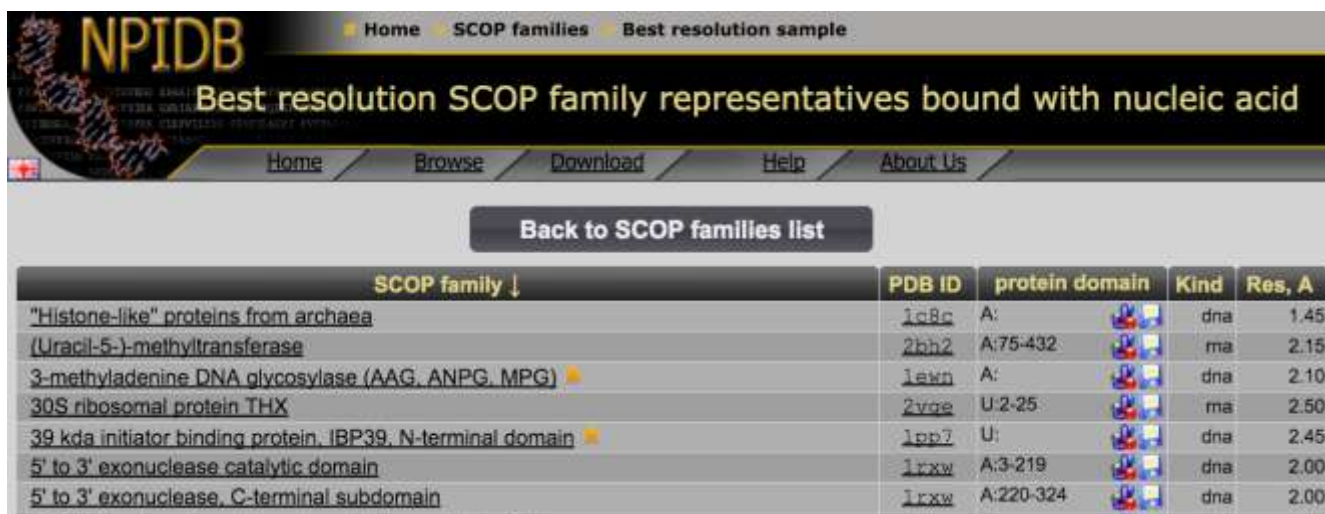
Hide all Show all CL = Class, CF = Fold, SF = Superfamily, FA = Family See SCOP-help for more information

SCOP: All list

- CL: All alpha proteins
 - CF: A DNA-binding domain in eukaryotic transcription factors
 - SF: A DNA-binding domain in eukaryotic transcription factors
 - FA: A DNA-binding domain in eukaryotic transcription factors
 - CF: alpha-alpha superhelix
 - SF: ARM repeat
 - FA: Pumilio repeat
 - SF: TROVE domain-like
 - FA: TROVE domain-like
 - CF: An anticodon-binding domain of class I aminoacyl-tRNA synthetases
 - SF: An anticodon-binding domain of class I aminoacyl-tRNA synthetases
 - FA: C-terminal domain of glutamyl-tRNA synthetase (GluRS)
 - CF: Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
 - SF: Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
 - FA: Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases
 - CF: ATP-dependent DNA ligase DNA-binding domain
 - SF: ATP-dependent DNA ligase DNA-binding domain
 - FA: ATP-dependent DNA ligase DNA-binding domain
 - CF: BRCA2 helical domain
 - SF: BRCA2 helical domain
 - FA: BRCA2 helical domain
 - CF: BRCA2 tower domain
 - SF: BRCA2 tower domain
 - FA: BRCA2 tower domain
 - CF: Cryptochrome/photolyase FAD-binding domain
 - SF: Cryptochrome/photolyase FAD-binding domain
 - FA: Cryptochrome/photolyase FAD-binding domain
 - CF: Cyclin-like
 - SF: Cyclin-like
 - FA: Transcription factor IIB (TFIIB), core domain ★ [a.74.1.2]
 - CF: DNA repair protein MutS, domain III
 - SF: DNA repair protein MutS, domain III
 - FA: DNA repair protein MutS, domain III ★ [a.113.1.1]
 - CF: DNA-glycosylase
 - SF: DNA-glycosylase
 - FA: DNA repair glycosylase, 2 C-terminal domains ★ [a.96.1.3]
 - FA: Endonuclease III ★ [a.96.1.1]
 - FA: Mismatch glycosylase ★ [a.96.1.2]

Рисунок 37. Страница со списком семейств белковых доменов SCOP, представленных в базе данных NPIDB.

Отдельно доступен список наилучших по разрешению представителей 350 семейств SCOP (рисунок 38).



SCOP family ↓	PDB ID	protein domain	Kind	Res, Å
"Histone-like" proteins from archaea	1c8c	A:	dna	1.45
(Uracil-5-)-methyltransferase	2bh2	A:75-432	ma	2.15
3-methyladenine DNA glycosylase (AAG, ANPG, MPG)	1ewn	A:	dna	2.10
30S ribosomal protein THX	2vge	U:2-25	ma	2.50
39 kda initiator binding protein, IBP39, N-terminal domain	1pp7	U:	dna	2.45
5' to 3' exonuclease catalytic domain	1rxw	A:3-219	dna	2.00
5' to 3' exonuclease, C-terminal subdomain	1rxw	A:220-324	dna	2.00

Рисунок 38. Страница со списком имеющих максимальное разрешение структур, относящихся к семействам белковых доменов SCOP, представленных в базе данных NPIDB.

На странице каждого семейства SCOP представлено место данного семейства в иерархии SCOP, а также список белковых доменов, входящих в данное семейство. Для каждого домена указано происхождение, идентификаторы структур, где этот домен встречается, сколько раз он встречается в структуре, а также сколько доменов в данной структуре связаны с НК (рисунок 39).



NPIDB

Home SCOP families SCOP family 140865

TROVE domain-like

Home Browse Download Help About Us

SCOP: [All list](#)

Class: [All alpha proteins](#)

Fold: [alpha-alpha superhelix](#)

Superfamily: [TROVE domain-like](#)

Family: [TROVE domain-like](#)

SCOP: [All list](#)

Class: [All alpha proteins](#)

Fold: [alpha-alpha superhelix](#)

Superfamily: [TROVE domain-like](#)

Family: [TROVE domain-like](#)

protein domain ↓	species	PDB ID	Instances	bonded
60-kda SS-aARo ribonucleoprotein	African clawed frog (<i>Xenopus laevis</i>) [TaxId: 8355]	2i91	2	2
60-kda SS-aARo ribonucleoprotein	African clawed frog (<i>Xenopus laevis</i>) [TaxId: 8355]	1vyp	2	2
Total:			4	4

© NPIDB team 2003 - 2016

Рисунок 39. Страница семейства Trove domain-like белковых доменов SCOP.

4.1 Расширение функционала базы данных NPIDB

4.1.1 Подробное описание семейств ДНК-связывающих доменов SCOP

76 семейств белковых доменов SCOP, взаимодействующих с ДНК и содержащих хорошо пространственно совмещаемые структуры, были описаны подробно. Стандартная страница семейства SCOP была дополнена следующими блоками (рисунок 40):

1. Общая информация:

- Класс взаимодействия данного доменов данного семейства с ДНК
- Метод, которым было получено пространственное совмещение доменов семейства (SSM/PDBeFold или CE)
- Экспертно оцененное качество пространственного совмещения
- Название семейства SCOP и его идентификатор
- Количество проанализированных комплексов
- Из них — количество комплексов с ДНК
- Количество найденных кластеров консервативных молекул воды
- Ссылки для загрузки файлов (в скобках указано расширение файлов):
 - Выравнивания последовательностей совмещенных доменов в формате FASTA (. fas) и MSF (. msf)
 - Выравнивание последовательностей с размеченной на них вторичной структурой (. fas)
 - Изображение выравнивания с отмеченными блоками вторичной структуры (. png)
 - Файл совмещенных структур доменов, где каждая структура представлена отдельной моделью (. pdb)
 - Список кластеров консервативных молекул воды в виде текста (. flat. txt) и таблицы (. pivot. txt)
 - Скрипт для визуализации кластеров консервативных молекул воды в программе RasMol (. spt)

2. Визуализация посредством Jmol совмещенных структур комплексов белкового домена данного семейства с ДНК с возможностью отображения:

- выборочных структур комплексов
- кластеров консервативных молекул воды

[Home](#)
[SCOP families](#)
[SCOP family 46748](#)

Paired domain

[Home](#)
[Browse](#)
[Download](#)
[Help](#)
[About Us](#)

SCOP: [All list](#)
Class: [All alpha proteins](#)
Fold: [DNA/RNA-binding 3-helical bundle](#)
Superfamily: [Homeodomain-like](#)
Family: [Paired domain](#)

SCOP: [All list](#)
Class: [All alpha proteins](#)
Fold: [DNA/RNA-binding 3-helical bundle](#)
Superfamily: [Homeodomain-like](#)
Family: [Paired domain](#)

protein domain ↓	species	PDB ID	instances	bonded
Paired protein (prd)	Fruit fly (<i>Drosophila melanogaster</i>) [Taxid: 7227]	1pdm	1	1
Pax-5	Human (<i>Homo sapiens</i>) [Taxid: 9606]	1mdm	2	2
Pax-5	Human (<i>Homo sapiens</i>) [Taxid: 9606]	1k78	5	5
Pax-6	Human (<i>Homo sapiens</i>) [Taxid: 9606]	6pax	2	2
Total:			4	10

Extended information about type of interaction with DNA:

Interaction class: [H-Bb](#) [H-M](#) [L-Bb](#) [L-Mn](#)

Method of 3-D alignment: [SSM](#)

Quality of 3-D alignment: [low](#)

Analyzed family: [Paired domain \[a.4.1.5\]](#)

Analyzed complexes: 4

Complexes with DNA: 4

Conserved water: 47

Download:

Sequence alignment: [\[a_4_1_5.fas\]](#)
Sequence alignment with secondary structure: [\[a_4_1_5_with_ss.fas\]](#)
Superposition in PDB format: [\[a.4.1.5.pdb\]](#)
Alignment in MSF: [\[a_4_1_5.msf\]](#)
Image of alignment with secondary structure marked: [\[a_4_1_5.PNG\]](#)
Water clusters: [\[a.4.1.5.flat.txt\]](#)
[\[a.4.1.5.pivot.txt\]](#)
RasMol script: [\[a.4.1.5.spt\]](#)

General

Jmol visualisation of conserved water

wat1	wat2	wat3	wat4	wat5	wat6	wat7	wat8	wat9																																																																		
<table> <thead> <tr> <th>PDB ID</th><th>Chain</th><th>Start residue</th><th>End residue</th><th>Model number</th><th>Show</th></tr> </thead> <tbody> <tr><td>1k78</td><td>A</td><td>19</td><td>81</td><td>1</td><td>ON</td></tr> <tr><td>1k78</td><td>A</td><td>82</td><td>142</td><td>2</td><td>ON</td></tr> <tr><td>1k78</td><td>E</td><td>19</td><td>81</td><td>3</td><td>ON</td></tr> <tr><td>1k78</td><td>E</td><td>82</td><td>142</td><td>4</td><td>ON</td></tr> <tr><td>1k78</td><td>I</td><td>84</td><td>141</td><td>5</td><td>ON</td></tr> <tr><td>1mdm</td><td>A</td><td>19</td><td>81</td><td>6</td><td>ON</td></tr> <tr><td>1mdm</td><td>A</td><td>82</td><td>142</td><td>7</td><td>ON</td></tr> <tr><td>6pax</td><td>A</td><td>1</td><td>68</td><td>8</td><td>ON</td></tr> <tr><td>6pax</td><td>A</td><td>69</td><td>133</td><td>9</td><td>ON</td></tr> <tr><td>1pdm</td><td>C</td><td>-</td><td>-</td><td>10</td><td>OFF</td></tr> </tbody> </table>	PDB ID	Chain	Start residue	End residue	Model number	Show	1k78	A	19	81	1	ON	1k78	A	82	142	2	ON	1k78	E	19	81	3	ON	1k78	E	82	142	4	ON	1k78	I	84	141	5	ON	1mdm	A	19	81	6	ON	1mdm	A	82	142	7	ON	6pax	A	1	68	8	ON	6pax	A	69	133	9	ON	1pdm	C	-	-	10	OFF								
PDB ID	Chain	Start residue	End residue	Model number	Show																																																																					
1k78	A	19	81	1	ON																																																																					
1k78	A	82	142	2	ON																																																																					
1k78	E	19	81	3	ON																																																																					
1k78	E	82	142	4	ON																																																																					
1k78	I	84	141	5	ON																																																																					
1mdm	A	19	81	6	ON																																																																					
1mdm	A	82	142	7	ON																																																																					
6pax	A	1	68	8	ON																																																																					
6pax	A	69	133	9	ON																																																																					
1pdm	C	-	-	10	OFF																																																																					

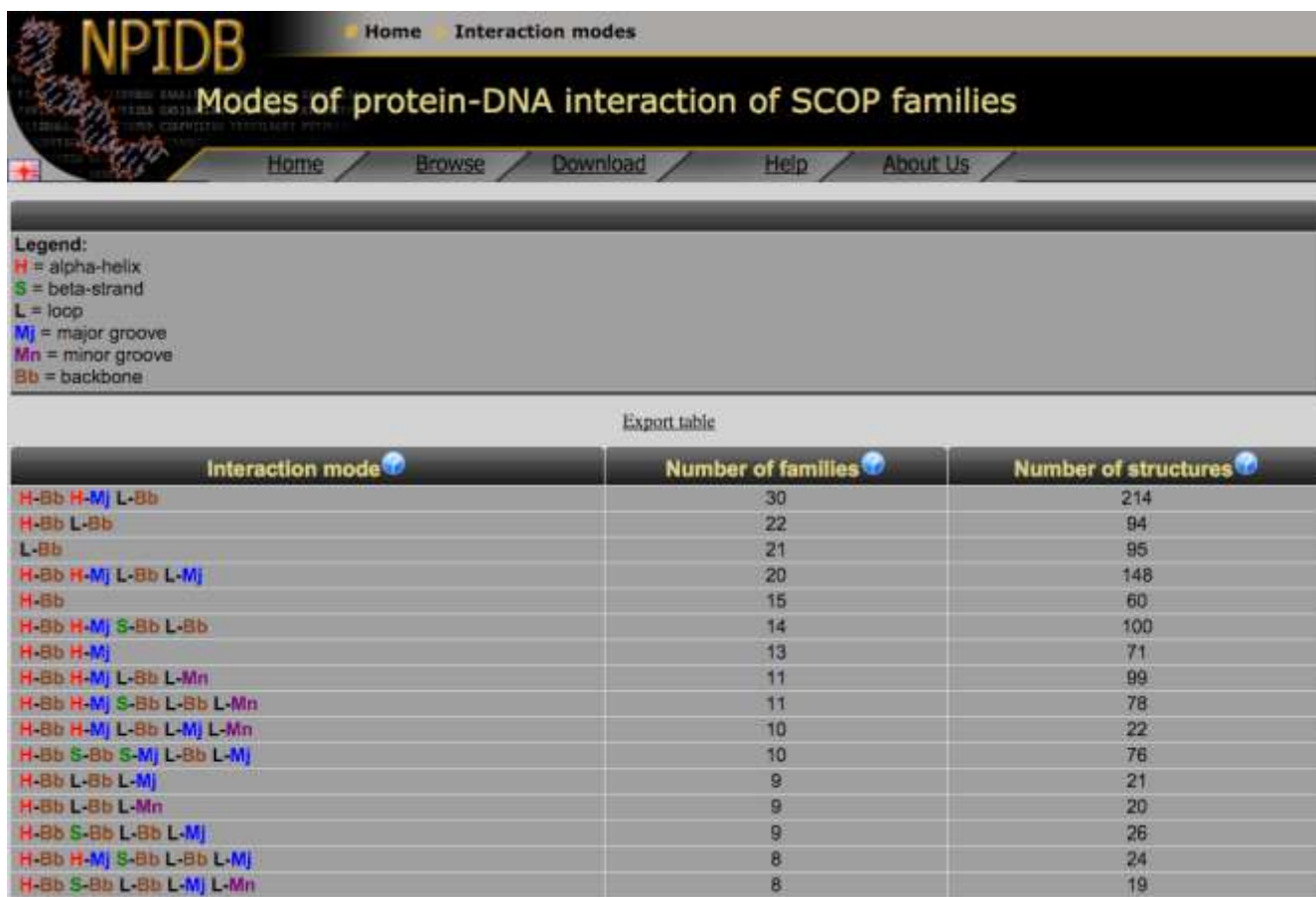
© NPIDB team 2003 - 2016

Рисунок 40. Страница семейства Paired domain белковых доменов SCOP.

4.1.2 Интеграция классификации ДНК-белковых взаимодействий в базу данных NPIDB

Представленная в данной работе классификация ДНК-белковых взаимодействий структурных доменов SCOP была интегрирована в базу данных NPIDB. Для этого данные о способах взаимодействия отдельных доменов SCOP, а также классах взаимодействия ДНК-контактирующих семейств SCOP были добавлены на соответствующие страницы.

Полный список способов взаимодействия (interaction mode) ДНК-контактирующих доменов SCOP представлен на отдельной странице. Для каждого из 97 обнаруженных способов взаимодействия приведено количество семейств SCOP и структур, в которых он встречается (рисунок 41).



NPIDB Home Interaction modes

Modes of protein-DNA interaction of SCOP families

Home Browse Download Help About Us

Legend:
H = alpha-helix
S = beta-strand
L = loop
Mj = major groove
Mn = minor groove
Bb = backbone

Export table

Interaction mode	Number of families	Number of structures
H-Bb H-Mj L-Bb	30	214
H-Bb L-Bb	22	94
L-Bb	21	95
H-Bb H-Mj L-Bb L-Mj	20	148
H-Bb	15	60
H-Bb H-Mj S-Bb L-Bb	14	100
H-Bb H-Mj	13	71
H-Bb H-Mj L-Bb L-Mn	11	99
H-Bb H-Mj S-Bb L-Bb L-Mn	11	78
H-Bb H-Mj L-Bb L-Mj L-Mn	10	22
H-Bb S-Bb S-Mj L-Bb L-Mj	10	76
H-Bb L-Bb L-Mj	9	21
H-Bb L-Bb L-Mn	9	20
H-Bb S-Bb L-Bb L-Mj	9	26
H-Bb H-Mj S-Bb L-Bb L-Mj	8	24
H-Bb S-Bb L-Bb L-Mj L-Mn	8	19

Рисунок 41. Страница со списком способов взаимодействия (interaction mode) ДНК-контактирующих доменов SCOP. Указано количество семейств и структур, для которых характерен данный способ взаимодействия.

На странице каждого из способов взаимодействия представлен подробный список семейств и структур, для которых характерен данный способ взаимодействия (рисунок 42).

NPIDB Home Interaction modes Families of interaction mode

Modes of protein-DNA interaction of SCOP families

Home Browse Download Help About Us

Interaction mode

Mode: **H-Bb H-Mn L-Bb L-Mj**

Description: Conserved protein-DNA contacts are provided by helix(es) (**H**) with DNA backbone atoms (**Bb**), helix(es) (**H**) with the minor groove of DNA (**Mn**), loop(s) (**L**) with DNA backbone atoms (**Bb**), loop(s) (**L**) with the major groove of DNA (**Mj**).

[Export table](#)

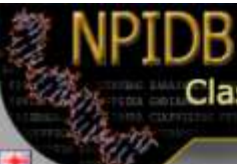
SCOP Domain Family	PDB entries	Number of structures
DNA repair glycosylase, 2 C-terminal domains (a.96.1.3)	3aw1 3ica	2
Single-strand selective monofunctional uracil-DNA glycosylase SMUG1 (c.18.1.3)	1aa5 1ca6	2
T-box (b.2.5.4)	1a6f	1
Restriction endonuclease NgoIV (c.52.1.10)	1a1u	4

© NPIDB team 2003 - 2016

Рисунок 42. Страница способа взаимодействия |H-Bb|H-Mn|L-Bb|L-Mj|.

Кроме того, на страницу с заданным способом взаимодействия можно перейти с соответствующих страниц записей PDB (рисунок 33) и семейств SCOP (рисунок 40). Таким образом, рассматривая конкретный ДНК-белковый комплекс или семейство ДНК-связывающих белковых доменов SCOP, можно найти другие комплексы, имеющие такой же способ взаимодействия, что и искомый.

Полный список **классов взаимодействий** (interaction class) семейств ДНК-связывающих доменов SCOP также представлен на отдельной странице базы данных NPIDB (<http://npidb.belozersky.msu.ru/contacttypesnew.html>). Для каждого класса взаимодействия указано количество семейств SCOP, а также количество структур и доменов внутри семейств с данным классом взаимодействия (рисунок 43).



Home Interaction classes

Classes of protein-DNA interaction of SCOP families

Home Browse Download Help About Us

Legend:
H = alpha-helix
S = beta-strand
L = loop
Mj = major groove
Mn = minor groove
Bb = backbone

[Export table](#)

Interaction class	Number of families	Number of structures	Number of domains
H-Bb H-Mj L-Bb	8	200	59
Miscellaneous	4	199	39
H-Bb H-Mj L-Bb L-Mn	4	121	18
H-Bb	3	46	13
H-Bb H-Mj	2	73	17
H-Bb H-Mj L-Bb L-Mj	2	29	8
H-Bb S-Bb L-Bb	1	255	23
L-Bb	1	127	9
S-Bb S-Mn L-Bb L-Mn	1	86	8
H-Bb L-Bb L-Mj	1	54	8
H-Bb H-Mj H-Mn L-Bb L-Mj	1	53	3
H-Bb S-Bb S-Mj L-Bb L-Mj	1	40	7
S-Bb L-Bb	1	27	4
H-Bb S-Bb L-Bb L-Mn	1	18	3
H-Bb H-Mn L-Bb	1	14	8
H-Bb H-Mj S-Bb L-Bb	1	9	4
H-Bb L-Bb	1	3	3
Total	34	1354	234

© NPIDB team 2003 - 2016

Рисунок 43. Страница со списком классов взаимодействия (interaction class) ДНК-контактирующих доменов SCOP. Указано количество семейств, для которых характерен данный класс взаимодействия, а также число доменов и структур в этих семействах.

Страницы, посвященные классам взаимодействия, содержат подробный список семейств белковых доменов SCOP, для которых характерен данный класс. Для каждого семейства указано число структур и доменов внутри семейства, ширина и глубина бороздок, изгиб ДНК, а также ссылки на способы взаимодействия доменов с ДНК, встречающиеся в данном семействе (рисунок 44).

Поскольку страницы классов взаимодействия и страницы соответствующих семейств SCOP связаны взаимными ссылками, то можно производить поиск семейств с тем же классом взаимодействия, что и данное.

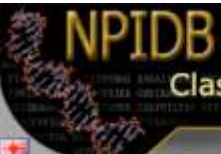
 Home Interaction classes Families of interaction class Classes of protein-DNA interaction of SCOP families Home Browse Download Help About Us								
Interaction class Class: H-Bb H-Mj L-Bb Description: Conserved protein-DNA contacts are provided by helix(es) (H) with DNA backbone atoms (Bb), helix(es) (H) with the major groove of DNA (Mj), loop(s) (L) with DNA backbone atoms (Bb). Export table								
SCOP Domain Family	Number of structures	Number of domains	Interaction modes	Ax-bend	MnG width	MnG depth	MjG width	MjG depth
Homeodomain (a.4.1.1)	58	19	H-Bb H-Mj H-Bb H-Mj L-Bb H-Bb H-Mj L-Bb L-Mn H-Bb H-Mj L-Mj H-Bb H-Mj L-Mn	1.1 (0.4–2.5)	5.9 (4.3–7.9)	4.8 (4.0–5.5)	12.5 (11.3–15.2)	4.5 (0.5–6.6)
Nuclear receptor (g.39.1.2)	46	12	H-Bb H-Mj H-Bb H-Mj H-Mn L-Bb H-Bb H-Mj L-Bb H-Bb H-Mj L-Bb L-Mj H-Bb H-Mj S-Bb L-Bb H-Bb H-Mj S-Bb L-Bb L-Mn H-Bb H-Mj S-Bb S-Mj L-Bb H-Bb H-Mj S-Bb S-Mj L-Bb L-Mn	0.9 (0.2–2.0)	6.5 (4.2–7.6)	4.8 (4.5–5.2)	12.3 (12.0–12.8)	4.0 (2.0–5.8)
Phage repressors (a.35.1.2)	24	5	H-Bb H-Mj L-Bb H-Bb H-Mj L-Bb L-Mj L-Bb L-Mn	1.8 (1.0–2.6)	7.1 (5.9–8.2)	4.6 (4.3–4.9)	12.5 (11.8–13.6)	4.2 (2.4–5.5)
ets domain (a.4.5.21)	20	7	H-Bb H-Mj L-Bb H-Bb H-Mj S-Bb L-Bb H-Bb H-Mj S-Bb L-Bb L-Mn L-Bb	0.9 (0.5–1.5)	8.2 (7.3–10.5)	3.8 (3.0–4.4)	13.6	6.3
Interferon regulatory factor (a.4.5.23)	18	3	H-Bb H-Mj L-Bb H-Bb H-Mj S-Bb L-Bb H-Bb H-Mj S-Bb L-Bb L-Mn	1.6 (0.3–2.8)	5.8 (3.6–7.6)	5.2 (5.0–5.4)	11.0 (10.5–11.4)	5.1 (4.9–5.3)
Myb/SANT domain (a.4.1.3)	14	6	H-Bb H-Mj H-Bb H-Mj L-Bb H-Bb H-Mj L-Bb L-Mj H-Bb L-Bb	0.5 (0.1–0.6)	6.4	4.5	–	–
PQI-specific domain (a.35.1.1)	12	3	H-Bb H-Mj L-Bb H-Bb H-Mj L-Bb L-Mj	0.8 (0.7–0.9)	9.8	0.6	11.4	1.2
Replication initiation protein (a.4.5.10)	8	4	H-Bb H-Mj L-Bb L-Mj H-Bb H-Mj S-Bb L-Bb H-Bb H-Mj S-Bb L-Bb L-Mn	0.7 (0.4–1.1)	7.5	4.6	12.5	5.6
All SCOP domain families with determined protein-DNA interaction classes								
© NPIDB team 2003 - 2016								

Рисунок 44. Страница класса взаимодействия (H – Bb), (H – Mj) | (L – Bb).

Глава 5. Поиск консервативных особенностей в семействах гомологичных белков

Используя функционал базы данных NPIDB можно проводить подробный анализ консервативных особенностей контакта с ДНК для семейств гомологичных белков. Рассмотрим подробнее примеры поиска консервативных особенностей в семействах ДНК-связывающих белков: TATA-box связывающего домена и LAGLIDADG хоминг эндонуклеаз.

5.1 TATA-box связывающий домен и белок

Два TATA-box связывающих домена (SCOP семейство TATA-box binding protein (TBP), C-terminal domain (d.129.1.1)) одной белковой цепи формируют TATA-box связывающий белок (TBP), участвующий в инициации транскрипции многих архей и эукариот в составе мультибелкового комплекса TFIID.

Каждый домен состоит из пяти β -тяжей антипараллельного β -листа, окаймленного двумя α -спиралям. Вместе они образуют общий β -лист, узнающий восемь пар оснований в малой бороздке ДНК. Узнаваемую последовательность T-A-T-A-a/t-A-a/t-N (N — любой нуклеотид) принято называть TATA-боксом [123], [124]. TATA-box связывающий белок обладает осью симметрии второго порядка и в литературе называется квазисимметричным (рисунок 45).

Связывание TBP с ДНК приводит к значительным конформационным изменениям ДНК. Изначально более узкая малая бороздка расширяется, а ДНК изгибается за счет встраивания остатков фенилаланина TBP в стекинг оснований [125]. Поверхность β -листа белка содержит много неполярных остатков, которые вносят значительный вклад в узнавание малой бороздки ДНК [126]. При образовании комплекса происходит удаление молекул воды с поверхности β -листа белка, что

значительно снижает энтропийную составляющую и способствует образованию комплекса [127]. Между белком и ДНК формируется сеть водородных связей [128] и гидрофобных взаимодействий [127]. Правильной ориентации ТВР относительно последовательности ТАТА-бокс способствуют вариации в специфических контактах между N- и С- доменом, а так же другие транскрипционные факторы, связывающиеся с ТВР. Последовательность ТАТА узнает С-концевой домен, а последовательность а/т-А-а/т-Н — N-концевой [129], [130].

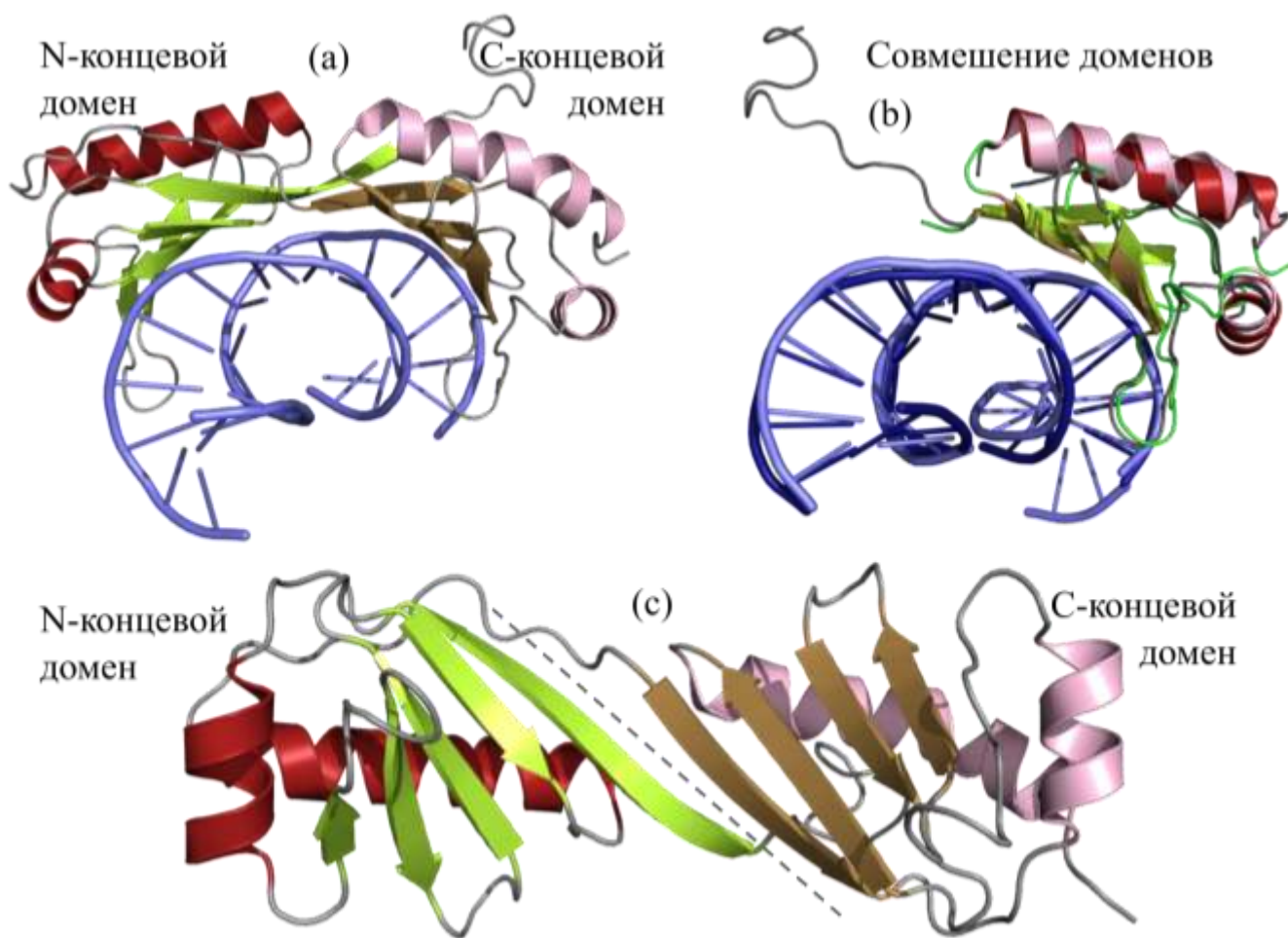


Рисунок 45. Структура ТАТА-бокс связывающего белка из *Arabidopsis thaliana* в комплексе с ТАТА-бокс последовательностью ДНК (5'-ТАТААААГ-3'), структура 1qpe. (a) Общий вид комплекса (b) Пространственное совмещение N- и С-концевых доменов с соответствующими фрагментами ДНК. (c) Вид белка со стороны β -листа, взаимодействующего с малой бороздкой ДНК. Ось симметрии обозначена пунктиром.

5.1.1 Структуры TATA-box связывающего домена

Для надежного выявления консервативных ДНК-белковых контактов были проанализированы все доступные структуры TATA-box связывающего домена. Среди имеющихся на данный момент 34 структур 25 представляют собой комплекс с ДНК, 29 содержат молекулы воды и еще две структуры не содержат ни ДНК, ни воды и в дальнейшем не рассматриваются. На основе этих структурных данных был произведен поиск прямых и опосредованных молекулами воды консервативных водородных связей и гидрофобных контактов между белком и ДНК. Были проанализированы как TATA-box связывающий белок в целом, так и входящие в его состав домены.

Всего было проанализировано 32 записи PDB, принадлежащих семи организмам, а именно: *Homo sapiens* (1c9b, 1cdw, 1jfi, 1nvp, 1tgh), *Sulfolobus acidocaldarius* (1mp9), *Arabidopsis thaliana* (1qn3, 1qn4, 1qn5, 1qn6, 1qn7, 1qn8, 1qn9, 1qna, 1qnb, 1qnc, 1qne, 1vto, 1vok, 1vol), *Saccharomyces cerevisiae* (1nh2, 1rm1, 1ytb, 1ytf, 1ngm), *Archaeon Pyrococcus woesei* (1ais, 1d3u, 1pcz), *Methanocaldococcus jannaschii* (2z8u), *Encephalitozoon cuniculi* (3eik, 3oc3, 3ocI). В ряде записей PDB встречается несколько одинаковых цепей белка/ДНК-белковых комплексов. Такие записи были разделены на несколько структур, и в итоге было проанализировано 45 структур TATA-box связывающего белка в комплексе с ДНК (с и без молекул воды) (группа #1), а также 16 структур TBP без ДНК, но с молекулами воды (группа #2).

Внутри каждой группы с помощью сервиса SSM было произведено пространственное совмещение структур и построено выравнивание последовательностей. Поиск водородных связей был произведен с помощью сервисов NPIDB. Гидрофобные кластеры были найдены с помощью программы CluD, также интегрированной в NPIDB. Поиск консервативных молекул воды был произведен программой wLake.

Для простоты сравнения аминокислотных последовательностей была введена единая нумерация: N-концевой домен имеет нумерацию, начиная с 1, а С-концевой — начиная с 1001. Также единая нумерация была введена и для ДНК: выделялась прямая (“d” — direct) и обратная (“r” — reversed) цепи ДНК. Прямая цепь содержит последовательность Т-А-Т-А-а/т-А-а/т-Н, а обратная — комплементарную ей последовательность Н'-т/а-Т-т/а-Т-А-Т-А. Нумерация идет попарно и начинается с пары Т (прямой цепи) и А (обратной цепи) (Т#d1•А#r1).

5.1.2 Консервативные особенности комплексов ТРВ с ДНК

На рисунке 46 приведено выравнивание последовательностей ТАТА-бокс связывающих белков из семи организмов: *A. thaliana* (1qne:A12-198), *H. sapiens* (1cdw:A155-333), *S. cerevisiae* (1ytf:A61-240), *E. cuniculi* (3eik:A20-197), *S. acidocaldarius* (1mp9:A5-197), *P. woesei* (1d3u:A1-181), *M. jannaschii* (2z8u:B3-177). В верхней панели приведено выравнивание N-концевого домена, в нижней - С-концевого; колонки в верхней и нижней панели выровнены по совмещению N- и С-концевых доменов.

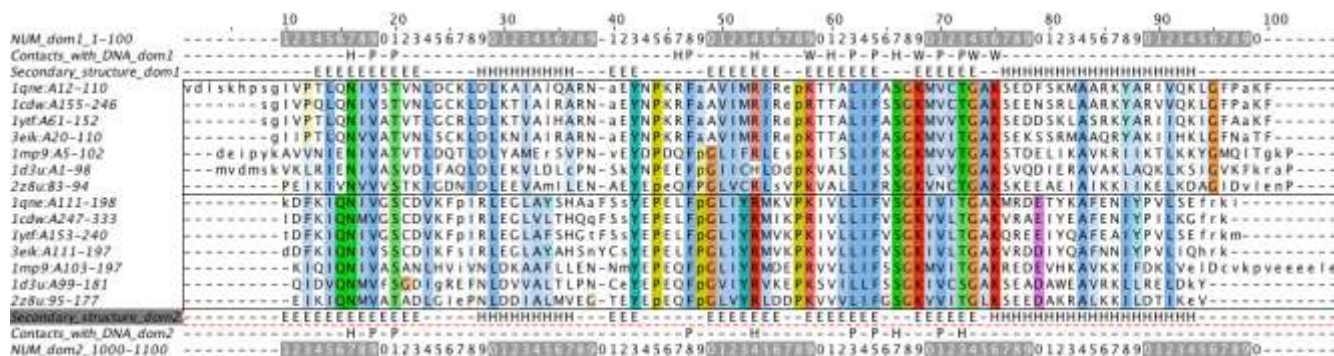


Рисунок 46. Выравнивание последовательностей ТАТА-бокс связывающих белков из семи организмов: *A. thaliana* (1qne:A12-198), *H. sapiens* (1cdw:A155-333), *S. cerevisiae* (1ytf:A61-240), *E. cuniculi* (3eik:A20-197), *S. acidocaldarius* (1mp9:A5-197), *P. woesei* (1d3u:A1-181), *M. jannaschii* (2z8u:B3-177). В верхней панели приведено выравнивание N-концевого домена, в нижней - С-концевого; колонки в верхней и нижней панели выровнены по совмещению N- и С-концевых доменов. В верхней и нижней строках приведена нумерация остатков выравнивания, согласно единой нумерации аминокислотных остатков ТРВ. Для каждой панели отмечены консервативные контакты с ДНК: Н — водородные связи, Р — гидрофобные кластеры, W — водородные связи, опосредуемые молекулами воды. Отдельно указана консенсусная вторичная структура: Н — α-спирали, Е — β-тяжи.

Как видно из выравнивания, порядка 30% аминокислотных остатков последовательностей из разных организмов идентичны, при этом структуры совмещаются с высокой точностью. Например в группе #1 из 175 C_α атомов аминокислотных остатков TBP выравниваются 165, а среднее квадратичное отклонение при попарном сравнении структур составляет в среднем 0.73 Å.

TATA-бокс связывающие домены SCOP имеют меньшее сходство по последовательности, чем белок в целом (~23% идентичных аминокислотных остатков из 80-100 остатков домена). При этом пространственное совмещение доменов между собой также весьма хорошее: RMSD составляет 1.08 Å, при этом совмещаются 67 C_α атомов из 80-100.

Октануклеотидные последовательности ДНК 5'-TATAAAAN-3', узнаваемые TBP, практически одинаковы у всех эукариот, отличия наблюдаются лишь в последней паре нуклеотидов, а также при целенаправленных мутациях. У архей наблюдается 2-3 замены (например, в комплексе 1d3u TBP связан с ДНК, имеющую последовательность 5'-TTTAAATA-3'). У *Arabidopsis thaliana* в 10 структурах были введены мутации в каждую из позиций 5'-TATAAAAG-3' и были измерены изменения транскрипционных активностей в таких комплексах [125].

5.1.3 Анализ прямых водородных связей для TBP и входящих в его состав доменов

Анализ водородных связей для TATA-бокс связывающих белков был проведен внутри группы #1. Были проанализированы структуры из *Homo sapiens* (9 структур), *Arabidopsis thaliana* (25 структур), *Saccharomyces cerevisiae* (9 структур) и *Archaeon Pyrococcus woesei* (2 структуры). Данные о водородных связях для TBP и входящих в его состав доменов были получены из базы данных NPIDB. Каждый аминокислотный остаток формирует от одной до трех водородных связей с атомами ДНК. Поскольку количество структур меняется от организма к организму, то

в начале был произведен поиск консервативных контактирующих аминокислотных остатков внутри организмов, а потом были найдены общие консервативные водородные связи для всех исследованных организмов.

Организм	<i>Homo sapiens</i>	<i>Arabidopsis thaliana</i>	<i>Saccharomyces cerevisiae</i>	<i>Pyrococcus woesei</i>
Пример структуры (PDB код, цепи белка и ДНК)	<u>1cdw</u> A – цепь белка B – прямая цепь ДНК C – обратная цепь ДНК	<u>1qne</u> A – цепь белка C – прямая цепь ДНК D – обратная цепь ДНК	<u>1ytf</u> A – цепь белка E – прямая цепь ДНК F – обратная цепь ДНК	<u>1d3u</u> A – цепь белка C – прямая цепь ДНК D – обратная цепь ДНК
Общая формула контакта согласно единой нумерации	Контакты с сахаро-фосфатным остовом ДНК			
Arg#44 /His#44 T#5:r	Arg199:A T108:C	Arg63:A T222:D	Arg105:A T4:F	His49:A T1435:D
Arg#1044 T#3:d, A#4:d	Arg290:A –, A8:B	Arg154:A T205:C, A206:C	Arg196:A –, A12:E	Arg140:A –, A1413:C
Ser#57 N#8:d	Ser212:A G12:B	Ser76:A G210:C	Ser118:A C16:E	Ser62:A A1417:C
Ser#1057 A#1:r	Ser303:A A112:C	Ser167:A A226:D	Ser209:A A8:F	Ser153:A A1439:D
Arg#37 T#6:r, T#7:r	Arg192:A T106:C, T107:C	Arg56:A T220:D, T221:D	Arg98:A T3:F, –	–
Thr#51 T#5:r	Thr206:A T108:C	Thr70:A T222:D	Thr112:A T4:F	–
Контакты с азотистыми основаниями ДНК				
Asn#7 T#4:r, T#5:r	Asn163:A –, T108:C	Asn27:A T222:D, T223:D	Asn69:A T4:F, T5:F	Asn13:A –, T1435:D
Asn#1007 A#4:d/A#5:d	–	Asn117:A – / A207:C	–	Asn104:A A1413:C / –
Thr/Ser#1063 T#3:d/A#4:d	Thr309:A – / A8:B	Thr173:A – / A206:C	Thr215:A – / A12:E	Ser159:A T1412:C / –

Таблица 2. Консервативные водородные связи между белком ТВР и ДНК. В левом столбце используется единая нумерация контактирующих аминокислотных остатков и нуклеотидов. Косая черта разделяет альтернативные контакты, не встречающиеся в одном комплексе; запятая разделяет совместно наблюдаемые контакты. В остальных столбцах используется нумерация соответствующей записи PDB. В строках указаны гомологичные аминокислотные остатки, формирующие водородную связь в одном и том же месте комплекса ТВР-ДНК. Прочерк обозначает, что в данной структуре соответствующего контакта нет. Водородные связи объединены в две группы: в начале указаны водородные связи с сахаро-фосфатным остовом, после — с основаниями ДНК со стороны малой бороздки. Аналогичные контакты N- и C-концевого домена приведены в соседних строках таблицы.

В таблице 2 приведен список консервативных контактов согласно единой нумерации, а также приведены примеры конкретных контактов для представителей каждого из четырех организмов.

Как видно из таблицы 2, с основаниями малой бороздки ДНК контактирует три аминокислотных остатка белка: (Asn#7, Asn#1007, Thr#1063). Вдвое больше, шесть связей образовано с сахаро-фосфатным скелетом ДНК: Arg#44, Arg#1044, Ser#57, Ser#1057, Arg#37, Thr#51. Часть контактов симметрична и присутствует в обоих доменах. Два аспарагина (Asn#7, Asn#1007), расположенных по центру β -листа, центрируют ТВР в малой бороздке. В вершинах ромба расположены пары аргининов (Arg#44, Arg#1044) и серинов (Ser#57, Ser#1057), фиксирующих положение белка на ДНК за счет контактов с сахаро-фосфатным остовом. Треонину (Thr#1063) С-концевого домена, образующему водородную связь с атомом азота (N3) четвертого аденина ТАТА-последовательности, соответствует треонин (Thr#63) N-концевого домена. Водородной связи он не образует, но участвует в формировании гидрофобного кластера. Симметрия контактов нарушается со стороны N-концевого домена, где присутствуют дополнительные контакты аргинина (Arg#37) и триптофана (Thr#51).

5.1.4 Гидрофобные контакты для ТРВ и входящих в его состав доменов

Анализ гидрофобных кластеров на интерфейсе ТАТА-box связывающего белка и ДНК для группы#1 позволил выявить два достаточно симметричных скопления гидрофобных остатков, формирующих в разных структурах от двух до восьми гидрофобных кластеров. Аналогично таблице 2, в таблице II приложения приведен список консервативных гидрофобных контактов согласно единой нумерации, а также приведены примеры конкретных контактов для представителей каждого из четырех организмов.

Из таблицы II видно, что гидрофобные контакты образуются только с неполярными группами атомов азотистых оснований со стороны малой бороздки. Гид-

гидрофобных контактов с сахаро-фосфатным остовом нет в силу его полярности. Гидрофобные контакты являются более симметричными по сравнению с водородными связями и присутствуют в обоих доменах. При связывании с ДНК образуются шесть пар эквивалентно контактирующих остатков: Leu#1053 и Leu#53; Phe#38 и Phe#1038; Val#9 и Val#1009; Val#61 и Val#1061; Phe#55 и Phe#1055; Thr#11 и Ser#1011 в 45% случаев они дополняются седьмой парой Thr#63 и симметричным Thr#1063, образующим гидрофобный контакт вместо водородной связи.

Контактирующие остатки разбиваются на две группы и формируют, в общем случае, два гидрофобных кластера — N- и C-концевого домена. В кластер первого домена входят остатки Val#9, Phe#38, Leu#53, Phe#55, Val#61, Thr#63 со стороны белка и нуклеотиды A#5:d, A#6:d, A/T#7:d, N#8:d, T#7:r. В кластер второго домена входят со стороны белка: Val#1009, Phe#1038, Leu#1053, Phe#1055, Val#1061; со стороны ДНК: T#1:d, A#2:d, T#3:d, A#4:d, A#5:; T#1:r, A#2:r, T#3:r. На рисунке 47 кластеры отмечены розовым и желтым. Эти гидрофобные кластеры не объединяются в один из-за наличия водородных связей аспарагинов (Asn#7, Asn#1007).

Для части гидрофобно контактирующих аминокислотных остатков известна их биологическая роль. Например, четыре фенилаланина (Phe#38 и Phe#1038; Phe#55 и Phe#1055) встраиваются между первым и вторым, а также седьмым и восьмым основаниями сайта связывания [126] изгибая таким образом ДНК. Связыванию нуклеотидных последовательностей, содержащих во второй и седьмой позициях G•C пару вместо A•T, стерически мешает пара лейцинов (Leu#1053 и Leu#53), стабилизирующих гидрофобное ядро фенилаланинов. За узнавание A•T пары в четвертой и пятой позициях сайта связывания [125] отвечает пара валинов (Val#9 и Val#1009). Роль остальных остатков в литературе не описана.

5.1.5 Симметрия второго порядка в β -листе ТВР

Несмотря на отличия в аминокислотных последовательностях N- и С-концевого домена ТВР, ключевые консервативные контакты с ДНК в основном расположены симметрично. На рисунке 47 приведена схема β -листа, на которой отмечены аминокислотные остатки, образующие водородные связи (в прямоугольниках) и гидрофобные контакты (в овалах). Контактирующие остатки расположены достаточно симметрично относительно оси симметрии второго порядка. Асимметрия в контактах проявляется для остатков аргинина (Arg#37) и треонина (Thr#51). Соответствующие им остатки в С-концевом домене консервативных водородных связей с ДНК не образуют.

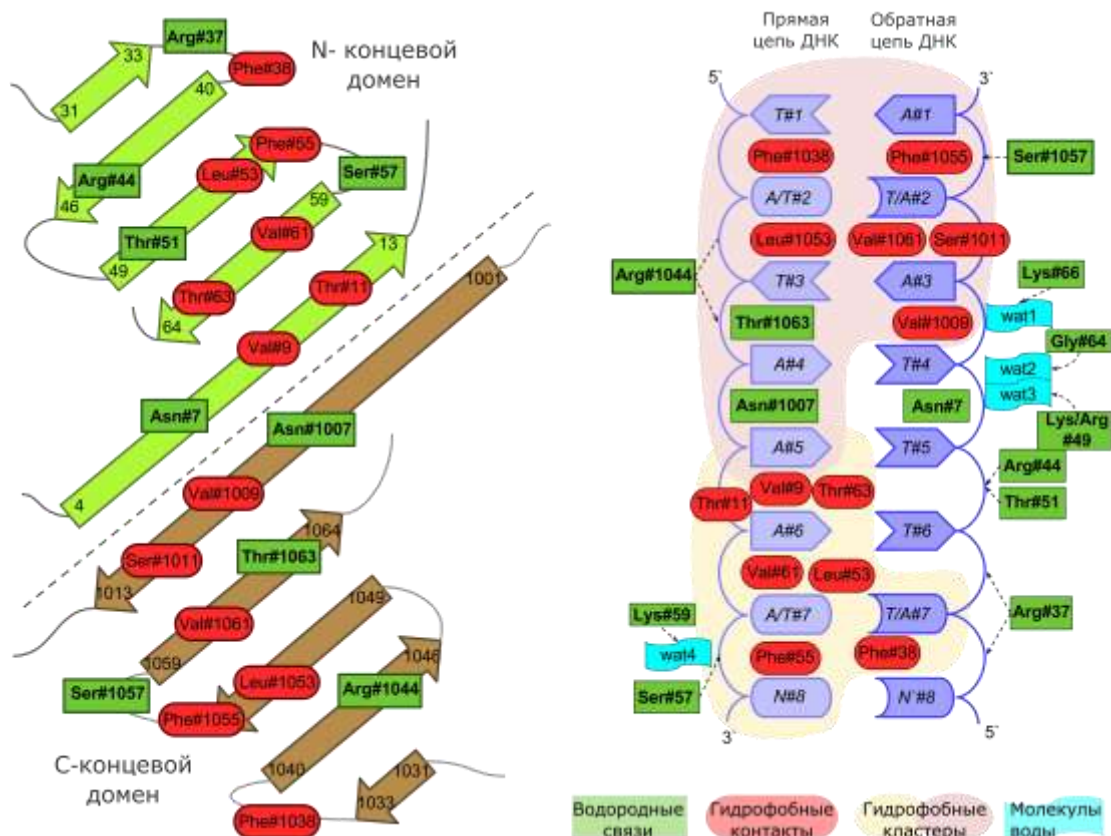


Рисунок 47. Общая схема β -листа структуры ТАТА-связывающего белка (слева) и контактов с ДНК (справа). В зеленом указаны остатки, образующие водородные связи, красным — остатки, образующие гидрофобные контакты с малой бороздкой ДНК, пунктир — ось симметрии второго порядка, разделяющая два домена ТВР. Цепи ДНК обозначены фиолетовым, водородные связи – пунктирными стрелками, а гидрофобные кластеры – розовой и желтой фигурами. Консервативные молекулы воды выделены голубым цветом.

5.1.6 Опосредованные водой водородные связи для ТАТА-box связывающего белка и входящих в его состав доменов

Для поиска функционально значимых молекул воды был произведен поиск консервативных молекул воды среди ДНК-белковых комплексов, содержащих молекулы воды. Из 45 структур группы #1 только 32 содержат молекулы воды, образующие водородные связи как с белком, так и с ДНК. Поскольку каждый организм представлен разным количеством структур ((*H. sapiens* (4 структуры), *A. thaliana* (24 структуры), *S. cerevisiae* (4 структуры) и *P. woesei* (2 структуры)) и молекулы воды могут быть пере/недопредставленными для этих организмов, то было введено понятие “консервативного кластера”. *Консервативный кластер* — это кластер, включающий в себя хотя бы одну молекулу воды из каждого организма. Молекулы воды, относящиеся к консервативному кластеру воды, могут как иметь водородные связи с белком, ДНК, другими молекулами воды, так и не иметь. Всего с помощью программы wLake на всей поверхности ТАТА-box связывающего белка было обнаружено 78 кластеров размером от 6 до 37 молекул воды в каждом. Из них 38 кластеров содержат хотя бы одну молекулу воды, опосредующую контакт белка с ДНК. Важно отметить, что непосредственно на поверхности β -листа, контактирующей с малой бороздкой ДНК, молекул воды нет — она удаляется при образовании гидрофобных кластеров между основаниями ДНК и аминокислотными остатками ТВР.

Если в консервативный водный кластер входили мостиковые молекулы воды из всех четырех организмов, то такой кластер назывался *консервативным кластером, опосредующим ДНК-белковую водородную связь (мостиковым кластером)*. Среди всех 38 консервативных кластеров на ДНК-белковом интерфейсе оказалось только 4 кластера, содержащих в себе хотя бы одну мостиковую молекулу воды из каждого организма. Эти консервативные мостиковые кластеры (wat1, wat2, wat3, wat4) содержат в себе от 17 до 32 молекул воды в каждом.

Для структуры *A. thaliana* (1qne) они соответствуют молекулам воды НОН2087А (wat1), НОН2083А (wat2), НОН2029D (wat3), НОН2027С (wat4) (таблица 3).

Организм	<i>Homo sapiens</i>	<i>Arabidopsis thaliana</i>	<i>Saccharomyces cerevisiae</i>	<i>Pyrococcus woesei</i>
Общая формула водного мостика согласно единой нумерации				
	Соответствующие водные мостики			
Lys#66 <u>Wat1</u> A#3:r	1cdw:	1qne:	1nh2:	1ais:
	LYS221:A	LYS85:A	LYS127:A	LYS71:A
	НОН1083:A	НОН2087:A	НОН24:F A6:F	НОН1585:E
Gly#64 <u>Wat2</u> T#4:r	A110:C	A224:D		A1425:E
	1nvp:	1qne:	1ytb:	1d3u:
	GLY223:A	GLY83:A	GLY125:A	GLY69:A
Arg/Lys#49 <u>Wat3</u> T#4:r	НОН345:A	НОН2083:A	НОН265:A	НОН241:A
	T6:F	T223:D	T5:F	T1436:D
	1cdw:	1qne:	1nh2:	1d3u:
Lys#59 <u>Wat4</u> G/A#8:d	ARG204:A	LYS68:A	LYS110:A	LYS54:A
	НОН1009:A	НОН2029:D	НОН248:A	НОН238:A
	T109:C	T223:D	T5:F	T1436:D
	1cdw:	1qne:	1ytb:	1d3u:
	LYS214:A	LYS78:A	LYS120:A	LYS64:A
	НОН1265:A	НОН2027:C	НОН432:C	НОН159:C
	G12:B	G210:C	A9:C	A1417:C

Таблица 3. Консервативные кластеры молекул воды, опосредующие водородные связи между белком ТВР и ДНК. В левом столбце используется единая нумерация контактирующих аминокислотных остатков и нуклеотидов. Косая черта разделяет альтернативные контакты, не встречающиеся в одном комплексе. В остальных столбцах приведены контакты с нумерацией в соответствующей записи PDB. В строках указаны гомологичные аминокислотные остатки и нуклеотиды, формирующие водный мостик в одном и том же месте комплекса ТВР-ДНК.

Расположение мостиковых кластеров в ДНК-белковом комплексе несимметрично (рисунок 48): три молекулы воды из кластеров wat1, wat2, wat3 находятся рядом, образуя контакты с сахаро-фосфатным остовом ДНК и Lys#66, Gly#64, Lys#49 соответственно. Наблюдаемая в белке человека аминокислотная замена Lys#49 на Arg не влияет на образование водного мостика через wat3, поскольку белок связывает эту молекулу воды остовным кислородом.

Четвертая молекула воды (wat4), также находится в N-концевом домене, но с другой стороны β -листа. Она опосредует контакт Lys#59 с сахаро-фосфатным остовом ДНК.

Среди симметрично расположенных аминокислотных остатков С-концевого домена (Gly#1064, Lys#1066, Lys#1049 и Lys#1059) только Gly#1064 образует водный мостик с ДНК. Кластер, в состав которого входит соответствующая молекула воды (НОН2014 для 1qne), является достаточно консервативным — он содержит в себе 13 молекул воды из структур белков трех организмов (всех, кроме архей). Таким образом, мостиковые молекулы воды нарушают симметричность узнавания ТВР ДНК. С N-концевым доменом образуется на одну прямую и три опосредованных консервативных водородных связи больше, чем с С-концевым доменом, взаимодействующим с последовательностью ТАТА.

Для того, чтобы проверить, были ли консервативные мостиковые молекулы воды иммобилизованы на поверхности ТВР до образования комплекса с ДНК, был произведен поиск водных кластеров на поверхности структур ТВР, закристаллизованных без ДНК. Было проанализировано 16 структур ТВР из 7 записей PDB, а именно: 1mp9_A/B (*S. acidocaldarius*), 1pcz_A/B (*P. woesei*), 1vok_A/B (*A. thaliana*), 2z8u_A/B/P/Q (*M. jannaschii*), 3eik_A/B, 3oc3_C/D, 3oci_A/B (*E. cuniculi*).

На всей поверхности ТВР было найдено 17 кластеров молекул воды, в состав которых входили молекулы воды из всех пяти организмов. Для 12 кластеров были найдены аналогичные кластеры в структурах ТВР, находящихся в комплексе с ДНК. Однако ни один из этих кластеров не содержит мостиковых молекул воды. Таким образом можно предположить, что в процессе частичной дегидратации ТВР при образовании комплекса с ДНК происходит перераспределение гидратной оболочки вокруг β -листа. Тем не менее, возможно, что иммобилизованные на белке консервативные молекулы воды участвуют в формировании водных мостиков при контакте с другими транскрипционными факторами.

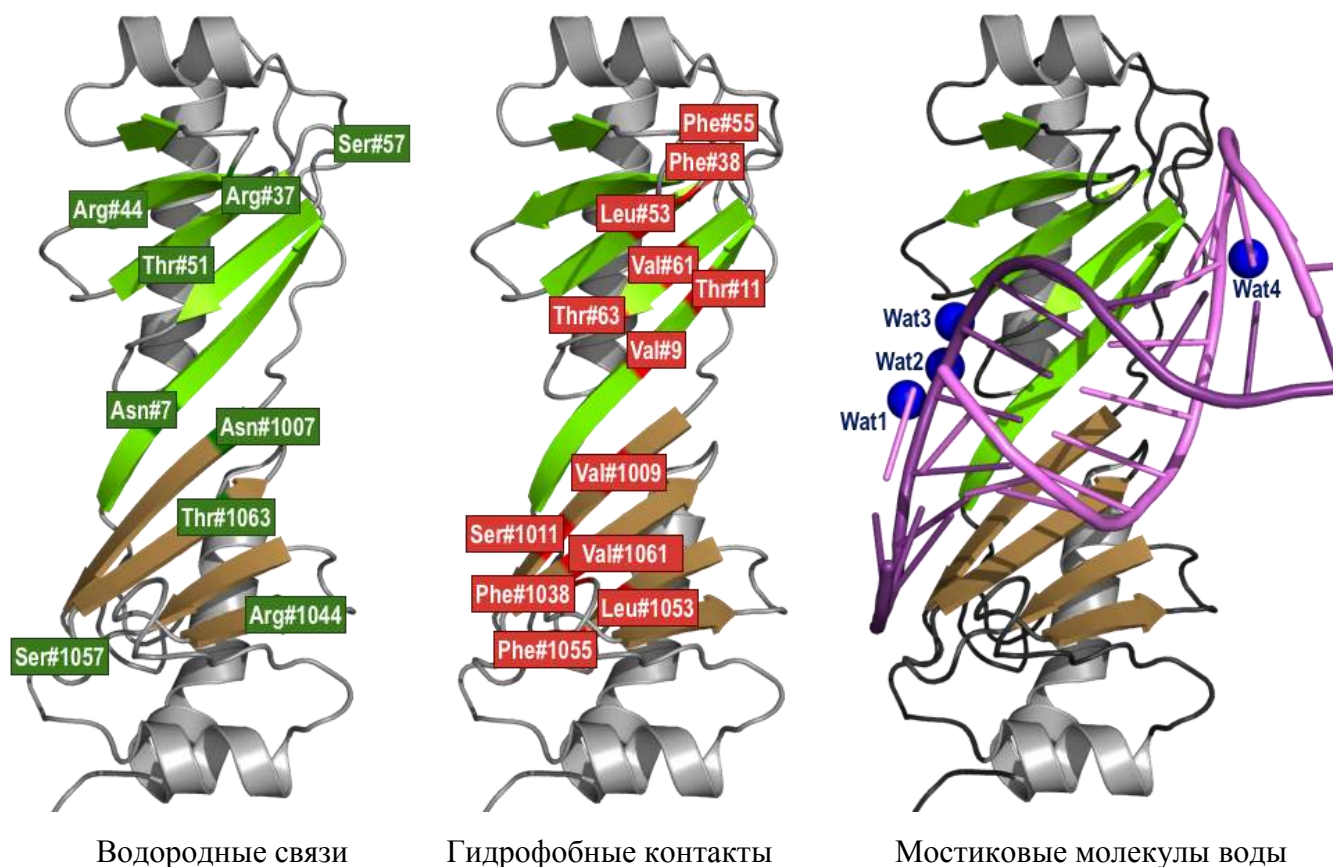


Рисунок 48. Расположение водородных связей, гидрофобных контактов и мостиковых кластеров в структуре ТАТА-бокс связывающего белка.

5.1.7 Консервативность контактов со стороны ДНК

Несмотря на то, что последовательности ТАТА-box ДНК в рассмотренных комплексах в ряде случаев отличаются по нуклеотидному составу от классической последовательности Т-А-Т-А-а/т-А-а/т-Н (Н — любой нуклеотид), можно выявить консервативные контакты со стороны ДНК. Со стороны сахара-фосфатного остова консервативными контактами можно считать: взаимодействие фосфата нуклеотида 4 прямой цепи с аргинином 1044 и симметричное взаимодействие фосфата нуклеотида 5 обратной цепи с аргинином 44 (с тем же фосфатом у эукариотических ТВА взаимодействует Thr#51), взаимодействие фосфата нуклеотида 8 прямой цепи с серином 57 и симметричное взаимодействие фосфата нуклеотида 1 обратной цепи с серином 1057 (см. таблицу 3 и рисунок 47). С фосфатом нуклеотида 6 по

обратной цепи во всех структурах комплексов ТВР эукариот и с фосфатом нуклеотида 7 в части таких структур взаимодействует аргинин 37.

Единственная полностью консервативная водородная связь азотистого основания образуется между пятым основанием обратной цепи и аспарагином 7. В большинстве случаев она дополняется взаимодействием того же аспарагина с четвертым основанием обратной цепи. Симметричные контакты четвертого и пятого оснований прямой цепи с аспарагином 1007 реализуются примерно в половине структур. Ещё один частично консервативный контакт образует четвертое основание прямой цепи с треонином 1063, у архей вместо него имеется контакт третьего основания прямой цепи с серином 1063.

В образовании двух консервативных гидрофобных кластеров принимают участие все основания прямой цепи и основания 1, 2, 3, 7 обратной цепи. При этом основания 5–8 прямой цепи и 7 обратной цепи образуют гидрофобный кластер (кластеры) на интерфейсе с N-концевым доменом ТВР, а основания 1–5 прямой цепи и 1–3 обратной цепи — на интерфейсе с С-концевым доменом ТВР.

Три из четырёх консервативных водных мостика связывают N-концевой домен ТВР с фосфатами третьего и четвертого нуклеотидов обратной цепи ДНК. Четвертый водный мостик связывает лизин 59 с фосфатом восьмого нуклеотида прямой цепи.

5.1.8 Биологическая роль консервативных контактов ТВР с ДНК

Рассмотренные контакты ТАТА-бокс связывающих белков с ДНК обладают высокой консервативностью как в различных структурах ДНК-белковых комплексов одного и того же белка, так и в структурах комплексов различных белков из разных организмов. Такая степень консервативности позволяет предположить, что выявленные прямые водородные связи, водородные связи, опосредуемые молекулами воды, и гидрофобные контакты являются ключевыми для узнавания ДНК ТАТА-бокс связывающими белками. Девять аминокислотных остатков, образу-

щих прямые водородные связи, 13 аминокислотных остатков, участвующих в формировании двух больших гидрофобных кластеров и 4 водных мостика образуют плотную сеть контактов, покрывающую большую часть β -листа, контактирующего с малой бороздкой ДНК. Распределение контактов на интерфейсе ДНК-ТВР соответствует квазисимметрии N- и C-концевых доменов белка. Хотя большинство контактов симметрично, N-концевой домен более плотно связан с ДНК, образуя дополнительные прямые и опосредованные водой водородные связи. При этом N-концевой домен контактирует с менее консервативной половиной узнаваемого участка ДНК, в то время как C-концевой домен, формирующий меньшее количество контактов, взаимодействует с консервативной половиной, имеющей последовательность TATA.

5.2 Семейство LAGLIDADG homing эндонуклеаз

Семейство LAGLIDADG хоминг эндонуклеаз (ХЭ) представляет собой редкостеpping ферменты, узнающие длинные последовательности (14-40 п.н.) закодированные в интронах группы I и II, а также в интронах архей, а также присутствующие в виде интеинов. Последовательности, кодирующие ХЭ часто встречаются в геномах бактерий, архей, митохондрий, пластид и ядерных геномов одноклеточных эукариот. Эти ферменты используют аппарат гомологичной рекомбинации хозяина чтобы способствовать размножению кодирующих их последовательностей *in vivo*. Если в клетке один аллель гена содержит, а другой не содержит ХЭ-кодирующий интрон/интеин, хоминг эндонуклеаза узнает и вносит разрыв в лишенный данного интрона аллель гена, и, таким образом, инициируют перенос интрона в разрезаемый участок ДНК [131].

В связи с чрезвычайной специфичностью эндонуклеаз, они были предложены в качестве инструмента точного узнавания генов и генной инженерии. Несмотря на значительное число охарактеризованных хоминг эндонуклеаз (135 согласно

базе эндонуклеаз ReBase [132]), количество разрезаемых последовательностей все же недостаточно для эффективного использования. Таким образом, необходимо разрабатывать подходы к дизайну новых эндонуклеаз с заданной специфичностью.

Из пяти семейств хоминг эндонуклеаз семейство LAGLIDADG является наиболее широко представленным и изученным. Название семейства происходит от единственного консервативного мотива, имеющего аминокислотную последовательность LAGLIDADG. Каталитически активная единица состоит из двух так называемых LAGLIDADG доменов, каждый из которых имеет копию мотива LAGLIDADG и узнает половину сайта связывания с ДНК [131], [133]. Каждый домен состоит из β -листа, взаимодействующего с большой бороздкой ДНК, и нескольких α -спиралей, расположенных над β -листом. Каждый из доменов вносит равный вклад в формирование активного сайта, расположенного рядом с поверхностью димеризации двух доменов [134]. Закодированные в интронах LAGLIDADG эндонуклеазы могут также содержать и одну копию мотива LAGLIDADG и действовать в виде гомодимеров, разрезая палиндромные или почти палиндромные последовательности ДНК. Если эти эндонуклеазы содержат две копии мотива LAGLIDADG, то они формируют псевдо-симметричные мономеры, без ограничений для последовательности узнаваемого сайта рестрикции. Закодированные в интронах эндонуклеазы этого семейства всегда содержат две копии мотива и действуют в виде мономера, однако кроме двух доменов LAGLIDADG эти ферменты также содержат дополнительные домены, например ДНК-связывающий или ДНК-сплайсирующий домен [133].

Большинство попыток изменения специфичности хоминг эндонуклеаз относится к семейству LAGLIDADG. Существуют два основных подхода: (1) изменение ДНК-белкового интерфейса для адаптации к новой последовательности ДНК и (2) совмещение в новых эндонуклеазах доменов из разных эндонуклеаз, то есть

комбинация узнаваемых последовательностей. Эти подходы не исключают друг друга и могут быть успешно совмещены [135], [136].

Далее приведён анализ консервативных особенностей PFAM семейства LAGLIDADG_1 (PF00961) хоминг эндонуклеаз. Поиск консервативных особенностей ДНК-белкового контакта позволяет идентифицировать наиболее важные для узнавания и связывания с ДНК остатки белка.

5.2.1 Структуры LAGLIDADG_1 хоминг эндонуклеаз

PFAM семейство LAGLIDADG_1 (PF00961) хоминг эндонуклеаз представлено на данный момент 48 пространственными структурами, относящимися к 10 организмам: *C. reinhardtii* (1bp7, 1g9y, 1g9z, 1n3e, 1mow, 1n3f, 1t9i, 1t9j, 1u0c, 1u0d, 2i3p, 2i3q, 2vbj, 2vbl, 2vbn, 2vbo, 2xe0, 3mx9, 3mxa, 3mxb, 4aab, 4aad, 4aae, 4aaf, 4aag, 4aqu, 4aqx), *L. truncatum* (3r7p, 4lq0), *A. nidulans* (1p8k, 2qoj, 3eh8), *C. moewusii* (2ex5), *F. graminearum* (4efj, 4z1x), *Monomastix sp. OKE-1* (1m5x, 2fld, 3fd2, 3ko2), *S. macrospora* (4lox), *O. novo-ulmi* (3qqy), *V. distributa* (3e54), *C. vulgaris* (5a72, 5a74, 5a77, 5a78), а также синтетические конструкции (3mir, 3mis), экспрессированные в *E. coli*.

Поскольку в каждой структуре присутствует как минимум два практически идентичных домена, а некоторые структуры представлены несколькими биологическими единицами, то для анализа доступно 116 структур LAGLIDADG_1 доменов в комплексе с ДНК. Первая попытка поиска консервативных контактов среди LAGLIDADG_1 хоминг эндонуклеаз [137] показала отсутствие таковых. Это связано с тем, что на момент выполнения работы было расшифровано всего пять структур, содержащих эндонуклеазу в контакте с ДНК. Возросшее количество структур позволило выявить особенности узнавания белковым доменом ДНК.

5.2.2 Консервативные контакты LAGLIDADG_1 хоминг эндонуклеаз с ДНК

Поиск консервативных контактов был произведен аналогично семейству TATA-box связывающего белка. Как и в случае с TBP для удобства анализа была введена единая нумерация аминокислотных остатков белковых доменов LAGLIDADG_1 хоминг эндонуклеаз. Несмотря на название семейства, мотив LAGLIDADG в нём присутствует не полностью, а в виде консенсусной последовательности LAGFVDGDG, с которой и начинается единая нумерация аминокислотных остатков.

Поскольку последовательности эндонуклеазы LAGLIDADG_1 в структурах, относящихся к одному и тому же организму часто идентичны, то в выравнивании на рисунке 49 представлены не все, а только отличающиеся последовательности первых 60 аминокислотных остатков доменов LAGLIDADG_1. Несмотря на невысокое сходство последовательностей, структуры хоминг эндонуклеаз LAGLIDADG_1 совмещаются с высокой точностью (порядка 60% C_{α} атомов аминокислотных остатков белкового домена выравниваются друг с другом, а среднее квадратичное отклонение при попарном сравнении структур составляет в среднем 1.26 Å).

В отличие от семейства TPB, в LAGLIDADG_1 хоминг эндонуклеазах консервативные водородные связи (как прямые, так и опосредованные молекулами воды) и гидрофобные кластеры образуются в основном на коротком участке первого тяжа β -листа белкового домена. При этом аминокислотные остатки, находящиеся в одной и той же позиции, могут образовывать как полярные (прямые водородные и опосредованные через воду), так и гидрофобные контакты с ДНК. Замена аминокислотного остатка в большинстве случаев не влияет на факт наличия консервативного контакта в данной позиции. Поэтому в случае с LAGLIDADG_1 хоминг эндонуклеазами можно говорить в первую очередь не о консервативных аминокислотных остатках, а о консервативных позициях в домене, тем или иным

образом контактирующих с ДНК. Поскольку хоминг эндонуклеазы обладают высокой специфичностью в узнавании последовательностей ДНК, то даже в рамках одного нельзя выделить консервативно узнаваемые нуклеотидные паттерны. Контактующий с ДНК участок β -листа взаимодействует с 5 -7 нуклеотидами подряд, как с основаниями, так и с сахаро-фосфатным остовом.

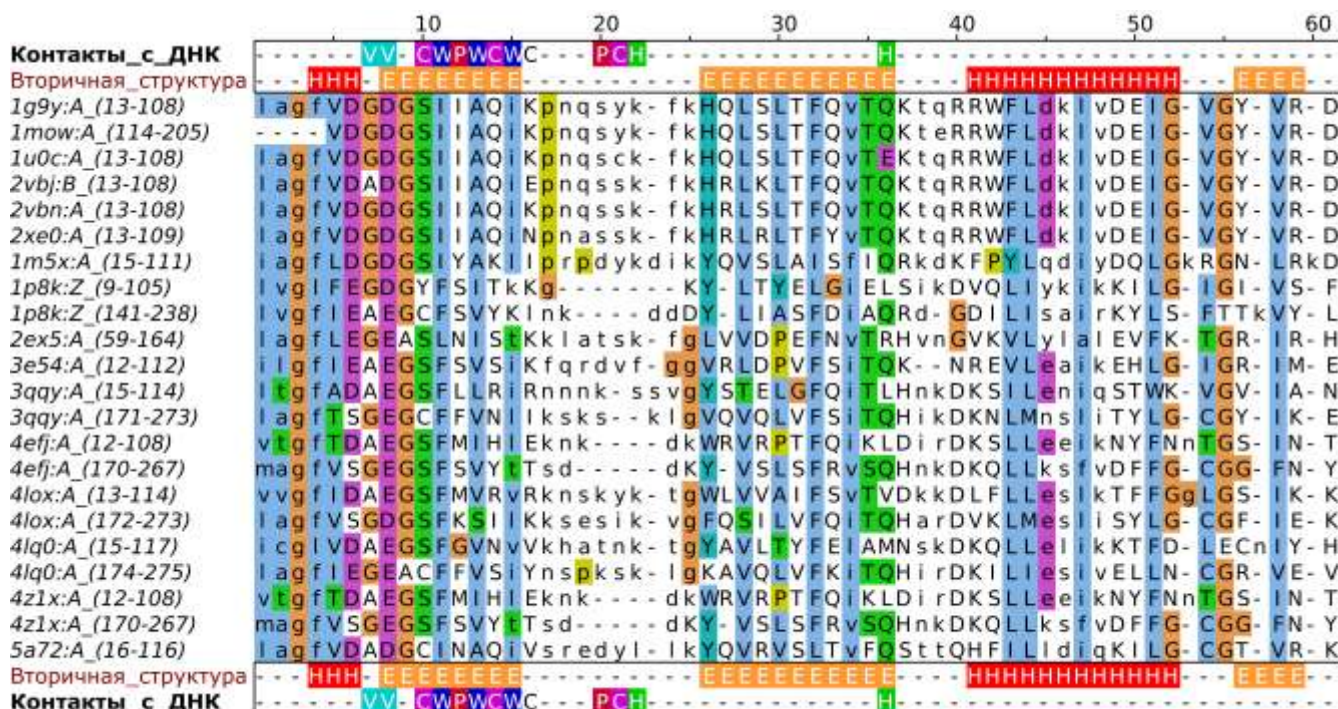


Рисунок 49. Выравнивание последовательностей доменов LAGLIDADG_1 хоминг эндонуклеаз из 10 организмов: *C. reinhardtii* (1g9y, 1mow, 1u0c, 2vbj, 2vbn, 2xe0), *L. truncatum* (4lq0), *A. nidulans* (1p8k), *C. moewusii* (2ex5), *F. graminearum* (4efj, 4z1x), *Monomastix sp. OKE-1* (1m5x), *S. macrospora* (4lox), *O. novo-ulmi* (3qqy), *V. distributa* (3e54), *C. vulgaris* (5a72). В верхней строке приведена единая нумерация аминокислотных остатков LAGLIDADG_1 хоминг эндонуклеаз. В верхней и нижней строках отмечены консервативные контакты с ДНК: Н — водородные связи, Р — гидрофобные кластеры, W — водородные связи, опосредуемые молекулами воды. Отдельно указана консенсусная вторичная структура: Н — α -спирали, Е — бета-тяжи

В таблице 4 приведен список консервативных контактирующих позиций в домене LAGLIDADG_1 с учетом единой нумерации. Указана природа контакта, аминокислотные остатки, встречающиеся в данной позиции, а также контактирующий элемент со стороны ДНК.

Природа контакта	Позиция в домене LAGLIDADG_1	Аминокислота	Контактирующий элемент ДНК
водородная связь / водный мостик	7	Gly/Ala/Ile	с/ф остов
водородная связь / водный мостик	8	Glu/Asp/Asn	с/ф остов
любой контакт	10	Ser / Cys / Gln	с/ф остов или большая бороздка
водный мостик	11	Phe / Ile	с/ф остов
гидрофобный	12	Ile / Lys / Met /Phe /Ser / Tyr	большая бороздка
водный мостик	13	Ala/Val/Ile/Leu	с/ф остов
любой контакт	14	Arg / Asn / Gln / His / Ile / Leu / Lys / Ser / Thr / Tyr	с/ф остов или большая бороздка
водный мостик	15	Ile / Glu/ Leu / Lys/ Thr/ Val	с/ф остов
любой контакт	16	Lys / Ala / Arg / Glu / Ile / Leu / Ser / Thr/ Tyr / Val	большая бороздка
гидрофобный	20	Ser / Asp / Thr / Lys	большая бороздка
любой контакт	21	Tyr/ Lys/ Cys/ Ser/ Val / Ile/ Asn / Arg/ Glu / His	с/ф остов или большая бороздка
водородная связь	22	Lys/ Ser/ Asn / Gln Tyr/ Leu	с/ф остов
водородная связь	35	Gln / Arg / Leu / Glu	с/ф остов
водородная связь	58	Arg / Ser/ Lys/ Asp	большая бороздка

Таблица 4. Консервативные контакты в домене LAGLIDADG_1 с учетом единой нумерации.

Рассмотрим расположение контактирующих с ДНК остатков на примере структуры LAGLIDADG_1 хоминг эндонуклеазы I-CreI (1t9i, цепь A13-108). На рисунке 50 контактирующие с ДНК остатки отмечены разными цветами в зависимости от природы контакта.

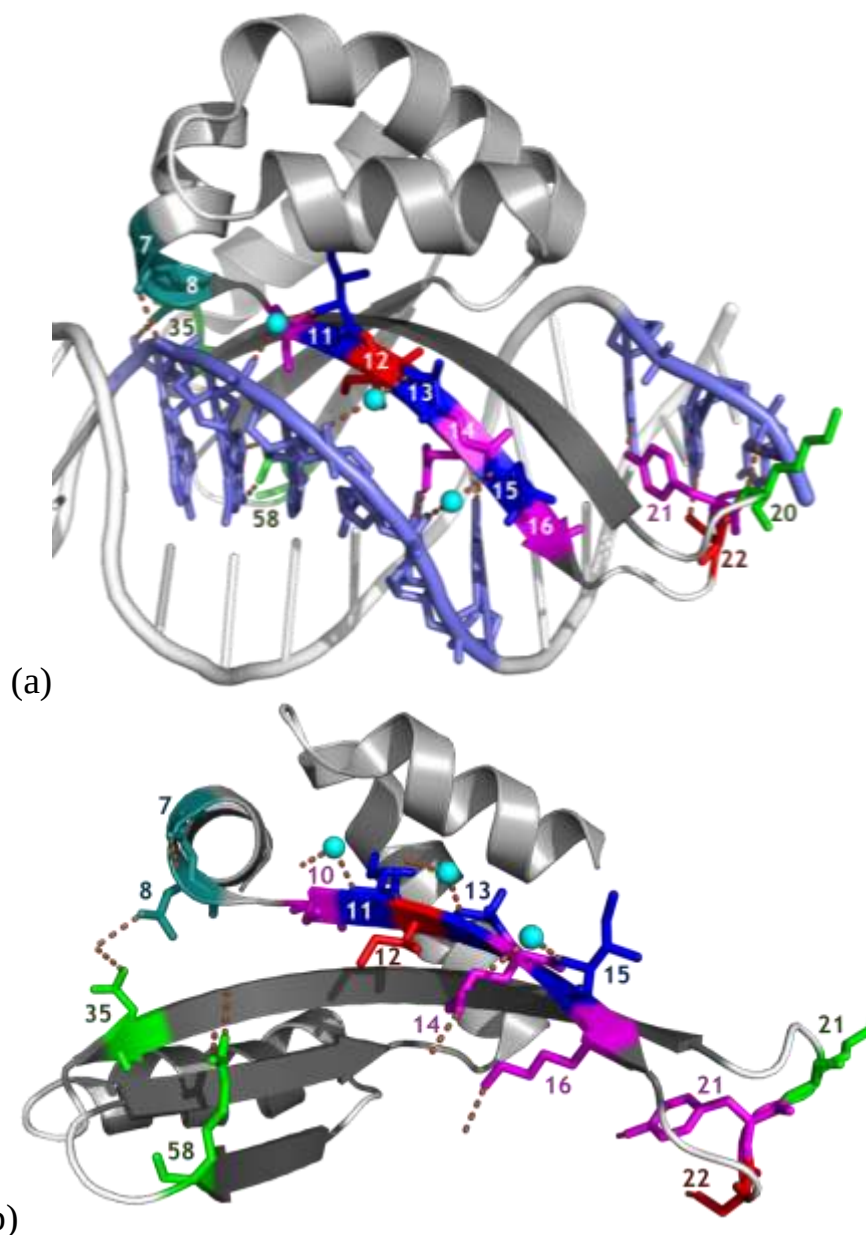


Рисунок 50. Структура домена LAGLIDADG_1 хоминг эндонуклеазы I-CreI (1t9i, цепь A: 13-108) (a) в комплексе с ДНК, (b) без ДНК. Контактующие с ДНК остатки обозначены согласно единой нумерации. Цвет контактирующего аминокислотного остатка зависит от природы контакта с ДНК: зеленый(№21,35,58) — водородные связи, синий(11, 13, 15) — водородные связи, опосредованные молекулами воды, голубые (№7, 8)— прямые или опосредованные водой водородные связи, красные (№ 12, 22) — гидрофобные контакты, розовые (10, 14, 16, 21) — контакты любой природы.

Как видно из таблицы 4 и рисунка 50 контактирующие с ДНК аминокислотные остатки расположены на первом тяжё β-листа, а также в крайних точках β-листа. Аминокислотные остатки β-листа, контактирующие с ДНК посредством боковых радикалов, взаимодействуют не только посредством прямых водородных

связей (позиции 35 и 58), но посредством гидрофобных контактов (позиция 12). В позициях 10, 14 и 16 консервативный контакт сохраняется, изменяется только его природа, которая может быть как полярной так и гидрофобной. Контакты, опосредованные молекулами воды, встречаются через один (позиции 11, 13, 15). Они характерны для тех остатков β -листа, у которых боковые радикалы направлены внутрь белка и не могут контактировать с ДНК. Вместо этого контакт обеспечивается молекулой воды, образующей водородную связь с белковым остовом и с ДНК. В целом, расположение консервативных контактирующих аминокислотных остатков совпадает с зоной наиболее плотного контакта β -листа с ДНК.

Из приведенного выше анализа видно, что в доменах семейства LAGLIDADG_1 хоминг эндонуклеаз хотя и невозможно выделить закономерности узнавания ДНК на уровне последовательностей, но узнавание ДНК на уровне пространственной структуры сохраняется. Этот результат согласуется с необходимостью узнавания различных последовательностей ДНК этими эндонуклеазами. Тем не менее видно, что при попытках изменения специфичности узнавания LAGLIDADG_1 хоминг эндонуклеаз необходимо в первую очередь рассматривать аминокислотные остатки, контактирующие с основаниями ДНК (позиции 10, 12, 14, 16, 20, 21, 58), а также учитывать ключевую роль первого тяжа β -листа каждого из доменов в узнавании ДНК.

Заключение

В данной работе была разработана процедура классификации как отдельных ДНК-связывающих белковых доменов, так и целых их семейств. Процедура была применена для анализа 1942 структур комплексов 314 белковых доменов SCOP с ДНК, а также для описания 34 семейств. В результате было выделено 97 способов взаимодействия белкового домена с ДНК в конкретных структурах комплексов и 17 классов взаимодействия, характеризующих консервативные контакты с ДНК внутри семейств.

В рамках предложенной классификации был произведен поиск закономерностей ДНК-белкового узнавания на уровне структурных доменов SCOP и их семейств. Характерные черты ДНК-белкового взаимодействия были проанализированы: для представителей одного и того же домена; для разных доменов, принадлежащих одному и тому же способу взаимодействия; внутри семейств белковых доменов; внутри одного класса взаимодействия. Кроме того описаны закономерности строения белковых доменов в рамках одного способа взаимодействия с ДНК и закономерности строения семейств белковых доменов, относящихся к одному и тому же классу взаимодействия.

Предложенная классификация была внедрена в базу данных НК-белковых взаимодействий NPIDB. В базе данных есть возможность поиска структур с одинаковым способом узнавания ДНК, а также поиск семейств белковых доменов, относящихся к одному классу взаимодействия с ДНК. Страницы семейств белковых доменов SCOP, содержащие большое количество структур, были дополнены развернутым описанием. Кроме того для 72 семейств ДНК-узнающих белковых доменов в структурах их комплексов с ДНК выявлены консервативно расположенные молекулы воды. Данные также были интегрированы в NPIDB.

Основываясь на разработанной классификации и используя функционал базы данных NPIDB, был проведен структурный анализ консервативных особенностей гомологичных белков. В частности, для семейства транскетолаз была рассмотрена роль консервативных молекул воды. Для семейств TATA-box связывающих белков и белков семейства LAGLIDADAG_1 были найдены и проанализированы консервативные контакты: прямые и опосредованные через воду водородные связи, а также гидрофобные кластеры.

Список публикаций

Статьи в научных журналах

- I. Aksianov E., Zanegina O., Grishin A., Spirin S., Karyagina A., Alexeevski A. Conserved water molecules in X-ray structures highlight the role of water in intramolecular and intermolecular interactions. // J. Bioinform. Comput. Biol. – 2008. – Vol. 6. – № 4. P. 775–788.
- II. Grishin A., Fonfara I., Alexeevski A., Spirin S., Zanegina O., Karyagina A., Alexeyevsky D., Wende W. Identification of conserved features of LAGLIDADG homing endonucleases. // J. Bioinform. Comput. Biol. – 2010. – Vol. 8. – № 3. P. 453–469.
- III. Kirsanov D., Zanegina O., Aksianov E., Spirin S., Karyagina A., Alexeevski A. NPIDB: Nucleic acid-Protein Interaction DataBase. // Nucleic Acids Res. – 2013. – Vol. 41. P. D517–D523.
- IV. Zanegina O., Kirsanov D., Baulin E., Karyagina A., Alexeevski A., Spirin S. An updated version of NPIDB includes new classifications of DNA-protein complexes and their families. // Nucleic Acids Res. Database issue. – 2015. – Doi: 10.1093/nar/gkv1339.

Тезисы научных конференций

- V. Занегина О. Анализ первичной структуры и моделирование пространственной структуры транскетолазы человека TKTL1. // Тезисы Международной конференции молодых ученых по фундаментальным наукам "ЛОМОНОСОВ-2006". – Москва. – 2006. – С. 40.
- VI. Aksianov E., Zanegina O., Alexeevski A., Karyagina A., Spirin S. A tool for comparative analysis of solvent molecules in PDB structures. // Proceedings of the fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'06). – Novosibirsk. – Vol. 1. – P. 226.
- VII. Aksianov E., Alexeevski A., Spirin S., Zanegina O., Karyagina A., Water-Mediated Interactions between Macromolecules. // Proceedings of the 3rd Moscow Conference on Computational Molecular Biology. – Moscow. – 2007. – P. 28.
- VIII. Zanegina O., Aksianov E., Alexeevski A., Karyagina A., Spirin S. Detecting conserved water molecules in protein-DNA complexes by comparative analysis of X-ray structures. // Proceedings of the sixth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'08). – Novosibirsk. – Vol. 1. – P. 264.

- IX. Zanegina O., Karyagina A., Alexeevski A., Spirin S. Contact-based approach to structural classification of protein-DNA complexes. Proceedings of the ninth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'14). – Novosibirsk. – Vol. 1. – P. 170.
- X. Zanegina O., Karyagina A., Alexeevski A., Spirin S. Structural classification of protein-DNA complexes and their families based on interacting elements. The 13th European Conference On Computational Biology (ECCB '14). – 2014. –Strasbourg. – Poster abstract E32.

Список литературы

1. Mirny L., Gelfand M.S. What evolution can tell us about protein-DNA interactions // NATO Sci. Ser. SUB Ser. I LIFE Behav. Sci. Citeseer, – 2001. – Vol. 333. – P. 211–226.
2. Kirsanov D.D. et al. NPIDB: Nucleic acid-Protein Interaction DataBase. // Nucleic Acids Res. – 2013. – Vol. 41. – P. D517–D523.
3. Singh J., Thornton J.M. Atlas of protein side-chain interactions. – Vols. I and II. // Acta Cryst. – 1993. – Vol. 49. – P. 355–356.
4. Dixit S.B., Jayaram B. Role of hydrogen bonds in protein-DNA recognition: a comparison of generalized born and finite difference Poisson-Boltzmann solvation treatments // J. Biomol. Struct.& Dyn. – 1998. – Vol. 16. – P. 237–242.
5. Coulocheri S. a. et al. Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity // Biochimie. – 2007. – Vol. 89. – № 11. – P. 1291–1303.
6. Mandel-Gutfreund Y., Schueler O., Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. // J. Mol. Biol. – 1995. – Vol. 253. – № 2. – P. 370–382.
7. Luscombe N.M., Laskowski R. a, Thornton J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. // Nucleic Acids Res. – 2001. – Vol. 29. – № 13. – P. 2860–2874.
8. Teeter M.M. Water-protein interactions: theory and experiment // /Annual review of biophysics and biophysical chemistry. – 1991. – Vol. 20. – № 1. – P. 577–600.
9. Otting G. et al. Protein Hydration in Aqueous Solution // Science. – 1991. – Vol. 254. – № 5034. – P. 974–980.
10. Westhof E. Water: an integral part of nucleic acid structure // /Annual review of biophysics and biophysical chemistry. – 1988. – Vol. 17. – № 1. – P. 125–144.
11. Eisenberg D., McLachlan A.D. Solvation energy in protein folding and binding. // Nature. – 1986. – Vol. 319. – № 6050. – P. 199–203.
12. Kuntz I.D., Kauzmann W. Hydration of Proteins and Polypeptides // Adv. Protein Chem. – 1974. – Vol. 28. – № C. – P. 239–345.
13. Otwinowski Z. et al. Crystal structure of trp repressor/operator complex at atomic resolution. // Nature. – 1988. – Vol. 335. – № 6188. – P. 321–329.

14. Joachimiak A, Haran T.E., Sigler P.B. Mutagenesis supports water mediated recognition in the trp repressor-operator system. // EMBO J. – 1994. – Vol. 13. – № 2. – P. 367–372.
15. Wilson I. a, Fremont D.H. Structural analysis of MHC class I molecules with bound peptide antigens. // Seminars in immunology. – 1993. – Vol. 5. – № 2. – P. 75–80.
16. Marc Lett C. et al. The role of a conserved water molecule in the redox-dependent thermal stability of iso-1-cytochrome c // J. Biol. Chem. – 1996. – Vol. 271. – № 46. – P. 29088–29093.
17. Fauman E.B. et al. Water-mediated substrate/product discrimination: the product complex of thymidylate synthase at 1.83 Å. // Biochemistry. – 1994. – Vol. 33. – № 6. – P. 1502–1511.
18. Likić V. a et al. A “structural” water molecule in the family of fatty acid binding proteins. // Protein Sci. – 2000. – Vol. 9. – № 3. – P. 497–504.
19. Mustata G., Briggs J.M. Cluster analysis of water molecules in alanine racemase and their putative structural role // Protein Eng. Des. Sel. – 2004. – Vol. 17. – № 3. – P. 223–234.
20. Karyagina A. et al. The role of water in homeodomain-DNA interaction // Bioinformatics of Genome Regulation and Structure II. Springer, – 2006. – P. 247–257.
21. Гришин А.В. et al. Консервативные особенности структур комплексов доменов семейства ETS с ДНК // Молекулярная биология. – 2009. – Vol. 43. – № 4. – P. 666–674.
22. Jayaram B., Jain T. The role of water in protein-DNA recognition. // Annu. Rev. Biophys. Biomol. Struct. – 2004. – Vol. 33. – P. 343–361.
23. Reddy C.K., Das A, Jayaram B. Do water molecules mediate protein-DNA recognition? // J. Mol. Biol. – 2001. – Vol. 314. – № 3. – P. 619–632.
24. Janin J. Wet and dry interfaces: The role of solvent in protein-protein and protein-DNA recognition // Structure. – 1999. – Vol. 7. – № 12.
25. Wilson D.S. et al. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. // Cell. – 1995. – Vol. 82. – № 5. – P. 709–719.
26. Tucker-Kellogg L. et al. Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. // Structure. – 1997. – Vol. 5. – № 8. – P. 1047–1054.
27. Schwabe J.W. The role of water in protein-DNA interactions. // Curr. Opin. Struct. Biol. – 1997. – Vol. 7. – № 1. – P. 126–134.
28. Ogata K., Wodak S.J. Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. // Protein Eng. – 2002. – Vol. 15. – № 8. – P. 697–705.
29. Bella J., Brodsky B., Berman H.M. Hydration structure of a collagen peptide. // Structure. – 1995. – Vol. 3. – № 9. – P. 893–906.

30. Poornima C.S., Dean P.M. Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. // *J. Comput. Aided Mol. Des.* – 1995. – Vol. 9. – № 6. – P. 521–531.
31. Bottoms C. a, Smith P.E., Tanner J.J. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. // *Protein Sci.* – 2002. – Vol. 11. – № 9. – P. 2125–2137.
32. Aksianov E. et al. Conserved water molecules in X-ray structures highlight the role of water in intramolecular and intermolecular interactions. // *J. Bioinform. Comput. Biol.* – 2008. – Vol. 6. – № 4. – P. 775–788.
33. Финкельштейн А.В., Птицын О.Б. Физика белка: Курс лекций с цветными и стереоскопическими иллюстрациями.// А.В. Финкельштейн, О.Б. Птицын – Москва: Книжный Дом“ Университет,” – 2002. – 63 с.
34. Волькенштейн М.В. Биофизика.// М.В. Волькенштейн – Москва: Наука – 1988. – 105 с.
35. Monera O.D. et al. Relationship of sidechain hydrophobicity and alpha-helical propensity on the stability of the single-stranded amphipathic alpha-helix. // *J. Pept. Sci.* – 1995. – Vol. 1. – № 5. – P. 319–329.
36. Alexeevski A., Spirin S., Alexeevski D. K.O., Ershova A., Titov M. K.A. CluD, a Program for the Determination of Hydrophobic Clusters in 3D Structures of Protein and Protein-Nucleic Acids Complexes. // *Biophys.* – 2004. – Vol. 48. – № (Suppl. 1). – P. 146–156.
37. Ha J.H., Spolar R.S., Record M.T. Role of the hydrophobic effect in stability of site-specific protein-DNA complexes. // *J. Mol. Biol.* – 1989. – Vol. 209. – № 4. – P. 801–816.
38. Norberg J. Association of protein-DNA recognition complexes: electrostatic and nonelectrostatic effects. // *Arch. Biochem. Biophys.* – 2003. – Vol. 410. – № 1. – P. 48–68.
39. Tolstorukov M.Y., Jernigan R.L., Zhurkin V.B. Protein-DNA Hydrophobic Recognition in the Minor Groove is Facilitated by Sugar Switching // *J. Mol. Biol.* – 2004. – Vol. 337. – № 1. – P. 65–76.
40. Zhitnikova M.Y., Boryskina O.P., Shestopalova A. V. Sequence-specific transitions of the torsion angle gamma change the polar-hydrophobic profile of the DNA grooves: implication for indirect protein-DNA recognition. // *J. Biomol. Struct. Dyn.* – 2013. – Vol. 1. – P. 37–41.
41. Karyagina A. et al. Analysis of conserved hydrophobic cores in proteins and supramolecular complexes. // *J. Bioinform. Comput. Biol.* – 2006. – Vol. 4. – P. 357–372.
42. Kabsch W., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. // *Biopolymers.* – 1983. – Vol. 22. – № 12. – P. 2577–2637.
43. Andersen C.A.F. et al. Continuum Secondary Structure Captures Protein Flexibility // *Structure.* – 2002. – Vol. 10. – № 2. – P. 175–184.

44. Humphrey W., Dalke A., Schulten K. VMD: Visual molecular dynamics // J. Mol. Graph. – 1996. – Vol. 14. – № 1. – P. 33–38.
45. McDonald I.K., Thornton J.M. Satisfying hydrogen bonding potential in proteins. // J. Mol. Biol. – 1994. – Vol. 238. – № 5. – P. 777–793.
46. Płochocka D., Zielenkiewicz P., Rabczenko A. Hydrophobic microdomains as structural invariant regions in proteins. // Protein Eng. – 1988. – Vol. 2. – № 2. – P. 115–118.
47. Heringa J., Argos P. Side-chain clusters in protein structures and their role in protein folding. // J. Mol. Biol. – 1991. – Vol. 220. – № 1. – P. 151–171.
48. Swindells M.B. A procedure for the automatic determination of hydrophobic cores in protein structures. // Protein Sci. – 1995. – Vol. 4. – № 1. – P. 93–102.
49. Zehfus M.H. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. // Protein Sci. – 1995. – Vol. 4. – № 6. – P. 1188–1202.
50. Tsai C.J., Nussinov R. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. // Protein Sci. – 1997. – Vol. 6. – № 1. – P. 24–42.
51. Kannan N., Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. // J. Mol. Biol. – 1999. – Vol. 292. – № 2. – P. 441–464.
52. Matthews B.W. No code for recognition // Nature. – 1988. – Vol. 335. – P. 294–295.
53. Suzuki M. A Framework for the DNA-Protein Recognition Code of the Probe Helix in Transcription Factors - the Chemical and Stereochemical Rules // Structure. – 1994. – Vol. 2. – № 4. – P. 317–326.
54. Suzuki M., Gerstein M. Binding geometry of alpha-helices that recognize DNA. // Proteins. – 1995. – Vol. 23. – № 4. – P. 525–535.
55. Choo Y., Klug A. Physical basis of a protein-DNA recognition code // Curr. Opin. Struct. Biol. – 1997. – Vol. 7. – № 1. – P. 117–125.
56. Pabo C.O., Nekludova L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? // J. Mol. Biol. – 2000. – Vol. 301. – № 3. – P. 597–624.
57. von Hippel P.H., Berg O.G. On the specificity of DNA-protein interactions. // Proc. Natl. Acad. Sci. U. S. A. – 1986. – Vol. 83. – № 6. – P. 1608–1612.
58. Seeman N.C., Rosenberg J.M., Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. // Proc. Natl. Acad. Sci. U. S. A. – 1976. – Vol. 73. – № 3. – P. 804–808.
59. Ohlendorf D.H. et al. The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. // Nature. – 1982. – Vol. 298. – P. 718–723.

60. Sarai A, Takeda Y. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. // *Proc. Natl. Acad. Sci. U. S. A.* – 1989. – Vol. 86. – № 17. – P. 6513–6517.
61. Suzuki M., Gerstein M., Yagi N. Stereochemical basis of DNA recognition by Zn fingers. // *Nucleic Acids Res.* – 1994. – Vol. 22. – № 16. – P. 3397–3405.
62. Nadassy K., Wodak S.J., Janin J. Structural features of protein-nucleic acid recognition sites. // *Biochemistry.* – 1999. – Vol. 38. – № 7. – P. 1999–2017.
63. Garvie C.W., Wolberger C. Recognition of specific DNA sequences // *Mol. Cell.* – 2001. – Vol. 8. – № 5. – P. 937–946.
64. Rohs R. et al. Origins of specificity in protein-DNA recognition. // *Annu. Rev. Biochem.* – 2010. – Vol. 79. – P. 233–269.
65. Shakked Z. et al. Determinants of repressor/operator recognition from the structure of the trp operator binding site. // *Nature.* – 1994. – Vol. 368. – № 6470. – P. 469–473.
66. Rohs R. et al. The role of DNA shape in protein-DNA recognition. // *Nature.* – 2009. – Vol. 461. – P. 1248–1253.
67. Spolar R.S., Record M.T. Coupling of local folding to site-specific binding of proteins to DNA. // *Science.* – 1994. – Vol. 263. – № 5148. – P. 777–784.
68. Oda M. et al. Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain. // *J. Mol. Biol.* – 1998. – Vol. 276. – № 3. – P. 571–590.
69. Gromiha M.M. et al. Role of inter and intramolecular interactions in protein-DNA recognition // *Gene.* – 2005. – Vol. 364. – P. 108–113.
70. Privalov P.L. et al. What Drives Proteins into the Major or Minor Grooves of DNA? // *J. Mol. Biol.* – 2007. – Vol. 365. – № 1. – P. 1–9.
71. Oda M., Nakamura H. Thermodynamic and kinetic analyses for understanding sequence-specific DNA recognition // *Genes to Cells.* – 2000. – Vol. 5. – № 5. – P. 319–326.
72. Seimiya M., Kurosawa Y. Kinetics of binding of Antp homeodomain to DNA analyzed by measurements of surface plasmon resonance. // *FEBS Lett.* – 1996. – Vol. 398. – № 2-3. – P. 279–284.
73. Livhitz M.A. et al. Equilibrium and kinetic aspects of protein-DNA recognition // *Nucleic Acids Res. Oxford Univ Press*, 1979. – Vol. 6. – № 6. – P. 2217–2236.
74. Murzin A.G. et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. // *J. Mol. Biol.* – 1995. – Vol. 247. – P. 536–540.
75. Sillitoe I. et al. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures // *Nucleic Acids Res.* – 2013. – Vol. 41. – P. 490–498.

76. Harrison S.C. A structural taxonomy of DNA-binding domains. // *Nature*. – 1991. – Vol. 353. – № 6346. – P. 715–719.
77. Pabo C.O., Sauer R.T. Transcription factors: structural families and principles of DNA recognition. // *Annu. Rev. Biochem.* – 1992. – Vol. 61. – P. 1053–1095.
78. Wintjens R.T., Rooman M.J., Wodak S.J. Automatic classification and analysis of alpha alpha-turn motifs in proteins. // *J. Mol. Biol.* – 1996. – Vol. 255. – № 1. – P. 235–253.
79. Wintjens R., Rooman M. Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. // *J. Mol. Biol.* – 1996. – Vol. 262. – P. 294–313.
80. Jones S. et al. Protein-DNA interactions: A structural analysis. // *J. Mol. Biol.* – 1999. – Vol. 287. – № 5. – P. 877–896.
81. Luscombe N.M. et al. An overview of the structures of protein-DNA complexes. // *Genome Biol.* – 2000. – Vol. 1. – № 1. – P. 1–37.
82. Malhotra S., Sowdhamini R. Re-visiting protein-centric two-tier classification of existing DNA-protein complexes // *BMC Bioinformatics*. – 2012. – Vol. 13. – № 1. – P. 165.
83. Ponomarenko J. V, Bourne P.E., Shindyalov I.N. Building an automated classification of DNA-binding protein domains. // *Bioinformatics*. – 2002. – Vol. 18 Suppl 2. – P. S192–S201.
84. Prabakaran P. et al. Classification of Protein-DNA Complexes Based on Structural Descriptors // *Structure*. – 2006. – Vol. 14. – № 9. – P. 1355–1367.
85. Sathyapriya R., Vijayabaskar M.S., Vishveshwara S. Insights into protein-DNA interactions through structure network analysis. // *PLoS Comput. Biol.* – 2008. – Vol. 4. – № 9. – P. e1000170.
86. Biswas S., Guharoy M., Chakrabarti P. Dissection, residue conservation, and structural classification of protein-DNA interfaces. // *Proteins*. – 2009. – Vol. 74. – № 3. – P. 643–654.
87. Schneider B. et al. Bioinformatic Analysis of the Protein/DNA Interface // *Biophys. J.* – 2011. – Vol. 100. – № 3. – P. 69a.
88. Kim R., Guo J. PDA: an automatic and comprehensive analysis program for protein-DNA complex structures. // *BMC Genomics*. – 2009. – Vol. 10 Suppl 1. – P. S13.
89. Berman H.M. et al. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. // *Biophys. J.* – 1992. – Vol. 63. – № 3. – P. 751–759.
90. Narayanan B.C. et al. The Nucleic Acid Database: New features and capabilities // *Nucleic Acids Res.* – 2014. – Vol. 42. – P. 114–122.
91. Sillitoe I. et al. CATH: comprehensive structural and functional annotations for genome sequences. // *Nucleic Acids Res.* – 2015. – Vol. 43. – № Database issue. – P. D376–D381.

92. Swindells M.B. A procedure for detecting structural domains in proteins. // Protein Sci. – 1995. – Vol. 4. – № 1. – P. 103–112.
93. Holm L., Sander C. The FSSP database of structurally aligned protein fold families. // Nucleic Acids Res. – 1994. – Vol. 22. – № 17. – P. 3600–3609.
94. Siddiqui A.S., Barton G.J. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. // Protein Sci. – 1995. – Vol. 4. – № 5. – P. 872–884.
95. An J., Nakama T., Kubota Y. Search Tool for the Structure, Function and Properties of Biomolecules // Bioinformatics. – 1998. – Vol. 14. – № 2. – P. 188–195.
96. Sigrist C.J. et al. New and continuing developments at PROSITE // Nucleic Acids Res. – 2013. – Vol. 41. – № D1. – P. 1–4.
97. Kawabata T., Ota M., Nishikawa K. The protein mutant database // Nucleic Acids Res. – 1999. – Vol. 27. – № 1. – P. 355–357.
98. Hoffman M.M. et al. AANT: the Amino Acid-Nucleotide Interaction Database. // Nucleic Acids Res. – 2004. – Vol. 32. – № Database issue. – P. D174–D181.
99. Kumar M.D.S. et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. // Nucleic Acids Res. – 2006. – Vol. 34. – № Database issue. – P. D204–D206.
100. Lee S., Blundell T.L. BIPA: A database for protein-nucleic acid interaction in 3D structures // Bioinformatics. – 2009. – Vol. 25. – № 12. – P. 1559–1560.
101. Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein-DNA complexes. // Nucleic Acids Res. – 2010. – Vol. 38. – № Database issue. – P. D91–D97.
102. Altschul S.F. et al. Basic local alignment search tool. // J. Mol. Biol. – 1990. – Vol. 215. – № 3. – P. 403–410.
103. Salgado H. et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. // Nucleic Acids Res. – 2013. – Vol. 41. – № Database issue. – P. D203–D213.
104. Matys V. et al. TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. // Nucleic Acids Res. – 2006. – Vol. 34. – № Database issue. – P. D108–D110.
105. Tkachenko M.Y. et al. ProtNA-ASA: Protein-nucleic acid structural database with information on accessible surface area // Int. J. Quantum Chem. – 2010. – Vol. 110. – № 1. – P. 230–232.
106. Norambuena T., Melo F. The Protein-DNA Interface database. // BMC Bioinformatics. – 2010. – Vol. 11. – P. 262.
107. Prakash T. et al. CoPS: Comprehensive Peptide Signature database // Bioinformatics. – 2004. – Vol. 20. – № 16. – P. 2886–2888.

108. Finn R.D. et al. The Pfam protein families database. // *Nucleic Acids Res.* – 2010. – Vol. 38. – P. D211–D222.
109. The Gene Ontology Consortium. Gene Ontology Consortium: going forward // *Nucleic Acids Res.* – 2015. – Vol. 43. – № D1. – P. D1049–D1056.
110. Krissinel E., Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions // *Acta Crystallogr. Sect. D Biol. Crystallogr.* – 2004. – Vol. 60. – № 12 I. – P. 2256–2268.
111. Shindyalov I.N., Bourne P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. // *Protein Eng.* – 1998. – Vol. 11. – № 9. – P. 739–747.
112. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. // *Nucleic Acids Res.* – 2004. – Vol. 32. – № 5. – P. 1792–1797.
113. Heinig M., Frishman D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins // *Nucleic Acids Res.* – 2004. – Vol. 32. – P. 500–502.
114. Waterhouse A.M. et al. Jalview Version 2-A multiple sequence alignment editor and analysis workbench // *Bioinformatics.* – 2009. – Vol. 25. – № 9. – P. 1189–1191.
115. Hanson R. M. et al. Jmol: an open-source Java viewer for chemical structures in 3D. // URL: www.jmol.sourceforge.net.net. – 2008.
116. DeLano W.L. The PyMOL Molecular Graphics System // Schrödinger LLC. – 2002. – Version 1. – URL: <http://www.pymol.org>.
117. Lu X.-J., Olson W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. // *Nucleic Acids Res.* – 2003. – Vol. 31. – № 17. – P. 5108–5121.
118. Tateno M. et al. DNA recognition by beta-sheets. // *Biopolymers.* – 1997. – Vol. 44. – № 4. – P. 335–359.
119. Davis A.M., Teague S.J., Kleywegt G.J. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design // *Angew. Chemie - Int. Ed.* – 2003. – Vol. 42. – № 24. – P. 2718–2736.
120. Schneider G., Lindqvist Y. Crystallography and mutagenesis of transketolase: mechanistic implications for enzymatic thiamin catalysis // *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* – 1998. – Vol. 1385. – № 2. – P. 387–398.
121. Nikkola M., Lindqvist Y., Schneider G. Refined structure of transketolase from *Saccharomyces cerevisiae* at 2.0 Å resolution. // *J. Mol. Biol.* – 1994. – Vol. 238. – № 3. – P. 387–404.
122. Rice P., Longden I., Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite // *Trends Genet.* – 2000. – Vol. 16. – № 1. – P. 276–277.

123. Kim Y. et al. Crystal structure of a yeast TBP/TATA-box complex. // *Nature*. – 1993. – Vol. 365. – № 6446. – P. 512–520.
124. Kim J.L., Nikolov D.B., Burley S.K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. // *Nature*. – 1993. – Vol. 365. – P. 520–527.
125. Patikoglou G. a. et al. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution // *Genes Dev*. – 1999. – Vol. 13. – P. 3217–3230.
126. Kim J.L., Burley S.K. 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. // *Nat. Struct. Biol*. – 1994. – Vol. 1. – P. 638–653.
127. Khrapunov S., Brenowitz M. Comparison of the effect of water release on the interaction of the *Saccharomyces cerevisiae* TATA binding protein (TBP) with “TATA Box” sequences composed of adenosine or inosine. // *Biophys. J*. – 2004. – Vol. 86. – № 1. – P. 371–383.
128. Pardo L. et al. Binding mechanisms of TATA box-binding proteins: DNA kinking is stabilized by specific hydrogen bonds. // *Biophys. J*. – 2000. – Vol. 78. – № 4. – P. 1988–1996.
129. Cox J.M. et al. Bidirectional binding of the TATA box binding protein to the TATA box. // *Proc. Natl. Acad. Sci. U. S. A*. – 1997. – Vol. 94. – № 25. – P. 13475–13480.
130. Grove A. et al. Affinity, stability and polarity of binding of the TATA binding protein governed by flexure at the TATA Box. // *J. Mol. Biol*. – 1998. – Vol. 282. – P. 731–739.
131. Stoddard B.L. Homing endonuclease structure and function. // *Q. Rev. Biophys*. – 2005. – Vol. 38. – № 1. – P. 49–95.
132. Roberts R.J. et al. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. // *Nucleic Acids Res*. – 2010. – Vol. 38. – № Database issue. – P. D234–D236.
133. Guhan N., Muniyappa K. Structural and functional characteristics of homing endonucleases. // *Crit. Rev. Biochem. Mol. Biol*. – 2003. – Vol. 38. – № 3. – P. 199–248.
134. Chevalier B. et al. Metal-dependent DNA cleavage mechanism of the I-CreI LAGLIDADG homing endonuclease // *Biochemistry*. – 2004. – Vol. 43. – № 44. – P. 14015–14026.
135. Arnould S. et al. Engineered I-CreI Derivatives Cleaving Sequences from the Human XPC Gene can Induce Highly Efficient Gene Correction in Mammalian Cells // *J. Mol. Biol*. – 2007. – Vol. 371. – № 1. – P. 49–65.
136. Fajardo-Sanchez E. et al. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. // *Nucleic Acids Res*. – 2008. – Vol. 36. – № 7. – P. 2163–2173.
137. Grishin A. et al. Identification of conserved features of LAGLIDADG homing endonucleases. // *J. Bioinform. Comput. Biol*. – 2010. – Vol. 8. – № 3. – P. 453–469.

Приложение

Таблица I. Список наблюдаемых способов взаимодействия. Указаны количества SCOP семейств, структур комплексов и структур доменов, для которых такие способы взаимодействия характерны. **H** - α -спираль, **S** - β -лист, **L** - петля/неструктурированный участок белка, **Mj** - большая бороздка ДНК, **Mn** - малая бороздка ДНК, **Bb** – сахаро-фосфатный остов ДНК.

Способ взаимодействия	Число семейств	Число записей PDB	Число доменов	Число структур
H-Bb H-Mj L-Bb	30	121	63	214
L-Bb	21	66	32	95
H-Bb H-Mj L-Bb L-Mj	20	84	41	146
H-Bb L-Bb	20	69	25	91
H-Bb	15	52	22	57
H-Bb H-Mj S-Bb L-Bb	14	67	35	100
H-Bb H-Mj	13	41	25	71
H-Bb H-Mj L-Bb L-Mn	11	69	34	97
H-Bb H-Mj S-Bb L-Bb L-Mn	11	50	15	78
H-Bb S-Bb S-Mj L-Bb L-Mj	10	43	17	76
H-Bb L-Bb L-Mn	9	20	10	20
H-Bb L-Bb L-Mj	9	17	12	21
H-Bb H-Mj L-Bb L-Mj L-Mn	9	15	12	21
H-Bb S-Bb L-Bb L-Mj	9	14	9	26
H-Bb H-Mj S-Bb L-Bb L-Mj	8	17	10	24
H-Bb S-Bb L-Bb L-Mj L-Mn	8	15	7	19
S-Bb S-Mj L-Bb L-Mj	8	13	9	15
H-Bb S-Bb S-Mj L-Bb L-Mj L-Mn	6	42	6	63
S-Bb L-Bb	6	28	9	40
H-Bb H-Mj S-Bb L-Bb L-Mj L-Mn	6	21	7	28
H-Bb S-Bb L-Bb L-Mn	5	15	10	20
H-Bb H-Mj H-Mn L-Bb	5	14	7	15
H-Bb S-Bb S-Mj L-Bb	5	12	6	29
L-Bb L-Mn	5	8	5	12
L-Bb L-Mj L-Mn	5	7	5	9
H-Bb S-Bb L-Bb	4	39	20	126
S-Bb	4	20	6	21
H-Bb L-Bb L-Mj L-Mn	4	18	5	26

Способ взаимодействия	Число семейств	Число записей PDB	Число доменов	Число структур
H-Bb S-Bb S-Mn L-Bb L-Mn	4	12	6	12
H-Bb S-Bb S-Mj	4	8	4	17
S-Bb L-Bb L-Mn	4	7	6	10
H-Bb H-Mn L-Bb L-Mj	4	6	4	9
S-Bb S-Mn L-Bb L-Mn	3	25	10	77
S-Bb S-Mn L-Bb	3	21	10	22
H-Bb H-Mj L-Mj	3	16	9	24
L-Bb L-Mj	3	10	4	16
H-Bb H-Mn L-Bb	3	8	4	8
S-Bb S-Mj L-Bb	3	4	3	4
S-Bb L-Bb L-Mj	3	3	3	6
H-Bb H-Mj S-Bb	3	3	3	4
H-Bb S-Mj	3	3	3	4
H-Bb S-Bb S-Mn L-Bb	2	11	4	12
H-Bb H-Mn L-Bb L-Mn	2	10	5	11
S-Bb S-Mj L-Bb L-Mj L-Mn	2	7	3	10
H-Bb H-Mj S-Bb S-Mn L-Bb L-Mn	2	7	4	7
H-Bb H-Mj S-Bb S-Mj L-Bb	2	6	4	11
S-Bb L-Bb L-Mj L-Mn	2	6	3	7
H-Bb S-Bb S-Mj L-Bb L-Mn	2	6	2	6
H-Bb H-Mn S-Bb S-Mj L-Bb L-Mj	2	3	3	4
H-Bb S-Bb	2	3	2	4
H-Bb S-Bb S-Mj L-Mj	2	3	2	3
H-Bb S-Mj L-Bb L-Mj	2	3	2	3
H-Bb H-Mn S-Bb S-Mj L-Bb L-Mj L-Mn	2	2	2	3
H-Bb H-Mj H-Mn L-Bb L-Mj L-Mn	2	2	2	2
H-Bb H-Mj H-Mn S-Bb L-Bb L-Mj L-Mn	2	2	2	2
H-Bb H-Mj L-Mn	2	2	2	2
H-Bb S-Mj L-Bb	2	2	2	2
H-Mn L-Bb	2	2	2	2
L-Mj	2	2	2	2
S-Bb S-Mn	2	2	2	2
H-Bb H-Mj H-Mn L-Bb L-Mj	1	8	3	10
H-Mj S-Bb L-Bb L-Mj	1	6	1	15
H-Bb H-Mn S-Bb S-Mj L-Bb	1	5	1	9

Способ взаимодействия	Число семейств	Число записей PDB	Число доменов	Число структур
S-Bb S-Mj L-Bb L-Mn	1	5	1	6
H-Bb H-Mj S-Mj L-Bb L-Mj	1	2	1	3
H-Bb H-Mj H-Mn S-Bb S-Mj L-Bb L-Mj	1	2	1	2
H-Bb H-Mj S-Bb S-Mj L-Bb L-Mj	1	2	2	2
H-Bb H-Mj S-Bb S-Mn L-Bb L-Mj L-Mn	1	2	1	2
H-Bb S-Bb S-Mn L-Bb L-Mj L-Mn	1	2	1	2
L-Mn	1	2	1	2
H-Bb H-Mj H-Mn S-Bb L-Bb L-Mj	1	1	1	2
H-Bb H-Mj H-Mn S-Bb L-Bb L-Mn	1	1	1	1
H-Bb H-Mj H-Mn S-Bb S-Mj L-Bb L-Mj L-Mn	1	1	1	1
H-Bb H-Mj H-Mn S-Bb S-Mj L-Bb L-Mj L-Mn	1	1	1	1
H-Bb H-Mj H-Mn S-Mj L-Bb L-Mj	1	1	1	1
H-Bb H-Mj S-Bb S-Mj L-Bb L-Mn	1	1	1	1
H-Bb H-Mj S-Bb S-Mj L-Mj	1	1	1	1
H-Bb H-Mj S-Bb S-Mn L-Bb L-Mj	1	1	1	1
H-Bb H-Mj S-Mn L-Bb L-Mn	1	1	1	1
H-Bb H-Mn L-Bb L-Mj L-Mn	1	1	1	1
H-Bb H-Mn L-Mj	1	1	1	1
H-Bb L-Mj	1	1	1	1
H-Bb L-Mn	1	1	1	1
H-Bb S-Mj L-Bb L-Mj L-Mn	1	1	1	1
H-Bb S-Mn L-Bb L-Mj L-Mn	1	1	1	1
H-Mj L-Bb	1	1	1	1
H-Mj L-Bb L-Mj	1	1	1	1
H-Mj L-Bb L-Mj L-Mn	1	1	1	1
H-Mj S-Bb S-Mj L-Bb L-Mj L-Mn	1	1	1	1
H-Mn L-Bb L-Mn	1	1	1	1
H-Mn S-Bb S-Mj L-Bb L-Mj	1	1	1	1
S-Bb L-Mn	1	1	1	1
S-Bb S-Mj	1	1	1	1
S-Bb S-Mj S-Mn L-Bb L-Mj	1	1	1	1
S-Mj L-Bb	1	1	1	1
S-Mj L-Bb L-Mj	1	1	1	1
S-Mn L-Bb	1	1	1	1

Таблица II. Консервативные гидрофобные контакты между ТАТА-бох связывающим белком и ДНК.

В левом столбце используется единая нумерация контактирующих аминокислотных остатков и нуклеотидов. Косая черта разделяет альтернативные контакты, не встречающиеся в одном комплексе; запятая разделяет совместно наблюдаемые контакты. В остальных столбцах используется нумерация соответствующей записи PDB. В строках указаны гомологичные аминокислотные остатки и нуклеотиды, формирующие гидрофобное взаимодействие в одном и том же месте комплекса ТВР-ДНК. Прочерк обозначает, что в данной структуре соответствующего контакта нет.

<i>Организм</i>			<i>Homo sapiens</i>			<i>Arabidopsis thaliana</i>			<i>Saccharomyces cerevisiae</i>			<i>Pyrococcus woesei</i>		
Пример структуры (PDB код)			1cdw			1qne			1ytf			1d3u		
цепь белка	прямая цепь ДНК	обратная цепь ДНК	цепь белка: A	прямая цепь ДНК: B	обратная цепь ДНК: C	цепь белка: A	прямая цепь ДНК: C	обратная цепь ДНК: D	цепь белка: A	прямая цепь ДНК: E	обратная цепь ДНК: F	цепь белка: A	прямая цепь ДНК: C	обратная цепь ДНК: D
<i>Общая формула контакта согласно единой нумерации</i>			<i>Соответствующие гидрофобные контакты</i>											
Leu#53	A#5:d, A#6:d, A/T#7:d, N#8:d	T/A#7:r	Leu208:A	A9:B, A10:B, A11:B, —	—	Leu72:A	A207:C, A208:C, A209:C, T220:D	—	Leu114:A	—, A14:E, A15:E, —	T2:F	Leu58:A	—, T1416:C, A1417:C	A1433:D
Leu#1053	T#1:d, A/T#2:d, T#3:d, A#4:d, A#5:d	A#1:r, T#2:r, A#3:r	Leu299:A	T5:B, A6:B, T7:B, A8:B, —	A110:C, T111:C, A112:C	Leu163:A	T203:C, A204:C, T205:C, A206:C, —	A224:D, T225:D, A226:D	Leu205:A	T9:E, A10:E, T11:E, A12:E, A13:E	A6:F, T7:F, A8:F	Leu149:A	—, T1411:C, —, —, —	A1439:D, A1438:D, —
Phe#38	A#5:d, A#6:d, A/T#7:d, N#8:d	T/A#7:r	Phe193:A	A9:B, A10:B, A11:B, —	—	Phe57:A	A207:C, A208:C, A209:C, —	T220:D	Phe99:A	—, A14:E, A15:E, —	T2:F	Phe43:A	—, T1416:C, A1417:C	A1433:D
Phe#1038	T#1:d, A/T#2:d, T#3:d, A#4:d, A#5:d	A#1:r, T#2:r, A#3:r	Phe284:A	T5:B, A6:B, T7:B, A8:B, —	A110:C, T111:C, A112:C	Phe148:A	T203:C, A204:C, T205:C, A206:C, —	A224:D, T225:D, A226:D	Phe190:A	T9:E, A10:E, T11:E, A12:E, A13:E	A6:F, T7:F, A8:F	Phe134:A	—, T1411:C, —, —, —	A1439:D, A1438:D, —
Val#9	A#5:d, A#6:d, A/T#7:d	T/A#7:r	Val165:A	A9:B, A10:B, A11:B	—	Val29:A	A207:C, A208:C, A209:C	T220:D	Val71:A	—, A14:E, A15:E	T2:F	Val15:A	A1414:C, A1415:C, —	—

<i>Организм</i>	<i>Homo sapiens</i>			<i>Arabidopsis thaliana</i>			<i>Saccharomyces cerevisiae</i>			<i>Pyrococcus woesei</i>		
Val#1009 T#1:d, A#1:r, A#2:d, T#2:r, T#3:d, A#3:r A#4:d, A#5:d	Val255:A	T5:B, A6:B, T7:B, A8:B, –	A110:C, T111:C, A112:C	Val119:A	T203:C, A204:C, T205:C, A206:C, –	A224:D, T225:D, A226:D	Val161:A	T9:E, A10:E, T11:E, A12:E, A13:E	A6:F, T7:F, A8:F	Val106:A	–, –, –, –, –	–, –, A1437:D
Val#61 A#5:d, T/A#7:r A#6:d, A/T#7:d	Val216:A	A9:B, A10:B, A11:B	–	Val80:A	A207:C, A208:C, A209:C	T220:D	Val122:A	–, A14:E, A15:E	T2:F	Val66:A	A1414:C, A1415:C, –	–
Val#1061 T#1:d, A#1:r, A/T#2:d, T#2:r, T#3:d, A#3:r A#4:d, A#5:d	Val307:A	T5:B, A6:B, T7:B, A8:B, –	A110:C, T111:C, A112:C	Val171:A	T203:C, A204:C, T205:C, A206:C, –	A224:D, T225:D, A226:D	Val213:A	T9:E, A10:E, T11:E, A12:E, A13:E	A6:F, T7:F, A8:F	Val157:A	–, –, –, –, –	–, A1438:D, A1437:D
Phe#55 A#5:d, T/A#7:r A#6:d, A/T#7:d, N#8:d	Phe210:A	A9:B, A10:B, A11:B, –	–	Phe74:A	A207:C, A208:C, A209:C, –	T220:D	Phe116:A	–, A14:E, A15:E, –	T2:F	Phe60:A	–, –, T1416:C, A1417:C	A1433:D
Phe#1055 T#1:d, A#1:r, A/T#2:d, T#2:r, T#3:d, A#3:r A#4:d, A#5:d	Phe301:A	T5:B, A6:B, T7:B, A8:B, –	A110:C, T111:C, A112:C	Phe165:A	T203:C, A204:C, T205:C, A206:C, –	A224:D, T225:D, A226:D	Phe207:A	T9:E, A10:E, T11:E, A12:E, A13:E	A6:F, T7:F, A8:F	Phe151:A	–, T1411:C, –, –, –	A1439:D, A1438:D, –
Thr#11 A#5:d, T/A#7:r A#6:d, A/T#7:d, N#8:d	Thr167:A	A9:B, A10:B, A11:B, –	–	Thr31:A	A207:C, A208:C, A209:C, –	T220:D	Thr73:A	–, A14:E, A15:E, –	–	–	–	–
Ser#1011 T#1:d, A#1:r, A/T#2:d, T#2:r, T#3:d, A#3:r A#4:d, A#5:d	Ser257:A	–, –, –, –, –	A110:C, –, –	Ser121:A	T203:C, A204:C, T205:C, A206:C, –	A224:D, T225:D, A226:D	Ser163:A	–, –, –, –, –, A6:F, –, –	–, –, –	Ser108:A	–, –, –, –, –	–, A1438:D, A1437:D
Thr#63 A#5:d, T/A#7:r A#6:d, A/T#7:d, N#8:d	Thr218:A	A9:B, A10:B, A11:B, –	–	Thr82:A	A207:C, A208:C, A209:C, –	T220:D	Thr124:A	–, A14:E, A15:E, –	T2:F	Thr68:A	A1414:C, A1415:C, –, –	–

