На правах рукописи

Занегина Ольга Николаевна

Сравнительная характеристика структур ДНК-белковых комплексов

03.01.09 – Математическая биология, биоинформатика

Автореферат диссертации на соискание ученой степени кандидата биологических наук

Москва – 2016

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет имени М.В.Ломоносова»

Научный руководитель	Кандидат физико-математических наук **Спирин Сергей Александрович** Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В.Ломоносова»
Официальные оппоненты	Доктор биологических наук, профессор РАН **Орлов Юрий Львович** Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук
	Кандидат физико-математических наук **Кулаковский Иван Владимирович** Федеральное государственное бюджетное учреждение науки Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук
Ведущая организация	Федеральное государственное бюджетное учреждение науки Институт математических проблем биологии Российской академии наук

Защита состоится _____на заседании диссертационного совета Д002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации имени А.А.Харкевича Российской академии наук по адресу: 127051, г. Москва Большой Каретный пер., д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН и на сайте www.iitp.ru

Автореферат разослан «__» _____ 201_г.
Ученый секретарь диссертационного совета
д.б.н., профессор
Рожкова Галина Ивановна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы

Развитие методов рентгеноструктурного анализа, ядерного магнитного резонанса и электронной микроскопии привело к экспоненциальному росту количества изученных 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 выяснять не только особенности ДНК-белкового взаимодействия в конкретных структурах, но и закономерности такого рода взаимодействий.

Понимание механизмов узнавания ДНК белком может помочь в предсказании специфичности ДНК-белкового взаимодействия, а также в направленном мутагенезе, особенно в тех случаях, когда пространственная структура ДНК-белкового комплекса недоступна.

К сожалению, к настоящему времени не существует однозначных подходов к предсказанию участков ДНК, узнаваемых данным белком. В большой степени это связано с тем, что на такое узнавание влияет множество факторов, и единого кода соответствия белку последовательности ДНК не существует.

Для поиска закономерностей ДНК-белкового узнавания важен анализ и систематизация взаимодействий, наблюдаемых в пространственных структурах ДНК-белковых комплексов. Был разработан ряд классификаций ДНК-белковых взаимодействий, рассматривающих все доступные на момент классификации структуры. Например, в работе Харрисона (Harrison S.C., 1991) было выделено четыре группы ДНК-связывающих доменов: спираль-поворот-спираль (HTH), цинк-связывающие домены, лейциновые молнии и домены, содержащие β-листы. Прабакаран с соавторами (Prabakaran P. et al., 2006) разделили 62 неродственных ДНК-белковых комплекса на 7 кластеров, учитывая число водородных связей между белком и ДНК, контакты по бороздкам и остову ДНК, глубину и ширину бороздок ДНК, изгиб ДНК, GC-состав ДНК и площадь ДНК-белкового контакта. Было показано, что сходство строения ДНК-узнающих мотивов не всегда обеспечивает сходство способа узнавания ДНК белком. Поэтому необходимо при построении классификаций ДНК-белковых взаимодействий рассматривать параметры, характеризующие взаимодействие в целом, а не только свойства ДНК-узнающего белка. Кроме того, классификация должна быть открытой к уточнению, дополнению и расширению при появлении новых пространственных структур. Имеющиеся классификации в большинстве случаев в дальнейшем не обновлялись новыми структурами.

Другим важным направлением в изучении ДНК-белковых взаимодействий является создание специализированных баз данных, предоставляющих информацию о структурах ДНК-белковых комплексов и о характеристиках ДНК-белковых взаимодействий, а также обладающих инструментами для анализа этих взаимодействий. Аналогично классификациям, часть существующих баз данных ДНК-белковых взаимодействий устарела и не обновляется автоматически. Например, база 3DinSight (An J., Nakama T., Kubota Y, 1998), в которой была возможность отображать на структуре белковые мотивы, значимые сайты и мутации, а также база AANT (Hoffman M. M. et al., 2004), в которой были охарактеризованы все контактирующие пары «аминокислота – нуклеотид», на данный момент неактивны.

Таким образом, представляется актуальным создание классификации ДНК-белковых взаимодействий, учитывающей взаимодействующие элементы как со стороны белка, так и со стороны ДНК, а также открытой для дополнения при появлении новых структур. Интеграция классификации в базу данных НК-белковых взаимодействий NPIDB (Kirsanov D et al., 2013) призвана облегчить навигацию по базе и поиск родственных по типу взаимодействия семейств ДНК-узнающих белков.

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.
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 подробный 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 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-box 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 [2, 3, 4, 9, 10] диссертантом выполнена работа по анализу ДНК-белковых взаимодействий. В публикациях [1, 5, 6, 7, 8] – работа по анализу консервативных молекул воды.

Структура и объём работы

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

СОДЕРЖАНИЕ РАБОТЫ

Введение содержит обоснование актуальности темы работы, формулировку цели и задач исследования, подтверждения его научной новизны и практической значимости. Сформулированы основные результаты, выносимые на защиту.

Глава 1 содержит обзор литературы. Рассмотрены физические основы ДНК-белковых взаимодействий: прямые и опосредованные водой водородные связи, а также гидрофобные кластеры. Дан обзор известных алгоритмов поиска водородных связей, гидрофобных кластеров и мостиковых молекул воды. Подробно рассмотрены описанные в литературе механизмы ДНК-белкового узнавания и обосновано отсутствие однозначного кода соответствия узнающей белковой последовательности и узнаваемой последовательности ДНК. Рассмотрено понятие специфичности ДНК-белкового узнавания с точки зрения последовательностей, структур и физико-химических параметров. Далее описаны одиннадцать существующих на данный момент классификаций ДНК-белковых взаимодействий. Все классификации разделены на белок-центрированные, ДНК-центрированные и комплекс-центрированные — в зависимости от того, которому из участников комплекса уделялось большее внимание. Представленная в данной работе классификация является комплекс-центрированной. Приведен обзор баз данных структур белков, ДНК и их комплексов.

Глава 2 описывает предложенную в данной работе классификацию ДНК-белковых взаимодействий.

В начале введены понятия: *структуры комплекса белкового домена с ДНК* как информации о координатах атомов, доступной из одной записи банка PDB и включающей координаты атомов фрагмента одной белковой цепи и контактирующего с ней фрагмента двойной спирали ДНК длиной не менее 10 пар нуклеотидов, при этом границы фрагмента белка определяются согласно границам некоторого структурного домена по базе SCOP*; белкового домена* как фрагмента конкретного белка, отвечающего структурному белковому домену по SCOP (один и тот домен может быть представлен в нескольких структурах комплексов, из одной или нескольких записей PDB); *семейства белковых доменов*, определяемого как семейство согласно классификации SCOP. Контактом белка с ДНК в данной структуре комплекса считалась либо водородная связь между атомами белка и ДНК, либо гидрофобный кластер, включающий одновременно атомы белка и атомы ДНК. *Взаимодействующим элементом со стороны ДНК* считался либо сахаро-фосфатный остов (Bb), либо группы атомов азотистых оснований, обращенных в большую бороздку (Mj), либо группы атомов азотистых основа-

ний, обращенных в малую бороздку (Mn). *Взаимодействующий элемент со стороны белка* определялся по вторичной структуре участка цепи, которому принадлежит контактирующий аминокислотный остаток: α-спираль или 3_{10}-спираль (H), β-лист (S), петля, поворот или неструктурированный участок (L) 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 (Karyagina A. et al., 2006).

Типы контакта описываются парой взаимодействующих элементов со стороны белка и ДНК. Взаимодействуя попарно, элементы образуют *9 типов контакта*, обозначаемые H–Bb, S–Bb, L–Bb; H–Mj, S–Mj, L–Mj; H–Mn, S–Mn, L–Mn.

Под *способом взаимодействия структуры* в данной работе понимается список типов контакта, наблюдаемых для данной структуры комплекса. Например, в принятых сокращениях запись “|H-Bb|H-Mj|S-Bb|S-Mn|L-Bb|L-Mn|” обозначает, что в структуре белок 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.

Под *способом взаимодействия белкового домена с ДНК* в заданной структуре комплекса понимается список типов контакта, наблюдаемых в данной структуре.

Тип контакта считается *характерным* для семейства белковых доменов SCOP, если он встречается хотя бы в одной структуре каждого домена данного семейства. *Класс взаимодействия* с ДНК *для семейства* SCOP — это список всех характерных для семейства типов контакта белкового домена с ДНК.

Процедура классификации состоит из двух этапов:

1. Определение способа взаимодействия для каждого комплекса белкового домена, контактирующего с ДНК.
2. Классификация семейств ДНК-контактирующих доменов SCOP по классу взаимодействия.

Определение способа взаимодействия в конкретном комплексе белкового домена с ДНК проводили согласно следующей процедуре:

1. Определяются водородные связи и гидрофобные кластеры на ДНК-белковом интерфейсе;
2. Определяются вторичная структура белка, а также структура ДНК, то есть выделяются пары комплементарных оснований;
3. Создается список из типов контакта, представленных хотя бы одной единицей контакта любой природы;
4. Список контактов корректируется при наличии интеркаляций, изгиба, выпетливания или выщепления азотистого основания, в результате которых аминокислотный остаток может контактировать с атомами противоположной бороздки ДНК.

В результате получается *способ взаимодействия* белка с ДНК в рассматриваемой структуре в виде сокращенной формулы, например, |H-Bb | H-Mj| S-Bb | S-Mn | L-Bb | L-Mn|.

Определение *класса взаимодействия* проводилось для семейств, представленных структурами трёх или более разных белковых доменов в комплексе с ДНК. Различие доменов определялось по различию белков, в состав которых они входят, или по различию положения в цепи одного и того же белка.

Класс взаимодействия определялся по следующей процедуре (рисунок 1):

1. Выявлялись все *типы контакта*, которые встретились хотя бы в одной структуре каждого белкового домена, представляющего данное семейство.
2. Семейству присваивался класс взаимодействия с ДНК, представляющий собой список всех таких типов контакта

Поскольку класс взаимодействия является пересечением списка типов контактов, наблюдаемых для белковых доменов данного семейства, то он может оказаться пустым. Семейства, для которых нельзя выделить общий контакт для входящих в его состав белковых доменов, отнесены к классу *Разнородных контактов*.

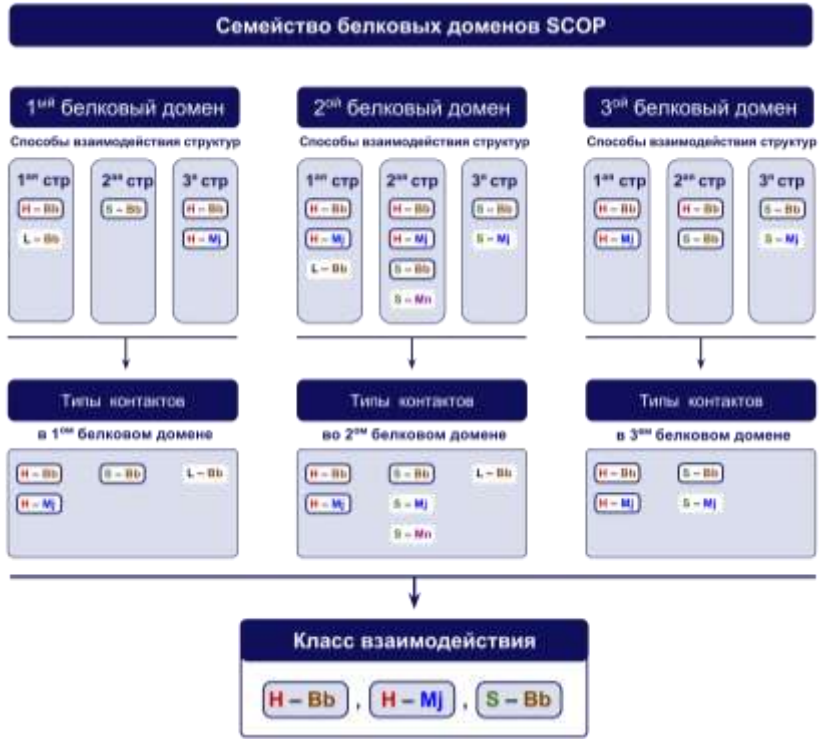


Рисунок 1. Процедура классификации семейств ДНК-белковых комплексов.

Результаты классификации индивидуальных комплексов. Для классификации были отобраны 748 записей PDB (включая комплексы разных субъединиц из одной записи PDB), содержащих двухцепочечную ДНК, возможно, с дополнительными одноцепочечными участками, и при этом ДНК-узнающий домен белка присутствовал в базе данных SCOP версии 1.75. Из этих записей PDB было извлечено 1942 структуры 314 различных белковых доменов (согласно SCOP) вместе с контактирующей ДНК. Эти домены представляют 115 семейств по классификации SCOP.

В результате классификации было выделено 97 различных способов взаимодействия белкового домена с ДНК (из 511 теоретически возможных). В таблице 1 приведены способы взаимодействия, присутствующие в четырех или более семействах. Для каждого способа взаимодействия указано количество семейств, записей PDB и белковых доменов, в которых этот способ взаимодействия встре-

чается. Все способы взаимодействия отсортированы по количеству семейств SCOP, внутри которых они встретились.

Со стороны ДНК чаще наблюдаются контакты по сахаро-фосфатному остову, что объясняется большей доступностью сахаро-фосфатного остова по сравнению с бороздками ДНК, а также тем, что практически все контакты белка с бороздками сопровождаются контактами с сахаро-фосфатным скелетом. Контакты с большой бороздкой более распространены, чем контакты с малой, вследствие различия их ширины и доступности.

Со стороны белка наиболее распространены контакты посредством петель и неструктурированных участков. Это объясняется несколькими факторами: (a) большей подвижностью неструктурированных участков; (b) тем, что контактирующие α-спираль или β-лист часто окаймлены петлями.

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 наблюдаются чаще, чем α-спирали. При этом связывание β-листов с малой бороздкой приводит к её расширению.

Анализ выявленных способов взаимодействия позволил выявить следующие закономерности ДНК-белкового взаимодействия:

1. Среди различных структур одного белкового домена примерно в половине случаев способ взаимодействия сохраняется. В остальных случаях разница в способе взаимодействия обусловлена подвижностью боковых цепей, различиями во определении вторичной структуры, а также разницей во взаимном расположении белка относительно ДНК.
2. В некоторых случаях (например, для способа взаимодействия |H-Bb|H-Mj|S-Bb|L-Bb|) можно выделить общие черты в архитектуре ДНК-узнающих доменов, даже если они принадлежат не только различным семействам, но и различным укладкам. В данном случае можно выделить лежащую внутри большой бороздки α-спираль и расположенный рядом (иногда непосредственно соединенный с α-спиралью) β-лист, контактирующий с сахаро-фосфатным остовом.

Способ взаимодействия	Количество семейств	Количество записей PDB	Количество доменов	Количество 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
|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

Таблица 1. Способы взаимодействия белковых доменов SCOP с ДНК, встречающиеся в четырех и более семействах SCOP. Обозначения: H – α-спираль, S – β-лист, L – петля/неструктурированный участок белка, Mj – большая бороздка ДНК, Mn – малая бороздка ДНК. Указано количество SCOP семейств, записей PDB, а также выделенных из них белковых доменов и структур, в которых встречается данный способ взаимодействия.

3. Гидрофобные контакты белковых доменов, как правило, повторяют и дополняют типы контактов, обеспечиваемых водородными связями. При этом основу взаимодействия все же составляют водородные связи. Структур, где количество аминокислотных остатков, образующих кластеры, значительно превышало бы количество аминокислотных остатков, образующих водородные связи, найдено не было.

4. Внутри семейства SCOP способы узнавания ДНК могут значительно варьировать. Это может быть связано с рядом причин: (a) подвижностью аминокислотных остатков белка – изменение конформации остатка белка может приводить к появлению/исчезновению контакта с ДНК; (b) аминокислотной заменой в месте контакта – остатки могут иметь разный размер или различаться другими свойствами, влияющими на возможность контакта; (c) разницей в расположении белка относительно ДНК; (d) различием в определении вторичной структуры белка – короткие неструктурированные участки белка при небольшом изменении геометрии белковой цепи могут определяться как участок α-спирали или β-листа.

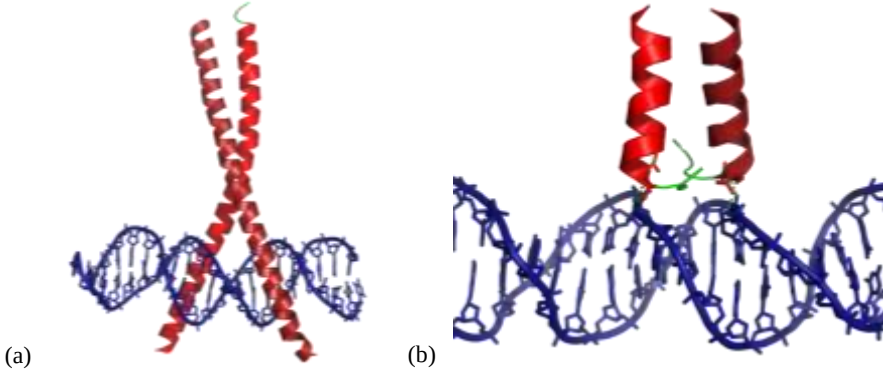
Результаты классификации семейств. Среди рассмотренных 115 семейств только в 34 есть достаточное для классификации количество данных, в остальных 81 семействе представлено менее трех различных белковых домена. Эти семейства могут быть классифицированы в будущем при появлении новых структур белков.

Все классифицированные семейства были разделены на 17 классов взаимодействия, которые представлены в таблице 2. Число классов взаимодействия меньше числа способов взаимодействия, поскольку класс отражает не все, а только консервативные контакты внутри семейства белковых доменов SCOP.

Процедура классификации является формальной: если в структурах хотя бы одного домена семейства отсутствует некоторый тип контакта, то класс взаимодействия всего семейства не будет содержать этот тип контакта. При появлении структур домена, содержащих недостающий тип контакта, класс взаимодействия расширится этим типом контакта. Семейства, для которых класс взаимодействия может быть расширен новыми типами контакта при появлении новых структур, отмечены в таблице 2 звездочкой.

Семейство лейциновых молний (Leucine zipper domain (h.1.3.1)) заслуживает отдельного внимания, поскольку оно было разделено на два подсемейства, каждое из которых классифицировано отдельно. Причина состоит в принципиально различном расположении белковых доменов относительно ДНК (рисунок 2). В первом подсемействе представлена классическая лейциновая молния, контакти-

рующая α-спиралью с сахаро-фосфатным остовом и большой бороздкой (класс взаимодействия (H – Bb) | (H – Mj)).



(a) (b)

Рисунок 2. Семейство Leucine zipper domain (h.1.2.1). (a) Контакт с остовом и большой бороздкой ДНК посредством α-спирали (1nwq), (b) контакт с остовом посредством α-спирали (1d66).

Второе подсемейство представлено белковыми доменами дрожжей, а именно: CD2-Gal4, PUT3 и HAP1. Короткая лейциновая молния связывает вместе два домена, а с сахаро-фосфатным остовом ДНК контактируют торцы α-спиралей и иногда дополнительно неструктурированные петли или β-листы. Класс взаимодействия для подсемейства: (H – Bb).

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. Наиболее распространенным является класс, включающий узнавание ДНК α-спиралью и неструктурированным участком (H – Bb), (L – Bb) | (H – Mj). Способы взаимодействия внутри каждого из восьми семейств данного класса варьируют, но неизменным остается наличие α-спирали, контактирующей с большой бороздкой и остовом ДНК, дополненное контактами неструктурированных участков и петель с сахаро-фосфатным остовом ДНК. Заметим, что, например, присутствующий в семействе Nuclear receptor (g.39.1.2) контакт β-листа с большой бороздкой ДНК консервативным не является. Из таблицы 2 нельзя сделать однозначный вывод о том, что какой-то из типов контакта склонен быть более консервативным, чем другие типы контактов. Для предсказания консервативности необходимо сравнивать контакты внутри семейства.

Класс взаимодействия	**Названия семейств (идентификатор SCOP)**
Разнородные контакты	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 2. Классы взаимодействия семейств белковых доменов SCOP с ДНК. Звездочками помечены семейства, чей класс взаимодействия может измениться при появлении новых структур. Обозначения те же, что в таблице 1.

Предложенная классификации ДНК-контактирующих белковых доменов SCOP может быть использована при решении ряда задач:

1. Для отделения наиболее консервативных и функционально значимых контактов от вариаций узнавания белка, присущих структурам данного семейства.
2. При предсказании по гомологии потенциальных ДНК-белковых контактов для новых белков.
3. При поиске белков, имеющих сходное строение в сайте связывания с ДНК.

Глава 3 описывает применение программы wLake (Aksianov E. et al., 2008) для поиска консервативных молекул воды в белковых и ДНК-белковых комплексах.

В структурах многих макромолекулярных комплексов присутствуют молекулы воды. В связи с техническими особенностями метода рентгеноструктурного анализа надежность координат молекул воды ниже, чем у атомов белка. В большинстве рентгеновских структур с разрешением 1,5˚A – 2,5˚A присутствуют молекулы воды, не связанные водородными связями ни с какими другими молекулами (Davis A. M., et al., 2003). Эти молекулы могут как реально присутствовать в структуре, так и являться артефактами метода.

Консервативными молекулами воды считаются молекулы воды, расположенные в одних и тех же местах и образующие водородные связи с соответствующими аминокислотными остатками, нуклеотидами, лигандами и ионами в структурах гомологичных белков и белковых комплексов. Можно предположить, что консервативные молекулы воды являются не случайными, а функционально значимыми и соответствуют сайтам гидратации макромолекул в растворе.

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

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.

Консервативные молекулы воды в структурах транскетолаз. Транскетолазы — это тимидин дифосфатзависимые ферменты, катализирующие перенос гидроксиацетильной группы с кетозы на альдозу. Они представляют собой гомодимеры и имеют два активных центра, в которых находятся молекулы тиаминдифосфата (ThDP) и иона двухвалентного металла с октаэдрической координацией (Mg^{2+}, Ca^{2+}, Mn^{2+}, Co^{2+}).

Поиск консервативных молекул воды производился среди всех доступных на момент исследования 34 структур 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. thermopilus HB8 TT0505, L. mexicana* и *Z. mays*. Каждая структура содержит в себе от одной до четырех субъединиц (всего 61 структура субъединиц из различных организмов). Субъединицы пространственно совмещались с помощью сервиса SSM (Krissinel E. et al., 2004), после чего программой wLake производился поиск консервативных молекул воды.

Всего было найден 221 кластер консервативных молекул воды, причем 11 кластеров находится в районе активного центра фермента. Молекулы воды в активном центре транскетолазы дрожжей были описаны еще в 1994 году в работе Никкола с соавторами (Nikkola M. et al., 1994). Однако на тот момент количество структур транскетолаз было малó и не позволяло отделить консервативные молекулы воды от неконсервативных. Из описанных в работе Никкола восьми молекул воды семь оказались консервативными и только одна, связанная единственной водородной связью с белком, – неконсервативной. Дополнительно были найдены еще четыре консервативных молекулы воды, одна из которых находится в середине активного центра. Как видно на данном примере, поиск консервативных молекул воды в родственных структурах помогает более точно определять функционально значимые молекулы в гидратной оболочке белка.

Консервативные молекулы воды в семействах ДНК-связывающих доменов. На страницах базы данных NPIDB для 72 семейств белковых доменов SCOP, взаимодействующих с ДНК и содержащих хорошо пространственно совмещаемые структуры, предоставлены данные о кластерах консервативных молекул воды на ДНК-белковом интерфейсе. Число кластеров варьирует от одного до 365 и в первую очередь зависит от количества молекул воды в каждой из совмещенных структур семейства и не коррелирует с количеством совмещенных структур. При наличии большого количества кластеров, расположенных рядом, можно наблюдать зоны гидратации, в которых могут находится несколько консервативных молекул воды.

Глава 4 описывает функционал базы данных NPIDB, который может быть использован для поиска консервативных особенностей ДНК-белкового взаимодействия. Для каждого комплекса белка с нуклеиновой кислотой доступны следующие сервисы: файл структуры (загрузка и визуализация), ссылки на другие базы данных, 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. На отдельной странице для каждого комплекса представлен список и визуализация водородных связей, водных мостиков и гидрофобных кластеров на ДНК-белковом интерфейсе.

Для каждого домена Pfam создана отдельная страница, включающая список структур, содержащих данный домен, а также интерактивное выравнивание всех структур семейства. Для семейств белковых доменов SCOP созданы отдельные страницы, где указано 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.

Функционал базы NPIDB был расширен представленной в данной работе классификацией. Данные о способах взаимодействия отдельных ДНК-контактирующих доменов SCOP, а также о классах взаимодействия их семейств были добавлены на соответствующие страницы. Полные списки способов взаимодействия (interaction mode) ДНК-контактирующих доменов SCOP и классов взаимодействий (interaction class) их семейств представлены на отдельных страницах. Страницы конкретных комплексов, семейств, способов взаимодействия и классов взаимодействия связаны гиперссылками. Таким образом, рассматривая конкретный ДНК-белковый комплекс или семейство ДНК-связывающих белковых доменов SCOP, можно найти другие комплексы или семейства, имеющие такой же способ взаимодействия или класс взаимодействия, что и искомый.

Глава 5 посвящена примерам использования функционала NPIDB для сравнительного структурного анализа консервативных ДНК-белковых контактов внутри белковых семейств.

TATA-box связывающий домен и белок. TATA-box связывающий белок (TBP) — это транскрипционный фактор, участвующий в инициации транскрипции многих архей и эукариот в составе мультибелкового комплекса TFIID. Этот белок состоит из двух доменов семейства TATA-box binding protein (TBP), C-

terminal domain (d.129.1.1)). Каждый домен состоит из пяти β-тяжей антипараллельного β-листа, окаймленного двумя α-спиралями. Вместе они образуют общий β-лист, узнающий 8 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.

Для выявления консервативных контактов были проанализированы 45 структур комплексов TBD с ДНК, часть из них содержит также молекулы воды (группа #1), а также 16 структур TBP без ДНК, но с молекулами воды (группа #2). Информация о структурах комплексов была получена из 32 записей PDB, относящихся к семи организмам (*Homo sapiens*, *Sulfolobus acidocaldarius*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Pyrococcus woesei*, *Methanocaldococcus jannaschii*, *Encephalitozoon cuniculi*).

Внутри каждой группы с помощью сервиса SSM было произведено пространственное выравнивание 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.

Было выявлено, что консервативные водородные связи образуют 9 аминокислотных остатков, а консервативные гидрофобные контакты формируются 13 аминокислотными остатками.

Также были найдены консервативные положения мостиковых молекул воды. Все четыре найденных кластера консервативных молекул воды расположены ближе к N-концевому домену, опосредуя контакт белка с сахаро-фосфатным остовом ДНК. Найденные консервативные контакты образуют плотную сеть, покрывающую большую часть β-листа, контактирующего с малой бороздкой ДНК. При этом имеющий дополнительные контакты 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, содержащей последовательность TATA.

Семейство LAGLIDADG хоминг эндонуклеаз. Семейство LAGLIDADG хоминг эндонуклеаз представляет собой редкощепящие ферменты, узнающие длинные последовательности ДНК (14–40 п.н.). Название семейства происходит от единственного консервативного мотива, имеющего аминокислотную последовательность LAGLIDADG. Каталитически активная единица состоит из двух так называемых LAGLIDADG доменов, каждый из которых узнает половину сайта

связывания с ДНК. Каждый домен состоит из β-листа, взаимодействующего с большой бороздкой ДНК, и нескольких α-спиралей, расположенных над β-листом. Домены вносят равный вклад в формирование активного сайта, расположенного рядом с поверхностью димеризации.

Для выявления консервативных особенностей ДНК-белкового узнавания было проанализировано Pfam семейство LAGLIDADG_1 (PF00961) хоминг эндонуклеаз. Оно представлено 116 пространственными структурами доменов LAGLIDADG_1 в комплексе с ДНК. Структуры относятся к 10 организмам: *Chlamydomonas reinhardtii*, *Leptographium truncatum*, *Aspergillus nidulans*, *Chlamydomonas moewusii*, *Fusarium graminearum*, *Monomastix sp. OKE-1*, *Sordaria macrospora*, *Ophiostoma novo-ulmi*, *Vulcanisaeta distributa*, *Chlorella vulgaris*, а также к синтетическим конструкциям, экспрессированным в *E. coli*.

Аналогично семейству TBP, в рассмотренном семействе LAGLIDADG_1 хоминг эндонуклеаз был произведен поиск консервативных водородных связей, гидрофобных контактов и мостиковых молекул воды. Консервативных аминокислотных остатков, контактирующих с ДНК, выявить не удалось. Это может объясняться невысоким сходством последовательностей хоминг эндонуклеаз. В то же время, структуры комплексов хоминг эндонуклеаз хорошо совмещаются внутри семейства, так что можно выделить консервативно контактирующие аминокислотные позиции в домене LAGLIDADG_1.

Всего найдено 14 консервативных позиций, аминокислотные остатки в которых так или иначе взаимодействуют с ДНК. Половина позиций сконцентрирована на первом β-листе, контактирующем с сахаро-фосфатным остовом и с большой бороздкой ДНК посредством прямых и опосредованных водой водородных связей, а также гидрофобных контактов. В части позиций природа контакта может варьировать от структуры к структуре, но положение контакта сохраняется. Вторая половина контактов находится в крайних точках тяжей β-листа, фиксируя расположение белкового домена относительно ДНК. С сахарофосфатным остовом ДНК образуются прямые и опосредованные молекулами воды водородные связи, а гидрофобные кластеры характерны для контактов по большой бороздке.

Полученные данные о консервативно контактирующих аминокислотных позициях в семействе LAGLIDADG_1 хоминг эндонуклеаз могут быть учтены в дизайне новых эндонуклеаз с заданной специфичностью.

Заключение содержит краткое перечисление и обсуждение полученных в диссертационной работе результатов.

ВЫВОДЫ

1. Разработана процедура классификации ДНК-связывающих белковых доменов SCOP, а также их семейств. Процедура применена для классификации 1942 структур, относящихся к 314 комплексам белковых доменов с ДНК, и 34 семейств. В результате выделено 97 способов взаимодействия структур белкового домена с ДНК и 17 классов взаимодействия, характеризующих консервативные контакты с ДНК внутри семейств.

2. Предложенная классификация внедрена в базу данных НК-белковых взаимодействий. В базе данных есть возможность поиска структур с данным способом узнавания ДНК, а также поиск семейств белковых доменов, относящихся к данному классу взаимодействия с ДНК. Страницы семейств белковых доменов SCOP, содержащие большое количество структур, дополнены развернутым описанием.

3. Для 72 семейств ДНК-узнающих белковых доменов в структурах их комплексов с ДНК выявлены консервативно расположенные молекулы воды. Данные интегрированы в NPIDB.

4. Для двух семейств ДНК-узнающих белков, TATA-box связывающих белков и белков семейства LAGLIDADG_1, выявлены консервативные особенности ДНК-белкового взаимодействия в структурах их комплексов с ДНК.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных журналах

1. 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.
2. 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.
3. Kirsanov D., Zanegina O., Aksianov E., Spirin S., Karyagina A., Alexeevski A. NPIDB: Nucleic acid-Protein Interaction DataBase. // Nucleic Acids Res. – 2013. – Vol. 41. – Database issue. – P. D517–D523.
4. 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. – 2016. – Vol. 44. – Database issue. – P. D133–B143.

Тезисы научных конференций

5. 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.
6. 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. –2006. – Vol. 1. – P. 226.
7. 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.
8. 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. –2008. – Vol. 1. – P. 264.
9. 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. –2014 – Vol. 1. – P. 170.
10. 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.