ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ “ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК”

На правах рукописи

ПОНОМАРЕНКО МИХАИЛ ПАВЛОВИЧ

КОМПЬЮТЕРНЫЙ АНАЛИЗ КОНТЕКСТНО-ЗАВИСИМЫХ КОЛИЧЕСТВЕННЫХ ХАРАКТЕРИСТИК СПЕЦИФИЧЕСКОЙ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ САЙТОВ В СОСТАВЕ ГЕНОМНОЙ ДНК

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание ученой степени доктора биологических наук

Научный консультант
академик РАН, д.б.н., проф. Колчанов Н.А.

Новосибирск
2017

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Актуальность. Одной из важнейших задач биоинформатики является выявление взаимосвязей между структурой геномной ДНК и фундаментальными молекулярно-генетическими процессами (репликация, транскрипция, мутагенез, репарация), механизмы которых основаны на специфических взаимодействиях определенных участков (сайтов) в составе ДНК с белками, РНК-белковыми комплексами и низкомолекулярными соединениями. В последнее десятилетие развитие технологий иммунопреципитации хроматина (ChIP, Chromatin ImmunoPrecipitation) в комбинации с высокопроизводительным секвенированием (NGS, Next-Generation Sequencing) позволило накопить гигантские объемы количественных экспериментальных данных по комплексным ДНК-белковым взаимодействиям на уровне геномов. Это открыло принципиально новые возможности по выявлению разнообразных аспектов таких мультикомпонентных взаимодействий, в том числе сайтов связывания белков упаковки геномной ДНК, регуляторными белками, ферментами модификации и репарации ДНК, которые функционируют в зависимости от состояния клеток в тканях в норме, при различных внешних воздействиях и при разных патологиях. Объемы и сложность имеющейся информации таковы, что ее принципиально невозможно использовать без привлечения нового поколения методов биоинформатики (Меркулова и др., 2013). До развития технологий ChIP и NGS большинство биоинформатических исследований структурно-функциональной организации геномной ДНК были нацелены на распознавание сайтов связывания регуляторных белков (Oshchepkov, Levitsky, 2011). При этом лишь в некоторых редких разрозненных работах по биоинформатике и математической биологии исследовали физико-химические механизмы, лежащие в основе количественных аспектов взаимодействий между регуляторными белками и ДНК, а также особенности контекста функциональных сайтов ДНК, которые отражают эти взаимодействия (например, обзор (Stormo et al., 1986)). Революционные успехи технологий ChIP и NGS (Cheng, Gerstein, 2012) открыли новые возможности для применения подходов биоинформатики для предсказания не

только локализации функциональных сайтов разных классов в геномной ДНК, но и количественных величин их специфической биологической активности. Соответственно, возникла необходимость создания адекватных методов компьютерного анализа корреляций между символьными последовательностями различных вариантов исследуемого функционального сайта и численными значениями активности каждого из этих его вариантов, измеренными в определенном эксперименте. Это представляет собой одну из наиболее актуальных задач современной математической биологии. В настоящее время доступны огромные объемы экспериментальных данных о нуклеотидных последовательностях геномов, о структуре генов, их изменчивости, параметрах экспрессии, регуляторных контурах и сигнальных путях. Для учета этой информации в контексте структурно-функциональной организации сайтов в составе геномной ДНК необходимо выявление молекулярных механизмов, связывающих дискретные последовательности нуклеотидов и количественные величины их функциональной активности. В частности, актуальным является выявление закономерностей молекулярного кодирования путем кооперативных взаимодействий нуклеотидов в рамках конформационных и физико-химических свойств B-формы двойной спирали ДНК, определяющих существенные эффекты модуляции степени эффективности ее взаимодействия с регуляторными белками.

В качестве примера одной из фундаментально-значимых научных задач можно выделить механизм формирования транскрипционного комплекса в геномах эукариот, первым этапом которого является связывание TATA-связывающего белка (TBP) с соответствующим ему сайтом (TATA-боксом), располагающимся чаще всего на расстоянии 30 п.о. выше старта транскрипции. С момента открытия TATA-бокса (Lifton et al., 1978) к настоящему времени по механизмам взаимодействия его с TBP было опубликовано более 7 тыс. научных статей, включая полногеномную карту его локализаций в промоторах 17181 генов человека, выборочно доказанных в эксперименте *ex vivo* (Yang et al., 2011). Помимо собственно базовой биологической значимости процесса транскрипции, изучение механизмов молекулярного распознавания и связывания TBP с TATA-боксом необходимо для предсказания локализации стартов транскрипции генов в геномной ДНК и

оценки последствий мутаций вблизи этих двух ключевых элементов промоторов (сайт связывания TBP и сайт старта транскрипции) для количественных характеристик экспрессии эукариотических генов, их эволюции и развития патологий.

К числу приоритетных задач постгеномной биоинформатики относится также создание методов количественной оценки влияния генетической изменчивости на функционирование регуляторных элементов, контролирующих транскрипцию генов. Крупнейший в мире научный проект “1000 геномов” (Colonna et al., 2014) характеризует полиморфизм человека в терминах 8.58 млрд всех потенциально возможных однонуклеотидных замен (SNP, Single-Nucleotide Polymorphism) относительно референсного генома как общепринятой нормы. Описания проявлений каждого SNP, которые наблюдали клинически *in vivo*, предсказывали биоинформатически *in silico*, детектировали экспериментально *ex vivo* или *in vitro*, собирали в базу данных dbWGFP (Wu et al., 2016). В частности, уже накоплены большие массивы экспериментальных данных, свидетельствующих о том, что даже такие одиночные нуклеотидные замены могут существенно нарушать функции генов и приводить к серьезным патологиям (Hamosh et al., 2005). В связи с этим весьма актуальным является создание компьютерных методов, позволяющих количественно оценивать изменения равновесных и кинетических констант комплексов ДНК с регуляторными белками при мутациях нуклеотидов.

В ряде фундаментальных работ (Соловьев и др., 1989; Rogozin et al., 1991) была установлена контекстная предетерминированность соматического мутагенеза, одного из важнейших факторов опухолевой гиперэкспрессии онкогенов, однако вопрос о прогнозах количественных величин частот возникновения предмутационных повреждений ДНК и физико-химических констант равновесия и скоростей их репарации остался открытым. В связи с этим является актуальным выявление контекстно-зависимых количественных характеристик генома, достоверно связанных с воздействием на него тех или иных мутагенов и оценка эффективности его защиты от мутаций.

Любые теоретические и информационно-компьютерные подходы приобретают значимость в научных исследованиях лишь тогда, когда они

обладают предсказательной силой и дают возможность планировать на этой основе эксперименты, позволяющие выявлять новые знания. Поэтому весьма актуальным является обнаружение ранее неизвестных биологических фактов в экспериментах, которые были спланированы на основе учета предсказания контекстно-зависимых количественных величин биологической активности сайтов в составе геномных ДНК.

Компьютерный анализ количественных характеристик геномов является одним из важнейших направлений информационной системной биологии, выявляющих значимые связи между определенными последовательностями нуклеотидов в геноме и величинами биохимических, физиологических и морфологических признаков организмов, характеризующими ту или иную форму реализации генетических программ, кодируемых этими последовательностями нуклеотидов. Эти достоверные взаимосвязи между определенным порядком нуклеотидов в гене и количественными величинами фенотипических признаков организма могут стать фундаментом экспериментально-компьютерного анализа геномов пациентов предиктивно-превентивной персонифицированной медицины. Поэтому исследования в рамках данного направления биоинформатики считают весьма актуальными. В этом направлении проводились исследования в рамках настоящей диссертации.

Цель работы. Выявление особенностей структурно-функциональной организации сайтов в составе геномной ДНК, определяющих количественные характеристики их специфической биологической активности, на основе использования методов компьютерного анализа и моделирования.

Задачи, поставленные для достижения указанной цели, включали:

1. Разработать комплекс компьютерных программ для выявления контекстно-зависимых конформационных и физико-химических свойств двойной спирали ДНК, определяющих взаимодействия сайтов в составе геномной ДНК со специфическими белками;

2. Выявить особенности структурно-функциональной организации ДНК промоторов эукариот, определяющие количественные величины сродства ТАТА-связывающего белка к ТАТА-боксам перед стартами транскрипции белок-кодирующих генов;
3. Продемонстрировать возможность предсказания количественных параметров взаимодействия между сайтами в составе геномной ДНК различных таксонов и соответствующими регуляторными белками на основе контекстно-зависимых конформационных и физико-химических свойств двойной спирали ДНК;
4. Выявить особенности структурно-функциональной организации ДНК-сайтов, определяющих предрасположенность различных районов генома к предмутационным повреждениям.

Научная новизна. На основе теории аддитивной полезности для принятия решений и нечетких множеств впервые предложен компьютерный подход к изучению контекстно-зависимых количественных характеристик специфической биологической активности сайтов в составе геномной ДНК, который использует для анализа контекстные и контекстно-зависимые конформационные и физико-химические характеристики B-формы двойной спирали ДНК и выявляет ограниченный набор характеристик, достоверно коррелирующих с количественными величинами специфической биологической активности сайтов в составе геномной ДНК. На основе этого подхода впервые разработана компьютерная система для анализа контекстно-зависимых конформационных и физико-химических свойств B-формы двойной спирали ДНК (bDNAvideo). С использованием системы bDNAvideo впервые обнаружена достоверная кластеризация транскрипционных факторов на две группы, первая из которых включает преимущественно основные и Zn-координируемые белки с локальным избытком электростатического заряда, вторая - белки с β-слоем и с гомеодоменом без локального избытка электростатического заряда. Впервые создана компьютерная система (Activity) для выявления контекстных и контекстно-зависимых

конформационных и физико-химических характеристик ДНК, достоверно коррелирующих с экспериментально измеренными уровнями специфической биологической активности сайтов в составе геномной ДНК и построения на их основе линейно-аддитивных регрессионных уравнений для предсказания количественных величин биологической активности регуляторных сайтов по их последовательностям ДНК. Впервые были построены регрессионные уравнения, достоверно предсказывающие количественные величины сродства таких регуляторных белков, как Cro-репрессор, активатор CRP, транскрипционных факторов USF, MEF2, YY1 к сайтам их связывания. Впервые был создан метод предсказания частот повреждений гуанинов в ДНК под действием лазерного ультрафиолетового излучения с длиной волны 193 нм, подтвержденный данными независимых экспериментов. Впервые выведены оценки константы Михаэлиса K_M и каталитической константы k_{CAT} фермента 8-оксогуанина ферментом 8-оксогуанин-ДНК гликозилаза (OGG1) человека при нарушении комплементарности ДНК вокруг 8-оксогуанина, подтвержденные на независимых данных. Обнаружены достоверные корреляции между константой равновесия K_D комплекса TATA-связывающего белка (TBP) с нитью ДНК и частотой динуклеотидов WR и TV в нити ДНК; в случае двунитевой ДНК - с частотой динуклеотида TA и шириной малой бороздки спирали ДНК. Это впервые позволило достоверно предсказать величины K_D комплекса “TBP/ДНК” для независимых экспериментов ($p<10^{-6}$).

Положения, выносимые на защиту:

1. Сочетание теории полезности для принятия решений с нечеткими множествами позволяет выявлять контекстные, а также контекстно-зависимые конформационные и физико-химические характеристики B-формы двойной спирали ДНК сайтов в составе геномной ДНК, величины которых статистически достоверно коррелируют с экспериментально измеренными величинами специфической биологической активности этих сайтов;

2. Контекстно-зависимые характеристики функциональных сайтов ДНК, коррелирующие с их активностью, адекватно отражают такие биологически значимые особенности генома как предрасположенность к предмутационным повреждениям, эффективность репарации этих повреждений и сродство транскрипционных факторов к промоторам генов;
3. Уравнения регрессии, построенные на основе биологически значимых контекстно-зависимых характеристик сайтов в составе геномной ДНК, позволяют достоверно предсказывать величины специфической биологической активности этих сайтов по их произвольным нуклеотидным последовательностям.

Теоретическая значимость работы. Разработан новый подход к изучению контекстно-зависимых количественных характеристик специфической биологической активности сайтов в составе геномной ДНК на основе использования теории аддитивной полезности для принятия решений и нечетких множеств, который позволяет: (1) учитывать консенсусы, позиционно-весовые матрицы, частоты встречаемости олигонуклеотидов в 15-буквенном коде IUPAC-IUB, конформационные и физико-химические характеристики B-формы двойной спирали ДНК в качестве контекстно-зависимых количественных характеристик сайтов в составе геномных ДНК; (2) генерирует и единообразно проверяет более миллиона вариантов таких характеристик для выборок последовательностей ДНК; (3) отбирает ограниченные наборы контекстно-зависимых количественных характеристик сайтов в составе геномной ДНК, величины которых статистически достоверно коррелируют с экспериментально измеренными величинами специфической биологической активности этих сайтов.

Научно-практическая значимость работы. Разработанная в диссертации компьютерная система bDNAvideo и выявленные с ее помощью контекстно-зависимые конформационные и физико-химические свойства сайтов в составе геномных ДНК нашли практическое применение при создании ряда современных компьютерных систем, в том числе: SITECON

(Россия), BiDaS (Греция), CRoSSeD (Бельгия), DISCOVER (США), а также FeatureScan, DiProDB, BioBayesNet, ProMapper (все: Германия). Разработанная в диссертации компьютерная система Activity имеет широкую область практического применения для построения регрессионных уравнений на основе выборок нуклеотидных последовательностей сайтов в составе геномной ДНК с экспериментально измеренными для них величинами специфической активности с целью предсказания этих величин при анализе природных геномных ДНК, их естественного генетического разнообразия, а также их искусственных синтетических аналогов. Это является наиболее важным при планировании экспериментов в области синтетической биологии для генно-инженерного конструирования новых вариантов сайтов в составе геномных ДНК с заданными количественными величинами их специфической биологической активности. Исследование влияния генетической изменчивости геномной ДНК человека на уровни специфической биологической активности сайтов в ней создает возможность для экспериментально-компьютерной реконструкции молекулярных механизмов патогенного проявления SNP, клинически связанных с наследственными заболеваниями.

Апробация работы. Результаты диссертационной работы были доложены или представлены на 23 международных конференциях, в том числе: Pacific Symposium on Biocomputing (USA, 1997, 1998), “Bioinformatics of Genome Regulation and Structure, BGRS” (Novosibirsk, 1998, 2000, 2004, 2006, 2008, 2010, 2012, 2014, 2016), “Intelligent Systems for Molecular Biology, ISMB” (Canada, 1998), “Bridging the Gap between Sequences and Functions” (Cold Spring Harbor, USA, 1999), “Genome Sequencing & Biology” (Cold Spring Harbor, USA, 2001), “EuroQSAR 2002” (UK, 2002); на 10 российских конференциях, в том числе: "Геном человека" (Черноголовка, 2000), на Московских конференциях по вычислительной молекулярной биологии MCCMB (2009, 2013); на III Московской международной конференции “Молекулярная филогенетика MolPhy-3” (2012).

Публикации. По материалам диссертации опубликовано 55 научных работ, из них – 30 статей в журналах из Перечня ВАК (все индексированы в РИНЦ, Scopus и Web of Science), в том числе за рубежом – 19. Все работы - в соавторстве. В ряде исследований приняли участие В.П. Валуев, Д.В. Воробьев, Д.А. Григорович, В.М. Ефимов, С.В. Зубова, Л.В. Катохина, А.Э. Кель, Ф.А. Колпаков, А.Н. Колчанова, Н.А. Колчанов, С.В. Лаврюшев, Г.В. Орлова, Е.Л. Перегоедова, О.А. Подколодная, Н.Л. Подколодный, П.М. Пономаренко, Ю.В. Пономаренко, Д.А. Рассказов, В.В. Суслов, И.И. Титов, А.С. Фролов, Д.П. Фурман (все - отдел системной биологии ИЦиГ СО РАН), В.Ф. Кобзев (сектор химии нуклеиновых кислот ИЦиГ СО РАН), Г.В. Васильев и Т.И. Меркулова (оба - лаборатория регуляции экспрессии генов ИЦиГ СО РАН), О.В. Аркова, Т.В. Аршинова, И.А. Драчкова, М.В. Лысова и Л.К. Савинкова (все - сектор молекулярно-генетических механизмов белок-нуклеиновых взаимодействий ИЦиГ СО РАН), С.Е. Пельтек (лаборатория молекулярных биотехнологий ИЦиГ СО РАН), О.О. Кирпота, А.В. Ендуткин, Д.О. Жарков и Г.А. Невинский (все - ИХБФМ СО РАН, г. Новосибирск), Н.Н. Втюрина и А.Б. Васильев (МГУ, Москва), С.Л. Гроховский и Ю.Д. Нечипуренко (ИМБ РАН, Москва), М.С. Гельфанд (ИППИ РАН, Москва), а также C. Overton (University of Pennsylvania, Philadelphia, USA), A.V. Mazin и S.C. Kowalczykowski (оба - University of California, Davis, USA), H. Karas, H. Sclenar и E. Wingender (все - Gesellschaft fur Biotechnologische Forschung mbH, Braunschweig, Germany), L. Milanesi (Istituto Di Tecnologie Biomediche Avanzate, Milano, Italy), A. Sarai (The Institute of Physical and Chemical Research, RIKEN, Tsukuba, Japan).

Личный вклад автора. Все представленные в диссертации результаты были получены автором самостоятельно. Роль автора в статьях, которые были включены в “Список публикации по теме диссертации”, тогда как он не был в них автором для переписки, первым или последним автором, была обозначена “компьютерный анализ данных” в соответствие с темой диссертационной работы “Компьютерный анализ контекстно-зависимых количественных

характеристик специфической биологической активности сайтов в составе геномной ДНК”.

Работы автора в разделе “Список литературы”, которые не были включены в “Список публикации по теме диссертации”, сделаны вне диссертационной работы в рамках научных исследований лаборатории эволюционной биоинформатики и теоретической генетики ФГБНУ ФИЦ ИЦиГ СО РАН.

Структура и объем работы. Диссертация включает введение, обзор литературы, а также четыре главы с описанием материалов, методов, результатов и обсуждений в соответствующих разделах этих глав, заключения, выводов, списка литературы (467 источников), списка обозначений и сокращений. Работа изложена на 310 страницах машинописного текста, включая 81 рисунок и 41 таблицу.

Благодарности. Автор искренне благодарит сотрудников отдела системной биологии, лаборатории регуляции экспрессии генов и сектора молекулярно-генетических механизмов белок-нуклеиновых взаимодействий ИЦиГ СО РАН. Особую признательность автор выражает академику РАН Н.А. Колчанову, который инициировал весь цикл исследований и поддерживал диссертационную работу на всех этапах ее выполнения.

ГЛАВА 1 ОБЗОР ЛИТЕРАТУРЫ

Регуляция молекулярно-генетических процессов с участием геномной ДНК: репликация, мутагенез, репарация, транскрипция и другие, - осуществляются путем взаимодействия между ее участками, которые общепринято называть функциональными или регуляторными сайтами в составе геномных ДНК, и соответствующими белками или РНК-белковыми комплексами (Neidle, 1994). Поэтому основополагающими понятиями диссертационной работы являются “нуклеотидная последовательность ДНК” и “количественная характеристика последовательности ДНК”. Под понятием “*нуклеотидная последовательность ДНК”* подразумевается *специфическое упрощенное представление молекул ДНК, их фрагментов или синтетических олигонуклеотидов ДНК в форме линейных последовательностей конечной длины из четырех типов символов “A” (или “a” для аденина), “T” (или “t” для тимина), “G” (или “g” для гуанина) и “C” (или “c” для цитозина) для обозначения четырех канонических нуклеотидов (указанных в скобках).* Под “*количественной характеристикой последовательности ДНК*” подразумевается *количественная величина, численное значение которой было измерено экспериментально или вычислено с использованием определенной последовательности ДНК, ассоциированной с этим значением характеристики*.

1.1 Компьютерные базы данных геномных последовательностей

Современное состояние математической биологии и биоинформатики характеризуется прежде всего документированием и аннотированием всех экспериментально установленных нуклеотидных последовательностей геномных ДНК в компьютерных базах данных.

Начало III тысячелетия н. э. было ознаменовано эпохальным достижением науки в области молекулярной биологии: расшифровкой генома

человека. В 2004 году было завершено секвенирование так называемого “референсного” (т.е. общепринятого стандарта) генома человека (The International Human Genome Sequencing Consortium, 2004). Это событие считается началом новой постгеномной эры наук о жизни. Для нее является характерной высокопроизводительная расшифровка индивидуальных геномов людей относительно референсного генома человека в качестве основы для предиктивно-превентивной персонализированной медицины с возможностью диагностики, терапии и мониторинга заболеваний с учетом генетической предрасположенности пациента и его индивидуальной чувствительности к лекарственным препаратам.

В этой связи интенсивно ведется статистическое выявление ассоциаций между риском заболеваний и вариациями генома, которые документируются в базе данных NHGRI GWAS (Hindorff *et al*., 2009). Это позволило, например, связать наличие ряда однонуклеотидных замен (SNP, Single Nucleotide Polymorphism, англ. яз.) с генетической предрасположенностью к раку кожи (Gerstenblith *et al*., 2010). Экспериментально установленные SNP относительно референсного генома человека документируются в базе данных dbSNP (NCBI Resource Coordinators, 2013), их ассоциации с патологиями – в базе данных OMIM (Hamosh *et al*., 2005). Официальным хранилищем референсного генома человека является база данных Ensembl (Flicek *et al*., 2011), границы регуляторных и кодирующих районов генов - ее раздел GENCODE (Harrow *et al*., 2012), границы регуляторных сайтов – раздел ENCODE (Raney *et al*., 2011), название которого является аббревиатурой от “Энциклопедия элементов ДНК” (англ. яз.).

Самая первая база данных по секвенированным ДНК (Dayhoff *et al*, 1981) была создана в 1981 г., а годом позже в лаборатории Лос-Аламос появилась “Библиотека последовательностей ДНК” (Kanehisa, 1982; Kanehisa *et al*., 1984), преобразованная еще через три года в GenBank (Burks *et al*., 1985) – официальное хранилище всех свободно доступных вариантов расшифровки

```
LOCUS       ECOLAC                  7477 bp    DNA     linear   BCT 05-MAY-1993
DEFINITION  E.coli lactose operon with lacI, lacZ, lacY and lacA genes.
ACCESSION   J01636 J01637 K01483 K01793
.....
KEYWORDS    acetyltransferase; beta-D-galactosidase; galactosidase; lac operon;
.....
ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1243 to 1266)
  AUTHORS   Gilbert,W. and Maxam,A.
  TITLE     The nucleotide sequence of the lac operator
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 70 (12), 3581-3584 (1973)
   PUBMED   4587255
.....
REFERENCE   37 (bases 5646 to 7477)
  AUTHORS   Hediger,M.A., Johnson,D.F., Nierlich,D.P. and Zabin,I.
  TITLE     DNA sequence of the lactose operon: the lacA gene and the
            transcriptional termination region
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 82 (19), 6414-6418 (1985)
   PUBMED   3901000
.....
FEATURES             Location/Qualifiers
     source          1..7477
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
.....
     misc_signal     1162..1199
                     /note="cap protein binding site"
     variation       1183..1186
                     /note="ttag in wild-type; aatt in strain UV5 [26]"
.....
     mRNA            1246..>4358
                     /note="lacZ mRNA [2],[5]"
     misc_signal     1246..1266
                     /note="lac repressor protein binding site"
.....
ORIGIN      HindII site [Nature 274, 762-765 (1978)].
        1 gacaccatcg aatggcgcaa acctttcgc ggtatggcat gatagcgccc ggaagagagt
       61 caattcaggg tggtgaatgt gaaaccagta acgttatacg atgtcgcaga gtatgccggt
.....
     1021 ggcaatcagc tgttgcccgt ctcactggtg aaaagaaaaa ccaccctggc gcccaatacg
     1081 caaaccgcct ctccccgcgc gttggccgat tcattaatgc agctggcacg acaggtttcc
     1141 cgactggaaa gcgggcagtg agcgcaacgc aattaatgtg agttagctca ctcattaggc
     1201 accccaggct ttacacttta tgcttccggc tcgtatgttg tgtggaattg tgagcggata
.....
     7381 actgatggcg acactgcgac gttcgctgac atgctgatga agccagcttc cggccagcgc
     7441 cagcccgccc atggtaacca ccggcagagc ggtcgac
//
```

Рисунок 1 - Документ базы данных GenBank (2015), в котором можно видеть между позициями 1123 и 1186 самый первый секвенированный фрагмент ДНК длиной 64 п.о. (выделен **жирным** шрифтом) из работы (Maxam, Gilbert, 1977). Строка точек – пропуск однотипных строк.

природных нуклеотидных последовательностей ДНК и РНК (Benson *et al*, 1996; GenBank, 2015).

В качестве иллюстративного примера на Рисунке 1 показан документ базы данных GenBank (2015), который между позициями 1123 и 1186 (**жирный шрифт**) содержит самый первый фрагмент ДНК длиной 64 п.о., секвенированный Максамом и Гильбертом (Maxam, Gilbert, 1977). Можно

видеть, что документ GenBank (2015) состоит из трех следующих последовательно одна за другой частей.

Верхняя часть Рисунка 1: (белый фон) является описанием всего документа в целом. На Рисунке 1 это – фрагмент ДНК lac-оперона *Escherichia coli* длиной 7477 п.о., аннотированный результатами 37 оригинальных статей. Первым его расшифрованным фрагментом был участок длиной 24 п.о., который Максам и Гильберт резали на части тремя сайт-специфическими нуклеазами и измеряли их относительные веса методом “задержки в геле”, что указало им порядок нуклеотидов (Gilbert, Maxam, 1973) еще до изобретения ими секвенирования. Средняя часть документа GenBank, “FEATURE” (Рисунок 1: светло серый фон) - аннотация биологических свойств нуклеотидной последовательности, документированной в нижней части, “ORIGIN” (Рисунок 1: темно серый фон), в виде строк по 60 нт с номерами позиций от ее первого нуклеотида.

На Рисунке 2 показаны данные GenBank (2015) о его ежегодном росте: ● – число документов, ○ – сумма длин нуклеотидных последовательностей, Δ – средняя длина последовательности в документе. Всего GenBank выпуска № 207.0 от 15 апреля 2015 г. содержал 182188746 документов, сумма длин ДНК и РНК в которых была 189739230107 нт (GenBank, 2015). Как можно видеть на этом рисунке, средняя длина нуклеотидной последовательности – это константа 1000 нт/документ, тогда как объем GenBank удваивался каждые 18 месяцев ($r=0.98$, $\alpha<10^{-39}$).

Вследствие столь стремительного секвенирования ДНК и РНК возникли вопросы, остро требующие урегулирования. Например, какие условия сделали бы правомерным патентование расшифровки генома живого организма (Ohlschlegel, 1981)? Как сделать “официальный” вариант генома ценным для прогноза предрасположенности пациента к патологиям (McKusick, 1982)? В каком порядке секвенировать хромосомы разных биологических видов, если ассоциированные с каждой патологией человека гены по-разному картируются на эти хромосомы (Shows, 1983; Marx, 1985) в зависимости от

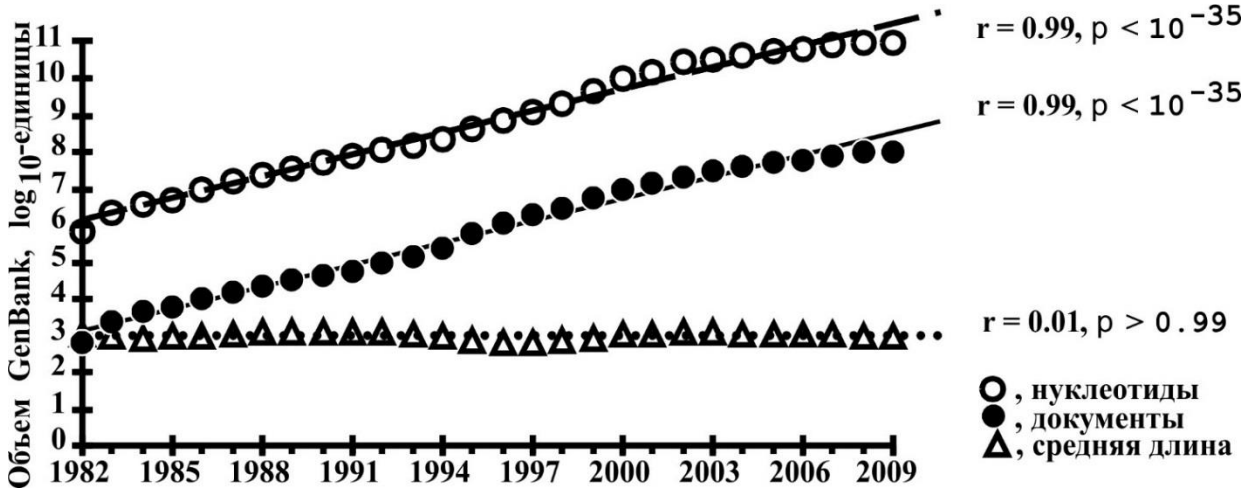

Рисунок 2 – Динамика базы данных GenBank (GenBank, 2015): с момента ее создания в 1982 (Kanehisa, 1982) объем удваивался каждые 18 месяцев как по количеству документов (●) и по суммарной длине всех нуклеотидных последовательностей ДНК и РНК (○) при сохранении постоянной средней длины нуклеотидной последовательности 1000 нт/документ (Δ). Коэффициент линейной корреляции r и его статистическая значимость p.

древа дивергенции человека и модельных животных от их общих предков (Bernardi, Bernardi, 1986)? В итоге взаимного согласования решений этих вопросов возник проект “Геном Человека” (Dulbecco, 1986), результат выполнения которого мог бы быть общедоступным в компьютерной базе данных. Эта информационно-технологическая (IT, сокращение от “**I**nformation **T**echnology”, англ. яз.) идея определила современное состояние компьютерного анализа регуляторных сайтов в составе геномных ДНК.

В результате успешного завершения проекта “Геном Человека” в 2004 г к настоящему времени было создано более тысячи компьютерных баз данных, из которых более 80 посвящены регуляторным районам геномов (Fernandez-Suarez *et al.*, 2014). Текущие состояния этих баз данных с 1982 г ежегодно публикуют спецвыпуски “Database issue” и “Web Server issue” журнала Nucleic Acids Research (Fernandez-Suarez *et al.*, 2014), а также такие

специализированные журналы как “Database (Oxford)” (например, описание базы данных dbWGFP (Wu *et al*., 2016)).

В Таблице 1, не претендующей на исчерпывающую полноту, можно видеть характерные примеры баз данных, которые используются для решения различных задач биоинформатики. В частности, база данных RepBase (Jurka *et al*., 2005) документирует мобильные генетические элементы, локализация и строение которых были идентифицированы методами биоинформатики для сравнения фрагментов геномов. В базе данных miRBase накапливаются сведения об экспериментально установленных микроРНК и об их биологических функциях (Griffiths-Jones *et al*., 2006).

Одним из хранилищ измеренных с помощью микрочипов величин количественных характеристик экспрессии генов в норме и при различных воздействиях на живые клетки в экспериментальных условиях *ex vivo* является база данных Genevestigator (Grennan, 2006). К числу наиболее полных массивов достоверных ассоциаций между генотипом и фенотипом организмов, включая метаболические сети, можно отнести базу данных KEGG (Kanehisa *et al*., 2008). Экспериментально установленные транскрипционные факторы и сайты их связывания на геномной ДНК эукариот - база данных TRANSFAC (Heinemeyer *et al*., 1999). Расшифрованные методами ядерного магнитного резонанса и рентгеноструктурного анализа пространственные (3D-) структуры ДНК, РНК и белков в форме декартовых координат их атомов были собраны в базе данных RCSB PDB (Dutta *et al*., 2009).

Следует также отметить, что существенный вклад в современное состояние баз данных по регуляторным сайтам в составе геномных ДНК был внесен ИЦиГ СО РАН, в котором автор выполнял настоящую диссертационную работу. В частности, мировым научным сообществом широко используется база данных TRRD (Kolchanov *et al*., 2002) по экспериментально установленным сайтам связывания транскрипционных факторов в регуляторных районах генов эукариот.

Таблица 1 – Примеры из числа 1170 банков данных, курируемых редакцией журнала Nucleic Acids Research (Fernandez-Suarez *et al.*, 2014).

База данных	объект	Информационное содержание	Ссылка
Ensembl	ДНК	Полные геномы 45 видов	(Hubbard *et al.*, 2009)
TRRD		Регуляции районы генов эукариот	(Kolchanov *et al.*, 2002)
TRANSFAC		Транскрипционные факторы и сайты их связывания на геномной ДНК эукариот	(Heinemeyer *et al.*, 1999)
dbSNP		Полиморфизм геномов	(Sherry *et al.*, 2001)
OMIM		Ассоциации генотип/фенотип	(Hamosh *et al.*, 2000)
rSNP_Guide		Полиморфизм регуляторных районов расшифрованных геномов	(Ponomarenko J. *et al.*, 2001b, 2002a,b, 2003)
GenBank	ДНК, РНК	Последовательности ДНК и РНК	(Benson *et al.*, 1996)
Genevestigator		Микрочип-измерения экспрессии генов	(Grennan, 2006)
Activity		Количественные измерения влияния мутаций на биологическую активность	(Ponomarenko J. *et al.*, 2001a)
RepBase		Мобильные элементы геномов	(Jurka *et al.*, 2005)
miRBase		Номенклатура миРНК: гены, мишени, формы	(Griffiths-Jones *et al.*, 2006)
SELEX_DB		Селекция олигоДНК и олигоРНК на заданную их биологическую активность *in vitro*	(Ponomarenko J. *et al.*, 2002c)
KEGG	ДНК, РНК, белки	Ассоциации генотип/фенотип	(Kanehisa *et al.*, 2008)
GeneNet		Генные сети – координация экспрессии генов	(Ananko *et al.*, 2005)
RCSB PDB		3D-структуры ДНК, РНК, белков и их комплексов	(Dutta *et al.*, 2009)
ASPD	белки	Селекция пептидов in vitro	(Valuev *et al.*, 2002)
PDBSite		Функциональные сайты белков	(Ivanisenko *et al.*, 2005)

В свою очередь, база данных GeneNet (Ananko *et al.*, 2005) содержит форматированную для автоматического компьютерного анализа информацию о генных сетях, представляющих собой наборы координированно экспрессируемых генов, их РНК и белковых продуктов, а также низкомолекулярных ко-факторов, субстратов, лигандов и метаболитов,

вовлеченных в функционирование этих макромолекул в определенных клеточных компартментах, тканях, органах и эукариотических организмах в целом. Кроме того, база данных PDBSite (Ivanisenko *et al*., 2005) является весьма часто используемым в мире источником информации об экспериментально известной локализации биологически активных центров в расшифрованных пространственных структурах глобулярных белков.

Наконец, ряд баз данных был создан с участием автора в рамках настоящей диссертации (Рисунок 3). Прежде всего, была создана база данных Activity (Колчанов и др., 1998; Ponomarenko J. *et al*., 2001a) по известным последовательностям ДНК сайтов связывания регуляторных белков, охарактеризованным экспериментальными количественными величинами их биологической активности. Она будет подробно представлена в главах 3 - 5.

Кроме того, в базе данных rSNP_Guide (Ponomarenko J. *et al*., 2001b, 2002a,b) был документирован полиморфизм регуляторной геномной ДНК, пример которого можно найти в разделе 5.3 настоящей диссертации. Наконец, база данных SELEX_DB (Ponomarenko J. *et al*., 2001a, 2002c) содержит результаты экспериментов по многократной циклической амплификации и селекции *in vitro* рандомизированных синтетических олигонуклеотидов ДНК (олигоДНК) на их высокое сродство к заданным регуляторным белкам.

Отличительной особенностью этих трех баз данных Activity (Колчанов и др., 1998; Ponomarenko J. *et al*., 2000a), rSNP_Guide (Ponomarenko J. *et al*., 2001b, 2002a,b, 2003) и SELEX_DB (Ponomarenko J. *et al*., 2001a, 2002c) является их информационное ядро, состоящее из двух оригинальных информационных разделов: SYSTEM об условиях экспериментов, результаты которых были документированы в этих базах данных, и CROSS_TEST о результатах перекрестного тестирования закономерностей, выявленных путем анализа данных одних опытов и, затем, подтвержденных с использованием данных независимых опытов. Пример документа из раздела SYSTEM показан на Рисунке 4, примеры информации, документированной раздела CROSS_TEST будут детально продемонстрированы в главе 5.

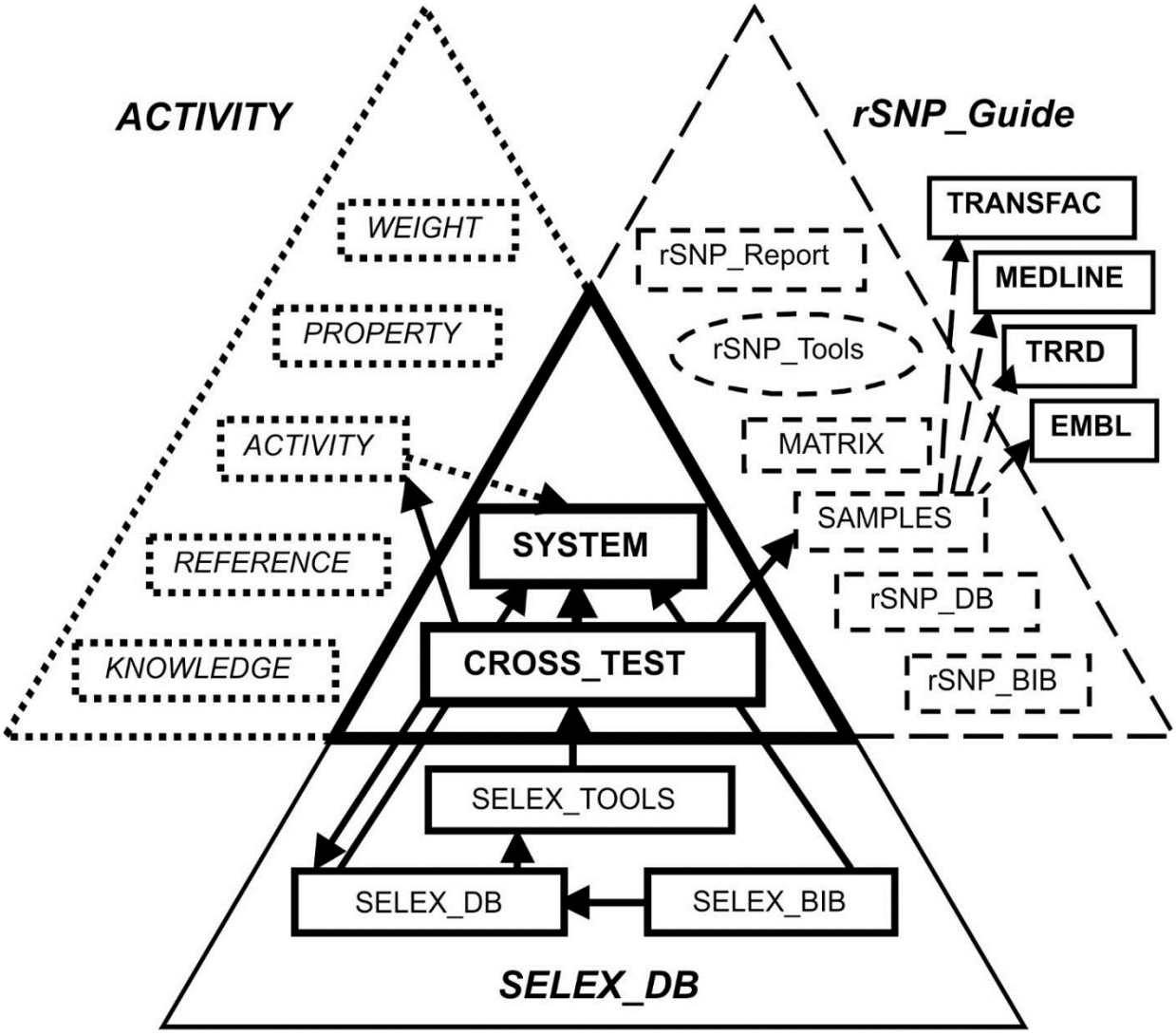

Рисунок 3 – Блок-схема системы баз данных по сайтам в составе геномных ДНК, созданной в настоящей диссертационной работе и включающей Activity (Колчанов и др., 1998; Ponomarenko J. *et al*., 2000a), rSNP_Guide (Ponomarenko J. *et al*., 2001b, 2002a,b, 2003) и SELEX_DB (Ponomarenko J. *et al*., 2001a, 2002c). □ - раздел; → - гиперссылки для биологически осмысленного перехода пользователя из одного документа в другой при поиске необходимой ему информации. ○ – компьютерные программы на Java Script. SYSTEM и CROSS_TEST – ядро из разделов “Условия опыта” и “Результаты перекрестных тестов закономерностей, выявленных на основе данных одних опытов и, затем, подтвержденных данными независимых опытов”. Рисунок автора из работы (Ponomarenko J. *et al*., 2002c).

Entry ID
T0SEL00J0031
Entry Name
Selection for YY1 binding sites
SRS-link
SCIENTIST; SCI00002
SELEX_BIB; RFSJ0021
SELEX_DB; S00J0031
Aim of an Experiment
To identify the range of DNA sequences to which YY1 can bind
Type of an Experiment
in vitro
Experimental Design and Materials
Six rounds of selection and amplification; GST-YY1 fusion protein bound to Glutathione-Sepharose beads;
Fusion protein was produced in E.coli lysate
Experimental Method
CASTing assay
Conclusion from an Experiment
The YY1 binding sites consensus was 5'-(C/g/a)(G/t)(C/t/a)CATN(T/a)(T/g/c)-3', with CGCCATTTT being the site captured most frequently; all of the single binding sites (47 clones out of 56) contain a core 5'-CCAT-3'
//

Рисунок 4 – Пример документа из раздела SYSTEM информационного ядра системы из трех баз данных Activity (Колчанов и др., 1998; Ponomarenko J. *et al*., 2000a), rSNP_Guide (Ponomarenko J. *et al*., 2001b, 2002a,b, 2003) и SELEX_DB (Ponomarenko J. *et al*., 2001a, 2002c). Общепринятая нотация GenBank (2015) из аббревиатур и ключевых слов для названий информационных полей (Рисунок 1) была заменена названиями (**жирный** шрифт, подчеркнуты) для так называемого “дружественного интерфейса”. Рисунок автора из его работы (Ponomarenko J. *et al*., 2002c).

Представленный краткий обзор не ставил своей целью исчерпывающего описания более тысячи современных баз данных по геномным последовательностям ДНК, а акцентировал внимание на двух ключевых проблемах: **неполнота** (т.е. в большинстве баз данных нет описания условий экспериментов: каким образом были получены эти данные) и **разрозненность** (т.е. отсутствие в большинстве баз данных сопоставления между результатами независимых экспериментов).

Обе эти проблемы были успешно преодолены автором путем создания информационного ядра для оригинальной системы баз данных (Рисунки 3 и 4), которую он разработал в рамках выполнения настоящей диссертационной работы и описал во всех последующих главах 2 – 5 диссертации.

1.2 Компьютерные системы анализа геномных последовательностей

Важной чертой современного состояния биоинформатики и математической биологии является широкое применение компьютерных систем и пакетов программ для анализа последовательностей геномных ДНК и РНК, собранных в компьютерных базах данных (как это было описано в предыдущем разделе).

Самый первый пакет компьютерных программ для поддержки секвенирования ДНК и РНК (Staden, 1977) был создан в 1977 г. Прежде всего, в нем был (Таблица 2) редактор SEQEDT ввода и редактирования

Таблица 2 – Самый первый пакет компьютерных программ для поддержки секвенирования нуклеотидных последовательностей геномных ДНК и РНК

Программа	Назначение	Ссылка
SEQEDT	редактор ввода и редактирования последовательности нуклеотидов	(Staden, 1977)
SEARCH	поиск заданной “маски” порядка нуклеотидов в последовательности	
SEQFIT	поиск совпадений между двумя последовательностями ДНК (РНК)	
TRANSQ	перекодирование ДНК (РНК) → белок	
BASTOT	подсчет количества каждого из канонических нуклеотидов	
HAIRPN	поиск палиндромов	(Staden, 1978)
DIAGON	построение dot-матриц парного выравнивания последовательностей	(Staden,1982a)
DBCOMP	непротиворечивая сшивка последовательности из shotgun-контигов	(Staden,1982b)
ANALYSEQ	поиск сайтов по позиционно-частотной матрице,	(Staden,1984a)
	поиск открытых рамок считывания	(Staden,1984b)

нуклеотидных последовательностей, который гарантировал надежность документирования результатов секвенирования. В пакете были также программы (Таблица 2) поиска сайтов ДНК и РНК по заданным “маскам” нуклеотидов в позициях этих сайтов (SEARCH), поиска совпадений между последовательностями (SEQFIT) и перекодирования нуклеотидной последовательности в аминокислотную (TRANSQ). Предложенные при этом (Staden, 1977) способ визуализации последовательностей ДНК и РНК по 60 нт в строке в виде 6 блоков длиной 10 нт и контрольный подсчет количества каждого из канонических нуклеотидов (BASTOT) стали стандартами баз данных по геномным последовательностям.

Возможности этого пакета были расширены (Staden, 1978) сначала поиском совершенных комплементарных палиндромов (“шпилек”, HAIRPN), затем (Staden, 1982a,b) - выравниванием пар нуклеотидных последовательностей (DIAGON) с использованием метода Нидлмана-Вунша (Needleman, Wunsch, 1970) и на этой основе непротиворечивой сшивки коротких элементарных единиц секвенирования, “shotgun”-контигов, в протяженные варианты расшифровки фрагментов геномов (DBCOMP). В более поздние выпуски этого пакета (Staden, 1984a,b) были добавлены возможности поиска сайтов ДНК и РНК с помощью позиционно-частотных матриц нуклеотидов и потенциальных открытых рамок считывания (ORF, **O**pen **R**eading **F**rame, англ. яз.) с учетом частот кодонов.

В свою очередь лаборатория Лос Аламос создала пакет SAS компьютерной поддержки секвенирования фрагментов геномов (Kanehisa, 1982), ставший впоследствии составной частью GenBank (Kanehisa *et al*., 1984; Burks *et al*., 1985).

Если самый первый спецвыпуск по биоинформатике журнала “Nucleic Acids Research” опубликовал в 1982 году 38 компьютерных баз данных и программ для анализа геномных последовательностей, то каждый ежегодный спецвыпуск “Web Server issue” этого журнала за предыдущие 10 лет

публиковал ≈100 компьютерных пакетов и систем программ, свободно доступных в сети Интернет.

В не претендующей на исчерпывающую полноту Таблице 3 можно видеть некоторые характерные примеры современных пакетов и систем компьютерных программ для решения различных задач биоинформатики. Несомненно, к числу самых важных из них, определяющих современное состояние биоинформатики и математической биологии в целом, следует отнести UCSC Genome Browser референсного генома человека (Karolchik *et al*., 2014). По-видимому, самыми используемыми компьютерными системами для манипулирования геномными последовательностями ДНК и РНК можно считать CLUSTAL (Sievers *et al*., 2011) и BLAST (Johnson *et al*., 2008).

Развитием символьного выравнивания последовательностей ДНК является система FeatureScan (Deyneko *et al*., 2006) для выравнивания количественных профилей конформационных и физико-химических свойств спирали ДНК вдоль последовательностей промоторов, которая основана на результатах главы 2 настоящей диссертации. Самый первый подход к компьютерному анализу контекстно-зависимых свойств спирали ДНК был предложен в системе CURVATURE (Shpigelman *et al*., 1993), реконструирующей 3D-ход оси спирали ДНК по ее заданной последовательности. Самыми используемыми для анализа районов регуляции транскрипции генов эукариот являются системы MATRIX SEARCH (Chen *et al*., 1995), MatInspector (Quandt *et al*., 1995) и TESS (Schug, Overton, 1997) для распознавания сайтов связывания транскрипционных факторов с помощью позиционно-частотных матриц, документированных в базе данных TRANSFAC (Heinemeyer *et al*., 1999).

В отделе системной биологии ИЦиГ СО РАН, где выполнялась диссертационная работа, была создана аналогичная система SiteGA (Levitsky *et al*., 2014), а также система SITECON (Oshchepkov *et al*., 2004), основанная на результатах главы 2 настоящей диссертации.

Таблица 3 – Примеры из более тысячи курируемых редакцией журнала Nucleic Acids Research пакетов и систем компьютерных программ для решения различных задач биоинформатики

Компьютерная система	Вид анализа	Ссылка
UCSC Genome Browser	Манипулирование геномом человека, hg19	(Karolchik *et al*., 2014)
CLUSTAL	Множественное выравнивание ДНК	(Sievers *et al*., 2011)
BLAST	Поиск контекстно сходных фрагментов геномной ДНК путем выравнивания	(Johnson *et al*., 2008)
FeatureScan	Поиск конформационно-сходных промоторов в геномах прокариот	(Deyneko *et al*., 2006)
CURVATURE	Реконструкция по последовательности ДНК пространственного хода оси B-спирали	(Shpigelman *et al*., 1993)
MATRIX SEARCH	Распознавание сайтов связывания транскрипционных факторов в составе геномных ДНК	(Chen *et al*., 1995)
MatInspector		(Quandt *et al*., 1995)
TESS		(Schug, Overton, 1997)
SiteGA		(Levitsky *et al*., 2014)
SITECON		(Oshchepkov *et al*., 2004)
Complexity	Оценка текстуальной сложности геномной ДНК	(Orlov, Potapov, 2004)
Leader_RNA	Дихотомический прогноз высокого или низкого уровня экспрессии генов белков по участку мРНК от кэп-сайта до стартового AUG-кодона	(Kochetov *et al*., 1999)
Likeness	Поиск конформационного сходства в пространственных структурах белков	(Пономаренко М. и др., 1999)
rSNP_Guide	Прогноз транскрипционного фактора, чей сайт связывания изменен мутацией с учетом данных метода “задержки в геле”	(Ponomarenko J. *et al*., 2002a, 2003)
bDNAvideo	Выявление статистически значимых для регуляторных сайтов свойств B-спирали ДНК	(Ponomarenko M. *et al*., 1997b)
Activity	Выявление количественных характеристик ДНК, которые коррелируют с измеренными величинами активности регуляторных сайтов	(Ponomarenko M. *et al*., 1997a)
GeneExpress	Комплексный анализ геномной ДНК	(Kolchanov *et al*., 1999)

Примером пакета программ, созданного в ИЦиГ СО РАН, является Complexity (Orlov, Potapov, 2004), который по заданной последовательности геномной ДНК оценивает широкий набор типов ее контекстной сложности (лингвистической и др.), рассмотренных в разделе 1.3.4 настоящей главы.

В числе программных продуктов отдела системной биологии ИЦиГ СО РАН имеется также ряд систем, созданных при участии автора вне рамок этой диссертации. Например, Leader_RNA (Kochetov *et al*., 1999) предсказывает высокий или низкий уровень экспрессии гена по последовательности от кэп-сайта до стартового AUG-кодона зрелой мРНК, продукта этого гена. Также, система Likeness (Пономаренко М. и др., 1999в) обобщает символьное выравнивание последовательностей ДНК, РНК и белков на случай 3D-выравнивания пространственных структур белков в форме декартовых координат их атомов, на основе использования результатов глав 2 и 3 настоящей диссертации. Наконец, в качестве развития результатов раздела 5.3 настоящей диссертации была создана система rSNP_Guide (Ponomarenko J. *et al*., 2002a, 2003), которая по данным опыта “задержки в геле” комплексов нормальной и мутантной ДНК с белковым экстрактом ядер из определенной клеточной линии предсказывает транскрипционный фактор, сайт связывания которого был изменен мутацией относительно нормы геномной ДНК.

Наконец, в рамках настоящей диссертации автор создал две компьютерные системы: bDNAvideo (Ponomarenko M. *et al*., 1997b) и Activity (Ponomarenko M. *et al*., 1997a), которые были интегрированы вместе со всеми программными продуктами отдела системной биологии ИЦиГ СО РАН в компьютерную систему GeneExpress (Kolchanov *et al*., 1999) для комплексного анализа геномов.

На Рисунке 5 представлен взятый из работы (Kolchanov *et al*., 1999) пример результатов такого комплексного анализа 500 участков ДНК [-300; +100] относительно главных стартов транскрипции негомологичных генов эукариот.

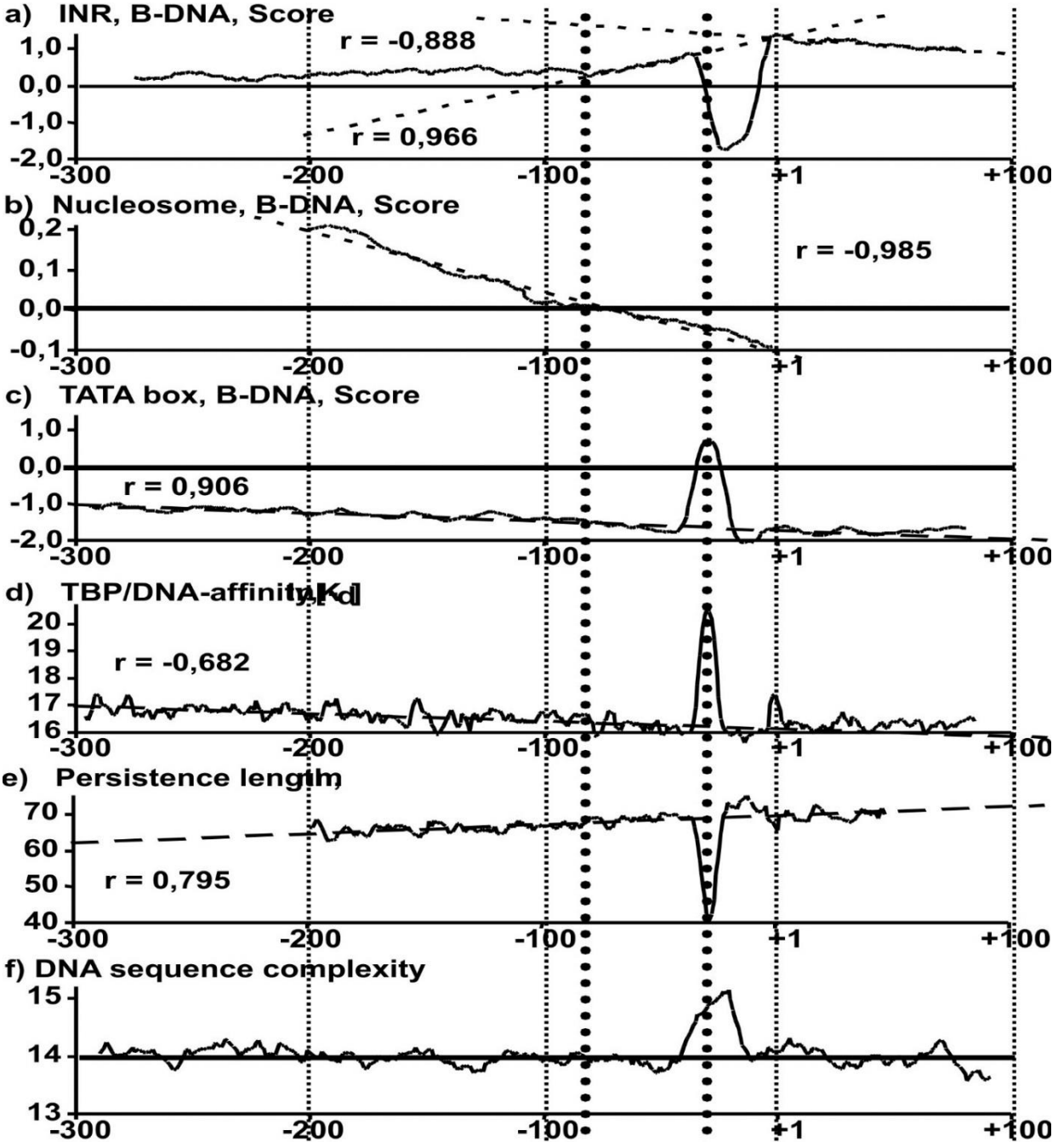

Рисунок 5 – Пример комплексного анализа 500 негомологичных промоторов генов эукариот (ось x) с помощью компьютерной системы GeneExpress (Kolchanov *et al*., 1999). Рисунок автора из его работы (Kolchanov *et al*., 1999).

Этот рисунок демонстрирует отличительную особенность систем bDNAvideo (Ponomarenko M *et al*., 1997b), Activity (Ponomarenko M *et al*., 1997a) и GeneExpress (Kolchanov *et al*., 1999), суть которой - осуществить как можно больше альтернативных вариантов анализа сайтов в составе геномных ДНК и взять из всех результатов лишь ограниченный взаимосогласованный

набор, все составляющие части которого достоверно соответствуют одни другим с целью синтеза на этой основе единой непротиворечивой картины регуляции экспрессии генов.

В представленном обзоре автор не ставил своей целью исчерпывающе полно описать сотни компьютерных систем анализа геномных ДНК, а фокусировал внимание на их отличительном свойстве в сравнении с пакетами разрозненных программ такого анализа: создание возможности получения нового интегрального результата, который выходит за рамки результатов отдельных модулей системы. На Рисунке 5 можно видеть, что таким интегральным результатом компьютерной системы GeneExpress (Kolchanov *et al*., 1999) может быть выявление достоверных корреляций между результатами анализа локальных окрестностей определенных сайтов в составе геномной ДНК с помощью различных модулей этой системы.

1.3 Методы компьютерного анализа геномных последовательностей

1.3.1 Методы статистического анализа геномных последовательностей

Самым первым методом биоинформатики является, по-видимому, статистический метод, с помощью которого открыли достоверную взаимосвязь между аминокислотными остатками и тринуклеотидами в природных белках и РНК, соответственно (Gamow, Ycas, 1955). Отдельно для каждого из 22 исследуемых белков определяли частоту p(k) встречаемости k-ого остатка в их упорядоченном списке по убыванию значений этих частот и, затем, усреднили p(k) для каждого k (то есть независимо от сходства или различия аминокислот с одинаковым k для разных белков). Результат этого усреднения (Gamow, Ycas, 1955) представлен темными кружками (●) на Рисунке 6 (ось y, в %), порядковые номера k - по горизонтали (ось x, в рангах).

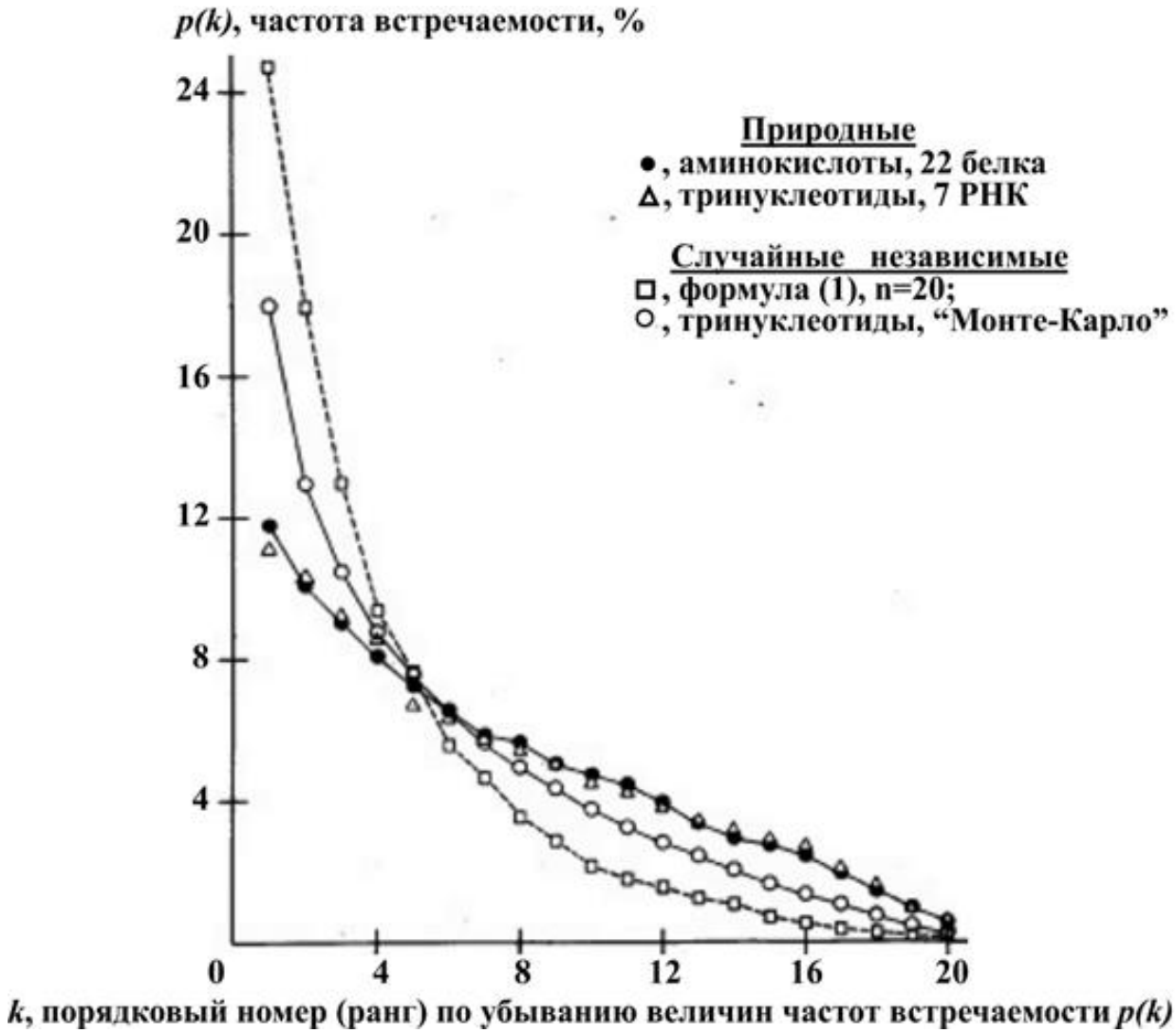

Рисунок 6 – Открытие (Gamow, Ycas, 1955) корреляции между частотами встречаемости p(k) аминокислотных остатков (●) и неперекрывающихся тринуклеотидов (Δ) в природных белках и РНК, соответственно. Оценки величин этих частот в случайных последовательностях независимых равновероятных символов (1) (○) и моделированию *in silico* методом Монте-Карло (□). Рисунок автора на основе работы (Gamow, Ycas, 1955).

В целях статистического анализа последовательностей независимых равновероятных случайных символов из алфавита объема n вывели (Gamow, Ycas, 1955: Appendix A) аналитическую формулу для ожидаемых таких величин p(k):

$$p(k|1 \le k \le n) = \frac{1}{n}\sum_{i=k}^{n}\frac{1}{n-i+1}. \quad (1)$$

Светлые кружки (○) на Рисунке 6 показывают предсказание формулы (1) для 20 канонических аминокислотных остатков (n=20). В качестве объяснения очевидного отличия природных белков (●) от случайных последовательностей (○) рассмотрели две гипотезы: (i) порядок нуклеотидов в белок-кодирующих РНК неслучаен и (ii) триплетные кодоны аминокислот не перекрываются в РНК (Gamow, Ycas, 1955). С целью проверки этих двух гипотез сгенерировали методом Монте-Карло (Metropolis, Ulam, 1949) последовательности из четырех независимых случайных равновероятных символов, группировали их триплеты в 20 непересекающихся наборов и за 3000 испытаний достигли предела сходимости p(k), показанного светлыми квадратами (□) на Рисунке 6.

На основании сходства результата компьютерного моделирования (□) с прогнозом формулы (1) (○), а не с экспериментальными данными о природных белках (●), сделали вывод о неслучайном порядке нуклеотидов в белок-кодирующих РНК (Gamow, Ycas, 1955). Чтобы понять биологический смысл неслучайных тринуклеотидов обработали 7 природных последовательностей РНК с использованием процедуры вычисления p(k) из модели Монте-Карло (Gamow, Ycas, 1955). Полученный результат показан светлыми треугольниками (Δ) на Рисунке 6. На этом рисунке можно видеть совпадение частот p(k) для аминокислотных остатков в природных белках (●) и для тринуклеотидов в природных РНК (Δ). Из этого совпадения сделали вывод о переносе генетической информации в живой клетке с нити РНК в нить белка в виде соответствия между аминокислотными остатками и неперекрывающимися тринуклеотидами (Gamow *et al*., 1956).

Через 12 лет, экспериментальное открытие генетического кода (Nirenberg *et al*., 1968) подтвердило этот самый первый результат биоинформатики, который был получен с помощью статистического метода компьютерного моделирования РНК и белков. Введенные Гамовым (Gamow, Ycas, 1955) частоты аминокислотных остатков и нуклеотидов до сих пор используются в качестве неотъемлемого атрибута последовательностей ДНК,

Таблица 4 – Примеры биологически значимых закономерностей, найденных с помощью методов статистического анализа.

Биологическое явление	объект	Статистическая закономерность	Ссылка
Соматический гипермутагенез	ДНК	Корреляция соматического гипермутагенеза с несовершенными повторами в ДНК	(Rogozin *et al*., 1991)
Сайты связывания транскрипционных факторов		Идентификация по данным ChIP-seq достоверных кластеров ДНК-сайтов связывания транскрипционных факторов	(Kuznetsov V *et al*., 2007)
Видовая специфичность геномов прокариот		Спектр частот динуклеотидов в геномах прокариот – “динуклеотидная подпись вида”	(Karlin, Burge, 1995)
Видовая специфичность промоторов эукариот		Спектр частот динуклеотидов в промоторах эукариот имеет видовую специфичность	(Колпаков и др., 1997)
Горизонтальный перенос генов		Противоречия древ филогении видов и частот использования нуклеотидов в их геномах	(Tamames, Moya, 2008)
Старт транскрипции гена эукариот		Частота гуанина вокруг стартов транскрипции генов смещена, как [p(G)-p(C)]/[p(G)+p(C)] > 0	(Tatarinova *et al*., 2003)
Интроны и экзоны гена эукариот	РНК	Белок-кодирующие экзоны отличаются от интронов по частотам гексануклеотидов	(Farber *et al*., 1992)
Альтернативные AUG-старты трансляции		Низкая частота триплетов AUG перед стартовым AUG-кодоном	(Kozak, 1989)

РНК и белков при их документировании в базах данных (GenBank, 2015), а также при компьютерном анализе расшифрованных геномов.

В Таблице 4 приведены некоторые характерные примеры биологических закономерностей, которые были обнаружены методами статистического анализа. Прежде всего, с помощью метода моделирования Монте-Карло (Metropolis, Ulam, 1949) была обнаружена достоверная корреляция соматического гипермутагенеза с несовершенными повторами в

геномах (Rogozin *et al*., 1991). Использование этого метода является неотъемлемым этапом экспериментально-компьютерной идентификации достоверных кластеров сайтов связывания транскрипционных факторов с использованием данных ChIP-seq по иммунопреципитации хроматина и высокопроизводительного секвенирования (Kuznetsov V. *et al*., 2007).

Авторы работы (Karlin, Burge, 1995; Karlin, 1998) обнаружили достоверное различие геномов бактерий по спектру частот динуклеотидов, который был назван ими “динуклеотидной подписью вида”.

С участием автора вне рамок настоящей диссертации было показано (Колпаков и др., 1997) аналогичное явление в случае промоторов генов эукариот. Достоверные отличия спектров частот динуклеотидов на протяженных участках генома от полногеномного такого спектра для определенного биологического вида, которые значимо коррелируют с полногеномным спектром частот динуклеотидов другого вида, используются для статистической проверки гипотезы о виде-доноре генетической информации при ее горизонтальном переносе (Tamames, Moya, 2008).

В свою очередь, в работе (Tatarinova *et al*., 2003) обнаружили достоверный признак окрестности старта транскрипции генов *Arabidopsis thaliana*, являющийся смещением частоты гуанина (G) вследствие адаптивного мутагенеза (Korogodin *et al*., 1991): многократно более частые мутации в регуляторных районах генов, чем мутации в кодирующих районах этих генов (Колчанов, 1989).

Авторы статьи (Farber *et al*., 1992) обнаружили достоверное отличие белок-кодирующих экзонов от интронов по частотам гексануклеотидов. Козак (Kozak, 1989) открыла достоверно низкую частоту триплетов “AUG” перед стартовым AUG-кодоном, названную впоследствии в ее честь “правило Козак”.

В целом, статистический анализ является неотъемлемой составной частью современных компьютерных подходов к исследованию регуляторных сайтов в составе геномных ДНК.

1.3.2 Контекстно-зависимые конформационные и физико-химические свойства геномной ДНК

По-видимому, один из самых первых результатов анализа контекстно-зависимых физико-химических свойств геномных ДНК был получен в работе (Vologodskii, Frank-Kamenetskii, 1978), когда оценили скользящим окном длиной L = 20 п.о. содержание нуклеотидов A и T вдоль генома бактериофага fd, $\{s_i\}_{i\leq 6383}$, длиной 6383 п.о. (Рисунок 7a, сверху):

$$[A+T]_L(i) = \frac{100\%}{L}\sum_{j=i-L/2}^{i+L/2}\Delta\big(s_j \in \{A,T\}\big), \quad (2)$$

где: Δ(истина)=1 и Δ(ложь)=0 – функция-индикатор.

Положения и высоты пиков содержания нуклеотидов A и T (формула 2) вдоль геномной ДНК бактериофага fd (Рисунок 7a, сверху) достоверно совпали с результатом независимого опыта по построению карты денатурации этого генома (Рисунок 7a, снизу) при температурах от 63 ^{O}C до 75 ^{O}C (Tachibana *et al*., 1978). Авторы этого опыта (Tachibana *et al*., 1978) оценили линейно-аддитивный вклад $T_M(s_is_{i+1}=\zeta\xi)$ каждого из 16 возможных динуклеотидов ζξ последовательности $\{s_i\}_{a\leq i\leq b}$ в температуру плавления ДНК между позициями a и b генома (Gotoh, Tagashira, 1981):

$$T_{M;[a;b]} = \frac{1}{b-a}\sum_{i=a}^{b-1} T_M(s_is_{i+1}), \quad (3)$$

где $T_M(\zeta\xi)$ – оценки линейно-аддитивного вклада каждого из всех 16 возможных динуклеотидов ζξ в температуры плавления ДНК, полученные (Gotoh, Tagashira, 1981) из данных опыта (Tachibana *et al*., 1978) и документированные автором в базе данных PROPERTY (Колчанов и др., 1998), созданной им в рамках настоящей диссертации и описанной в главе 2.

Благодаря этой серии работ (Vologodskii, Frank-Kamenetskii, 1978; Tachibana *et al*., 1978; Gotoh, Tagashira, 1981) возникла идея о возможном влиянии нуклеотидной последовательности на физико-химические свойства

б)

```
MI P0000022
MN Physico-chemical
RN Gotoh & Tagashira , Biopolymers, 1981
PN Melting temperature of B-helical DNA
PU Celsius degree
AA 54.50
AT 57.02
AG 58.42
AC 97.73
TA 36.73
TT 54.50
TG 54.71
TC 86.44
GA 86.44
GT 97.73
GG 85.97
GC 136.12
CA 54.71
CT 58.42
CG 72.55
CC 85.97
```

Рисунок 7 – Совпадение (а) между построенным в работе (Vologodskii, Frank-Kamenetskii, 1978) содержанием нуклеотидов A и T в геноме бактериофаге fd (сверху) и (снизу) экспериментальной температурной картой денатурации этого генома (Tachibana *et al*., 1978); (б) документ созданной в рамках настоящей диссертации базы данных “PROPERTY” (Колчанов и др., 1998) со значениями температуры плавления ДНК для всех 16 возможных динуклеотидов, оцененных авторами статьи (Gotoh, Tagashira, 1981) на основе температурной карты из статьи (Tachibana *et al*., 1978). Рисунок автора на основе работы (Vologodskii, Frank-Kamenetskii, 1978).

ДНК, которая вывела ДНК за рамки пассивного хранилища генетической информации в форме идеальной спирали ДНК (Watson, Crick, 1953).

В это же самое время авторы опыта (Klug *et al*., 1979) предположили возможность контекстно-зависимых отклонений спирали ДНК от ее идеальной формы (Watson, Crick, 1953), чтобы объяснить обнаруженные ими различия связывания между олигонуклеотидами $d\{AT\}_n$ с разным контекстом химически модифицированных T и lac-репрессором *E. coli*. В подтверждение этой гипотезы все три самые первые расшифрованные рентгеноструктурным анализом 3D-структуры коротких олигонуклеотидов $d\{CG\}_3$ (Wang *et al*, 1979), $d\{CG\}_2A_2T_2\{CG\}_2$ (Wing *et al*., 1980) и $d\{G_2C_2\}_2$ (Wang *et al*., 1982)

б)

```
MI P0000018
MN Conformational
RN Kabsch, Sander, Trifonov, NAR, 1982
PN Twist regressed linearly from X-ray
PU Degree
AA 35.62
AT 31.50
AG 27.70
AC 34.40
TA 36.00
TT 35.62
TG 34.50
TC 36.90
GA 36.90
GT 34.40
GG 33.67
GC 40.00
CA 34.50
CT 27.70
CG 29.80
CC 33.67
```

Рисунок 8 - Профиль (а) угла кручения τ "*twist*" спирали ДНК (сверху) вдоль додекамера $d\{CG\}_2A_2T_2\{CG\}_2$ (Wing *et al*., 1980) и его отклонения δτ от инвариантного значения этого угла в идеальной спирали ДНК (Watson, Crick, 1953) зависят от порядка нуклеотидов, "правила Калладина" (Calladine, 1982); б) документ базы данных "PROPERTY" (Колчанов и др., 1998) с величинами угла "*twist*" для всех 16 динуклеотидных шагов спирали ДНК (Kabsch *et al*., 1982). Рисунок автора на основе работы (Calladine, 1982).

соответствовали трем разным формам двойной спирали ДНК: левозакрученной Z-форме, правозакрученной B-форме с наклоном комплементарных пар к оси спирали ДНК и A-форме без какого-либо их наклона. Калладин (Calladine, 1982) эвристически увидел в них закономерность отклонений углов кручения "*twist*" и раскрытия "*roll*" соседних пар оснований от инвариантных значений идеальной двойной спирали ДНК (Watson, Crick, 1953; Wilkins *et al*., 1953; Franklin, Gosling, 1953), названную "правилами Калладина" (Рисунок 8а).

Благодаря кристаллографическим данным об оптических свойствах 40 олигоДНК с известными последовательностями были оценены методом линейной регрессии величины трех углов Эйлера сферической системы координат: кручение *twist* (Kabsch *et al*., 1982), наклоны *wedge* и *direction*

(Shpigelman *et al*., 1993), - для всех 16 динуклеотидов. На этой основе создали отмеченную выше (раздел 1.2) систему CURVATURE (Shpigelman *et al*., 1993) для реконструкции 3D-хода оси спирали ДНК на основе использования произвольной последовательности ДНК.

Величины этих конформационных углов спирали ДНК были впоследствии уточнены для свободной ДНК и для комплексов ДНК/белок по отдельности (Suzuki *et al*., 1996) посредством их усреднения по всем расшифрованным пространственным структурам ДНК в базе данных RCSB PDB (Dutta *et al*., 2009). В диссертационной работе автор документировал (Рисунок 8б) все варианты величин углов спирали ДНК в базе данных “PROPERTY” (Колчанов и др., 1998).

В свою очередь, с использованием 3D-структуры комплекса ДНК с репрессором фага 434 были получены (Hogan, Austin, 1987) оценки линейно-аддитивных вкладов всех возможных 16 динуклеотидов в персистентную длину ДНК, характеризующую такое физико-химическое свойство спирали ДНК, как жесткость ее изгиба. Кроме того, по 177 участкам ДНК длиной 145 п.о. экспериментально установленных контактов с октамером гистонов нуклеосом курицы были оценены частоты контактов гистонов с каждым из всех 16 динуклеотидов (Satchwell *et al*., 1986).

Следующим шагом изучения влияния геномных последовательностей ДНК на локальную конформацию спирали ДНК стало введение ее номенклатуры (Dickerson *et al*., 1989), которая выборочно представлена на Рисунке 9.

С помощью методов компьютерного моделирования молекулярной динамики (Karas *et al*., 1996) биоинформатически были оценены величины ряда свойств конформации ДНК всех 4096 возможных гексануклеотидов. Аналогично, авторы статьи (Sugimoto *et al*., 1996) оценили величины вкладов каждого из 16 динуклеотидных шагов спирали ДНК в ее энтропию, энтальпию и энергию Гиббса как дополнения к экспериментально установленным свойствам спирали ДНК.

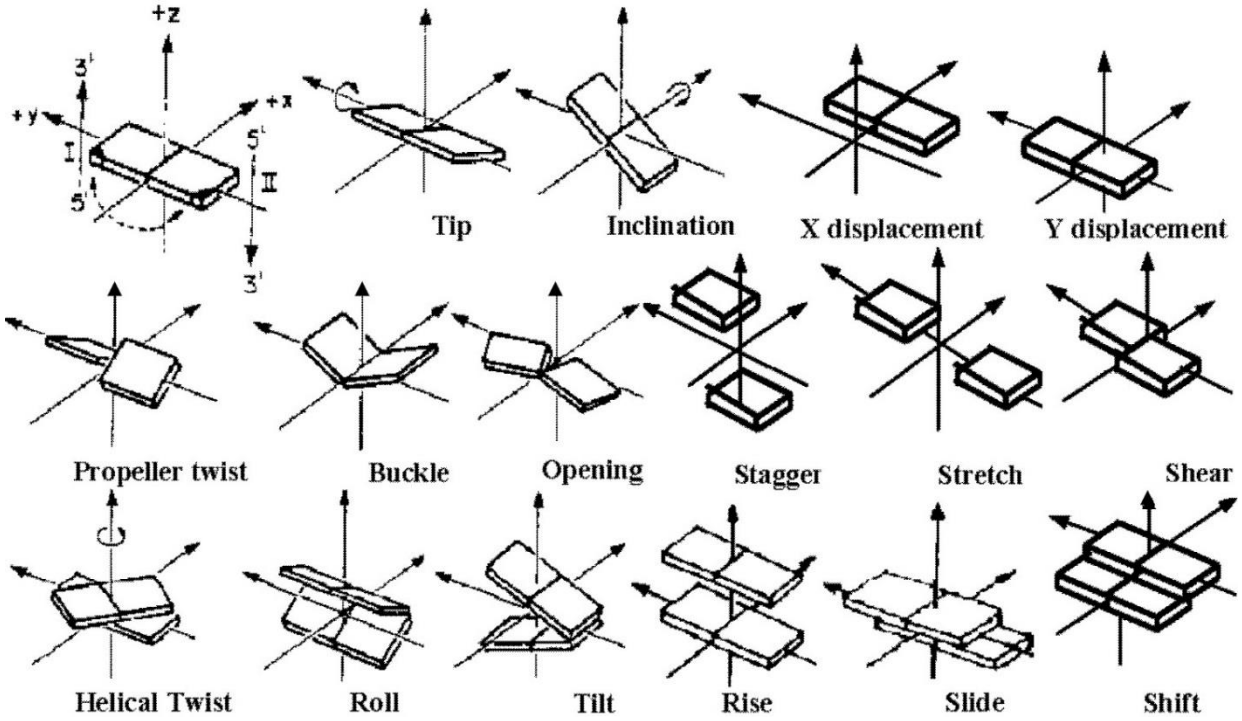

Рисунок 9 - Номенклатура конформационных количественных характеристик для B-формы спирали ДНК (Dickerson *et al*., 1989). Рисунок автора на основе иллюстраций из работы (Dickerson *et al*., 1989).

Результаты всех этих и многих других аналогичных исследований были документированы автором в базе данных PROPERTY (Колчанов и др., 1998), которая содержит величины 38 свойств спирали ДНК. Использованию этой базы данных в компьютерном анализе регуляторных сайтов в составе геномных ДНК посвящены все остальные главы настоящей диссертации. Вслед за базой данных PROPERTY (Колчанов и др., 1998) за рубежом создали базу данных DiProDB (Friedel *et al*., 2009) по 115 количественным свойствам ДНК и РНК.

1.3.3 Методы анализа периодичностей в геномных последовательностях

Началом изучения периодичности регуляторных сайтов в составе геномных ДНК был, по-видимому, анализ (Trifonov, Sussman, 1980) контактов

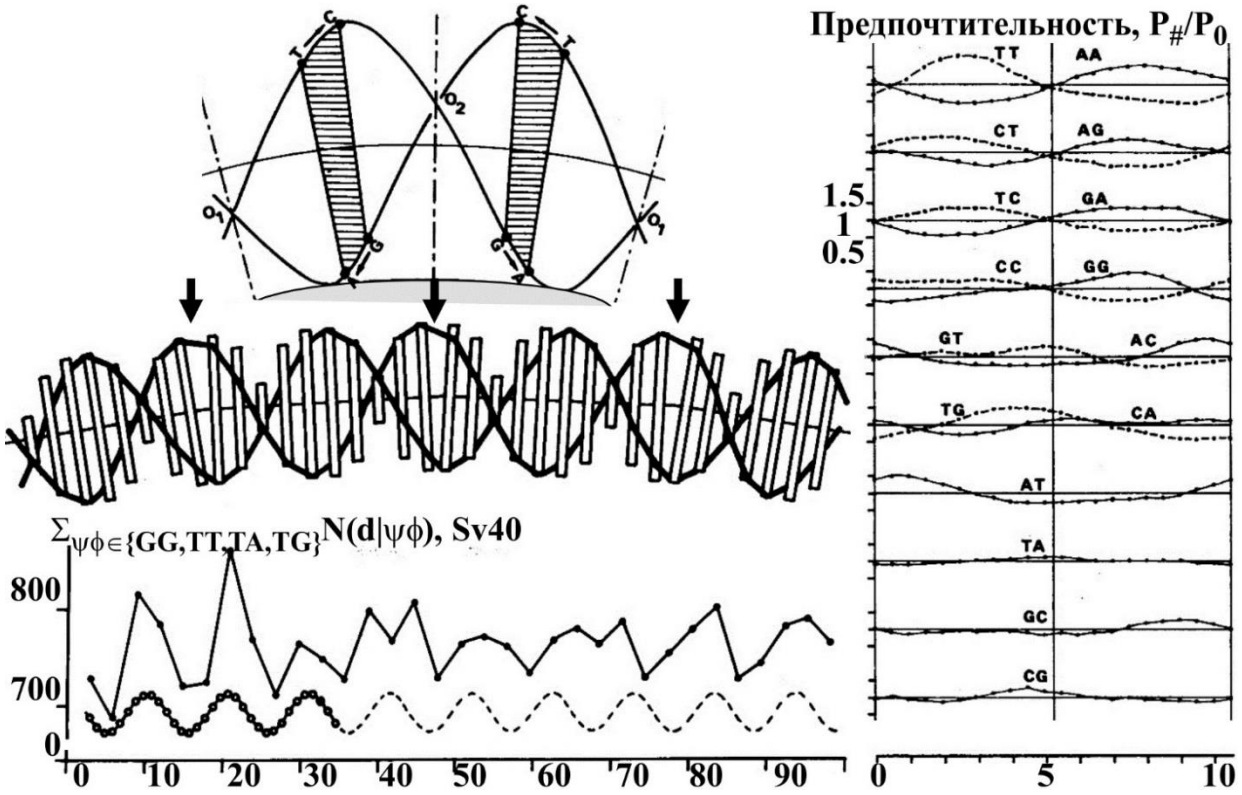

Рисунок 10 - Код периодичности нуклеосомной ДНК (Trifonov, Sussman, 1980). Рисунок автора на основе работы (Trifonov, Sussman, 1980).

октамера гистонов с ДНК на удалении 10.5 п.о. один от другого, равном витку спирали ДНК (Рисунок 10, сверху):

$$N(d|\varphi\psi) = \sum_{i=1}^{n-d-1} \Delta(\varphi\psi = s_i s_{i+1})\, \Delta(\varphi\psi = s_{i+d} s_{i+d+1}). \qquad (4)$$

здесь: ψ, φ ∈{A, T, G, C}; d – заданное расстояние в нити ДНК длины n, 1<d<<n.

С помощью формулы (4) подсчитывали однотипные динуклеотиды ψφ на фиксированном расстоянии между ними в геноме вируса SV40. В итоге для суммарного количества таких пар четырех динуклеотидов GG, TA, TT и TG была обнаружена периодичность с шагом 10.5 п.о., равным витку спирали ДНК (Рисунок 10, внизу). Поэтому в работе (Trifonov, 1980) разбивали геномы аденовируса и вируса полиомы, гены овальбумина, фиброина и гистонов H2B и H3 на фрагменты длиной 11 п.о. с лучшим совпадением динуклеотидов и оценили их частоты $P_{\#}$ в каждой позиции фрагмента. Эти оценки,

нормированные на среднюю частоту P_0, доказали факт кодирования предрасположенности ДНК эукариот к хроматину на уровне порядка динуклеотидов (Рисунок 10: справа), названного “кодом нуклеосомной ДНК” (Trifonov, Sussman, 1980).

В свою очередь, в работе (Shepherd, 1981) обобщили формулу (4) для произвольных олигонуклеотидов ψ и φ длиной k и m, соответственно, из 16-символьного алфавита {A, T(U), G, C, W=A+T(U), S=G+C, R=A+G, Y=C+T(U), M=A+C, K=G+T(U), V=A+G+C, B=C+T(U)+G, H=A+T(U)+C), D=A+T(U)+G, N=A+T(U)+G+C} (IUPAC-IUB, 1971):

$$N(d|\varphi;\psi) = \sum_{i=1}^{n-MAX(k;m)-d+1} \Delta(s_i s_{i+1} \in \varphi)\,\Delta(s_{i+d} s_{i+d+1} \in \psi). \qquad (5)$$

С помощью усовершенствованной формулы (5) обнаружили у вирусов ϕX174, td, SV40 и у бактериальной плазмиды pBR322 общую очевидную для всех без исключения природных нуклеотидных последовательностей ДНК и РНК периодичность с шагом кратным 3 нт, которая получила название “реликтовый генетический код триплетов без запятых” (Shepherd, 1981).

Наконец, учет периодичности регуляторных сайтов в составе геномных ДНК был (Makeev, Tumanyan, 1996) формализован в самом общем виде на основе теории преобразований Фурье для набора Ψ олигонуклеотидов $\{\psi,\ldots,\phi\}$ длин $\{m_\psi,\ldots,m_\phi\}$ с матрицей $L_{\psi\phi}$ сходства между ними - например, матрицей эволюционной взаимозаменяемости аминокислот (Dayhoff *et al.*, 1983), - в рамках которой спектром мощности Фурье для кода с периодом ω в ДНК $\{s_j\}_{1\le j\le n}$ длины n было:

$$Z(\omega|\{s_i|1 \le i \le n\}) = \sum_{\psi,\varphi\in\Psi} L_{\psi\varphi}\,\delta_\psi(\omega)\delta_\varphi(\omega); \qquad (6)$$

здесь: $\delta_\xi(\omega)$ – Фурье-образ $\Delta(s_i\ldots s_{i+m-1}\in\xi)$ функции-индикатора олигонуклеотида ξ длины m; которая введена в подписи к формуле (2).

Благодаря этому обобщению было обнаружено, что эвристически замеченная ранее (Hofmann *et al.*, 1980) периодичность 39 а.к.о. (117 п.о.) в

Таблица 5 – Примеры биологически значимых закономерностей, выявленных с помощью методов анализа периодичностей в геномных последовательностях

Биологическая закономерность	объект	Достоверная периодичность в нуклеотидной последовательности	Ссылка
Слабоизогнутая B-форма спирали ДНК старта репликации (ORI)	ДНК	Расположение динуклеотидных трактов $(A+T)_{3-6}$ с периодом 10 п.о. вблизи старта репликации (ORI)	(Eckdahl, Anderson, 1990)
Отличие ядерных геномов от внеядерных геномов		Во внеядерных ДНК нет периодичности 10.5 п.о. динуклеотида AA, как в ядерных ДНК	(Tomita *et al*., 1999)
Положение гена на хромосоме		Периоды 110±20 и 400±50 п.о. для G/C – это влияние G/C на экспрессию гена	(Nicolay *et al*., 2004)
Связывание ДНК генома с белками упаковки		Олигонуклеотиды длиной 20 п.о. имеют достоверную периодичность с шагом 50 п.о. в геномах про- и эукариот	(Larsabal, Danchin, 2005)
Достоверное отличие интронов от экзонов	РНК	Избыток динуклеотидных $(\xi\psi)_{15-35}$ трактов (кроме $(CG)_n$) в интронах, но не в экзонах	(Konopka et al., 1987)
Мутагенез гемаглютенина		Ген гемаглютенина имеет достоверную периодичность мутаций в нем	(Wu, Yan, 2005)

коллагене - это инвариант всех уровней кодирования генетической информации в ДНК, в РНК и в белке (Makeev, Tumanyan, 1996). Эта инвариантность подтвердила ранее высказанную гипотезу (Makeev *et al*., 1995) об эволюционном происхождении коллагена путем дупликации пра-коллагенового пептида длины 39 а.к.о., равной шагу периодичности третьей (синонимической) позиции кодонов глицина в нем.

В Таблице 5 представлены характерные примеры биологически значимых закономерностей, выявленных методами анализа периодичностей в регуляторных геномных последовательностях. В целом, форма спирали ДНК и факт связывания ДНК с белками упаковки или с регуляторными белками нашли отражение в различных периодичностях нуклеотидов.Например,

слабый изгиб спирали ДНК вокруг старта репликации ORI предопределяет расположение трактов $(A+T)_{3\text{-}6}$ с периодом 10 п.о.: шаг спирали ДНК (Eckdahl, Anderson, 1990). Внеядерные геномы отличаются от внутриядерных геномов отсутствием у них этой периодичности для динуклеотида AA (Tomita *et al*., 1999). Более протяженные периодичности от 50 п.о. (Nicolay *et al*., 2004) до 400 п.о. (Larsabal, Danchin, 2005) соответствуют кластерам скоординированно экспрессирующихся генов со сходными наборами сайтов связывания транскрипционных факторов в промоторах этих генов. Избыток несовершенных динуклеотидных трактов длиной от 30 нт до 70 нт (кроме CpG-тракта) в ничего не кодирующих интронах позволяет отличать их от фланкирующих белок-кодирующих экзонов в геномах (Konopka *et al*., 1987). Наконец, “горячие точки” мутагенеза в гене гемагглютинина вируса гриппа также имеют достоверную периодичность (Wu, Yan, 2005).

1.3.4 Методы анализа сложности геномных последовательностей

Бурное развитие в 70-х годах XX века методов рестрикционного анализа (Караванов, Иорданский, 1971; Pater *et al*., 1979) для сравнения геномов разных биологических видов по насыщенности ДНК повторами было обобщено (Lipman, Maizel, 1982) в измерение энтропии H в битах для последовательности $\{s_i|\ 1\le i\le n\}$ из n нуклеотидов $\{e_k|\ 1\le k\le 4\}$:

$$\begin{cases} H_{max}(\{e_k|1\le k\le 4\} \stackrel{\text{def}}{=} \{A,T(U),G,C\}) = -\sum_{k=1}^{4}\frac{1}{4}log_2\left(\frac{1}{4}\right) = 2; \\ H_1(\{s_i|1\le i\le n\}) \stackrel{\text{def}}{=} -\sum_{k=1}^{4}\left[\frac{N(e_k)}{n}log_2\left(\frac{N(e_k)}{n}\right)\right]; \\ H_2(\{s_i|1\le i\le n\}) \stackrel{\text{def}}{=} -\sum_{k=1}^{4}\sum_{m=1}^{4}\left[\frac{N(e_ke_m)}{n}log_2\frac{N(e_ke_m)}{N(e_k)}\right]; \\ \Delta H_0(\{s_i|1\le i\le n\}) \stackrel{\text{def}}{=} H_{max}(\{s_i|1\le i\le n\}) - H_2(\{s_i|1\le i\le n\}); \\ \Delta H_1(\{s_i|1\le i\le n\}) \stackrel{\text{def}}{=} H_{max}(\{A,T(U),G,C\}) - H_1(\{s_i|1\le i\le n\}); \\ \Delta H_2(\{s_i|1\le i\le n\}) \stackrel{\text{def}}{=} H_1(\{s_i|1\le i\le n\}) - H_2(\{s_i|1\le i\le n\}); \end{cases} \quad (7)$$

здесь: $N(\xi)$ – количество олигонуклеотидов ξ в последовательности $\{s_i|\ 1\le i\le n\}$ из n нуклеотидов.

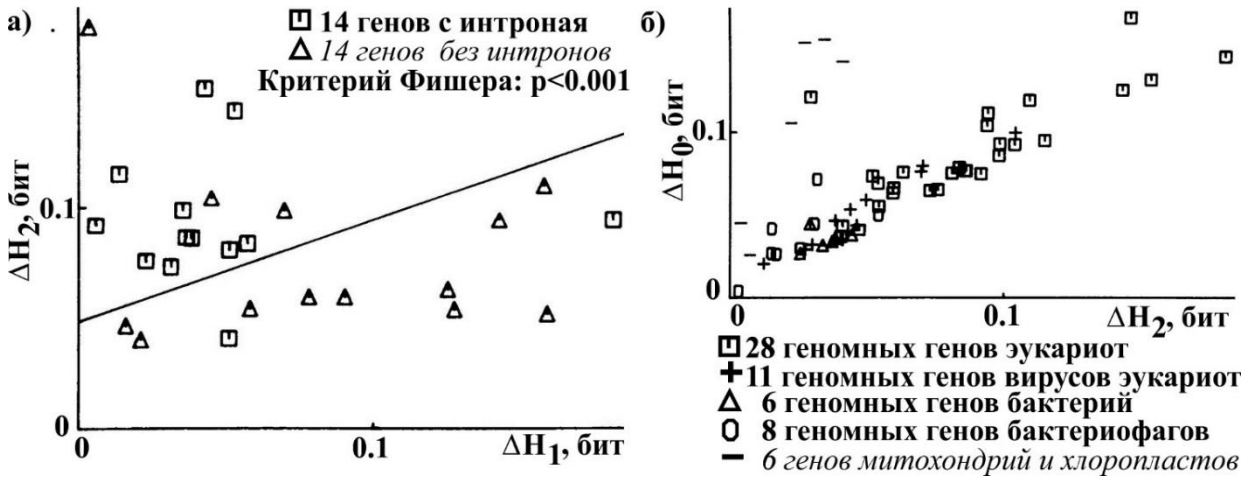

Рисунок 11 - Достоверные различия (а) между белок-кодирующими ДНК (кДНК) генов с интронами (□) и без интронов (Δ), а также (б) между ядерной геномной ДНК и внеядерной геномной ДНК органелл (Lipman, Maizel, 1982).

Рисунок автора на основе работы (Lipman, Maizel, 1982).

С помощью формулы (7) были открыты (Рисунок 11) достоверные различия между генами эукариот с интронами и без интронов, вне- и внутриядерных геномов по белок-кодирующей ДНК (кДНК). При этом в рамках теории информации (Shannon, 1948) ввели “энтропийную сложность” для скользящего окна $\{s_{i+j-\omega/2-1}\}_{1\le j\le \omega << n}$ длины ω в позиции i последовательности $\{s_i\}_{1\le i\le n}$:

$$EC_\omega\left(i|\left\{s_i|1\le\frac{\omega}{2}\le i\le n-\frac{\omega}{2}\right\}\right)=$$

$$=-\sum_{k=1}^{4}\left[\frac{N\left(e_k\epsilon[i-\frac{\omega}{2};i+\frac{\omega}{2}]\right)}{\omega}log_2\left(\frac{N\left(e_k\epsilon[i-\frac{\omega}{2};i+\frac{\omega}{2}]\right)}{\omega}\right)\right]. \quad (8)$$

Следующий шаг в развитии сложностного анализа последовательностей ДНК был сделан в работе (Trifonov, 1990), где определили лингвистическую сложность:

$$LC_\omega(i|\{s_i|1\le i\le m\le\omega\le n\})=\prod_{k=1}^{\omega}\frac{\sum_{m=1}^{4^k}\Delta\left(e_k^{m-1}\in\{s_{i+j-1}\}\right)}{MIN(4^k;\,n-k+1)}. \quad (9)$$

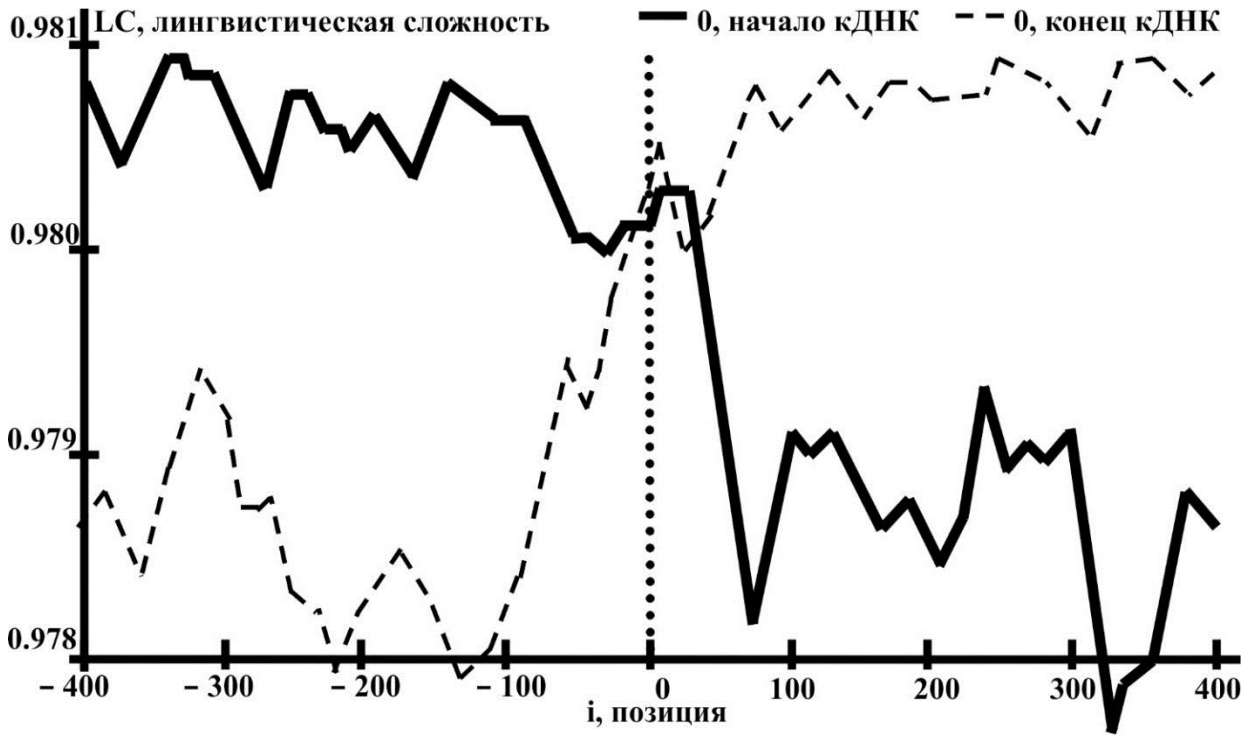

Рисунок 12 – Лингвистическая сложность (Trifonov, 1990) кДНК *E. coli* (формула 9) на 3’- и 5’-концах (Gabrielian, Bolshoy, 1999; Troyanskaya *et al*., 2002). Рисунок автора на основе работы (Troyanskaya *et al*., 2002).

С помощью формулы (9) было обнаружено (Gabrielian, Bolshoy, 1999) контекстное различие между белок-кодирующей и некодирующей ДНК бактерий, которое заключается в достоверном изменении лингвистической сложности при переходе границ между ними (Рисунок 12).

Далее, авторами работы (Гусев и др, 1991) был применен алгоритм сжатия данных Лемпела-Зива (Lempel, Ziv, 1976) для оценки алгоритмической сложности геномных последовательностей в единицах минимально возможного числа шагов их генерации *de novo* (из “пустой последовательности” Ø) с помощью элементарных операций (φ) вставить (копировать, удалить и пр.) нуклеотид в текущую последовательность:

$$AC_{\omega;n}(i|\varnothing \Rightarrow \{s_i|1 \le i \le \omega \le n\}) =$$

$$= \frac{1}{\omega} MIN_{\psi_\xi \in \Psi}(K|\{s_i|1 \le i \le \omega \le n\} \stackrel{\text{def}}{=} \prod_{k=1}^{K} \psi_k(\varnothing)). \quad (10)$$

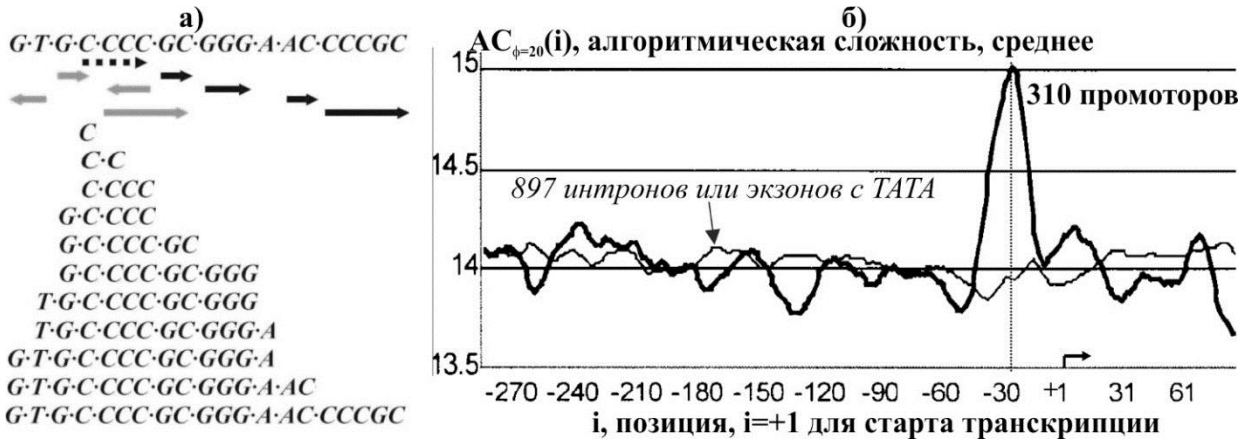

Рисунок 13 – Вычисление по формуле (10) алгоритмической сложности (а) на примере ДНК gtgccccgcgggaaccccgc длиной 20 п.о., $AC_{\omega=1;n=20}$=11 из (Гусев и др, 1991) и (б) достоверный пик алгоритмической сложности ТАТА-бокса промоторов генов эукариот (Babenko *et al*., 1999). Рисунок автора на основе иллюстраций из оригинальных статей (Гусев и др, 1991; Babenko *et al*., 1999)

.

На Рисунке 13а показан пример вычисления с помощью формулы (10) оценки $AC_{\omega=1;n=20}$=11 алгоритмической сложности для ДНК “gtgccccgcgggaaccccgc” длины 20 п.о (Гусев и др, 1991).

С помощью формулы (10) было найдено (Babenko *et al*., 1999) достоверное отличие (Рисунок 13б) экспериментально доказанных ТАТА-боксов в промоторах генов эукариот от результатов их прогноза общепринятым критерием ТАТА-бокса (Bucher, 1990) в экзонах и в интронах. Пик сложности экспериментально доказанного ТАТА-бокса отражает асимметрию его флангов, которые являются консервативным и вариабельным сайтами связывания транскрипционного фактора TFIIB, взаимная ориентация дает направление транскрипции.

Затем для нуклеотидных последовательностей геномной ДНК были введены еще два типа ее сложности: композиционная сложность ее нуклеотидного состава (Wootton, Federhen, 1996):

$$\mathrm{CC}_{\omega}(i|\{s_i|1 \le i \le \omega \le n\}) = -\frac{1}{\omega} log_4 \left(\prod_{k=1}^{4} \left[\sum_{j=1}^{n} \frac{\Delta(s_j=e_k)}{\omega}\right]!/\omega!\right), \qquad (11)$$

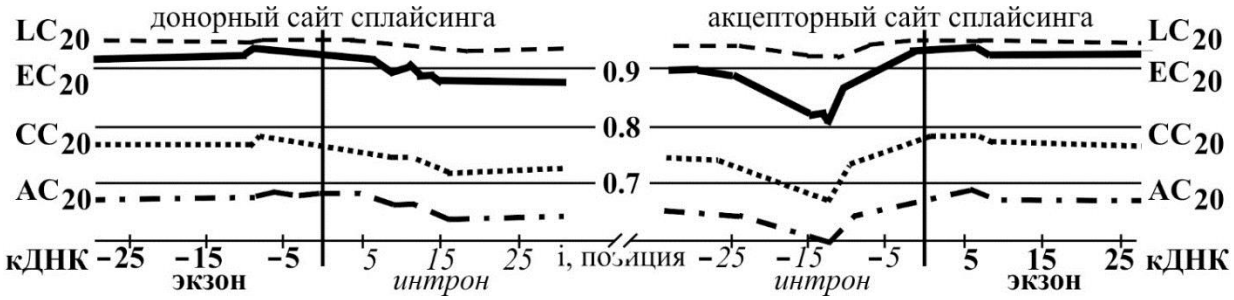

Рисунок 14 - Достоверное изменение сложности (8-12) нуклеотидной последовательности при переходе в сайтах сплайсинга от белок-кодирующей к некодирующей ДНК (Orlov, Potapov, 2004), обобщившее результат (Рисунок 12) работы (Gabrielian, Bolshoy, 1999) на случай эукариот. Рисунок автора на основе работы (Orlov, Potapov, 2004).

а также сложность Маркова порядка m взаимосвязи m позиций-соседей (Jimenez-Montano *et al*., 2002):

$$MC_{\omega}(i|\{s_i|1 \le i \le m \le \omega \le n\}) = \\ = -\left\{\sum_{k=1}^{4^m} \Delta\left(e_k^m \in \{s_{i+j-1}\}\right) \times \sum_{i=1}^{n}\left[\frac{\Delta\left(e_k^m \in \{s_{i+j-1}\}\right)}{\omega - m + 2} log_4 \sum_{j=1}^{n} \frac{\Delta\left(e_k^m \in \{s_{i+j-1}\}\right)}{\omega - m + 2}\right]\right\} + \\ + \left\{\sum_{k=1}^{4^{m-1}} \Delta\left(e_k^{m-1} \in \{s_{i+j-1}\}\right) \times \sum_{i=1}^{n}\left[\frac{\Delta\left(e_k^{m-1} \in \{s_{i+j-1}\}\right)}{\omega - m + 2} log_4 \sum_{j=1}^{n} \frac{\Delta\left(e_k^{m-1} \in \{s_{i+j-1}\}\right)}{\omega - m + 2}\right]\right\}. \quad (12)$$

Наконец, для комплексного анализа всех этих разнообразных оценок сложности геномных нуклеотидных последовательностей с помощью формул (8-12) был создан пакет компьютерных программ Complexity (Orlov, Potapov, 2004). С его помощью было, в частности, обнаружено (Orlov, Potapov, 2004) достоверное изменение всех типов сложности геномных ДНК при переходе в сайтах сплайсинга генов эукариот от некодирующих интронов к белок-кодирующим экзонам (Рисунок 14), обобщившее выводы (Рисунок 12) работы (Gabrielian, Bolshoy, 1999) на геномы эукариот.

Не претендуя на исчерпывающую полноту информации, в Таблице 6 представлен ряд характерных примеров биологически значимых закономерностей, которые были найдены с помощью методов сложностного анализа регуляторных геномных последовательностей.

Таблица 6 – Примеры биологически значимых закономерностей, найденных с помощью методов анализа сложности регуляторных сайтов в составе геномных ДНК

Биологическое явление	объект	Сложностная закономерность	Ссылка
Мозаика промотора генов-ортологов	ДНК	Промоторы ортологов: мозаики зон разной алгоритмической сложности	(Chuzhanova *et al.*, 2000)
InDel-мутации и их связь с болезнями человека		Ассоциированные болезням InDel мутации более частые в районах низкой сложности	(Chuzhanova *et al.*, 2003a)
“Горячие точки” транслокаций		“Горячие точки” транслокаций более часты в районах геномов с низкой сложностью	(Chuzhanova *et al.*, 2003b)
Консервативные сайты вне кДНК		Консервативный $(G/C)_k(A/T)_m(G/C)_n$ паттерн границы между участками разной сложности вне кДНК	(Abnizova *et al.*, 2007)
Мозаика неоднородной сложности генома	ДНК и РНК	Геномы состоят из районов значимо разной энтропийной сложности	(Oliver *et al.*, 1999)
		Геномы имеют иерархию районов с характерными частотами нуклеотидов	(Ouyang *et al.*, 2005)
		Динуклеотиды достоверно более информативны, чем нуклеотиды	(Cheng et al., 2007)
Контекст промоторов генов эукариот		Низкая сложность сигналов транскрипции, чаще всего, палиндромов, повторов и трактов	(Orlov *et al.*, 2006)
Рост сложности геномов в процессе их эволюции		Рост сложности генома в эволюции из-за дупликаций и, затем, дивергенции копий	(Adami *et al.*, 2000)
Мутационное старение геномов		Низкая сложность фрагмента генома означают его “важность” и “предковость”	(Costa *et al.*, 2005)
Соответствие между генотипом и фенотипом		Рост сложности генома с усложнением организмов: дупликации → дивергенции	(Takeuchi, Hogeweg, 2008)

Прежде всего, промоторы генов были описаны в виде мозаик регуляторных районов с постоянной алгоритмической сложностью (Chuzhanova *et al*., 2000, 2003a, 2003b). На границах этих районов оказались кластеры “горячих точек” замен, вставок, делеций, дупликаций и транслокаций геномной ДНК (Costa *et al*., 2005). Высокие оценки сложности (8-12) геномных последовательностей вокруг этих “горячих точек” соответствуют уникальным порядкам нуклеотидов в их ближайшем локальном окружении (Adami *et al*., 2000). В отличие от высокой степени сложности “горячих точек” мутагенеза, низкая сложность кластеров сайтов связывания транскрипционных факторов соответствует их строению, чаще всего, в виде палиндромов (Orlov *et al*., 2006), повторов или трактов (Abnizova *et al*., 2007).

Аналогично, мозаичное строение всех геномов в целом из протяженных районов последовательностей ДНК с характерной средней сложностью каждого такого района и с достоверным ее скачком на границах было показано независимо на уровне нуклеотидов (Ouyang *et al*., 2005), на уровне динуклеотидов, оказавшихся наиболее информативными (Cheng *et al*., 2007), и на самом общем уровне энтропийной (формула 8) сложности геномов (Oliver *et al*., 1999).

Наконец, все эти частные контекстные закономерности в целом могут быть отражением соответствия между фенотипом и генотипом: по мере усложнения живых организмов в эволюции растет сложность их геномов за счет мутагенеза путем дупликации и последующей дивергенции копий (Takeuchi, Hogeweg, 2008).

1.3.5 Методы контекстного анализа геномных последовательностей

В 1983 г. авторы статьи (Karlin *et al*., 1983) констатировали факт превышения суммарной длины всех секвенированных геномных последовательностей в базах данных величины в один миллион нуклеотидов,

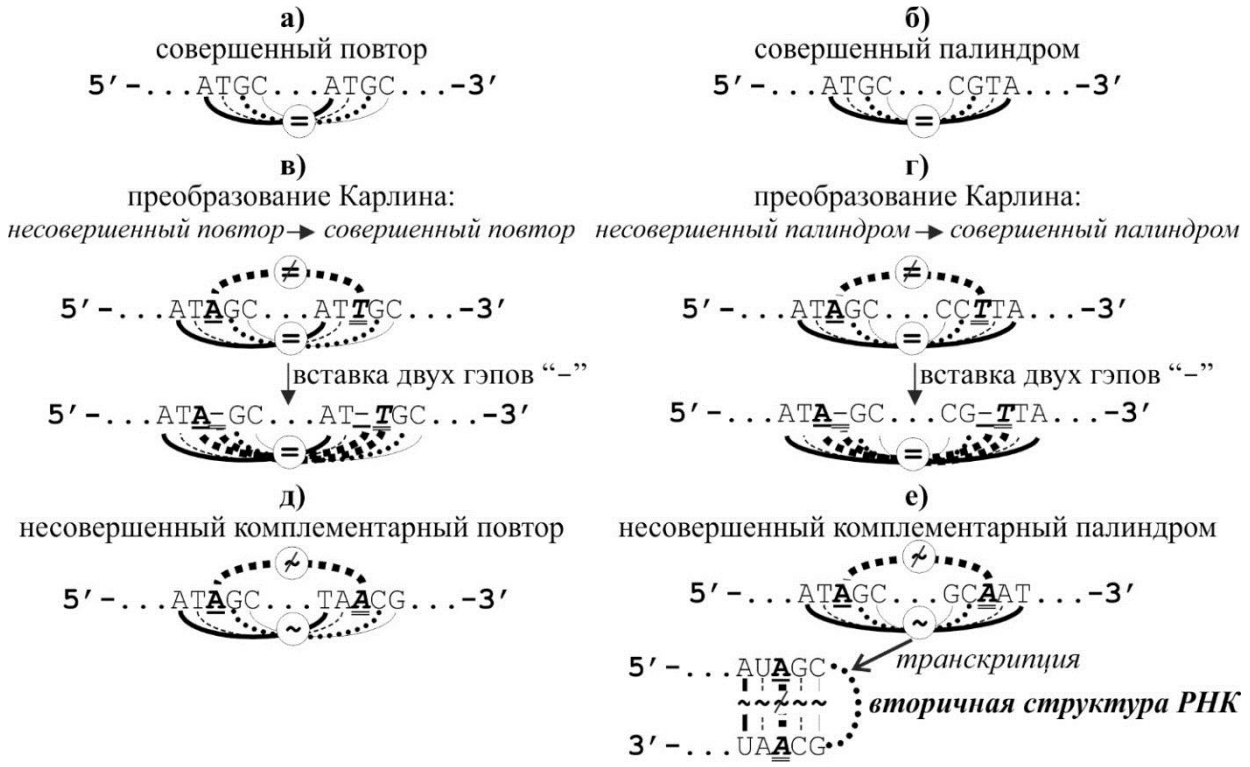

Рисунок 15 - Объекты контекстного анализа геномных последовательностей, совершенные (а) повтор и (б) палиндром без несовпадений соответствующих нуклеотидов, и их обобщения (Karlin *et al*., 1983) на случай (в, г) несовпадений нуклеотидов и на случай (д, е) комплементарности нуклеотидов A ~ T(U) и G ~ C.

что было уже достаточным для введения нового направления – контекстного анализа ДНК и РНК. В качестве первоначальных объектов анализа были взяты совершенные повтор (Рисунок 15а) и палиндром (Рисунок 15б): пары неперекрывающихся контекстно-сходных сегментов в случайной последовательности из n независимых символов $\{s^k_i | 1\le i\le n\}$ алфавита $\{s^k | 1\le k\le r\}$ объема r с частотами $\{p_k>0 \mid 1\le k\le r\}$, их нормированием $\Sigma_k p_k\equiv 1$ и ожидаемой частотой $\lambda=\Sigma_k p_k^2$ парных совпадений. В этих обозначениях были выведены (Karlin *et al*., 1983) асимптотические оценки ожидаемой длины $L_0(n)$ такого сегмента, традиционно называемого "плечем повтора или палиндрома", и ее вариансы $Var(L_0(n))$:

$$\begin{cases} L_0(n) = \dfrac{2\log(n)}{-\log(\lambda)} - \left(1 + \dfrac{0.5772 + \log(1-\lambda)}{\log(\lambda)}\right) + \dfrac{\log(2)}{\log(\lambda)}; \\ Var(L_0(n)) = \dfrac{1.645}{\log^2(\lambda)}. \end{cases} \quad (13)$$

Формула (13) описывает логарифмическое увеличение длины L_0 сегмента палиндрома (повтора) при увеличении длины n исследуемой последовательности и независимость вариансы длины сегмента от этого n (Karlin *et al*., 1983).

Например, в последовательности из n=5000 равновероятных нуклеотидов четырех канонических типов можно ожидать один совершенный палиндром (повтор) длины L_0=12±1 нт. Эта оценка L_0 была в согласии с длинами совершенных повторов и палиндромов в расшифрованных геномах фага λ и митохондрий, но она оказалась достоверно меньшей таковых длин в случае геномов вирусов SV40, BKV, полиомы, папилломы и генов иммуноглобулинов (Karlin *et al*., 1983). При этом самые длинные совершенные повторы и палиндромы были обнаружены в биологически значимых районах ДНК: вблизи от начал репликации ORI и в белок-кодирующих районах генов (Karlin *et al*., 1983). Поэтому было эвристически предложено характеризовать расшифрованные фрагменты геномной ДНК в терминах “совершенный повтор (палиндром) наибольшей длины” и впоследствии–реализовано в словаре кодирования геномов “Gnomic” (Trifonov, Brendel, 1987).

Поскольку большинство достоверно длинных повторов и палиндромов оказались в белок-кодирующей ДНК, то предположили, что насыщенность кДНК повторяющимися мотивами может быть следствием взаимосвязи между нуклеотидами-соседями в кодонах генетического кода, и, в этой связи, ввели термин “марковость белок-кодирующей ДНК” (Karlin *et al*., 1983). Методом компьютерного моделирования Монте-Карло (Metropolis, Ulam, 1949) были оценены вклад частот использования кодонов (Wain-Hobson *et al*., 1981) в повторы и палиндромы белок-кодирующих районов генов (Karlin *et al*., 1989). Достигнутые в результате этого моделирования “размножение → мутация →

отбор" пределы распределений количества и длин комплементарных палиндромов (Рисунок 15е: "шпилек") достоверно соответствовали таковым распределениям, наблюдаемым в расшифрованных генах природных белков. На этой основе была выдвинута гипотеза о связи частот использования кодонов со "шпильками" мРНК (Рисунок 15е), ответственными за устойчивость мРНК к деградации. Однако в случае прямых совершенных повторов результаты моделирования не имели сходства с природными ДНК (Karlin *et al*., 1989).

Для несовершенных повторов и палиндромов (Рисунки 15в,г) с нарушением совпадений нуклеотидов расширили алфавит нуклеотидов путем введения символа "-" ("гэп", "gap" англ. яз.), {A, T, G, C, "-"}, совпадавший со всеми нуклеотидами и не совпадавший с таким символом "-" в другой позиции (Рисунки 15в,г). Это дополнение обобщило формулу (13) на случай несовпадений. Учет ожидаемых частот несовпадений, $p_{r+1}=p_{"-"}>0$, их вкладов в нормирование $\Sigma_{1\le k\le r+1} p_k\equiv 1$ и в ожидаемые частоты $\lambda=\Sigma_{1\le k\le r+1} p_k^2$ парных совпадений, увеличил оценки $L_0(n)$ и ее вариансы от n, что получило общепринятое название "штраф за несовпадение" ("gap penalty", англ. яз.).

Наконец, случай отдельного фрагмента нуклеотидной последовательности длины n был остроумно обобщен (Karlin *et al*., 1983) на случай двух независимых фрагментов равной длины n/2 в задаче поиска с помощью формулы (13) в каждом из них по одному из двух сегментов "их общего повтора (палиндрома)". Эта находка дала оценку доверительных границ ожидаемого случайного сходства независимых нуклеотидных последовательностей, ставшей фундаментом самых используемых в настоящее время компьютерных систем CLUSTAL (Higgins, Sharp, 1988) и BLAST (Altschul *et al*., 1990) поиска потенциальных генов во вновь расшифрованных геномах на основе их кусочно-непрерывных совпадений с известными генами в аннотированных геномах.

В это же самое время, авторы статьи (Колчанов и др., 1983) оценили вероятность $p\{N(L, m, n)\}$, величину $N_0(L, m, n)$ и доверительные границы

$N_\alpha(L,m,n)$ при значимости α для числа повторов (палиндромов) длины L с m несовпадениями в последовательности ДНК длины n:

$$\begin{cases} p\{N(L,m,n)\} = C^{N(L,m,n)}_{(n-2L+1)(n-2L+2)/2}\left(C_L^m\lambda^{L-m}(1-\lambda)^m\right)^{N(L,m,n)} \times \\ \qquad\qquad \times\left(1-C_L^m\lambda^{L-m}(1-\lambda)^m\right)^{N(L,m,n)(n-2L+1)(n-2L+2)/2}; \\ N_0(L,m,n) = (n-2L+1)(n-2L+2)C_L^m\lambda^{L-m}(1-\lambda)^m/2\,; \\ \sum_{N(L,m,n)=0}^{N_\alpha(L,m,n)-1} p\{N(L,m,n)\} < 1-\alpha \le \sum_{N(L,m,n)=0}^{N_\alpha(L,m,n)} p\{N(L,m,n)\}. \end{cases} \quad (14)$$

Формула (14) единообразно описала все совершенные, несовершенные, комплементарные повторы и палиндромы, изображенные на Рисунке 15 при частотах встречаемости $\lambda=\Sigma_k\Sigma_q p_k p_q \Delta(s_k\sim s_q)$ комплементарных/совпадающих ("~") нуклеотидов (Колчанов и др., 1985). С ее помощью в гене σ-субъединицы РНК-полимеразы *E. coli* было достоверно ($\alpha<0.05$) найдено (Колчанов и др., 1983) 8 совершенных повторов длины 10 п.о. (m=0, N_0=2, $N_{5\%}$=6), 8 несовершенных повторов длины 20 п.о. с 5 несовпадениями (N_0=2, $N_{5\%}$=4) и 7 повторов длины 29 п.о. с 10 несовпадениями (N_0=1, $N_{5\%}$=3). Аналогичные результаты были получены для 30 генов белков про- и эукариот (Колчанов и др., 1985), а также для комплементарных палиндромов (Кель и др., 1988), которые общепринято связывать со "шпильками" мРНК (Рисунок 15e), экспрессируемых с этих генов.

Так как частоты использования кодонов (Wain-Hobson *et al*., 1981) не позволили объяснить прямые повторы в кДНК (Karlin *et al*., 1989), то в работе (Кель и др., 1988) дополнили модель Монте-Карло (Metropolis, Ulam, 1949) учетом дупликаций случайных фрагментов ДНК, их дивергенции в силу независимости мутаций в каждой копии и отбором кодонов аминокислот в пользу кодирования β-нитей или α-спиралей белка на уровне ДНК (Chou, Fasman, 1974; Kabsch, Sander, 1983). После 580 шагов "размножение → мутация → отбор" в 1000 последовательностях случайных равновероятных нуклеотидов ДНК был достигнут предел компьютерного моделирования Монте-Карло (Metropolis, Ulam, 1949) для распределения прямых несовершенных повторов, который оказался в достоверном согласии с

таковым распределением для расшифрованных природных белок-кодирующих районов генов (Колчанов и др., 1988).

В не претендующей на исчерпывающую полноту Таблице 7 можно видеть некоторые типичные примеры биологически значимых закономерностей, которые были найдены методами контекстного анализа геномных нуклеотидных последовательностей. Прежде всего, был установлен (Blaisdell, Karlin, 1988) асимптотический критерий нижней доверительной границы ω_α при уровне значимости α для длины ω неслучайного непрерывного совершенного тракта $\{s_i \in \psi \mid 1 \le i \le \omega << n\}$ однотипных нуклеотидов ψ с оценкой p_ψ частоты их встречаемости:

$$\begin{cases} \omega - 1 \le \omega_\alpha = \varepsilon(\alpha) - \dfrac{\ln(n)}{ln(p_\psi)} \le \omega; \\ \alpha = 1 - exp\left(-\left(1 - p_\psi\right){p_\psi}^{\varepsilon(\alpha)}\right). \end{cases} \quad (15)$$

Затем, благодаря использованию биномиального распределения и неравенства Бонферрони, была введена к-статистика несовершенного кластера $\{s_i \in \psi \mid 1 \le i \le \omega << n\}$ длины ω с количеством $N_\omega(\psi)$ символов ψ и рекомендовано ее критическое значение $\kappa_{\alpha<0.001}=4.5$ для $30 \le \omega \le 50$ и $n>1000$:

$$\kappa_\alpha > abs\left|N_\omega(\psi) - \omega p_\psi\right| / \sqrt{\omega p_\psi \left(1 - p_\psi\right)}. \quad (16)$$

В свою очередь, авторы статьи (Nussinov *et al.*, 1988) обнаружили достоверно частое прерывание трактов $\psi_{\omega>>3}$ их короткими комплементами.

В это же время, были открыты достоверные пониженные частоты дупликаций (несовершенных прямых повторов) в геномах прокариот и их повышенные частоты в геномах эукариот в сравнении с частотами делеций в соответствующих геномах (Кель и др., 1989). Одним из проявлений этого фундаментального различия между геномами про- и эукариот стало обнаружение масштабной инвариантности геномов бактерий (Audit, Ouzounis, 2003).

Таблица 7 – Примеры биологически значимых закономерностей, выявленные с использованием методов контекстного анализа геномных ДНК.

Биологическое явление	объект	Контекстная закономерность	Ссылка
Мононуклеотидные тракты и кластеры	ДНК и РНК	к-тест (16) с квантилями для фрагментов генома	(Blaisdell, Karlin, 1988)
Локальное окружение трактов у эукариот	ДНК	Тракты длиной >3 п.о. имеют комплемент на флангах	(Nussinov et al., 1988)
Обедненность геномов прокариот повторами		Неправильное спаривание повтора дает делеции чаще дупликаций	(Кель и др., 1989)
Геномы бактерий масштабно-инвариантны		У бактерий нет промежуточного уровня между опероном и геном	(Audit, Ouzounis, 2003)
Специфическое локальное окружение кДНК генов		Самые перепредставленные олигонуклеотиды – границы фланкируют кДНК	(Hampson et al., 2002)
Соматический гипер-мутагенез генов иммуноглобулинов		Соматические мутации в генах иммуноглобулинов чаще всего в палиндромах	(Соловьев и др., 1989)
Неравномерность частот замен в эволюции		Вне кДНК замены CpG↔TpG между геномами в 18 раз чаще других замен	(Lunter, Hein, 2004)
Реликтовая иммунная система (Ig) челюстных		Транспозон *Transib* – предок гена *RAG1* и V(D)J-рекомби-нации иммуноглобулинов	(Kapitonov, Jurka, 2005)
Строение гена миРНК чаще всего палиндром		Прогноз кандидатных генов миРНК в геномах	(Huang et al., 2007)
Вторичная структура мРНК препятствует ее деградации	РНК	Конкуренция несовместимых шпилек (несовершенных палиндромов)	(Миронов и др., 1984)
Скорость элонгации трансляции мРНК		Палиндромы снижают скорость трансляции	(Likhoshvai, Ma-tushkin, 2002)
Вторичная структура и конформация мРНК		пара G:A достоверно частая в несовершенных палиндромах	(Villescas-Diaz, Zacharias, 2003)
Вирус подавляет иммунную систему организма-хозяина		МикроРНК вируса герпеса 4 репрессируют цитокины и хемокины В-клеток	(Pfeffer et al., 2004)
Посттрансляционная регуляция экспрессии		Мишени посттрансляционной регуляции мРНК – сайты ее комплемента к микроРНК	(Zhang, 2005)

Напротив, самые избыточные олигонуклеотиды достоверно часто фланкируют белок-кодирующие районы геномов эукариот, что сейчас общепринято ассоциировать с присутствиtv в геномах останков реликтовых пра-сигналов регуляции транскрипции генов (Hampson *et al*., 2002).

Открытую авторами статьи (Соловьев и др., 1989) контекстную предетерминированность соматического гипермутагенеза в настоящее время уже принято считать общеизвестным фактом (Lunter, Hein, 2004).

Благодаря методам контекстного анализа открыли (Kapitonov, Jurka, 2005) транспозон *Transib* и обосновали его ключевую роль в эволюционном возникновении иммунной системы у челюстноротых.

В случае РНК, авторы работы (Миронов и др., 1984) впервые оценили вторичную структуру РНК (Рисунок 15e) с учетом конкуренции взаимно несовместимых комбинаций “шпилек” (перекрывающихся несовершенных комплементарных палиндромов).

Авторами работы (Likhoshvai, Matushkin, 2002) была обнаружена достоверная корреляция между скоростью элонгации трансляции и самокомплементарностью мРНК. Аналогичным путем было установлено (Villescas-Diaz, Zacharias, 2003), что самокомплементарность мРНК достоверно часто нарушается лишь в единственной п.о. “G≠A” (здесь, символ “≠” обозначает отсутствие комплементарности в п.о.).

Наконец, несовершенные комплементарные палиндромы в первичных транскриптах генов микроРНК (миРНК) задают биогенез их созревания (Huang *et al*., 2007), комплементарные мишени на мРНК для заданной зрелой миРНК - способ (ингибирование или разрезание), адресность (Pfeffer *et al*., 2004) и эффективность (Zhang, 2005) воздействия этой миРНК на мРНК.

В целом, с использованием контекстного анализа было установлено, что нуклеотидные последовательности геномных ДНК кодируют не только последовательности РНК и белков, но и физико-химические и конформационные свойства спирали ДНК, ее упаковку в хроматин, регуляцию экспрессии генов, устойчивость РНК к деградации, конформационные и

физико-химические свойства белков, дивергенцию современных биологических видов от их общих предков и др.

Все эти частные закономерности были обобщены (Trifonov, 1989) в гипотезу о многообразии генетических кодов и интерференции всех этих кодов в потенциально возможные варианты последовательностей ДНК, к числу которых относятся природные геномы, несущие в себе всю совокупность многократно дублированной, перекрывающейся, избыточной и, возможно, в чем-то уже неполной генетической информации о структуре, функции и эволюции ДНК, РНК, белков, клеток, тканей, органов, организмов, таксонов и всей биосферы в целом.

1.3.6 Статистическая механика связывания белков с геномной ДНК

Вероятно, одним из самых первых применений статистической механики к анализу регуляторных сайтов в составе геномных ДНК было заимствование из биосистематики понятия "*консенсус*" (Kolmer *et al*., 1917) и приспособление его к анализу контекста: *гипотетическая последовательность определенного сайта ДНК, составленная из достоверно частых нуклеотидов в каждой позиции всех экспериментально доказанных вариантов этого сайта* (Hawley, McClure, 1983). Для построения самого первого консенсуса (Рисунок 16), был применен критерий Пуассона для сравнения между наблюдаемыми оценками частот встречаемости канонических нуклеотидов в районе [-50; +10] относительно стартов транскрипции 168 промоторов *E. coli* и ожидаемой оценкой p_0=0.25 этих частот для модельных последовательностей из случайных независимых равновероятных нуклеотидов. На Рисунке 16 заглавным шрифтом выделены достоверно частые нуклеотиды при уровне статистической значимости $\alpha<10^{-9}$ (чаще 54%), строчным – при $\alpha<10^{-3}$ (чаще 39%). Анализ всех 98 экспериментально известных на тот момент мутаций в промоторах *E. coli*

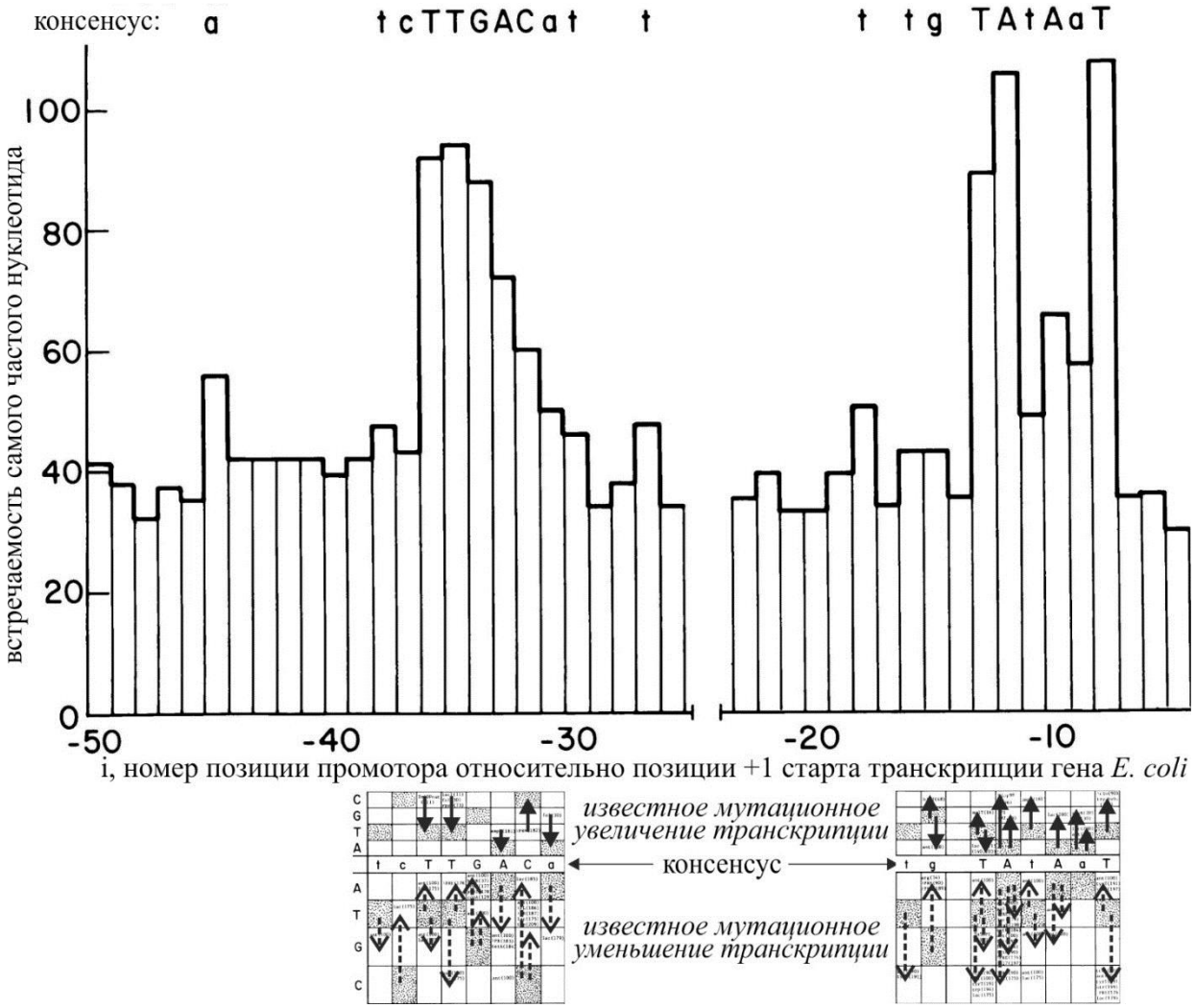

Рисунок 16 - Введение авторами работы (Hawley, McClure, 1983) модели консенсуса, взятой из биосистематики (Kolmer *et al*., 1917) и приспособленной к анализу геномных ДНК в качестве *списка достоверно частых нуклеотидов в каждой позиции исследуемого сайта*, - на примере промоторов *Escherichia coli* и демонстрация корреляции консенсуса со всеми известными случаями изменения транскрипционной активности генов *E. coli* в результате мутаций (стрелки, "→"). Рисунок автора на основе иллюстраций из работы (Hawley, McClure, 1983).

показал, что замене консенсусного нуклеотида соответствует снижение транскрипции, а появлению консенсусного нуклеотида в результате мутации – увеличение транскрипции, как показано внизу на Рисунке 16.

Бериков и Рогозин (Berikov, Rogozin, 1999) обобщили понятие консенсуса в регрессионное древо значимо частых нуклеотидов для предсказания мутационных спектров геномной ДНК при воздействии на нее определенного мутагена. В рамках выполнения настоящей диссертационной работы автор также обобщил понятие консенсуса на случай достоверно редких нуклеотидов в каждой позиции сайта. Это позволило впервые достоверно предсказать частоты предмутационных повреждений гуанинов при воздействии лазерного ультрафиолетового излучения с длиной волны 193 нм (Втюрина и др., 2011), как это описано в разделе 4.1 настоящей диссертации. Консенсус был дополнен позиционно-частотной матрицей $f_{i;\zeta}$ с мерой Hs "сходства по гомологии" ("homology score", англ. яз.), нормированной среднеарифметическим $M_0(Hs) \stackrel{def}{=} 0$ в качестве порога распознавания потенциальных промоторов в геноме *E. coli* (Mulligan *et al.*, 1984):

$$Hs(s_{-45} \ldots s_{-4}) = 100\% \times \frac{\sum_{i=-45}^{-30} f_{i;s_i} + \varepsilon_{\text{"}-35\text{"}\leftrightarrow\text{"}-10\text{"}} + \sum_{i=-17}^{-4} f_{i;s_i} - \varepsilon_0}{\sum_{i=-45}^{-30} MAX_{\xi}(f_{i;\xi}) + \sum_{i=-17}^{-4} MAX_{\xi}(f_{i;\xi}) - \varepsilon_0}, \quad (17)$$

здесь: $\varepsilon_{\text{"-35"}\leftrightarrow\text{"-10"}}$ и ε_0 – две эвристические поправки на размер спейсера между каноническими боксами "-35" и "-10" промотора *E. coli* и на среднеарифметическое $M_0(Hs) \stackrel{def}{=} 0$, соответственно.

Вычисленные с помощью формулы (17) по известным последовательностям ДНК промоторов *E. coli* величины Hs достоверно коррелировали (Рисунок 17: $r = 0.83$; $p < 0.01$) с экспериментально измеренными величинами транскрипционной активности этих промоторов.

Понятия "консенсус" и "позиционно-частотная матрица" были обобщены автором вне рамок настоящей диссертации (Ponomarenko M *et al.*, 1999b) на случай олигонуклеотидов в расширенном 15-символьном коде IUPAC-IUB (IUPAC-IUB, 1971). При этом для распознавания регуляторных сайтов в произвольной ДНК было показано (Пономаренко М и др., 1999a), что с ростом количества N обобщенных консенсусов и позиционно-частотных матриц уменьшаются ошибки I (недопредсказание) и II (перепредсказание)

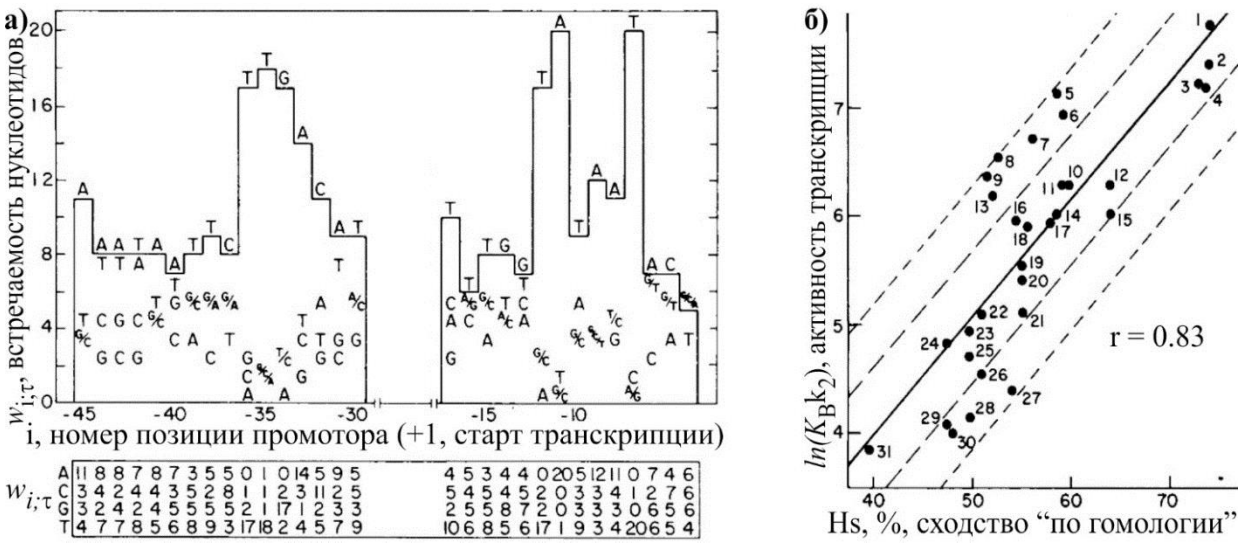

Рисунок 17 - Расширение авторами работы (Mulligan *et al*., 1984) модели консенсуса (Рисунок 16) до более детальной модели позиционно-частотной матрицы $\{w_{i;\tau}\}$ (а) с мерой сходства "по гомологии" (формула 17), предсказания которой достоверно ($\alpha<0.01$) коррелируют с экспериментально измеренной транскрипционной активностью генов *E. coli* (б). Рисунок автора на основе иллюстраций из работы (Mulligan *et al*., 1984).

рода, которые асимптотически стремятся соответственно к минимуму и к нулю, как $N^{-1/2}$, в согласии с центральной предельной теоремой. На основе этих обобщенных мер сходства и результатов раздела 5.3 настоящей диссертации с участием автора была создана система rSNP_Guide (Ponomarenko J *et al*., 2001b, 2002a,b), которая по данным опыта "задержки в геле" комплексов нормальной и мутантной ДНК с белковым экстрактом ядер из определенной клеточной линии достоверно прогнозирует транскрипционный фактор, сайт связывания которого изменен в мутантной ДНК относительно нормы.

В свою очередь, на основе теории информации (Shannon, 1948) обобщили (Schneider *et al*., 1986) эвристическую формулу (17) в количество информации благодаря учету оценок *a priori* ожидаемой f_ζ и наблюдаемой $f_{i;\zeta}$ частот нуклеотида ζ в позиции i сайта (Рисунок 18):

$$Ic(s_a \dots s_i \dots s_b) = \sum_{i=a}^{b} w_{i;s_i} \stackrel{\text{def}}{=} \sum_{i=a}^{b} f_{i;s_i} log_2\left(f_{i;s_i}/f_{s_i}\right). \qquad (18)$$

Рисунок 18 – Мера “количество информации” (формула 18) на примере сайта связывания белка-рецептора цикло-АМФ у *E. coli* (Stormo, Hartzell, 1989). Рисунок автора на основе работы (Stormo, Hartzell, 1989).

Дополнительно к прогностическим способностям формулы (17), количество информации (формула 18) позволяет оценивать позиции сайта ДНК (РНК), что иллюстрирует пример (Stormo, Hartzell, 1989) сайта связывания белка-рецептора цикло-АМФ у *E. coli* на Рисунке 18.

В качестве следующего шага применения статистической механики к исследованию комплексов ДНК с регуляторными белками ввели (Stormo *et al.*, 1986) позиционно-весовую матрицу коэффициентов $w_{i;\xi}$ множественной линейной регрессии для минимизации функции потерь σ^2 линейно-аддитивной аппроксимации количественных измерений $\{\Psi_j|\ 1\leq j\leq N\}$ биологической активности N заданных фрагментов ДНК (РНК) $\{s_{i;j}|\ 1\leq i\leq n;\ 1\leq j\leq N\}$ длины n со среднеарифметической оценкой $M_0(\Psi)$ этих измерений активности при $N\geq 3n+1$:

$$\sigma^2 = MIN_{w_{\zeta\xi}}\{\sum_{j=1}^{N}(Act_j - (M_0(Act) + \sum_{i=a}^{b} w_{i;s_{i;j}}))^2\}, \quad (19)$$

где: в каждой позиции i ровно один из $w_{i;A}$, $w_{i;T}$, $w_{i;G}$, $w_{i;C}$ установлен тождественно равным “0”. Пример применения формулы (19) из статьи (Stormo *et al.*, 1986) показан на Рисунке 19.

Кроме того, был найден способ применения (Stormo *et al.*, 1982) простейшей нейронной сети перцептрон Минского (Минский, 1971) для минимизации функции потерь σ^2 (формула 19) за конечное число шагов:

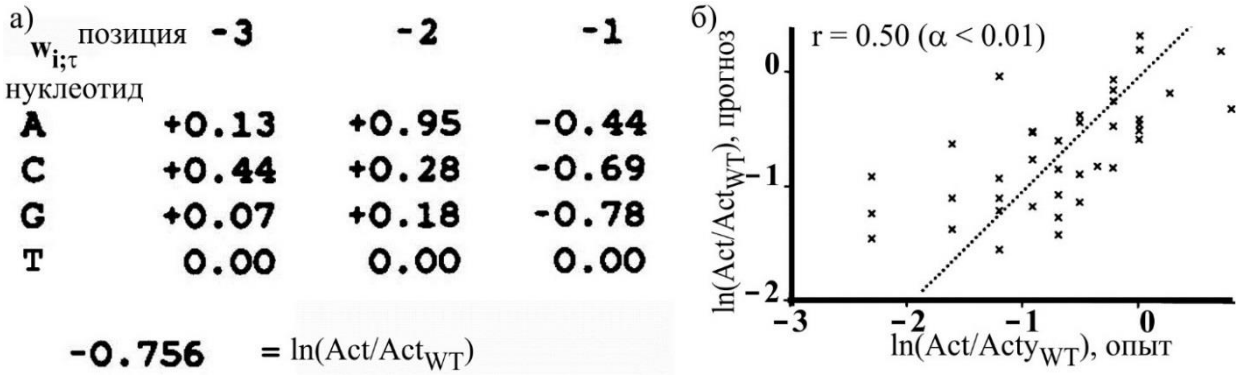

Рисунок 19 – Позиционно-весовая матрица вкладов нуклеотидов в позициях сайта в величину биологической активности этого сайта на основе линейной регрессии (Stormo, Hartzell, 1989) на примере тринуклеотида перед стартовым AUG-кодоном мРНК *E. coli* (а), прогноз которой достоверно коррелирует с экспериментальными измерениями активности трансляции (б). Рисунок автора на основе работы (Stormo, Hartzell, 1989).

$$w_{\zeta;\xi} = lim_{\substack{k\to N \\ w^0_{\zeta;\xi}\neq\varnothing}} \left\{ w^k_{\zeta;\xi} \stackrel{\text{def}}{=} w^{k-1}_{\zeta;\xi} + 2\sum_{j=1}^{N} \Delta\left(\xi = s_{\zeta;j}\right)\left[\Delta\left(\sum_{i=1}^{n} w^{k-1}_{i;s_{i;j}}\right) - 0.5\right]\right\}. \quad (20)$$

На Рисунке 20 показан авторский пример (Stormo *et al.*, 1982) применения формулы (20) в случае сайта мРНК для связывания рибосомы *E. coli*. Как можно видеть, позитивные позиционные веса $w_{i;\xi}$ перцептрона (Рисунок 20а, обведены кружками, ○) и пики их вариансы (Рисунок 20б, стрелки, →) соответствуют двум общеизвестным сигналам мРНК *E. coli* на флангах этого сайта: бокс Шайн-Далгарно (SD) ARRRGGGA (Shine, Dalgarno, 1974) и стартовый AUG-кодон.

С целью исчерпывающего анализа этого сайта были синтезированы (Barrick *et al.*, 1994) 185 олигоДНК длиной по 15 нт каждый, которые вставляли в конструкт на основе использования гена-репортера β-галактозидазы *E. coli*, и была измерена активность трансляции для каждого такого конструкта. На этих данных он независимо оптимизировал для сайта связывания рибосомы *E. coli* три варианта позиционно-весовой матрицы $w_{i;\xi}$: перцептрон (20), множественную линейную регрессию (19) для всех 185 проб,

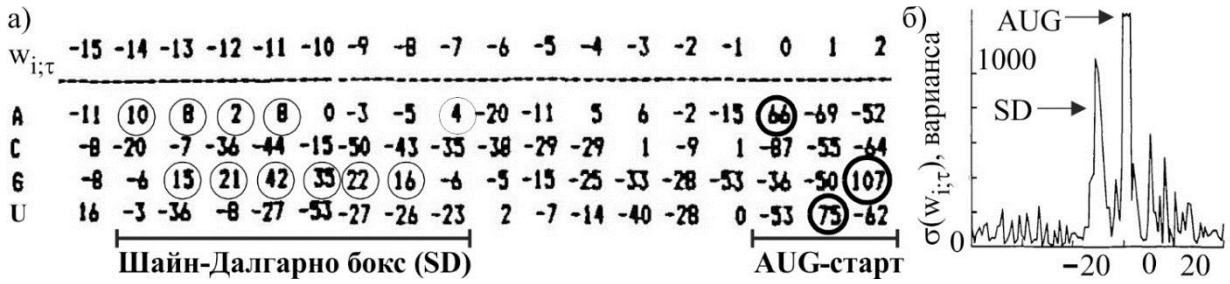

Рисунок 20 - Введение авторами работы (Stormo *et al*., 1982) на основе конечно-шаговой оптимизации (Минский, 1971) нейронной сети перцептрона (а) позиционно-весовой матрицы на примере сайта на мРНК для связывания рибосомы *E. coli* и (б) демонстрация пиков вариансы этой матрицы ($w_{i;\zeta}$), указывающих на общеизвестные составные части этого сайта: бокс Шайн-Далгарно (SD) ARRRGGGA (Shine, Dalgarno, 1974) и стартовый AUG кодон. Рисунок автора на основе иллюстраций из работы (Stormo *et al*., 1982).

а также отдельно для 158 проб с каноническим стартовым AUG-кодоном, - и сравнил их все с позиционно-частотной матрицей $f_{i;\zeta}$ “сходства по гомологии” (формула 17) для всех 1055 экспериментально доказанных природных таких сайтов. Оказалось, что все четыре матрицы достоверно отличались одни от других, хотя результаты их прогнозов были достоверно неразличимыми: коэффициенты корреляции r между предсказанными и экспериментальными величинами были от 0.83 до 0.89 при уровне значимости $p < 10^{-17}$ (Barrick *et al*., 1994). Поэтому было сделано заключение, что позиционно-весовые матрицы $w_{i;\xi}$ нельзя интерпретировать в терминах влияния нуклеотидов в позициях сайта на биологическую активность из-за известной неустойчивости результатов минимизации функции потерь σ^2 к неоднородностям исходных экспериментальных данных $\{\Psi_j, \{s_{i;j}| \, 1 \le I \le n\} \mid 1 \le j \le N\}$ (Barrick *et al*., 1994).

Все эти биоинформатические нововведения были интегрированы теорией статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987). В рамках этой теории все четыре варианта нуклеотида в каждой из n позиций контакта ДНК/белок рассматриваются независимыми при их равновероятных эволюционных заменах, а также независимы все

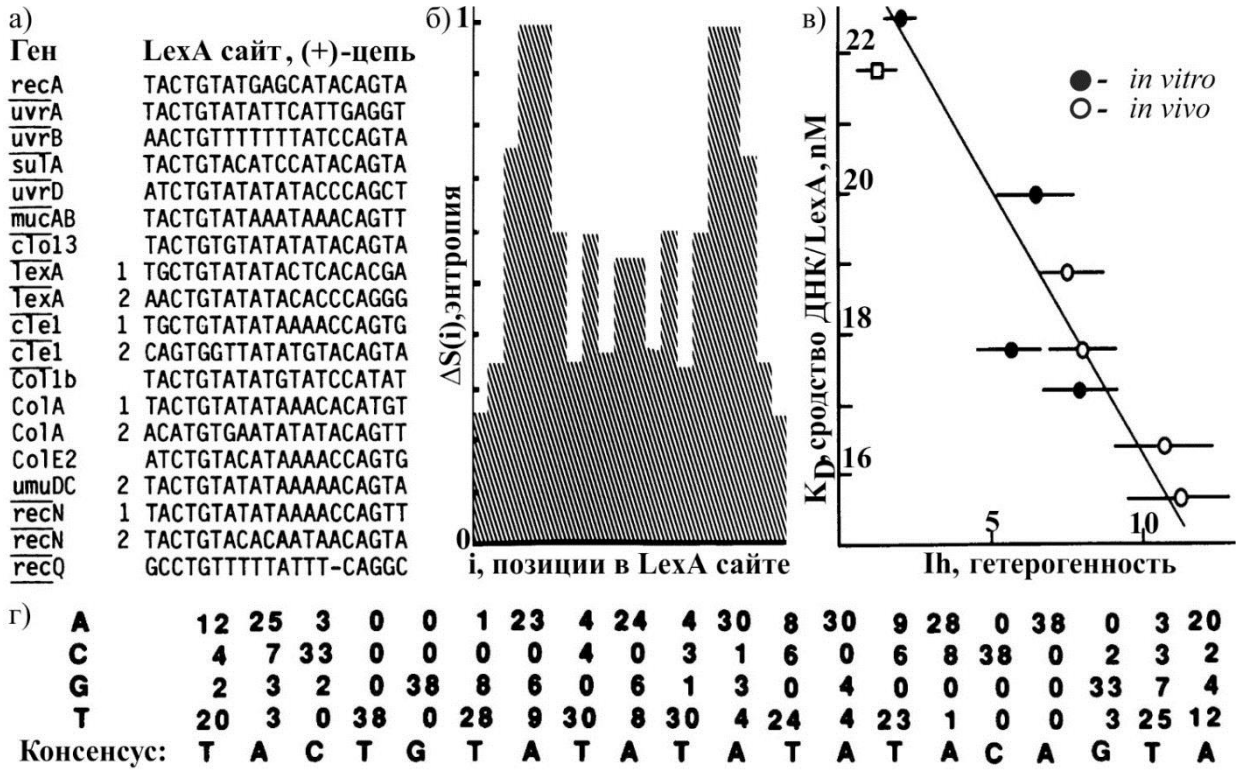

Рисунок 21 - Теория (Berg, von Hippel; 1987; Berg, 1988) статистической механики связывания ДНК с регуляторными белками (формулы 21-22) на примере (а) 19 сайтов связывания белком LexA ДНК *E. coli*, (б) оценки энтропии каждой позиции этого сайта, (в) корреляция между прогнозом этой теории и данными эксперимента, (г) позиционно-весовая матрица сайта связывания LexA. Рисунок автора на основе работы (Berg, 1988).

позиции ДНК при ее взаимодействии с белком. На Рисунке 21 можно видеть пример (Berg, 1988) применения этой теории (Berg, von Hippel; 1987) в случае анализа 19 сайтов связывания ДНК с белком LexA, активатором SOS-системы репарации разрывов хромосомы *Escherichia coli*.

Для учета поправок Байеса на неполноту исходных данных $\{\Psi_j | 1 \le j \le N << 4^n\}$ ввели (Berg, 1988) две позиционно-весовые матрицы сайта $\{s_1 \ldots s_i \ldots s_n\}$ связывания регуляторного белка: функциональная $w^0_{i;\xi}$ и структурная $w^{\#}_{i;\xi}$. В этих обозначениях оценке сродства ДНК/белок, $K_D(s_1 \ldots s_i \ldots s_n)$, относительно его гипотетически возможного максимума $K_{D;MAX}$ при самом частом $N_{i;MAX}$ нуклеотиде в каждой позиции i сайта длины n среди

N его вариантов соответствует стат-сумма Ih, тогда как мере естественной неоднородности исходных данных об этом сайте (включая неполноту) - оценка разности энтропий, ΔS, этих его вариантов:

$$\begin{cases} Ih \stackrel{\text{def}}{=} ln\left(\frac{K_D(s_1 \dots s_i \dots s_n)}{K_{D;MAX}}\right) \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_{i;s_i}^{0} \stackrel{\text{def}}{=} \sum_{i=1}^{n} ln\left(\frac{N_{i;s_i}+0.5}{N_{i;MAX}+0.5}\right); & (21) \\ \Delta S \stackrel{\text{def}}{=} \sum_{\xi \in \{A,T(U),G,C\}} w_{i;\xi}^{\#} \stackrel{\text{def}}{=} \sum_{\xi \in \{A,T(U),G,C\}} \left(\frac{N_{i;\xi}+1}{N+4} ln\left(\frac{4\left(N_{i;\xi}+1.5\right)}{N+4.5}\right)\right). & (22) \end{cases}$$

В свою очередь, для сайта связывания TATA-связывающего белка ("**T**ATA-**B**inding **P**rotein", TBP, англ. яз.) в промоторах эукариот заменили функцию потерь σ^2 (формулы 17 - 22) на критерий χ^2 сходимости позиционно-частотной матрицы $f_{i;\xi}$ (Bucher, 1990) методом Монте-Карло (Metropolis., Ulam, 1949) "… → оценка $f^{(k)}_{i;\xi}$ → k-ый шаг распознавания TATA-боксов →переоценка $f^{(k+1)}_{i;\xi}$ по результату распознавания TATA-боксов →…" для встречаемости $N_{i;\xi}$ каждого нуклеотида ξ в каждой позиции i распознанных TATA-боксов при ожидаемых вероятностях $p_{\xi\zeta}$ всех 16 возможных динуклеотидов ξζ и эвристических параметрах ε_0 и $\{\varepsilon_i\}$ "настройки " ("refinement" англ. яз., см. (Bucher, 1990)) для учета неполноты и неоднородности данных обо всех 502 известных тогда промоторах позвоночных:

$$ln\left(f_{i;\xi}\right) = \ln\left(\frac{N_{i;\xi}}{\sqrt{\left(\sum_{\zeta \in \{A,T,G,C\}} N_{i-1;\zeta} p_{\zeta\xi}\right)\left(\sum_{\zeta \in \{A,T,G,C\}} N_{i+1;\zeta} p_{\xi\zeta}\right)}} + \varepsilon_0\right) + \varepsilon_i. \quad (23)$$

В результате обработки 79% из 502 промоторов эукариот была получена позиционно-частотная $f_{i;\xi}$ матрица (Bucher, 1990), элементы которой были выражены в натуральных логарифмических единицах (здесь и далее: ln-ед.). Контекстно-зависимая сумма этих элементов вдоль заданной последовательности ДНК стала общепризнанной как критерий TATA-содержащих промоторов генов эукариот, Рисунок 22.

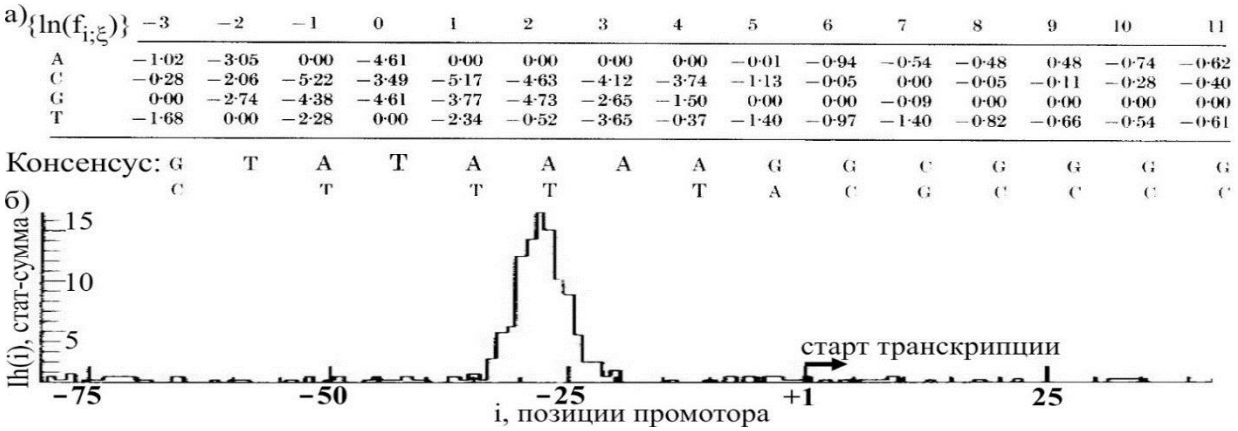

Рисунок 22 - Позиционно-частотная матрица $f_{i;\xi}$ (Bucher, 1990), общепризнанный критерий ТАТА-бокса эукариот (а), и (б) идентификация 346 ТАТА-содержащих из всех 502 промоторов в качестве предела сходимости опыта *in silico* при пороге −8.16 натуральных логарифмических единиц (здесь и далее: ln-ед.) между ТАТА-содержащими и ТАТА-несодержащими промоторами генов (Bucher, 1990). Рисунок автора на основе работы (Bucher, 1990).

Соответственно, оставшиеся 21% из этих 502 промоторов были названы “ТАТА-несодержащими”. Существенно, что эта позиционно-частотная $f_{i;\zeta}$ матрица была построена для распознавания ТАТА-боксов, и ее автор (Bucher, 1990) рекомендовал не применять ее для оценки количественных величин транскрипционной активности гена, поскольку при ее построении он исключил 21% промоторов.

В главе 3 настоящей диссертации можно найти результаты автора по выявлению контекстно-зависимых конформационных характеристик B-формы спирали ДНК ТАТА-боксов и содержания динуклеотидов вокруг него, объединение которых с позиционно-частотной матрицей ТАТА-бокса (Рисунок 22) на основе оригинальных данных (Савинкова и др., 2007) впервые позволило достоверно предсказать как абсолютные ($\alpha < 10^{-6}$), так и относительные ($\alpha < 0.01$) величины сродства TBP/ДНК (Пономаренко П. и др.,

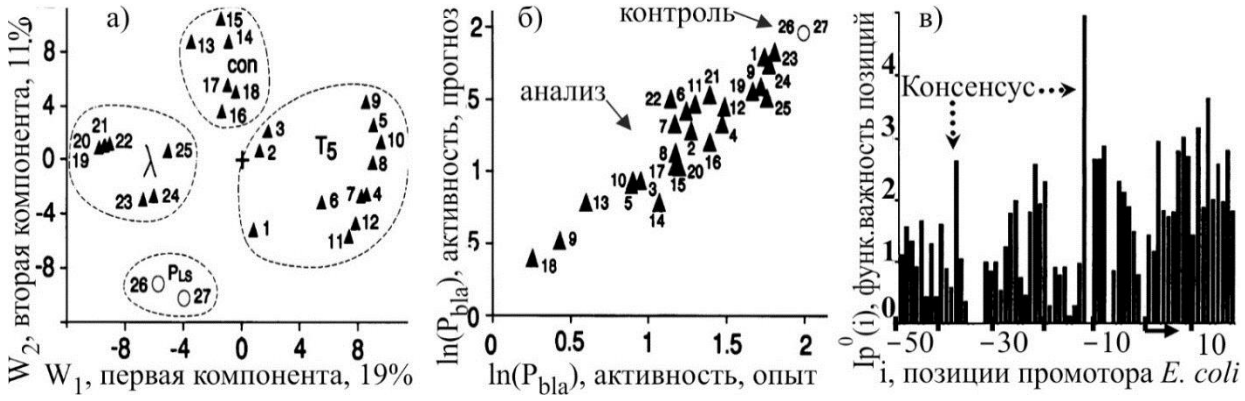

Рисунок 23 - Пример авторов работы (Jonsson *et al*., 1993) анализа (формулы 24 и 25) силы $\ln(P_{bla})$ 27 промоторов *E. coli*, натуральный логарифм отношения уровней транскрипции с промоторов заданного гена и гена β-галактозидазы (“promoter strength”, англ. яз.): (а) две достоверные главные компоненты W_1 и W_2 учли неоднородность 27 промоторов *E. coli*, представлявших в разной степени четыре различные группы координировано регулируемых генов, благодаря чему авторами работы (Jonsson *et al*., 1993) был преодолен (б, в) этот главный дефект применения функций потерь σ^2. Рисунок автора на основе работы (Jonsson *et al*., 1993).

2008), подтвержденные как в равновесных (Savinkova *et al*., 20133), так и в неравновесных условиях (Drachkova *et al*, 2014) опытов *in vitro*.

В работе (Jonsson *et al*., 1993) остроумно преодолели негативное влияние неоднородностей исходных экспериментальных данных на минимизацию функции потерь σ^2 благодаря учету лишь достоверных главных компонент представления последовательности ДНК длины n в 4n-мерном булевом пространстве с 4-битным кодом нуклеотидов {A=1000, T=0100, G=0010, C=0001}. При этом сначала нашли (Рисунок 23) частоты $f_{k;i;\xi}$ нуклеотида ξ в i-ой позиции сайта и позиционно-весовые матрицы $\{w_{k;i;m}| 1\le m\le 4; 1\le i\le n; 1\le k\le K<<N\}$ для каждой k-ой достоверной главной компоненты отдельно и затем объединили их все коэффициентами β_k множественной регрессии для

количественных величин $\{\Psi_j|\ 1\leq j\leq N\}$ активности N фрагментов ДНК $\{s_{i;j}|\ 1\leq i\leq n;\ 1\leq j\leq N\}$:

$$\sigma^2 = \max_{\{\beta_k|1\leq k\leq K\}}\left\{\sum_{j=1}^{N}\left(\Psi_i - M_0(\Psi) - \sum_{k=1}^{K}\beta_k\left(\sum_{i=1}^{n} f_{i;s_{i;j}} + \sum_{m=1}^{4} w_{k;i;m}\,\Delta_{k;m;s_{ij}}\right)\right)^2\right\}. \quad (24)$$

При этом были также впервые введены (Jonsson *et al.*, 1993) эмпирические оценки для функциональной $Ip^0(i)$ и для структурной $Ip^{\#}(i)$ важности позиций регуляторного сайта в составе геномной ДНК:

$$\begin{cases} Ip^0(i) = \sum_{k=1}^{K}\sum_{m=1}^{4}\dfrac{abs\left|\beta_k w_{k;i;m}\right|}{4K-1}; \\ Ip^{\#}(i) = \sum_{k=1}^{K}\sum_{m=1}^{4}\dfrac{abs\left|w_{k;i;m}\right|}{4K-1}. \end{cases} \quad (25)$$

На Рисунке 23 показан пример (Jonsson *et al.*, 1993) анализа с помощью формул (24 - 25) экспериментально измеренных величин $\ln(P_{bla})$ силы 27 промоторов *E. coli*, выраженной в натуральных логарифмических единицах отношения активности транскрипции с исследуемого промотора к фиксированному варианту промотора перед репортерным геном β-галактозидазы (“promoter strength”, англ. яз.). Как можно видеть, учет лишь достоверных главных компонент, действительно, позволил преодолеть негативное влияние неоднородности 27 промоторов *E. coli* на минимизацию функции потерь σ^2.

Идея множественности позиционно-весовых матриц $w_{i;\xi}$ определенного сайта связывания регуляторных белков была развита в дальнейшем экспериментами (Takeda *et al.*, 1989; Sarai, Takeda, 1989) по измерению в единицах свободной энергии Гиббса, ΔΔG, (Рисунок 24) каждого элемента матрицы $w_{i;\xi}$ для регуляторного сайта O_R1 переключения между стадиями литического цикла фага λ путем связывания O_R1 с λ– (Sarai, Takeda, 1989) или с Cro- (Takeda *et al.*, 1989) репрессором, соответственно. В этих опытах синтезировали олигоДНК со всеми возможными одиночными заменами нуклеотидов в природном варианте WT сайта O_R1 (общепринятое обозначение

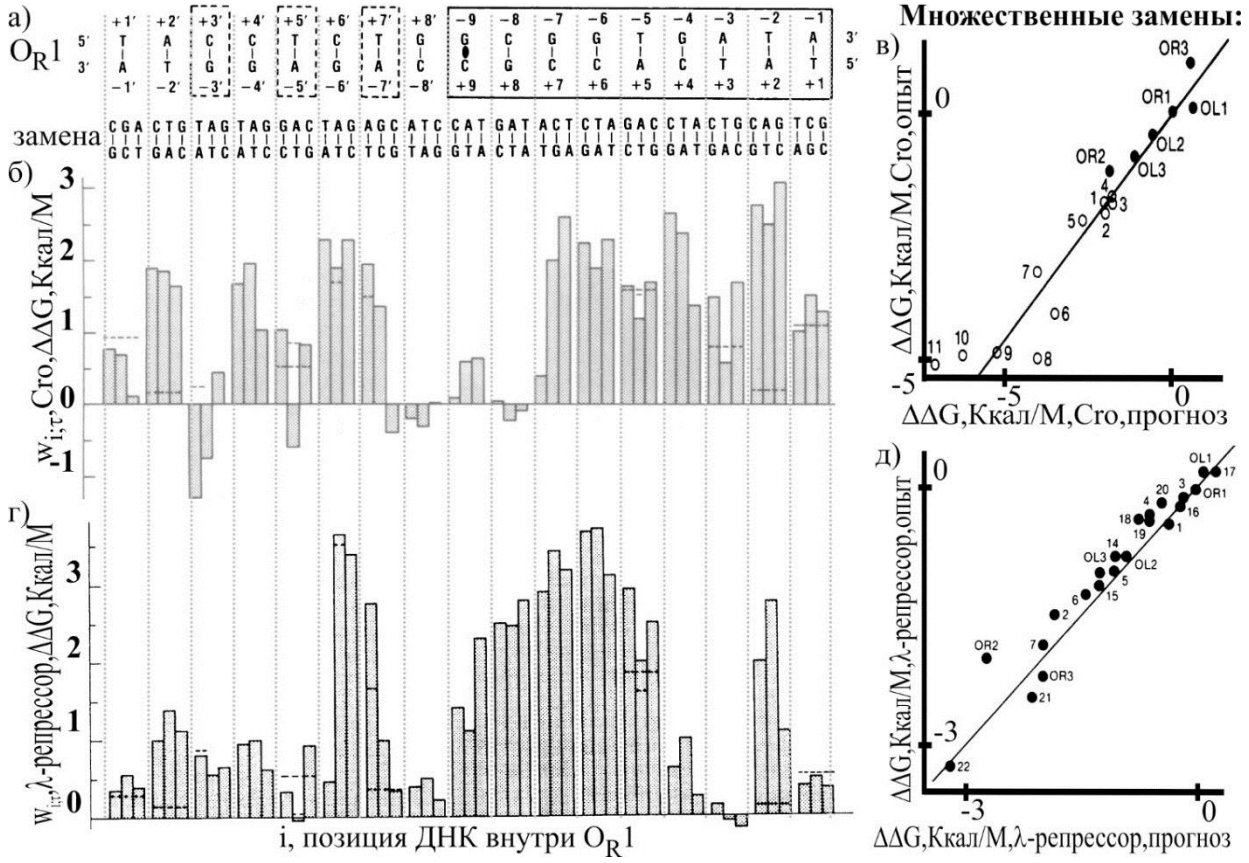

Рисунок 24 - Введение авторами работ (Takeda *et al*., 1989; Sarai, Takeda, 1989) экспериментального измерения элементов позиционно-весовой матрицы $w_{i;\xi}$ сайта связывания регуляторного белка на примере (а) сайта O_R1 переключения между стадиями литического цикла фага λ, соответствующим его связыванию с (б) Cro-репрессором и (г) с λ-репрессором, а также (в, д) корреляций между прогнозами с помощью этих матриц и данными независимых опытов. Рисунок автора на основе иллюстраций из работ (Takeda *et al*., 1989; Sarai, Takeda, 1989).

“WT”, “дикий тип”, “**W**ild **T**ype”, англ. яз.), измеряли равновесную константу диссоциации K_D комплексов ДНК/белок и на основе этих измерений экспериментально определяли каждый элемент позиционно-весовых матриц $w_{i;\xi}$:

$$w_{i;\xi} \stackrel{\text{def}}{=} \Delta\Delta G_{i;\xi} = ln\left(K_{D;\xi}/K_{D;WT}\right) \times [-0.546 \text{ Ккал/моль}], \quad (26)$$

здесь: -0.546 Ккал/моль - постоянная Больцмана.

Результат этих измерений (Takeda *et al*., 1989; Sarai, Takeda, 1989) показан на Рисунке 24, включая нумерацию позиции обеих нитей ДНК (а). Можно видеть очевидное различие между позиционно-весовыми матрицами $w_{i;\xi}$ сайта-переключателя O_R1, соответствующими связыванию с репрессорами Cro (б) и λ (г), а также достоверные прогнозы влияния мутаций в O_R1 на сродство к каждому из этих белков (в, д). Это свидетельствует, что один и тот же сайт может иметь разные молекулярные механизмы взаимодействия с разными регуляторными белками и, в этой связи, функциональная важность нуклеотидов в позициях сайта может быть разной для его связывания с разными белками.

Учет подобных различий особенно важен для транскрипционных факторов эукариот, которые связывают на регуляторной геномной ДНК как специфичные к ним сайты, так и боксы с широким спектром специфичности к белкам (например, Е-бокс, CCAAT-бокс), а также композиционные элементы в транскрипционных машинах, состав которых может варьировать для одного и того же гена в зависимости, например, от внутриклеточных условий.

Такие боксы, композиционные элементы и транскрипционные машины выходят за рамки теории статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987), основанной на гипотезе о связи между биологической активностью регуляторного сайта, сродством этого сайта к соответствующему белку и частотами нуклеотидов в позициях этого сайта.

На Рисунке 25 в качестве иллюстративного примера показано различие между позиционно-частотными матрицами (а) природных (“*in vivo”*) сайтов связывания геномной ДНК *E. coli* с белком Lrp регуляции ответа на лейцин (**L**eucine-responsive **r**egulatory **p**rotein) и (б) селектированных *in vitro* Lpr-аффинных олигоДНК, найденных авторами работы (Shulzaberger, Schneider, 1999) в числе подобных различий для 51 регуляторных белков *E. coli*. Соответствия между природными сайтами связывания регуляторных белков и селекцией *in vitro* высокоаффинных олигоДНК к этим белкам, а также между

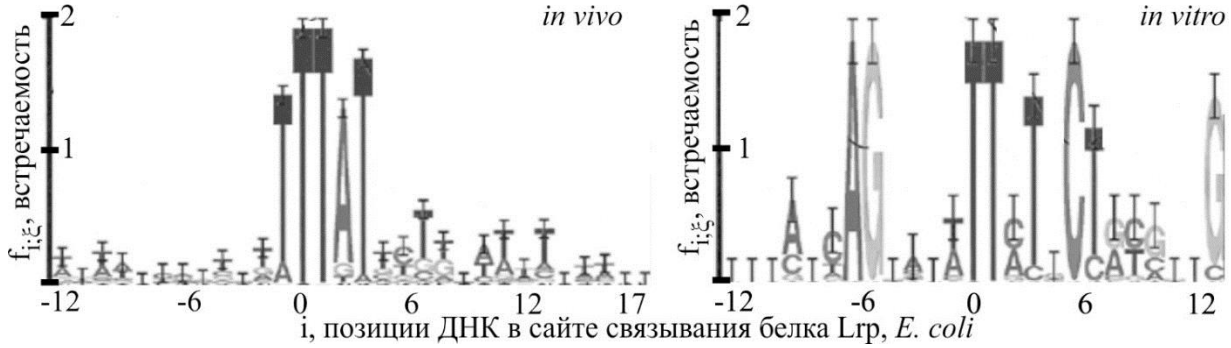

Рисунок 25 – Позиционно-частотные матрицы $f_{i;\xi}$, соответствующие природным сайтам связывания регуляторного белка Lpr у *E. coli* (слева, селекция *in vivo*) и Lpr-высокоаффинным олигоДНК (справа, селекция *in vitro*) имеют очевидные достоверные различия (Shulzaberger, Schneider, 1999). Рисунок автора на основе работы (Shulzaberger, Schneider, 1999).

регуляторной активностью сайтов и эффективностью их связывания с белками исследуются автором (Ponomarenko J *et al*., 2000a, 2002c) в главе 5 настоящей диссертации.

В свою очередь, в работе (Roulet *et al*., 1998) измерили сродство фактора транскрипции CTF-1/NF-1 к 26 известным сайтам его связывания и сравнили их с прогнозами трех самых используемых систем MATRIX SEARCH (Chen *et al*., 1995), MatInspector (Quandt *et al*., 1995) и TESS (Schug, Overton, 1997) для распознавания этих сайтов (Рисунок 26). Как можно видеть на Рисунке 26а, прогнозы этих трех компьютерных систем достоверно коррелировали между собой (Рисунок 26а) вопреки различию между ними по критериям оптимизации позиционно-весовых матриц $w_{i;\xi}$, но не коррелировали с экспериментальными величинами сродства CTF-1/NF-1 к природным сайтам связывания (Рисунки 26б,в,г: коэффициент τ ранговой корреляции Кендалла, τ, был от 0.2 до 0.4.) вопреки теории статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987).

Это несоответствие было впоследствии устранено (Roulet *et al*., 2000) вне рамок теории статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987) путем дополнительного экспериментально-

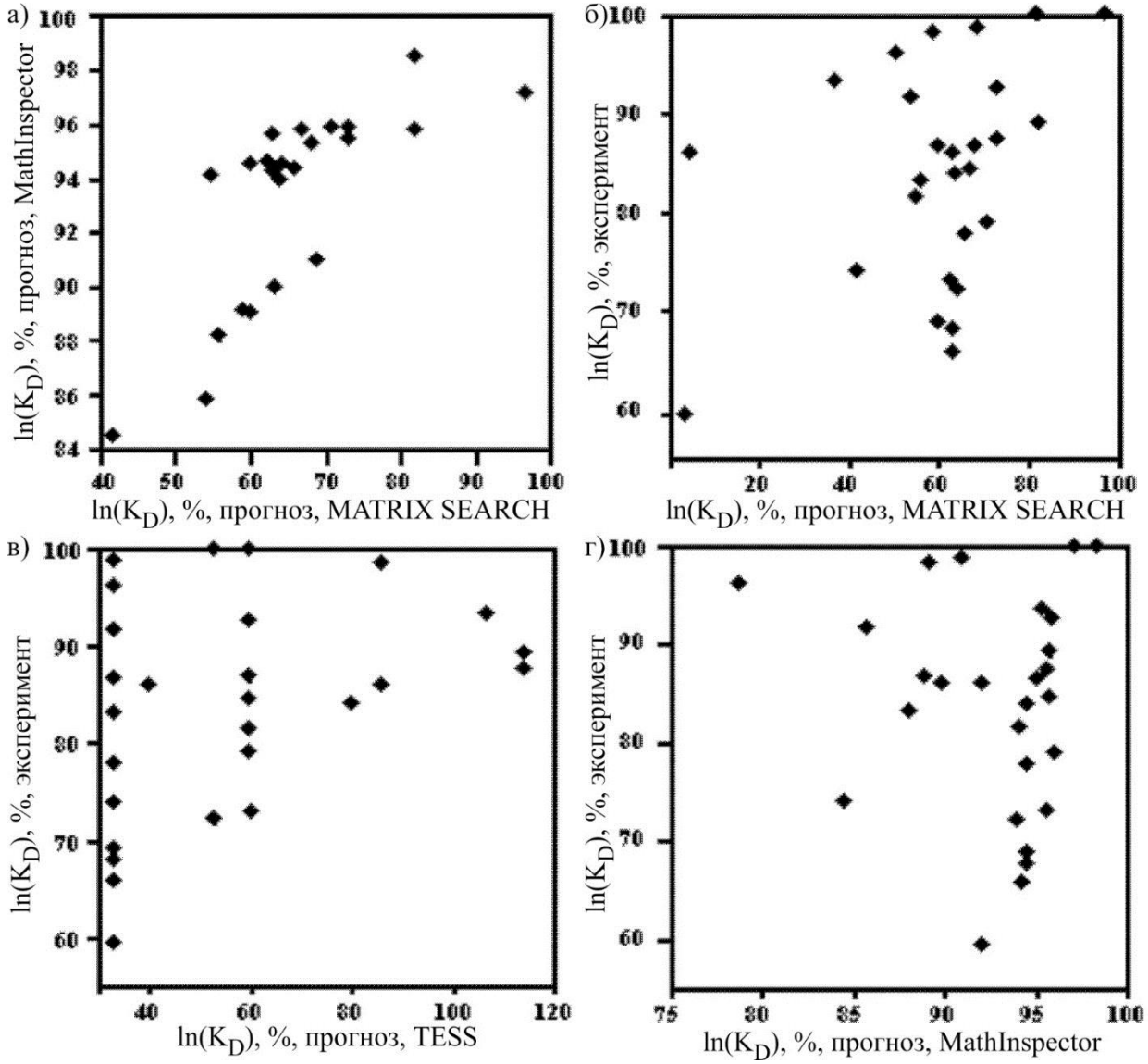

Рисунок 26 - Прогнозы трех позиционно-весовых матриц $w_{i;\zeta}$, оптимизированные наиболее часто используемыми методами MATRIX SEARCH (Chen *et al*., 1995), MatInspector (Quandt *et al*., 1995) и TESS (Schug, Overton, 1997), для сайтов связывания ДНК с транскрипционным фактором CTF-1/NF-1 достоверно (а) коррелируют между собой (б, в, г), но не руют с экспериментально измеренным сродством между ДНК и CTF-1/TF-1 (Roulet *et al*., 1998). Рисунок автора по материалам статье (Roulet *et al*., 1998).

компьютерного исследования влияния замен нуклеотидов на гибкость спирали ДНК и на размер зоны контакта между ДНК и CTF-1/NF-1 в их комплексе.

Наконец, измерение влияния замен нуклеотидов в инициаторном элементе, Inr, старта транскрипции генов млекопитающих (Javahery *et al*., 1994) на концентрацию преинициирующих комплексов и на активность транскрипции обнаружило независимость их друг от друга. Также, активность репрессии флюоресценции в случае использования селектированных *in vitro* YY1-аффинных олигоДНК в плазмиде pGL2 вместо природного сайта YY1 промотора к репортерному гену LUC не коррелировали ни между собой в клетках линий HeLa и PYS-2, ни со сродством ДНК/YY1 (Hyde-DeRuyscher *et al*., 1995).

В разделе 5.3 настоящей диссертации можно найти результаты исследования автором (Ponomarenko J *et al*., 2001a) в рамках диссертационной работы всех этих несоответствий как между теорией статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987) и экспериментами, так и между разными экспериментами. Представленные выше и многие другие несоответствия между результатами компьютерных расчетов и экспериментов, а также между результатами различных независимых экспериментов были обобщены в рамках “гипотезы молекулярной мозаики” (“jigsaw puzzle hypothesis”, англ. яз.) (Johnson, McKnight, 1989). Согласно этой фундаментальной гипотезе (Johnson, McKnight, 1989), в зависимости от внутриклеточных условий на одной и той же регуляторной геномной ДНК может собираться широкое разнообразие различных регуляторных молекулярных машин, которые способны конкурентно ингибировать друг друга на основе механизмов межмолекулярного распознавания ДНК/белок и белок/белок, как “ключ/замок”. В рамках этой гипотезы, каждый регуляторный сайт в составе геномной ДНК может связывать различные регуляторные белки, конкурирующие за связывание с этим сайтом в зависимости от внутриклеточных условий. Так “гипотеза молекулярной мозаики” (Johnson, McKnight, 1989) требует выхода за рамки статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel; 1987).

ЗАКЛЮЧЕНИЕ ПО ОБЗОРУ ЛИТЕРАТУРЫ

Анализ современного состояния литературы показал, что до развития технологий NGS большинство биоинформатических исследований структурно-функциональной организации геномной ДНК были нацелены на распознавание в ней сайтов связывания регуляторных белков (Oshchepkov, Levitsky, 2011) и лишь в ряде разрозненных биоинформатических работ исследовались физико-химические механизмы, лежащие в основе количественных аспектов ДНК-белковых взаимодействий и связанных с ними особенностей работы функциональных сайтов (например, обзор (Stormo *et al.*, 1986)). К моменту начала диссертационной работы в фокусе биоинформатики были символьные объекты (повтор, периодичность, палиндром, тракт, кластер, консенсус и т.п.) в силу их адекватности символьным последовательностям геномной ДНК. Однако, количественные характеристики геномных последовательностей ДНК - температура плавления, жесткость, изгиб оси, кручение, энтропия, размеры большой и малой бороздок спирали ДНК, - могут влиять на связывание регуляторных белков с геномной ДНК.

Общепризнанные соответствия между символьным и количественным представлением регуляторных районов геномной ДНК укладывались в узкие рамки простых корреляций Пирсона между температурой плавления ДНК и содержанием нуклеотидов “A+T”, между чередованием пуринов/пиримидинов и отклонением спирали ДНК от идеальной формы (правила Калладина), между влиянием регуляторного белка на биологическую активность гена и числом совпадений сайта связывания этого белка с консенсусом этого сайта. Однако ряд опытов продемонстрировал отсутствие корреляций между результатами как количественных величин регуляторной активности одних и тех же сайтов в разных клеточных линиях, так и различных количественных характеристик определенного сайта. Это означало, что связи между символьными объектами и количественными

характеристиками сайтов в составе геномных ДНК выходят за рамки простых корреляций Пирсона. Поэтому оставалась неясной бездна вопросов, которые можно коротко резюмировать так: “Какие количественные характеристики сайтов в составе геномных ДНК могут быть биологически значимыми для экспрессии гена (если они, конечно, есть)? Как найти такие характеристики сайтов в составе геномных ДНК путем анализа данных об экспрессии определенного гена? Как на этой основе оценить влияние мутаций на экспрессию этого гена в том же опыте или иных генов в иных опытах (если это, конечно, было бы возможным)?”

Но расшифровка референсного генома человека спровоцировала массовое секвенирование индивидуальных геномов пациентов, геномов ценных животных и растений, а также микроорганизмов. Количественные отличия экспрессии генов у индивидов в рамках одного биологического вида сделали вышеперечисленные вопросы актуальными как для прикладных (медицина, селекция), так и для фундаментальных (генетика) исследований. Настоящая диссертация обобщает оригинальные статьи автора, в которых он нашел ряд ответов на подобные вопросы в границах области применимости математической биологии и биоинформатики.

ГЛАВА 2 КОМПЬЮТЕРНАЯ СИСТЕМА BDNAVIDEO: КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ДНК САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

Настоящая глава описывает компьютерную систему bDNAvideo, которая была создана в рамках настоящей диссертации и благодаря которой были обнаружены статистически достоверные конформационные и физико-химические характеристики B-формы спирали ДНК сайтов связывания транскрипционных факторов (Пономаренко М. и др., 1997в; Колчанов и др., 1999; Ponomarenko M. *et al*., 1997b; Kolchanov *et al*., 1998, 1999; Ponomarenko J. *et al*., 1999a,b).

К моменту выполнения этой части диссертационной работы в конце 90-ых годов XX века самыми используемыми методами компьютерного анализа регуляторной ДНК были консенсус (Hawley, McClure, 1983) и позиционно-весовая матрица (Mulligan *et al*., 1984), как было описано в предыдущей главе. Они были основаны на символьном выравнивании последовательностей ДНК экспериментально доказанных сайтов связывания определенных регуляторных белков (Needleman, Wunsch, 1970; Lawrence *et al*., 1993). Кроме того, теория статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel, 1987) установила соответствие между этими методами и линейно-аддитивным приближением взаимодействия ДНК/белок в виде сумм независимых вкладов взаимодействий всего белка в целом с каждым нуклеотидом отдельно. Благодаря этим методам был достигнут впечатляющий успех в компьютерном анализе регуляторных районов генов прокариот (например, (Berg, 1988)), однако в случае эукариот они оказались менее эффективными вследствие, по-видимому, влияния нуклеосомной ДНК геномов эукариот, существенно ограничивающей независимость нуклеотидов. Даже самый консервативный из сайтов связывания транскрипционных факторов эукариот, TATA-бокс, был успешно распознан общепризнанным для него критерием (Bucher, 1990) только лишь для 79% экспериментально

доказанных на тот момент его локализаций. В постгеномную эру высокопроизводительное секвенирование/иммунопреципитация хроматина (ChIP-Seq) сайтов на геномной ДНК для связывания ТАТА-связывающего белка (TBP) снизили эту оценку в 4 раза с 79% до 20% (Tirosh *et al*., 2006) для канонических ТАТА-боксов, удовлетворяющих критерию (Bucher, 1990), тогда как на трансгенных мышах $TBP^{(-)(-)}$ было доказано обязательное присутствие сайта связывания TBP во всех без исключения промоторах генов эукариот (Martianov *et al*., 2002). Поэтому возник вопрос: "Как учесть кооперативные взаимодействия нуклеотидов в процессе связывании регуляторных белков с ДНК?". Решению этого вопроса на примере среднеарифметических оценок физико-химических и конформационных свойств спирали ДНК для сайтов связывания транскрипционных факторов посвящена эта глава настоящей диссертационной работы.

Поскольку к началу диссертационной работы в литературе были свидетельства как в пользу прогноза величин количественных характеристик специфической биологической активности сайтов в составе геномных ДНК с помощью ее символьных характеристик на основе частот нуклеотидов в позициях консенсуса (17), так и против таких прогнозов (Рисунки 24-26), то был предложен компромисс, линейно-аддитивное приближение (Колчанов и др., 1998):

$$F_{\text{опыт}}(Seq_{\text{ДНК}}) = F_0 + \sum_{n=1}^{N} \varphi_n X_n(Seq_{\text{ДНК}}) \qquad (27)$$

здесь: $F_{\text{опыт}}(Seq_{\text{ДНК}})$ – экспериментальная величина количественной характеристики специфической биологической активности некоторых сайтов в составе геномных ДНК; F_0 - базовый уровень этой активности, общий для всех таких сайтов независимо от их контекста; $X_n(Seq_{\text{ДНК}})$ - контекстно-зависимая количественная характеристика ДНК, значения которой коррелируют с экспериментальными величинами $F_{\text{опыт}}(Seq_{\text{ДНК}})$ заданной активности; φ_n и N – коэффициенты регрессии и число характеристик.

Таким образом, с помощью формулы (27) была переформулирована целевая задача предсказания количественных величин $F_{опыт}(Seq_{ДНК})$ специфической биологической активности сайтов в составе геномных ДНК по нуклеотидной последовательности этих сайтов в новую задачу поиска контекстно-зависимых количественных характеристик $\{X_n(Seq_{ДНК})\}_{1\le n\le N}$ ДНК, величины которых статистически достоверно линейно коррелируют с количественными величинами $F_{опыт}(Seq_{ДНК})$ заданной активности этих сайтов. В этой главе диссертации в качестве примера количественных характеристик X контекста ДНК будут изучены контекстно-зависимые свойства B-формы двойной спирали ДНК.

2.1 Количественные характеристики спирали ДНК ТАТА-боксов эукариот

Прототип системы bDNAvideo был создан с помощью электронной библиотеки из 9 конформационных свойств гексануклеотидных шагов спирали ДНК (Karas *et al*., 1996), которые профессор Скленар (Германия) оценил методом компьютерного моделирования молекулярной динамики гетеродуплексов ДНК и любезно предоставил автору в рамках их совместной работы (Пономаренко М. и др., 1997в). Пробной содержательной задачей для оценки работоспособности биоинформатических новшеств этого прототипа был анализ с его помощью ТАТА-боксов трех групп эукариотических организмов: позвоночные, беспозвоночные и дрожжи, а также, в качестве независимого контроля, Прибнов-бокса (синоним “ТАТА-бокс”) *Escherichia coli*. В результате были найдены участки ДНК вокруг всех этих ТАТА-боксов, которые достоверно отличались от случайных нуклеотидных последовательностей по среднеарифметическим оценкам кручения, шага, ширины малой и большой бороздок спирали ДНК (Пономаренко М. и др., 1997в). Была обнаружена достоверная корреляция между положением организмов в эволюционном ряду (в рамках данной работы это было эвристически обозначено термином “эволюционный ранг”) и протяженностью

окрестности ТАТА-бокса, статистически значимой для его отличия от случайной ДНК по среднеарифметическим оценкам указанных свойств ДНК: “чем выше ранг, тем короче окрестность”. Это соответствовало общепринятому представлению об эволюционном усложнении транскрипционных машин в ряду “бактерии → дрожжи → беспозвоночные →позвоночные”.

В геномах эукариот наиболее представлены гены, транскрибируемые РНК-полимеразой II, RNAPII (Latchman, 1995). Сборка ее преинициаторного комплекса начинается с распознавания ТАТА-связывающим белком (TBP) сайта его связывания, ТАТА-бокса. Это включает реорганизацию нуклеосом (Godde *et al*., 1995); скольжение TBP вдоль ДНК (Coleman, Pugh, 1995) благодаря их неспецифическому сродству (Hahn *et al*., 1989); остановку TBP на ТАТА-боксе в силу специфического сродства между ними (Hahn *et al*., 1989) и стабилизацию комплекса “TBP/TATA” изгибом оси спирали ДНК с 19^{O} до 90^{O} (Powell *et al*., 2002). Было установлено (Starr *et al*., 1995), что изгиб спирали ДНК в комплексе с TBP зависит от нуклеотидного контекста и коррелирует как со сродством TBP/ДНК, так и с активностью транскрипции. Однако, вопрос о роли B-формы спирали ДНК в процессе связывания TBP с ТАТА-боксом оставался открытым.

Экспериментальный подход к решению этой проблемы был основан на расшифровке 3D-структур (A+T)-богатых олигоДНК методами рентгеноструктурного анализа додекамеров (Wing *et al*., 1980) и привел к формулированию правил Калладина для коррекции идеальной спирали ДНК Уотсона-Крика с учетом последовательности нуклеотидов (Calladine, 1982). Компьютерный подход был основан на методах конформационного анализа молекулярной механики и динамики (Neidle, 1994), с использованием которых искали низкоэнергетические конформации дуплексов для заданной последовательности ДНК (Lavery *et al*., 1982). В этих рамках проф. Скленар (Karas *et al*., 1996) вычислил свойства гексануклеотидных дуплексов ДНК, использованные в работе.

2.1.1 TATA-боксы промоторов генов эукариот (введение)

Голдберг и Хогнесс открыли (A+T)-богатый инвариант длиной 8 п.о. в промоторах генов гистонов дрозофилы (Lifton *et al*., 1978). Поскольку его консенсус оказался TATA(t/a)A(t/a)g, то он был назван “TATA-бокс” (синонимы: “ATA бокс”, бокс Голдберга-Хогнесса, Хогнесс-бокс, (Ponomarenko M. *et al*., 2013c)). Затем на TATA-боксе промотора гена кональбумина был открыт стабильный комплекс ДНК/белок, формирование которого предшествовало связыванию РНК полимеразы II (RNAPII) с промотором этого гена (Davison *et al*., 1983). В опыте (Parker, Topol, 1984) на промоторе гена hsp 70 теплового шока дрозофилы из числа инициаторных белков в составе анкерного комплекса ДНК/белок для RNAPII был идентифицирован инициаторный транскрипционный фактор TFIID, который связывал TATA-бокс (Ponomarenko M. *et al*., 2013e). В свою очередь было замечено (Fire *et al*., 1984), что в результате связывания RNAPII с анкерным комплексом TATA-бокса на этой основе начинается процесс сборки другого предшествующего транскрипции стабильного комплекса ДНК/белок, который был назван “активированным”, но затем переименован в “преинициаторный”. Наконец, у дрожжей нашли ген, который кодирует TATA-связывающий полипептид ДНК-связывающей субъединицы транскрипционного фактора TFIID (Schmidt *et al*., 1989), переименованный впоследствии в TATA-связывающий белок, TBP.

В настоящее время TATA-бокс является одним из самых изученных регуляторных сайтов на геномной ДНК эукариот: о нем опубликовано более 7000 научных статей. Прежде всего, хотя опыт (Wieczorek *et al*., 1998) с антителами против TBP продемонстрировал *in vitro* возможность транскрипции без этого белка (“TBP-free”, англ. яз.), до сих пор не было найдено ни одного гена эукариот с такой инициацией транскрипции *in vivo*. Более того, опыт *in vivo* (Martianov *et al*., 2002) с искусственным нокаутным вариантом гена *TBP*$^{(-)}$ у трансгенных мышей доказал нежизнеспособность

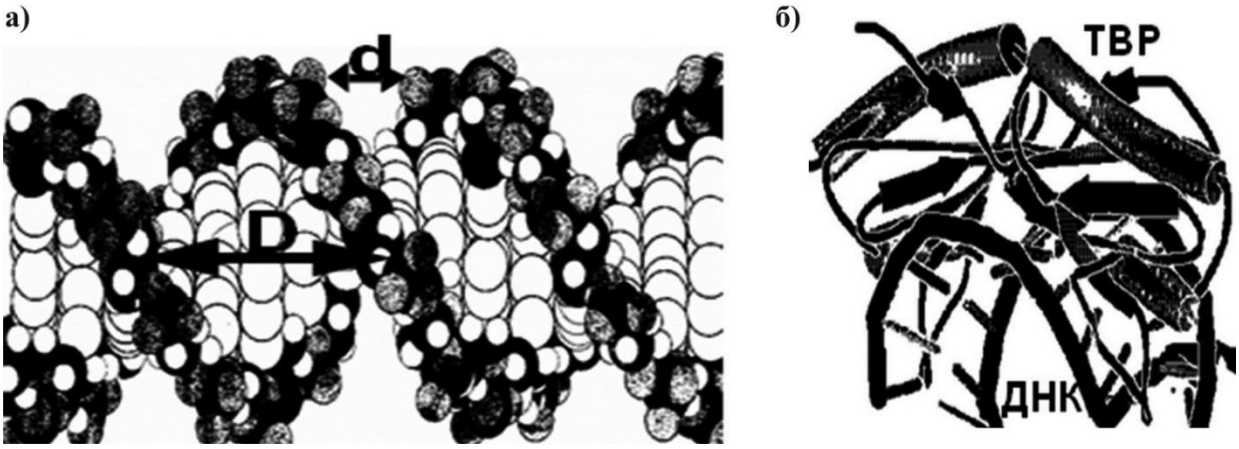

Рисунок 27 - Две конформации двойной спирали ДНК ТАТА-бокса: (а) слабо изогнутая (угол 19^{O}) В-форма спирали свободной ДНК (Drew *et al*., 1981) и (б) сильно изогнутая (угол 90^{O}) спираль ДНК в комплексе с TBP (Kim J *et al*., 1993; Kim Y *et al*., 1993). D и d – большая и малая бороздки ДНК, соответственно.

бластулы гомозигот $TBP^{(-)(-)}$ вследствие разбавления запаса материнского TBP ниже концентрации, необходимой для инициации транскрипции генов. ПоэтомуTBP/TATA-комплекс общепринято считать обязательным якорем на ДНК для связывания RNAPII и преинициирующего комплекса транскрипции (Auble, 2009). Наконец, для 17181 генов человека построили (Yang *et al*., 2011) полногеномную карту потенциальных ТАТА-боксов, выборочно доказанных экспериментально.

В свою очередь, с помощью рентгеноструктурного анализа для ТАТА-бокса были установлены трехмерные структуры как свободной (A+T)-богатой В-спирали ДНК (Drew *et al*., 1981), так и TBP/TATA-комплексов (Kim J. *et al*., 1993; Kim Y. *et al*., 1993). Этим двум состояниям ТАТА-бокса (Рисунок 27): (а) свободная ДНК и (б) комплекс ДНК/белок, - соответствуют экспериментально измеренные величины 10^{-5} М и 10^{-9} М сродства TBP/ДНК (Hahn *et al*., 1989): первое - неспецифическое, второе – специфическое TBP/TATA-сродство. Стало общепринятым разделять все промоторы эукариот для RNAPII на два класса: ТАТА-содержащие (≈20%) и ТАТА-

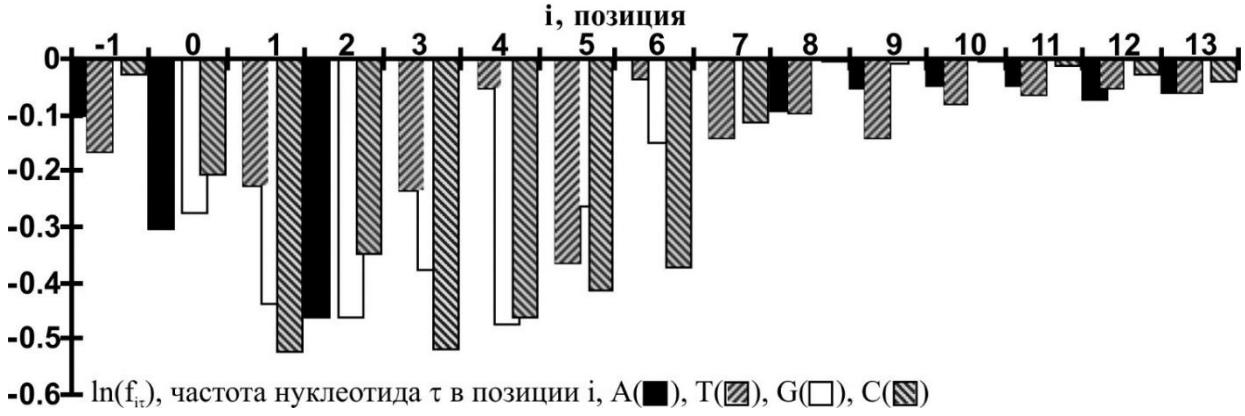

Рисунок 28 - Графическое представление матрицы $f_{i\eta}$ частот нуклеотидов η в позиции i "окна" длиной 15 п.о. с первым "T" в позиции i=0 консенсуса TATA-бокса (Bucher, 1990). Рисунок автора на основе таблицы численных величин $\ln(f_{i\eta})$, которые были опубликованы в статье (Bucher, 1990).

несодержащие (≈80%), - с помощью общепринятого критерия (Bucher, 1990) на основе позиционно-частотной матрицы $f_{i\eta}$ для "скользящего окна" длиной 15 п.о. с номерами позиций от -1 до +13 относительно первого "T" в консенсусе TATA-бокса (Рисунок 28):

$$\begin{cases} PWM_{TATA}(\eta_{-1}\eta_0\eta_{+1}\eta_{+2}\dots\eta_{+13}) = \sum_{i=-1}^{13} \ln\left(f_{i\eta_i}\right); \\ \max\limits_{-50\le i\le -20}\left[PWM_{TATA}(\eta_{i-1}\dots\eta_{i+13})\right] \begin{cases} > -8, & TATA-\textit{содержащий}; \\ \textit{иначе} & TATA-\textit{несодержащий}. \end{cases} \end{cases} \quad (28)$$

Такая классификация TATA-содержащих и TATA-несодержащих промоторов генов эукариот попала в фокус постгеномных исследований, в результате которых было обнаружено более 20 принципиальных биологических различий между ними, которые резюмированы в Таблице 8.

Оказалось, что эти два класса промоторов различаются вовсе не лишь по характерным для них величинам TBP/ДНК-сродства (Hahn *et al*., 1989) и/или по связыванию TBP с кор-промотором самостоятельно или при содействии других белков (Ou *et al*, 1994; Penner, Davie, 1994; Sax *et al*, 1995; Godde *et al*, 1995; Zhao *et al*, 1999; Wolner, Gralla, 2001; Wright *et al*, 2006; Hsu *et al*, 2008*).*

Таблица 8 – ТАТА-содержащие и ТАТА-несодержащие промоторы

свойство / промотор	ТАТА-содержащий	ТАТА-несодержащий	Литература
критерий (28)	> −8.16	≤ -8.16	Bucher, 1990
сродство TBP/ДНК	≈ 10^{-9} M	≈ 10^{-5} M	Hahn *et al*., 1989
TBP связывает ДНК в [-90; -20], чаще в 30 п.о. перед стартом транскрипции (TSS)	самостоятельно	в присутствии TFIIA и TFIIB, а также Sp1, NC2, Mot1, PC4, AP4, SAGA, TAF4, GATA и некоторых других регуляторных белков	Ou *et al*, 1994; Penner, Davie, 1994; Sax *et al*, 1995; Godde *et al*, 1995; Zhao *et al*, 1999; Wolner, Gralla, 2001)
консенсус DPE	R_{24}-A-GATC$^{\#}$	T_{17}G-----G--RGWYGT	Kutach, Kadonaga, 2000
консенсус INR	CA_1-GT	TCA_1NKTY	
	YYA_1-WYY	GSCGCCA_1TYTTG	Yarden *et al*., 2009
TSS	одиночный	множественные	Lin *et al*., 2001
предпочтительные генные сети	ответ на стимулы, развитие, органогенез, морфогенез	фотосинтез, энергия, “домашнее хозяйство”, дефосфорилирование, рост и деление клеток, транспорт, секреция, эндоцитоз, биосинтез, метаболизм	Nakamura *et al*., 2002; Yang *et al*., 2007; Dong *et al*., 2011
состав	(A+T)-богатый	(G+C)-богатый	
ответа на цАМФ	консенсус	вариабельный	Conkright *et al*,2003
ответ на индуктор	стохастически	детерминирован	Raser, O'Shea, 2004
давление отбора	направленное	нейтральное	Basehoar *et al*., 2004
реорганизация хроматина	TBP-зависимая	гистон-зависимая	
локализация	прителомерная	повсеместная	
спираль ДНК	жесткая	гибкая	Tirosh *et al*., 2007
промоторы генов гистонов	все, кроме гистона H1	только у гистона H1	Isogai *et al*., 2007; Cavalieri *et al*., 2009
регуляция	пластичная	постоянная	Lin *et al*., 2010
транскрипция	индуцибельная	конститутивная	Tora, Timmers, 2010
частоты кодонов	частые чаще	редкие чаще	Tatarinova *et al*., 2010
тРНК~мРНК при трансляции	“wobble”-пары чаще	канонические пары чаще	
экспрессия	дифферен-циальная	единообразная	Carreto *et al*., 2011
предпочтительные гены	микроРНК	белок-кодирующие	Long *et al*., 2011

Прежде всего, TATA-содержащие промоторы отличаются от TATA-несодержащих структурно: нуклеотидным составом (Yang *et al*., 2007), наличием главного старта транскрипции, TSS, (Lin *et al*., 2001), консенсусами нижележащего DPE и инициаторного INR элементов (Kutach, Kadonaga, 2000; Yarden *et al*., 2009), гибкостью спирали ДНК (Tirosh *et al*., 2007) и предпочтением к локализации в прителомерных районах хромосом (Basehoar *et al*., 2004). Кроме того, гены с TATA-содержащими промоторами отличаются от генов с TATA-несодержащими промоторами способом реорганизации нуклеосомы кор-промотора (Basehoar *et al*., 2004), индуцибельностью (Tora, Timmers, 2010) и пластичностью регуляции (Lin *et al*., 2010), чувствительностью к цикло-АМФ при наличии совпадений с консенсусом сайта CRE ответа на этот стимул (Conkright *et al*., 2003).В свою очередь, гены с TATA-содержащими и TATA-несодержащими промоторами характеризуются повышенной и пониженной встречаемости у них дифференциальной экспрессии (Carreto *et al*., 2011), соответственно, повышенным и пониженным давлением естественного отбора на белок-кодирующие районы ДНК этих генов (Basehoar *et al*., 2004), предпочтением в использовании частых и редких кодонов, использованием “wobble”- и канонических пар кодон:антикодон (мРНК:тРНК) в процессе трансляции (Tatarinova *et al*., 2010), стохастическим и детерминированным ответом на индукторы (Raser, O'Shea, 2004), а также специфичностью к определенным генным сетям (Nakamura *et al*., 2002; Yang *et al*., 2007; Isogai *et al*., 2007; Cavalieri *et al*., 2009; Dong *et al*., 2011). Доля генов с TATA-содержащими промоторами варьирует от вида к виду в границах от 10% до 30%, например, 29% у арабидопсиса (Molina, Grotewold, 2005). Список таких генов общепринято считать “геномной подписью эукариотического вида” (Tirosh *et al*., 2006). Интересно, что доля TATA-содержащих промоторов генов микроРНК у человека превышает 60% (Long *et al*., 2011).

Наконец, сравнительный анализ всех расшифрованных геномов эукариот установил TATA-бокс (вместе с нижележащим DPE, B-

регуляторным BREu, и инициаторным INR элементами кор-промоторов) в качестве универсального общего регуляторного сайта геномной ДНК грибов, растений, животных и вирусов (Florquin *et al*., 2005; Yamamoto *et al*., 2007). Более того, распознавание TATA-бокса в расшифрованных геномах часто используется (Abeel *et al*., 2008) в качестве “отправной точки” на ДНК, в окрестности примерно ±30 п.о. относительно которой ищут гораздо менее консервативный потенциальный старт транскрипции для неизвестного гена.

Структуры TATA-бокса. Прежде всего, было установлено (Imbalzano *et al*., 1994), что в хроматине TATA-бокс упакован специфической нуклеосомой кор-промотора с центром в 43 п.о. перед стартом транскрипции, TSS (Ioshikhes *et al*., 1999). Положение TATA-бокса в 3D-структуре этой нуклеосомы определяет, доступен TATA-бокс для TBP или нет. Важно, что в известной 3D-структуре нуклеосомы (Richmond, Davey, 2003) были найдены два (A+T)-богатых сайта длиной 5 п.о. для связывания гистоновых димеров H3-H4 на расстоянии ±13 п.о от ее центра, один из которых соответствует оптимальному положению TATA-бокса в 30 п.о. перед TSS, 43-13=30.

Поэтому для возможности инициации транскрипции гена (Luk *et al*., 2010), гистон-модифицирующие ферменты, белок-Mediator, TFIIA, TFIIB и TBP реорганизуют нуклеосому кор-промотора (Hsu *et al*., 2008; Wright *et al*., 2006; Contreras-Levicoy *et al*., 2008). В частности, связывание TBP с TATA-боксом в присутствии TFIIA снижает сродство октамера гистонов к ДНК (Godde *et al*., 1995). В свою очередь, присоединение транскрипционного фактора TFIIB к TBP/TATA-комплексу необратимо сдвигает равновесие от нуклеосомы кор-промотора, сначала, к анкерному комплексу TFIIB(TBP/TATA)TFIIA и, после присоединения RNAPII к нему (Davison *et al*., 1983), к преинициаторному комплексу транскрипции генов эукариот (Wolner, Gralla, 2001).

В известной 3D-структуре додекамера $d(CGCGAATTCGCG)_2$ (Drew *et al*., 1981) спираль свободной ДНК имеет слабый (19°) изгиб оси (Рисунок 27a), что было подтверждено десятками расшифрованных (A+T)-богатых

олигоДНК и соответствовало изгибу оси спирали ДНК в 3D-структуре нуклеосомы (Richmond, Davey, 2003). Напротив, сильный (90°) изгиб оси спирали ДНК был найден лишь в 3D-структуре (Kim J *et al*., 1993; Kim Y *et al*., 1993) комплекса TBP/TATA-бокс (Рисунок 27б). Компьютерная модель молекулярной динамики (Flatters, Lavery, 1998) показала, что невозможно изменить изгиб спирали ДНК от 19° до 90° без денатурации гетеродуплекса TATA-бокса. Это было подтверждено флюориметрически в опыте *in vitro* (Powell *et al*., 2002). Авторы недавней работы (Zanegina et al., 2016) выявили во всех известных пространственных структурах комплекса TBP/TATA-бокс человека, арабидопсиса, грибов и архей консервативные водородные связи между аминокислотными остатками и нуклеотидами, а также контакты между кластер-образными интерфейсами на поверхности глобулы TBP и на спирали ДНК, опосредованные молекулами воды.

Наконец, было экспериментально открыто (Coleman, Pugh, 1995) скольжение TBP вдоль спирали ДНК, обусловленное неспецифическим сродством TBP/ДНК (10^{-5} M), вплоть до остановки TBP на TATA-боксе из-за их тысячекратно большего специфического TBP/TATA-сродства (10^{-9} M) (Hahn *et al*., 1989). В терминах теории статистической механики связывания регуляторных белков с ДНК (Berg, von Hippel, 1987), критерий (формулы 27-28) для дискриминации между TATA-содержащими и TATA-несодержащими промоторами (Bucher, 1990) соответствует стадии экспериментально известной остановки скольжения TBP на TATA-боксе (Coleman, Pugh, 1995).

Функция TATA-бокса. В случае реорганизации нуклеосомы кор-промотора с участием гистон-модифицирующих ферментов и белка Mediator (Wright *et al*., 2006; Hsu *et al*., 2006; Contreras-Levicoy *et al*., 2008) TBP неспецифически связывает ДНК (Hahn *et al*., 1989), скользит вдоль нее (Coleman, Pugh, 1995) до своей остановки на TATA-боксе (Berg, von Hippel, 1987; Bucher, 1990) в силу специфического (Hahn *et al*., 1989) связывания между ними в TBP/TATA-комплекс (Kim J *et al*., 1993; Kim Y *et al*., 1993). Существенно, что был описан (Schroer, 2010) общий TATA-бокс двух

противоположно направленных одновременно транскрибируемых генов *RhoA* и *TCTA* человека, старты транскрипции которых на расстоянии 90 п.о. и 28 п.о. от этого TATA-бокса, соответственно, были границами их общего двунаправленного промотора длиной 112 п.о. Это подтвердило *in vivo* способность TBP/TATA-комплекса к инициации транскрипции обеих нитей ДНК, показанную ранее *in vitro* (Wang *et al*., 1996). Выбор нити для транскрипции, 5'→BREu→(TBP/TATA)→BREd→3', завершает сборку "якоря" на ДНК для RNAPII (Davison *et al*., 1983) присоединением TFIIA и TFIIB к TBP/TATA-комплексу, поскольку BREu – более консервативный из двух сайтов связывания TFIIB, BREu и BDEd, непосредственно перед и после TATA-бокса, соответственно (Lagrange *et al.,* 1998; Tsai, Sigler, 2000; He *et al.*, 2016).

Вследствие взаимно однозначной ориентации анкерного комплекса TFIIB(TBP/TATA)TFIIA и гетеродимера RNAPII/TFIIF они мгновенно связываются (Juven-Gershon *et al*., 2008) в начале сборки преинициаторного комплекса. Затем идет длительная аллостерическая перегруппировка всех белковых субъединиц (Yakovchuk *et al*., 2010), в результате которой возникает комплекс ДНК/белок, сходный с "закрытым комплексом" на промоторах *E. coli* (Yakovchuk *et al*., 2010). В свою очередь, присоединение к нему транскрипционного фактора TFIIE вызывает локальное плавление гетеродуплекса ДНК и фиксацию отдельных ее нитей (Lee, Young, 2000), что сходно с "открытым комплексом" промоторов *E. coli*. Наконец, последним к ним присоединяется TFIIH, субъединицы циклин H и циклин-зависимая киназа 7 которого фосфорилируют RNAPII и снижают ее сродство к ДНК (Lee, Young, 2000). Нормальным итогом вышеперечисленных шагов является преинициаторный комплекс, ответственный за базальную транскрипцию (Fire *et al*., 1984). Для генов ответа на стресс было описано (Xing *et al*., 2005) одновременное ингибирование диссоциации преинициаторного комплекса и базальной транскрипции для мгновенного ответа на стресс с появлением соответствующего активатора. Это явление было названо остановленный

(“bookmarked”, англ. яз.) преинициаторный комплекс (Xing *et al*., 2005). Аналогичная остановка преинициаторного комплекса была экспериментально доказана (Vernoux *et al*., 2011) также для ауксин-зависимых генов растений в отсутствие ауксина, главного гормона-морфогена растений.

Первый нуклеотид считываемой РНК соответствует нуклеотиду ДНК, который был назван сайтом старта транскрипции (TSS) и который расположен в 5’→BREu→(TBP/TATA)→BREd→3’направлении на расстоянии от 20 п.о. до 90 п.о. (чаще 30 п.о.) от первого “**T**” в консенсусе **T**ATA(t/a)A(t/a)G канонического TATA-бокса (Juven-Gershon, Kadonaga, 2010; Tsai, Sigler, 2000). Были установлены достоверные корреляции активности транскрипции с отклонением расстояния между TATA-боксом и стартом транскрипции от оптимума 30 п.о. (Ponjavic *et al*., 2006) и с TBP/TATA-сродством (Mogno *et al*., 2010).

Эволюция TATA-бокса. На генах дрожжей были обнаружены (Bazykin, Kondrashov, 2006) переходы промоторов между классами TATA-содержащие ↔ TATA-несодержащие, туда и обратно, но биологический смысл этого явления до сих пор остается неясным из-за слабой изученности таких генов. Был описан ряд случаев специфического замещения TBP другим TBP-подобным белком из их семейства, дивергировавшего от общего предка (Akhtar, Veenstra, 2011).

Было установлено (van Werven *et al*., 2009), что TATA-бокс является единственным регулятором оборота TBP в клетке *in vivo*. В свою очередь, при повреждениях TATA-бокса ультрафиолетовым излучением, TBP вместе с циклином H и циклин-зависимой киназой 7 - субъединицы TFIIH (Lee, Young, 2000), - модулирует восстановление репарационной системой этого поврежденного TATA-бокса к его норме (Aboussekhra, Thoma, 1999). Это объясняет самую высокую консервативность TATA-бокса из всех известных в настоящее время сайтов ДНК для связывания транскрипционных факторов эукариот (Florquin *et al*., 2005; Yamamoto *et al*., 2007).

Как можно видеть из рассмотренного выше, структура и функции ТАТА-бокса были достаточно полно изучены, в то время как сведения о его эволюции остаются разрозненными и фрагментарными. Поэтому выборки его нуклеотидных последовательностей из геномов организмов, стоящих на разных ступенях эволюции, были использованы в качестве модельных данных для компьютерного выявления контекстно-зависимых характеристик, отражающих эволюционно-значимые особенности структуры и функции регуляторных сайтов в составе геномных ДНК.

2.1.2 Исследуемые последовательности ДНК ТАТА-боксов эу- и прокариот

Исследовали три выборки нуклеотидных последовательностей длиной 70 п.о. фрагментов ТАТА-содержащих промоторов дрожжей, беспозвоночных и позвоночных, взятых из базы данных EMBL Data Library, вып.42 (Rice *et al*., 1993), по ключевым словам “promoter” и “primary transcript”, Таблица 9. Примеры анализируемых ТАТА-содержащих ДНК представлены на Рисунке 29.

В качестве контроля аналогичным способом была получена выборка Прибнов-боксов (синоним “ТАТА-бокс”) *Escherichia coli*, охарактеризованная в Таблице 9. Она была взята в силу отдаленного функционального сходства между Прибнов- и ТАТА-боксами, например, по “закрытому” и “открытому” комплексам РНК-полимераз на них обоих. Недавний (Brindefalk *et al*., 2013) филогенетический анализ суперсемейства белков, содержащих TBP-подобный

Таблица 9 - Выборки анализируемых последовательностей ДНК

группа организмов	объем	длина	Старт ТАТА-бокса	Критерий ТАТА-бокса
позвоночные	486	70 п.о.	−30 п.о.	критерий ТАТА-бокса эукариот (Bucher, 1990)
беспозвоночные	158			
дрожжи	75			
Escherichia coli	135		−10 п.о.	Консенсус TATAAAA
Случайные ДНК	500		нет	нет

```
EMBL ID ......-30.......-20.......-10.........0........
HSTUBB2 agggaggg TATATAA gcgttggcggacggtcggttgtagcactctgc
GGAC01  ggcggccc TATAAAAa gcgaagcgcgcggcgggcgggagtcgctgcg
GGMYHE  tcagagcc TATAAAA ggaccttagggtcagtgtgtcttgtccttctt
...     . . . . . . . . . . . . . . . . . . . . . . . . .
HSAPOA2 acagg TATATA gccccttcctctccagccagggcaggcacagac
RNALBPR tagagaag TATATTA gagcgagtttctctgcacacagaccacctttc
```

Рисунок 29 - Примеры TATA-содержащих промоторов генов эукариот из числа входных данных для bDNAvideo (Ponomarenko J. *et al*., 1999)

домен, у бактерий, архей и эукариот подтвердил правомерность сравнения Прибнов- и TATA-боксов.

Наконец, поскольку наиболее часто используемые на тот момент методы консенсуса (Hawley, McClure, 1983) и позиционно-весовой матрицы (Mulligan *et al.,* 1984) были основаны на достоверном отличии сайтов связывания регуляторных белков от последовательностей равновероятных независимых случайных нуклеотидов, то каждая из четырех выборок TATA-боксов сравнивалась с выборкой из 500 таких случайных ДНК, охарактеризованных в последней строке Таблицы 9.

2.1.3 Прототип компьютерной системы bDNAvideo: анализ TATA-бокса

Каждую анализируемую последовательность ДНК $\{s_1 \ldots s_i \ldots s_{70}\}$ характеризовали оценкой среднеарифметического $X_{k[a;b]}$ значения k-ого из девяти конформационных свойств ($1 \le k \le 9$) гексануклеотидных дуплексов (Таблицы 10 и 11) на участке [a; b], $1 \le a < b < 65$, вычисляемых как:

$$X_{k[a;b]}(S = \{s_1 \ldots s_a \ldots s_i \ldots s_b \ldots s_{70}\}) = \\ = \sum_{i=a}^{b-5} X_k(s_i s_{i+1} s_{i+2} s_{i+3} s_{i+4} s_{i+5})/(b-a-5). \quad (29)$$

Всего для каждого из всех 4^6=4096 возможных гексануклеотидных гетеродуплексов ДНК в электронной библиотеке (Karas *et al*., 1996) было девять конформационных свойств (Таблицы 10 и 11): углы раскрытия по длинной tip и короткой inclination осям п.о., кручения twist и изгиба оси bend

Таблица 10 – Примеры оценок девяти конформационных свойств B-формы спирали ДНК для трех гексануклеотидных дуплексов, взятые из электронной библиотеки (Karas *et al.*, 1996)

Свойство спирали ДНК	шкала	AGGGAT	TTTTTT	AAGCCT
кручение спирали twist	градус	34.14	38.90	38.81
раскрытие tip по длинной оси п.о.		-2.44	0	0.59
раскрытие inclination по короткой оси п.о.		0.03	0	1.90
изгиб оси спирали bend		2.44	0	1.99
шаг спирали rise	Å	3.97	3.19	3.73
ширина малой бороздки		4.22	5.16	4.89
глубина малой бороздки		9.46	9.09	9.40
ширина большой бороздки		17.36	9.70	14.44
глубина большой бороздки		8.11	8.91	9.12

спирали ДНК; а также длина шага спирали rise, ширина и глубина ее большой и малой бороздок. Названия, шкалы и способы измерения свойств соответствуют номенклатуре (Dickerson *et al.*, 1989), представленной в главе 1 (Рисунок 17).

В Таблице 10 показан пример данных для трех из 4096 гексануклеотидов, в Таблице 11 представлены диапазоны значений свойств. Всего по формуле (29) было получено $9 \times 65 \times (65\text{-}1) / 2 = 18720 \approx 10^5$ вариантов $X_{k[ab]}$. Необходимость рассмотрения столь большого их количества была вызвана отсутствием какой-либо информации о влиянии спирали ДНК на функционирование TATA-боксов.

Поэтому идея прототипа bDNAvideo была в оценке достоверности дискриминации между природными TATA-боксами (любая из четырех выборок $\{S^+\}$ в верхней части Таблицы 9) и случайными ДНК (выборка $\{S^-\}$ в нижней строке Таблицы 9) посредством сравнения количественных величин $\{X_{k[ab]}(S^+)\}$ и $\{X_{k[ab]}(S^-)\}$, вычисленных для каждой выборки по формуле (29) для выявления лучших из них.

Таблица 11 – Характеристики конформационных свойств B-формы спирали ДНК для гексануклеотидных дуплексов из электронной библиотеки (Karas *et al*., 1996)

Свойство спирали ДНК	шкала	минимум	средн.±ст.откл	максимум
кручение, twist	градус	30.48	36.14±3.91	42.54
шаг, rise	Å	2.99	3.53±0.33	4.83
раскрытие по длинной оси п.о., tip	градус	-8.12	1.34±2.96	12.15
раскрытие по короткой оси п.о., inclination	градус	-4.31	0.0±1.44	4.31
изгиб оси, bend	градус	0.0	3.03±1.86	12.15
ширина малой бороздки	Å	3.81	5.18±0.76	7.50
глубина малой бороздки	Å	7.75	9.00±0.44	10.37
ширина большой бороздки	Å	8.44	13.51±1.63	18.52
глубина большой бороздки	Å	7.28	8.92±0.56	10.53

Для иллюстративного примера на Рисунке 30 показана автоматическая оценка достоверности α соответствия между выборочным $p(X_{k[ab]}(S^{@}))$ и нормальным $N[M_0(X_{k[ab]}(S^{@})); \delta(X_{k[ab]}(S^{@}))]$ распределениями со среднеарифметическим M_0 и стандартным отклонением δ (здесь и далее: $@ \in \{+;-\}$). Для этого, в рамках метода bootstrap (Efron, 1979), каждая из выборок $S^{@}$ разделялась случайным образом на две $S^{@\$}$ и $S^{@\#}$ неперекрывающиеся подвыборки 50% объема. По одной из них оценивали параметры M_0 и δ нормального распределения $N[M_0(X_{k[ab]}(S^{@\$})); \delta(X_{k[ab]}(S^{@\$}))]$, по другой независимой подвыборке $S^{@\#}$ - достоверность соответствия между этим нормальным $N[M_0(X_{k[ab]}(S^{@\$})); \delta(X_{k[ab]}(S^{@\$}))]$ и выборочным $p(X_{k[ab]}(S^{@\#}))$ распределениями по критерию χ^2. Кроме того, с помощью *t*-теста Стьюдента на тех же парах обучающих и контрольных подвыборок оценивалась достоверность α отличия между оценками среднеарифметических $M_0(X_{k[ab]}(S^{\$+}))$ и $M_0(X_{k[ab]}(S^{\#-}))$, а также $M_0(X_{k[ab]}(S^{\$-}))$ и $M_0(X_{k[ab]}(S^{\#+}))$. Наконец, по критериям Фишера и χ^2 была оценена достоверность α различия между контрольными подвыборками, $p(X_{k[ab]}(S^{\#+}))$ и $p(X_{k[ab]}(S^{\#-}))$, с помощью

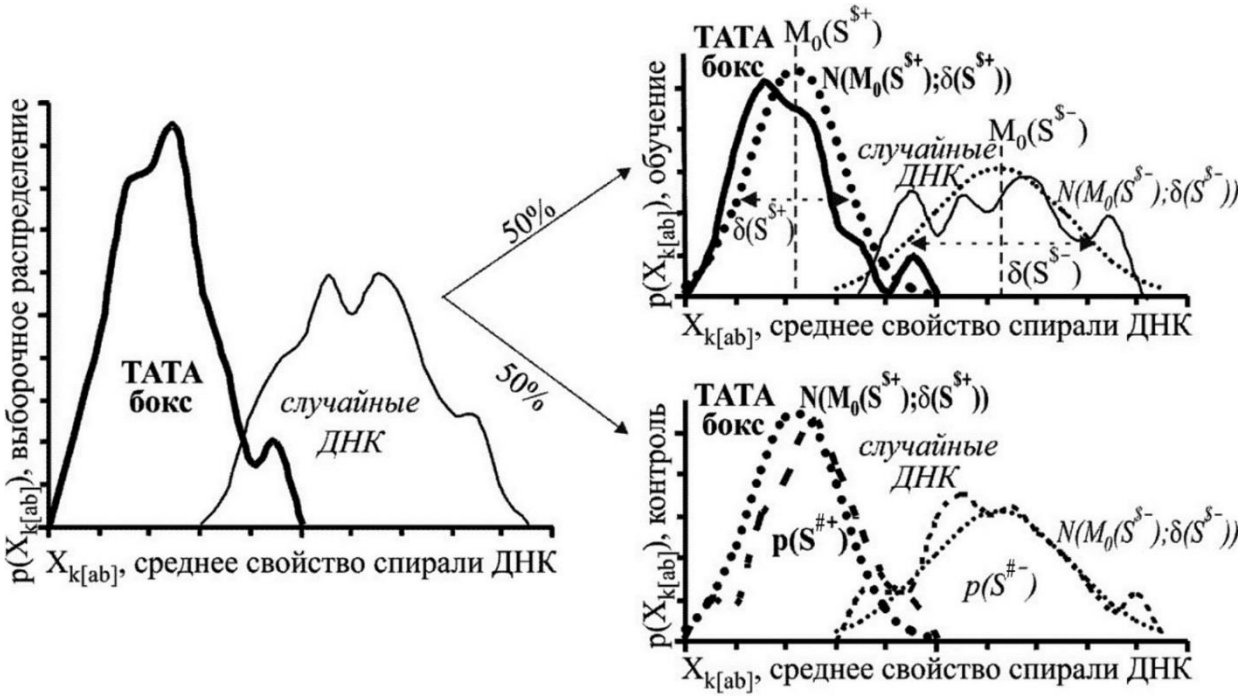

Рисунок 30 - Оценка достоверности соответствия между выборочными распределениями $p(X_{k[ab]}(S^+))$ и $p(X_{k[ab]}(S^-))$ и нормальными распределениями $N[M_0(X_{k[ab]}(S^+)); \delta(X_{k[ab]}(S^+))]$ и, $N[M_0(X_{k[ab]}(S^-)); \delta(X_{k[ab]}(S^-))]$. M_0 и δ - среднее и стандартное отклонение; $S^{@}$, $S^{@\$}$ и $S^{@\#}$ - выборка S^+ или S^- и ее случайное разбиение на неперекрывающиеся полувыборки в рамках метода bootstrap (Efron, 1979) для оценки $N[M_0(X_{k[ab]}(S^{@\$})); \delta(X_{k[ab]}(S^{@\$}))]$ по $S^{\$}$ и проверке критерием χ^2 его соответствия по выборочному распределению $p(X_{k[ab]}(S^{\#}))$.

общепринятого порога $p = (M_0(X_{k[ab]}(S^{\$+})) + M_0(X_{k[ab]}(S^{\$-}))) / 2$, который был установлен по обучающим подвыборкам.

Всего было проверено шесть перечисленных критериев в 100 независимых bootstrap-испытаниях. Это давало $6 \times 100 = 600$ оценок достоверности α_{qz} соответствия выборочных распределений $p(X_{k[ab]}(S^+))$ и $p(X_{k[ab]}(S^-))$ каждому q-ому критерию при z-ом их разбиении на 50%-обучение и 50%-контроль.

Соответственно, в рамках теории полезности для принятия решений (Fishburn, 1970) и нечетких множеств (Zadeh, 1965) качественные оценки

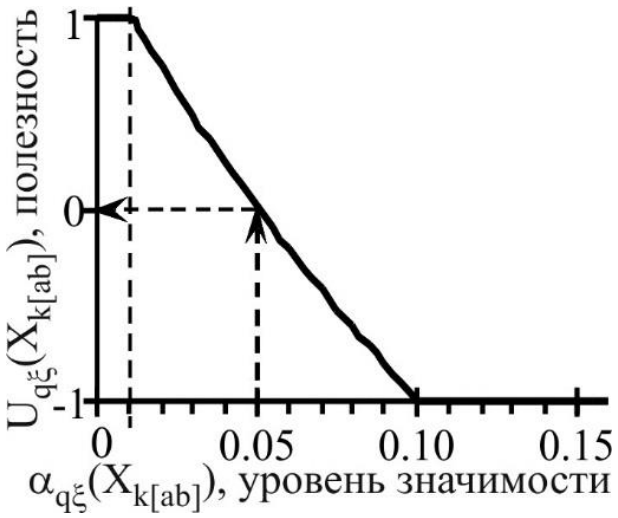

Рисунок 31 - Шкала полезности $U_{q\xi}(X_{k[ab]})$ среднеарифметической оценки $X_{k[ab]}$ для k–го свойства конформации на участке [a; b] спирали ДНК (Karas *et al*., 1996) для отличия ТАТА-боксов от случайных ДНК при достоверности $\alpha_{q\xi}(X_{k[ab]})$ q-го критерия в ξ-ом испытании, (формула 30).

достоверности α_{qz} были перемасштабированы к единой для всех них количественной шкале так называемой полезности:

$$U_{q\xi}(X_{k;a;b}) = \begin{cases} 1, & \text{ЕСЛИ} \quad \alpha_{q\xi} \le 0.01; \\ 1.3 - 28.3\alpha_{q\xi} + 55.6\alpha_{q\xi}^2, & \text{ЕСЛИ} \quad 0.01 \le \alpha_{q\xi} \le 0.1; \\ -1, & \text{ЕСЛИ} \quad 0.1 \le \alpha_{q\xi}; \end{cases} \qquad (30)$$

здесь: λ_0=1.3, λ_1=−28.3, λ_2=55.6 – коэффициенты квадратичного сплайна, проходящего через общепринятый порог достоверности {U_{qz}≡0 при α_{qz}≡0.05} и непрерывного в двух его крайних точках {U_{qz}≡−1 при α_{qz}≡0.1} и {$U_{qz}(X_k)$≡1 при $\alpha_{qz}(X_k)$≡0.01}, как это показано на Рисунке 31.

Шкала полезности (30) показана на Рисунке 31: позитивная полезность $U_{qz}(X_{kab}) > 0$ соответствует достоверному q-ому критерию в z-ом bootstrap-испытании, $\alpha_{qz}(X_{kab}) < 0.05$; негативная $U_{qz}(X_{kab}) < 0$ – недостоверному, $\alpha_{qz}(X_{kab})>0.05$, $U_{qz}(X_{kab})=0$ - общепринятому порогу достоверности, $\alpha_{qz}(X_{kab}) = 0.05$. Поэтому в рамках теории аддитивной полезности для принятия решений (Fishburn, 1970) интегральной оценкой

анализируемой количественной характеристики X_{kab} было среднеарифметическое всех 600 частных оценок ее полезности (μ=6; ν=100):

$$U\left(X_{k;[a;b]}\right) = \frac{1}{\mu\nu}\sum_{q=1}^{\mu}\sum_{\xi=1}^{\nu} U_{q\xi}\left(X_{k;[a;b]}\right). \quad (31)$$

Согласно теории аддитивной полезности для принятия решений (Fishburn, 1970), вычисленная по формуле (31) оценка $U(X_{k[ab]})$ обладает двумя асимптотическими свойствами:

ЕСЛИ $U(X_{k'[a'b']})\leq 0$, **ТО** $X_{k'[a'b']}$ бесполезна для отличия (32)

ЕСЛИ $U(X_{k[ab]})>U(X_{k'[a'b']})>0$, **ТО** $X_{k[ab]}$ полезнее $X_{k'[a'b']}$ (S^+) от (S^-). (33)

Свойство (32) позволяет исключить все $X_{k'[a'b']}$ с негативными $U(X_{k'[a'b']}) \leq 0$, тогда как свойство (33) позволяет выявить для каждого k-го из девяти конформационных свойств гексануклеотидных гетеродуплексов ДНК один вариант X^0_k с наибольшей позитивной полезностью $U(X^0_k)=MAX_{ab}\{U(X_{k[ab]})\}>0$. Их список $\{X^0_1,..., X^0_k,\ldots, X^0_{38}\}$ был результатом анализа природных TATA-боксов с помощью прототипа компьютерной системы bDNAvideo. Случай $\{X^0_k\} = Ø$ соответствовал отсутствию достоверных различий между природными TATA-боксами и случайными ДНК по среднеарифметическим конформационным свойствам гексануклеотидных гетеродуплексов B-формы спирали ДНК.

Существенно, что вследствие выбора $\{X^0_k\}$ из гигантского массива $\approx 10^5$ вариантов $X_{k[ab]}$ нельзя исключать вероятность случайности такого выбора, которую можно оценить. Согласно критерию Пуассона, в 600 испытаниях можно ожидать в качестве верхней 5%-ной доверительной границы ≈ 89 случайных позитивных исходов. Им соответствуют 600 – 89 = 511 негативных исходов, благодаря чему (89 << 511) формула (31) гарантирует негативную оценку $U(X_{k'[a'b']})\leq 0$. Для получения $U(X_{k[ab]})>0$ необходимо более 50% позитивных из 600 исходов при достоверности $\alpha < 0.05$. Биномиальное распределение оценивает вероятность такого события величиной:

$$P_{X_{k;[a;b]}}\left(U\left(X_{k;[a;b]}\right) > 0\right) = \\ = \sum_{q=1+600/2}^{600} C_{600}^{q} \times 0.05^{q}(1-0.05)^{600-q} < 10^{-45} \quad (34)$$

Тогда неравенство Бонферрони ограничивает искомую вероятность ее верхней оценкой:

$$P\left(U\left(X_{k;[a;b]}\right) > 0\right) < 10^{5} P_{X_{k;[a;b]}}\left(U\left(X_{k;[a;b]}\right) > 0\right) < 10^{5} \times 10^{-45} < 10^{-40}. \quad (35)$$

Это означает высокую степень достоверности ($\alpha < 10^{-45}$) взаимосвязи между выявленной с помощью формул (30 - 35) контекстной характеристикой $X_{k[a;b]}$ и проанализированной при этом "обучающей" парой контрастных выборок S^{+} и S^{-} последовательностей ДНК, которое, тем не менее, необходимо дополнительно подтвердить на независимых экспериментальных данных.

2.1.4 Результаты прототипа системы bDNAvideo в случае TATA-боксов

Для каждой пары выборок природных TATA-содержащих и случайных ДНК было проанализировано $\approx 10^5$ вариантов $X_{k[a;b]}$ ($1 \le k \le 9$; $1 \le a < b \le 65$). Полученные результаты (Пономаренко М. и др., 1997в; Ponomarenko M. *et al.*, 1997b) представлены в Таблице 12. Всего для TATA-содержащих ДНК позвоночных было выявлено четыре значимых конформационных отличия от случайных ДНК.

Прежде всего, наибольшую оценку $U(X_{2[-32;\ -22]}) = 0.887$ получил (формула 47) шаг спирали rise, усредненный по району от -32 п.о. до -22 п.о. относительно старта транскрипции. У промоторов позвоночных он оказался достоверно короче ($\alpha<10^{-40}$, критерий χ^2) шага B-формы спирали случайной ДНК, как это можно видеть на Рисунке 32. Кроме того, малая бороздка спирали ДНК района [-32; -24] промоторов позвоночных 5.80 ± 0.03 Å была шире, чем в случайной ДНК, 5.15 ± 0.05 Å (Рисунок 33а), а большая 11.98 ± 0.4 Å – более узкая, 13.37 ± 1.30 Å.

Таблица 12 – Оценки средних свойств гексануклеотидных шагов спирали ДНК (Karas *et al*., 1996) ТАТА-содержащих ДНК, найденных (Пономаренко М. и др., 1997в; Ponomarenko M. *et al*., 1997b) в качестве достоверных избытков (↑) и недостатков (↓) в сравнении со случайной ДНК

природные последовательности ДНК			прототип bDNAvideo				
группа организ-мов	свойство B-формы спирали ДНК	район [a, b]#	$U(X_{k[a;b]})$	среднее ± ст.ош.средн. ТАТА-бокс	среднее ± ст.ош.средн. случайные	↓↑	α
позво-ночные	шаг спирали rise	[-32;-22]	0.887	3.18±0.01	3.48±0.01	↓	10^{-40}
	кручение twist	[-34;-25]	0.632	34.48±0.11	36.02±0.17	↓	10^{-12}
	ширина малой бороздки	[-32;-24]	0.692	5.80±0.03	5.15±0.05	↑	10^{-13}
	ширина большой бороздки	[-32;-24]	0.671	11.98±0.04	13.37±011	↓	10^{-13}
беспо-звоно-чные	шаг спирали rise	[-35;-25]	0.837	3.18±0.01	3.49±0.01	↓	10^{-40}
	кручение twist	[-34;-25]	0.664	33.75±0.11	36.02±0.12	↓	10^{-14}
	ширина малой бороздки	[-34;-22]	0.834	6.06±0.04	5.16±0.04	↑	10^{-14}
	ширина большой бороздки	[-32;-20]	0.976	11.79±0.08	13.33±0.09	↓	10^{-13}
дрожжи	шаг спирали rise	[-51;-16]	0.979	3.39±0.01	3.50±0.01	↓	10^{-7}
	кручение twist	[-34;-25]	0.940	34.01±0.15	36.02±0.13	↓	10^{-7}
	ширина малой бороздки	[-34;-23]	0.857	5.98±0.06	5.15±0.04	↑	10^{-6}
	ширина большой бороздки	[-40;-21]	0.958	12.38±0.09	13.41±0.06	↓	10^{-7}
E. coli	шаг спирали rise	[-39;-8]	0.571	3.47±0.01	3.50±0.01	↓	10^{-8}
	кручение twist	[-46;-3]	0.637	36.36±0.07	36.05±0.03	↑	10^{-8}
	ширина малой бороздки	[-11;2]	0.562	5.36±0.07	5.14±0.03	↑	10^{-7}
	ширина большой бороздки	[-33;1]	0.644	12.98±0.09	13.41±0.04	↓	10^{-8}

Это согласуется с независимыми данными рентгеноструктурного анализа комплекса TBP/TATA (Kim J. *et al*., 1993; Kim Y. *et al*., 1993): ширина от 7.4 Å до 9.9 Å малой бороздки ДНК ТАТА-бокса достоверно шире малой бороздки идеальной спирали ДНК (Watson, Crick, 1953) от 4.2 Å до 6.7 Å и малой бороздки додекамеров 3.8 Å (Dickerson, Drew, 1981), как тест на независимых экспериментальных данных по рекомендациям к формуле (35).

Более широкая малая бороздка ДНК в случае ТАТА-бокса является важной для внедрения в нее боковых групп фенилаланинов TBP (Kim J. *et al*.,

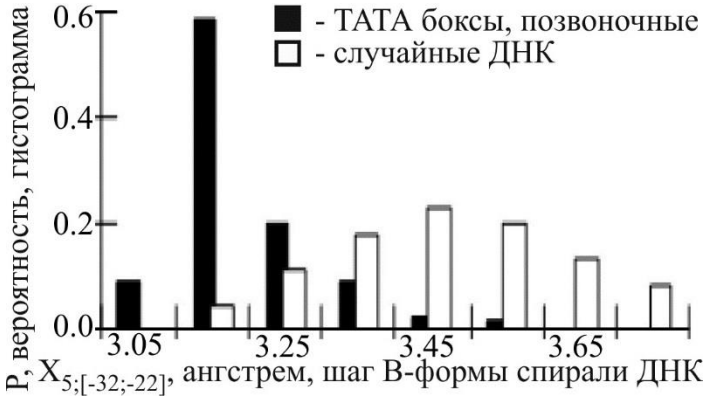

Рисунок 32 - Гистограмма $P(X_{2[-32;-22]})$ свойства "шаг rise"спирали гексануклеотидных дуплексов ДНК (Karas *et al*., 1996), усредненной на участке [-32; -22] для ТАТА-боксов позвоночных (■) и случайных ДНК (□).

1993; Kim Y. *et al*., 1993) при его специфическом связывании с ТАТА-боксом (Hahn *et al*., 1989). Соответственно, большая бороздка спирали ДНК в случае ТАТА-бокса, 11.98 ± 0.04^{O}, район [-32;-20] промоторов генов позвоночных, оказалась достоверно более узкой, чем в случайной ДНК, 13.37 ± 0.11^{O} (Таблица 12).Наконец, кручение спирали twist 34.48 ± 0.11^{O} района [-34;-25] промоторов позвоночных в окрестности ТАТА-бокса было достоверно меньше случайного 36.02 ± 0.12^{O}, $\alpha < 10^{-12}$ (Рисунок 34а).

Как можно видеть в Таблице 12, для ТАТА-содержащих промоторов беспозвоночных и дрожжей значимыми были те же четыре отличия конформации их спиралей ДНК от случайной ДНК: укороченный шаг, меньшее кручение, широкая малая и узкая большая бороздки спирали ДНК. Это согласуется с данными рентгеноструктурного анализа TBP/TATA-комплекса дрожжей (Kim Y. *et al*., 1993): кручение спирали ДНК не превышало 22.85^{O} и было достоверно меньше кручения 36^{O} идеальной спирали ДНК Уотсона-Крика (Watson, Crick, 1953). При этом, единственное отличие Прибнов-бокса *E. coli* от этой инвариантной картины всех ТАТА-содержащих ДНК было в большем кручении спирали ДНК, 36.36±0.07, чем таковое для случайной ДНК, 36.05±0.03, (Таблица 12 и Рисунок 34г).

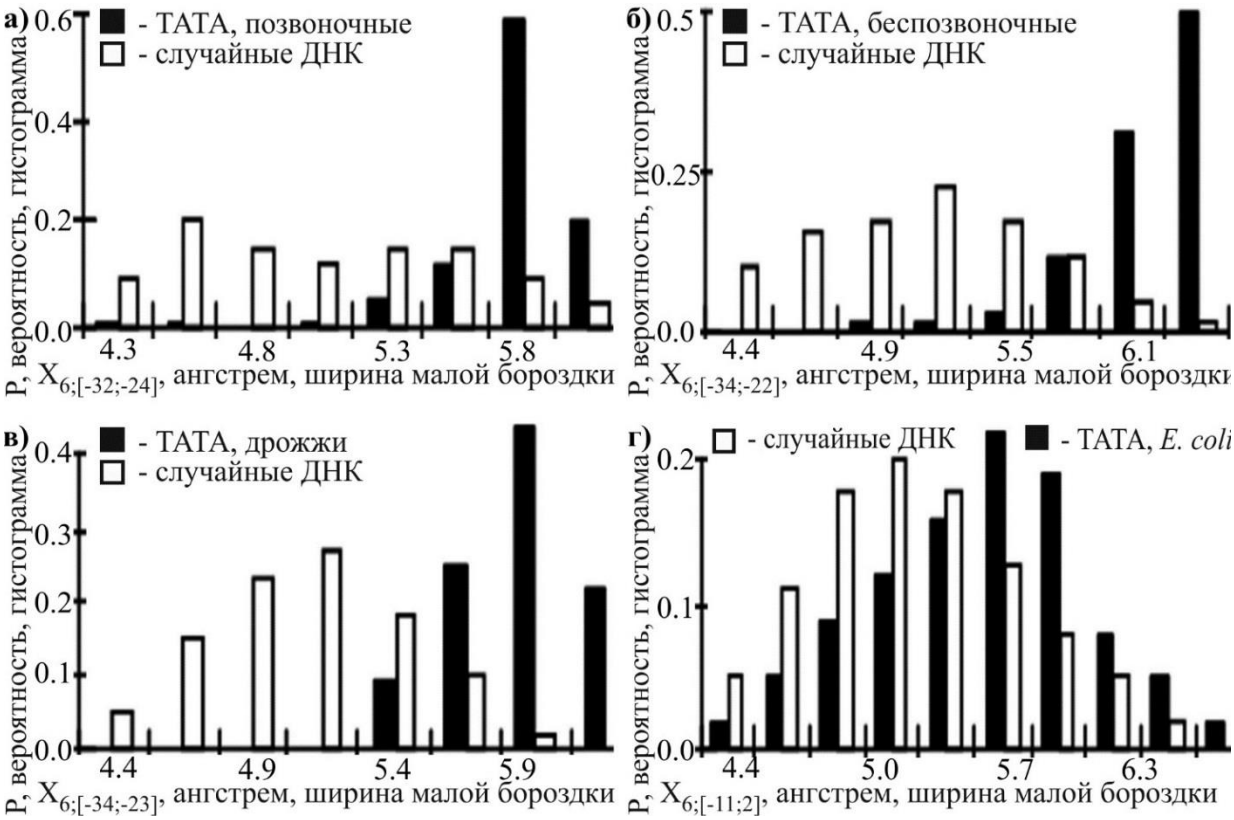

Рисунок 33 – Гистограммы контекстно-зависимых количественных оценок ширины малой бороздки спирали ДНК (Karas *et al*., 1996) случайных ДНК (□) и ТАТА-боксов (■) промоторов генов из групп организмов: а) позвоночные; б) беспозвоночные; в) дрожжи; г) *E. coli*.

В этой связи интересно отметить, что проведенный впоследствии вне рамок настоящей диссертации независимый анализ (Levitsky *et al*., 1999) 39 сайтов связывания октамеров гистонов нуклеосом из базы данных (Ioshikhes, Trifonov, 1993) идентифицировал большее кручение $36.50^{O} \pm 0.09^{O}$ спирали нуклеосомной ДНК, чем у случайных ДНК $36.05^{O} \pm 0.02^{O}$ (Рисунок 35: $\alpha<10^{-7}$, χ^2-тест). Как можно видеть, это является биологически значимым различием между ТАТА-боксами промоторов про- и эукариот, которые, действительно, отличаются одни от других отсутствием/наличием у них нуклеосомной ДНК.

Таким образом было обнаружено, что достоверно малое кручение спирали ДНК делает ТАТА-бокс промотора гена эукариот антагонистом большому кручению спирали нуклеосомной ДНК для снижения сродства октомера гистонов к ДНК кор-промоторов (Godde *et al*., 1995).

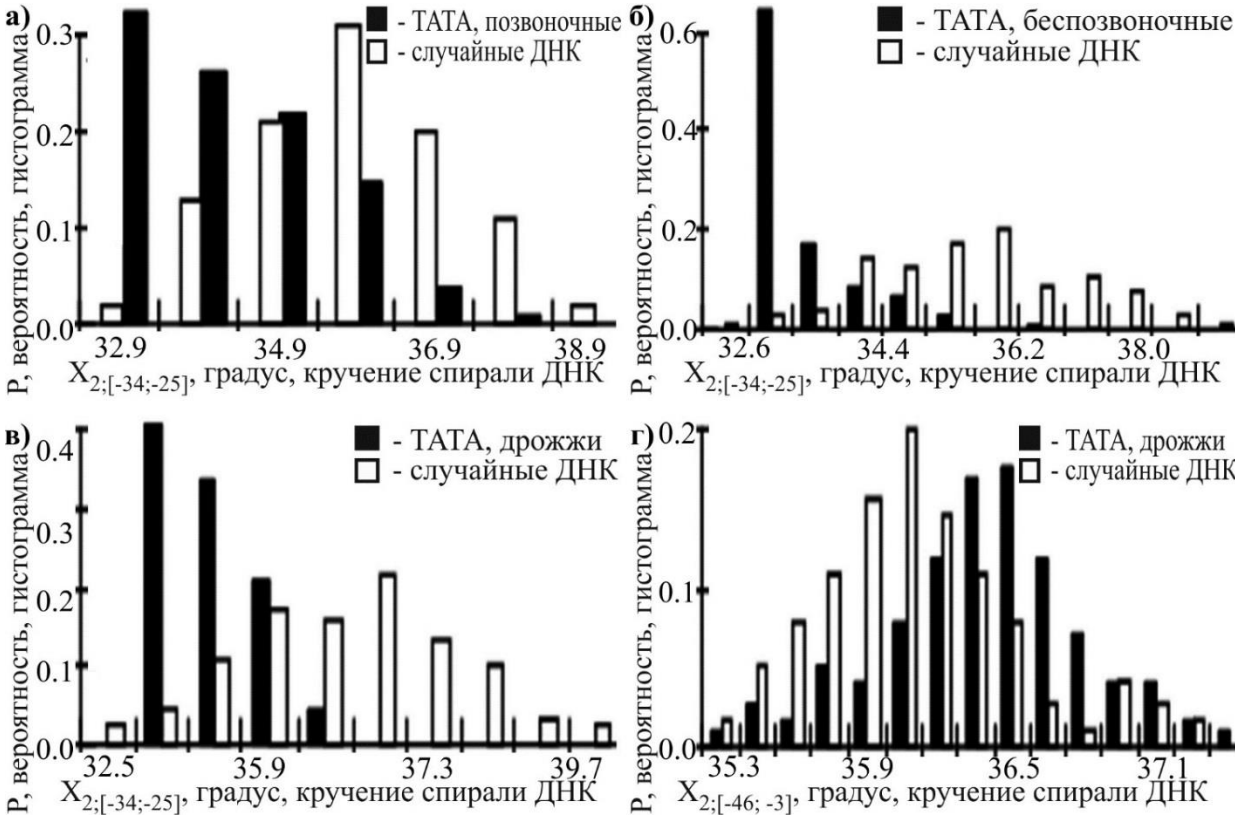

Рисунок 34 - Распределения кручения спирали ДНК (Karas *et al*., 1996) для случайных ДНК (□) и для промоторов (■): а) позвоночные, район [-34; -25]; б) беспозвоночные, [-34; -25]; в) дрожжи, [-34; -25]); г) *E. coli*, [-46; -3].

Кроме того, бактерии имеют гистон-подобные белки HU, которые являются гомологами гистонов эукариот и которые пакуют геномную ДНК бактерий в нуклеосом-подобные структуры (Киселева и др., 1986, 1988а) как и эволюционные гомологи HU-белков в митохондриях и хлоропластах растений (Киселева и др., 1988б; Salganik *et al*., 1990, 1991). Дефектные по белкам HU линии *E. coli* отличаются аномально высокой частотой повреждений ДНК ультрафиолетовым излучением (Li, Waters, 1998). Поэтому повышенное кручение ДНК у Прибнов-бокса *E. coli* и у нуклеосомной ДНК соответствует их повышенной защите от повреждений UV-излучением. Этот результат согласуется с консервативностью Прибнов-бокса в сравнении с другими районами промоторов *E. coli*. В свою очередь, пониженное кручение ДНК у ТАТА-боксов эукариот соответствует их меньшей защите от повреждений UV-излучением. Впоследствии было экспериментально

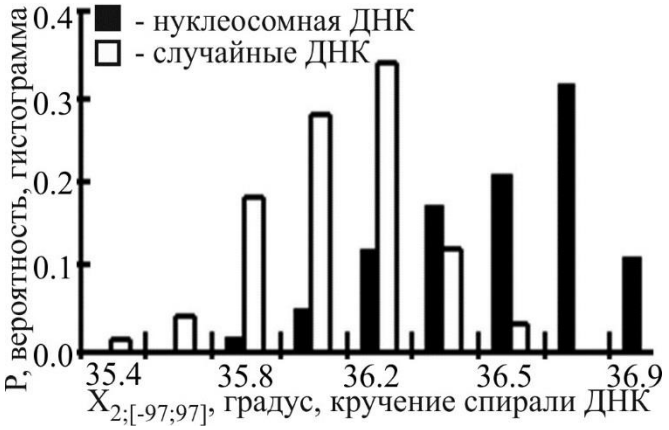

Рисунок 35 - Распределения кручения B-спирали ДНК (номенклатура, Dickerson, 1989) для нуклеосом (■) и для случайных ДНК (□), согласно независимой работе (Levitsky *et al*., 1999).

обнаружено (Aboussekhra, Thoma, 1999), что TBP и две субъединицы транскрипционного фактора TFIIH (циклин H и циклин-зависимая киназа 7, CDK7) модулируют работу репарационной системы для TATA-боксов эукариот в качестве их дополнительной защиты от UV-повреждений. Вероятно, благодаря "своей собственной" TBP-зависимой репарационной машине TATA-бокс вместе с нижележащим DPE, B-регуляторным и инициаторным INR элементами кор-промоторов оказались среди самых общих регуляторных сайтов в составе геномной ДНК всех эукариот (Florquin *et al*., 2005).На Рисунке 36 показана впервые обнаруженная (Ponomarenko M. *et al*., 1997b) достоверная линейная ($r= -0.71$, $\alpha<0.0025$) и ранговая Кендалла ($\tau= -0.67$; $\alpha<0.0005$) корреляция между эволюционным рангом и протяженностью L участка достоверного отличия TATA-боксов от случайных ДНК по конформационным свойствам гексануклеотидных шагов B-спирали ДНК (Karas *et al*., 1996).

Эта достоверная негативная корреляция соответствовала общепринятому представлению об усложнении транскрипционных машин в эволюционном ряду "*Escherichia coli* → дрожжи → беспозвоночные → позвоночные". На ее основе была

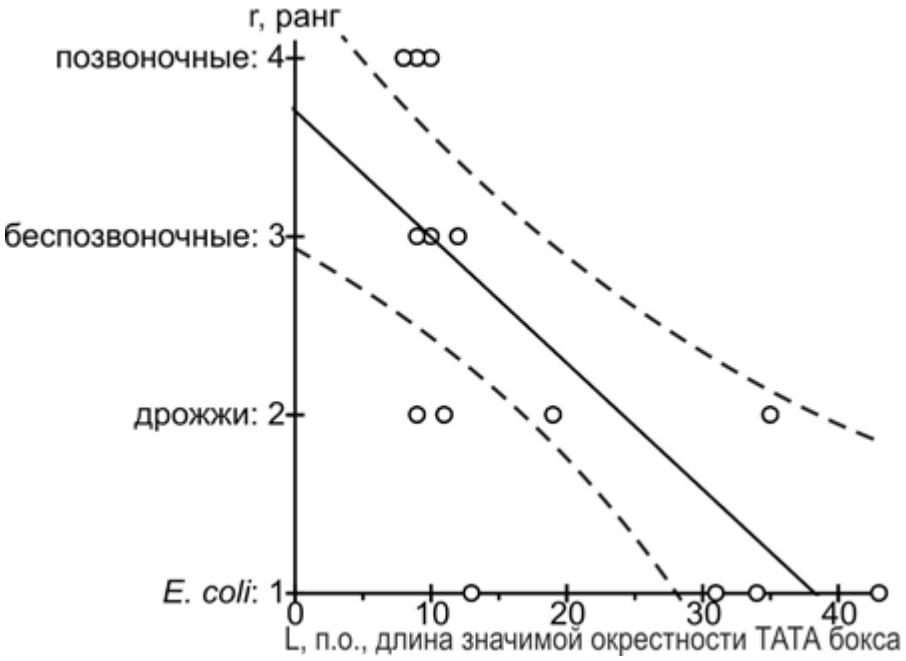

Рисунок 36 - Достоверная линейная ($r = -0.71$, $\alpha<0.0025$) и ранговая Кендалла ($\tau = -0.67$, $\alpha<0.0005$) корреляция между эволюционным рангом и длиной L участка значимых свойств спирали ДНК вокруг ТАТА-боксов (Ponomarenko M. *et al*., 1997b). Пунктир – границы 95%-доверительных интервалов (здесь и далее: построенных с использованием общепринятого пакета статистических программ Statistica (Statsoft^{TM}, Tulsa, USA)).

сформулирована гипотеза (Ponomarenko M. *et al*., 1997b), что эволюционное усложнение транскрипционных машин могло сопровождаться укорачиванием регуляторных районов геномной ДНК, вовлеченных в сборку на ТАТА-боксе анкерного комплекса, присоединение к которому РНК полимеразы II запускает формирование преинициаторного комплекса транскрипции на промоторах генов эукариот.

Таким образом, использование количественных характеристик регуляторных сайтов в составе геномных ДНК в форме среднеарифметических оценок конформационных свойств B-формы спирали ДНК позволило обнаружить неизвестную на момент выполнения настоящей диссертационной работы особенность эволюции ТАТА-боксов.

2.1.5 Верификация прототипа системы bDNAvideo для TATA-бокса

Поскольку эволюционная закономерность TATA-боксов (Рисунок 36) была обнаружена на основе использования результатов компьютерного моделирования молекулярной динамики ДНК (Lavery *et al*., 1981; Karas *et al*., 1996), то было важно проверить ее на экспериментальных данных.

С этой целью были собраны литературные данные (Kabsch *et al*., 1982; Shpigelman *et al*., 1993; Suzuki, Yagi, 1995) об экспериментальных оценках конформационных углов всех 16 возможных динуклеотидных шагов спирали ДНК (Таблица 13). Всего было 9 наборов пяти типов конформационных углов, характеризовавших свободные ДНК и комплексы ДНК/белок. Они были документированы в базе данных PROPERTY (Колчанов и др., 1998).

Для сравнения четырех выборок TATA-содержащих ДНК со случайными ДНК (Таблица 9) на основе конформационных углов, указанных в Таблице 13, была модифицирована только лишь формула (29):

Таблица 13 – Углы (в градусах) динуклеотидных шагов спирали ДНК для проверки корреляции между эволюционным рангом и протяженностью района значимых особенностей конформации спирали ДНК TATA-боксов (Рисунок 36).

Угол, номенклатура (Dickerson, 1989)	min	ср. ± ст.откл.	max	Литература
twist, кручение, все ДНК	27.7	33.93 ± 3.22	40	Kabsch *et al*., 1982
twist, кручение, свободные ДНК	29.5	34.72 ± 2.16	37.5	Suzuki, Yagi, 1995
twist, кручение, комплекс ДНК/белок	29	34.66 ± 2.50	39.5	Suzuki, Yagi, 1995
наклон direction, все ДНК	-154	1.25 ± 100.69	180	Shpigelman *et al*,1993
наклон wedge, все ДНК	0.9	4.40 ± 2.55	8.4	Shpigelman *et al*,1993
раскрытие roll, свободная ДНК	-3	2.81 ± 2.38	5	Suzuki, Yagi, 1995
раскрытие roll, комплекс ДНК/белок	-2.5	2.66 ± 2.61	6.5	Suzuki, Yagi, 1995
раскрытие tilt, свободная ДНК	-0.5	0.38 ± 0.54	1	Suzuki, Yagi, 1995
раскрытие tilt, комплекс ДНК/белок	0	0.69 ± 0.61	1.5	Suzuki, Yagi, 1995

$$X_{k[a;b]}(S = \{s_1 \dots s_a \dots s_i \dots s_b \dots s_{70}\}) = \sum_{i=a}^{b-1} X_k(s_i s_{i+1})/(b-a). \qquad (36)$$

При этом все остальные формулы (30 - 33) прототипа компьютерной системы bDNAvideo (Пономаренко М. и др., 1997в; Kolchanov *et al*., 1998) остались без изменений. В результате была создана пилотная версия компьютерной системы bDNAvideo (Ponomarenko J. *et al*., 1999a,b).

Выявленные с помощью компьютерной системы bDNAvideo (Ponomarenko J. *et al*., 1999a,b) особенности конформации динуклеотидных шагов спирали ДНК ТАТА-бокса показаны в Таблице 14 и на Рисунке 37. Для каждой из четырех групп организмов наибольшие позитивные оценки (формула 31) полезности U от 0.72 до 0.99 имело различие между ТАТА-содержащими и случайными ДНК по значимо меньшему в случае ТАТА-бокса

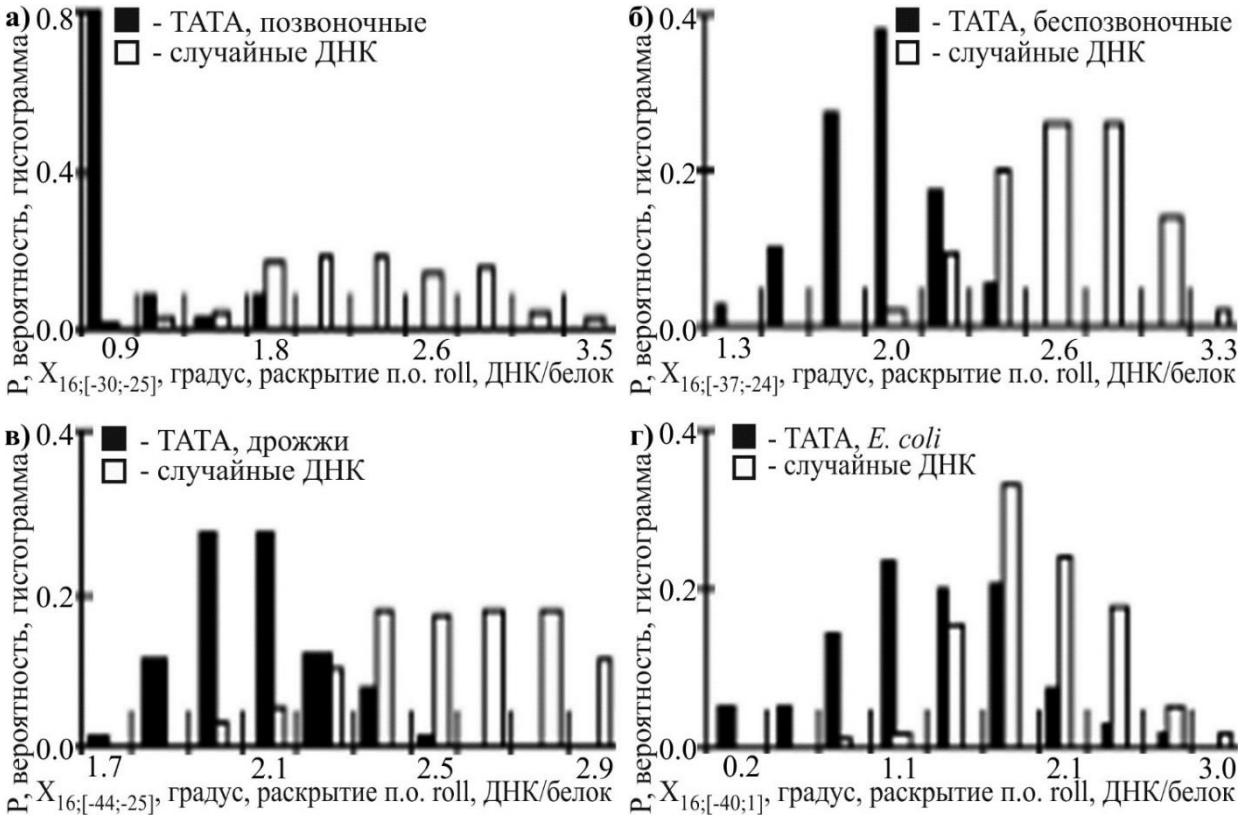

Рисунок 37 - Распределения значений раскрытия п.о. roll динуклеотидных шагов спирали ДНК в комплексах с белками (Suzuki, Yagi, 1995) для случайных (□) ДНК и для ТАТА-содержащих промоторов (■) генов (а) позвоночных, (б) беспозвоночных, (в) дрожжей и (г) *E. coli*.

Таблица 14 – Раскрытие roll динуклеотидного шага спирали ДНК в комплексе с белком (Suzuki, Yagi, 1995) было лучшим отличием ТАТА-содержащих промоторов от случайных ДНК.

группа организмов	район [a, b]#	bDNAvideo $U(X_{k[a;b]})$	среднее±ст.ош.средн.		α
			ТАТА-бокс	случайные ДНК	
позвоночные	[-30;-25]	0.99	0.88±0.02	2.22±0.05	10^{-40}
беспозвоночные	[-37;-24]	0.90	1.93±0.04	2.67±0.02	10^{-40}
дрожжи	[-44;-25]	0.93	1.97±0.06	2.53±0.02	10^{-7}
E. coli	[-40; 1]	0.72	2.35±0.04	2.60±0.01	10^{-7}

среднему углу раскрытия roll по короткой оси динуклеотидных шагов спирали ДНК в комплексе с белком (Suzuki, Yagi, 1995).

Это согласуется с независимыми данными рентгеноструктурного анализа (A+T)-богатых додекамеров (Dickerson, Drew, 1981), которые имеют слабый (19^{O}) изгиб оси, так называемую “бананообразную” форму B-спиралей ДНК (“banana form”, англ. яз., (Suzuki *et al*., 1996)).

На Рисунке 38 можно видеть достоверную ($r=-0.96$, $\alpha<0.05$ и $\tau=-1.00$, $\alpha<0.05$) корреляцию между эволюционным рангом и протяженностью участка усреднения угла roll, выявленного bDNAvideo (Ponomarenko J. *et al*., 1999a,b) в качестве лучшего различия между ТАТА-боксами и случайными ДНК. Эта корреляция (Рисунок 38) является независимым подтверждением в случае использования экспериментальных данных (Таблица 13) для эволюционной закономерности (Рисунок 36), найденной с использованием данных компьютерного моделирования молекулярной динамики (Таблицы 10 и 11).

В целом, обнаружение на основе новой методологии анализа контекстно-зависимых количественных характеристик ДНК, достоверно высокого у *E. coli* и достоверно низкого у эукариот кручения twist B-формы спирали ДНК ТАТА-боксов, а также корреляции между протяженностью биологически значимого участка усреднения свойств конформации спирали ДНК ТАТА-боксов и эволюционным рангом организмов (Пономаренко М и др., 1997б) показали возможность поиска неизвестных закономерностей

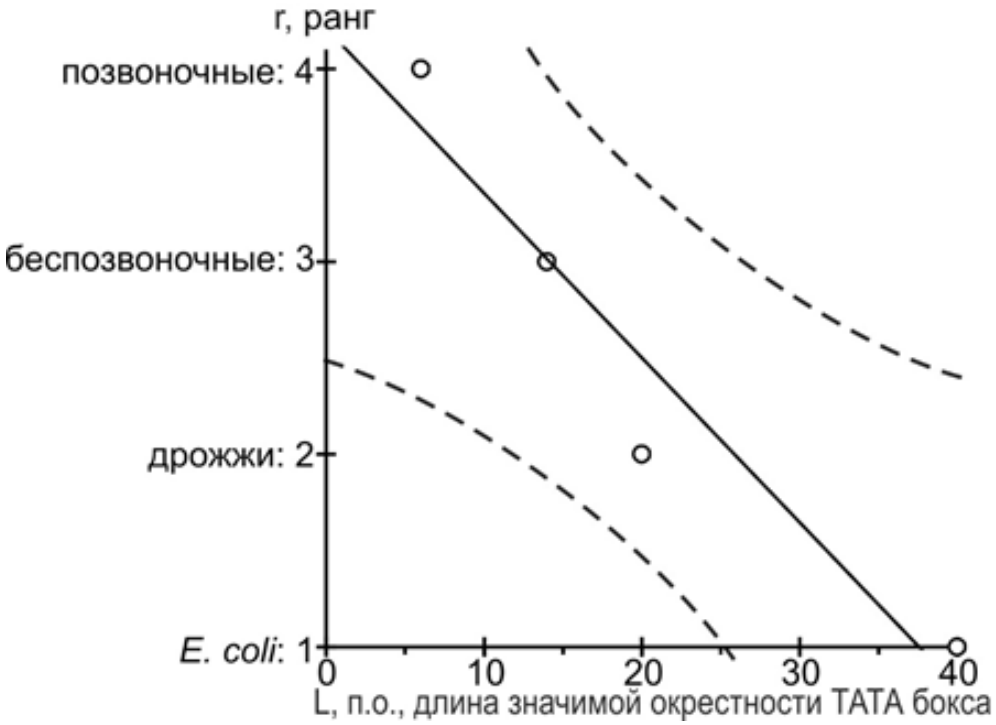

Рисунок 38 – Достоверная (r= –0.96, $\alpha<0.05$ и $\tau=-1.00$, $\alpha<0.05$) корреляция между эволюционным рангом и протяженностью участка усреднения раскрытия roll динуклеотидного шага спирали ДНК (Suzuki, Yagi, 1995). Пунктир – границы 95%-доверительных интервалов.

структурно-функциональной организации и эволюции геномов, адекватно соответствующих общепринятым представлениям наук о жизни.

2.1.6 Распознавание TATA-боксов путем усреднения значимых контекстно-зависимых количественных характеристик промоторов генов эукариот

Следующим шагом создания компьютерной системы bDNAvideo стала база данных PROPERTY (Колчанов и др., 1998) о физико-химических и конформационных свойствах динуклеотидных шагов спирали ДНК. Всего было собрано 38 таких свойств с экспериментально измеренными или теоретически оцененными величинами для всех 16 возможных динуклеотидов. Их примеры показаны в Таблице 15. В сравнении с девятью

Таблица 15 – Примеры свойств динуклеотидных шагов спирали ДНК, собранных в базе данных PROPERTY (Колчанов и др., 1998), которая была создана в рамках этой диссертации.

Свойства спирали ДНК	Единицы	Диапазон		Литература
		min	max	
Физико-химические свойства:				
Температура плавления, T_M	^{O}C	36.7	136.1	(Gotoh, Tagashira, 1981)
Персистентная длина	п.о.	20	130	(Hogan, Austin, 1987)
Частота контакта с гистонами	%	1	18	(Satchwell *et al*., 1986)
Гибкость по малой бороздке	ln-ед.	1.02	1.27	(Gartenberg, Crothers, 1988)
Гибкость по бол. бороздке		0.99	1.18	
Энтропия ΔS	cal/mol/K	-28.4	-15.2	(Sugimoto *et al*., 1996)
Энтальпия ΔH	Kcal/mol	-11.8	-5.6	
Свобод. энергии Гиббса, ΔG		-2.8	-0.9	
Конформационные свойства:				
Кручение *twist*	градус	27.7	40.0	(Shpigelman *et al*, 1993)
Перекрест *propeller* п.о.		-17.3	-6.7	(Gorin *et al*., 1995)
Наклон *tip* короткой оси		-1.64	6.7	(Karas *et al*., 1996)
Наклон *inclination*, длин. ось		-1.43	1.43	
Изгиб *bend* оси		2.16	6.74	
Раскрытие *tilt*, короткая ось		-0.7	2.8	(Suzuki, Yagi, 1995)
Раскрытие *roll* по длинной оси		-2.0	6.5	
Угол Эйлера *wedge* поворота		1.1	8.4	(Shpigelman *et al*., 1993)
Угол Эйлера *direction*		-154	180	
Сдвиг *slide* по длинной оси	ангстрем	-0.37	1.46	(Gorin *et al*., 1995)
Шаг *rise* спирали вдоль оси		3.16	4.08	(Karas *et al*., 1996)
Ширина малой бороздки		4.62	6.40	
Глубина малой бороздки		8.79	9.11	
Ширина большой бороздки		12.1	15.5	
Глубина большой бороздки		8.45	9.60	

свойствами гексануклеотидов для прототипа bDNAvideo (Таблицы 10 и 11) и пятью типами углов спирали ДНК для его верификации (Таблица 13) это существенно усилило информативность (Таблица 15) финальной версии bDNAvideo.

С ее помощью были исследованы 500 TATA-содержащих промоторов негомологичных генов эукариот из базы данных EPD (Perier *et al*., 1999). В результате было выявлено (Ponomarenko J. *et al*., 1999a) четыре контекстно-

Таблица 16 – Результат системы bDNAvideo для сравнения 500 TATA-содержащих промоторов эукариот с из базы данных (Perier *et al*., 1999) с 500 случайными последовательностями ДНК

Среднеарифметическое свойства спирали ДНК, X_{kab} (формула 37)				U ф-ла	TATA-бокс	Случайные ДНК
k	Свойство, *обозначение*	Ед., ф-а	[a; b]	(31)	ср.±ст.ош.ср	ср.±ст.ош.ср
24	Гибкость малой бороздки	ln-ед.	[−8; 6]	0.924	1.101±0.002	1.138±0.002
3	Изгиб оси спирали, *bend*	градус	[−8; 4]	0.874	3.47 ±0.03	3.00±0.04
4	Наклон к короткой оси, *tip*	градус	−10; 7	0.818	2.05±0.04	1.33±0.04
22	Температура плавления, T_M	^{O}C	−10; 6	0.766	65.87±0.52	73.95±0.50
Среднее отличие		(37)	−10; 7		0.93±0.10	−1.03±0.10

зависимых свойства спирали ДНК, достоверно дискриминирующих TATA-боксы от случайной ДНК (Таблица 16, Рисунок 39). Наибольшую оценку U=0.924 (формула 31) получила достоверно ($\alpha<0.0005$) низкая жесткость изгиба спирали ДНК (Gartenberg, Crothers, 1988) от -38 до -24 позиции относительно старта транскрипции в сравнении с жесткостью случайной ДНК (Рисунок 39а). Этот результат bDNAvideo соответствует известным пространственным структурам TBP/TATA-комплекса (Kim J. *et al*., 1993; Kim Y. *et al*., 1993; Juo *et al*., 1996), где ось двойной спирали ДНК, действительно, согнута под углом 90^{O}.

Кроме того, в качестве значимых для B-спирали ДНК TATA-бокса были найдены большие изгиб оси bend (U=0.874, $\alpha<0.0005$) и угол раскрытия tip (U=0.817, $\alpha<0.0005$), соответствующие указанным 3D-структурам TBP/TATA (Kim J. *et al*., 1993; Kim Y. *et al*., 1993; Juo *et al*., 1996). Наконец, достоверной количественной характеристикой TATA-боксов была также низкая температура плавления ДНК (U=0.766, $\alpha<0.0005$). Это согласуется с независимыми выводами (Flatters, Lavery, 1998) модели молекулярной динамики додекамеров $d(GCGTATATAAAACGC)_2$ и $d(GCGTAAAAAAAACGC)_2$ о высокой эластичности этих спиралей ДНК, которой, тем не менее, было недостаточно для изменения изгиба оси от 19° в

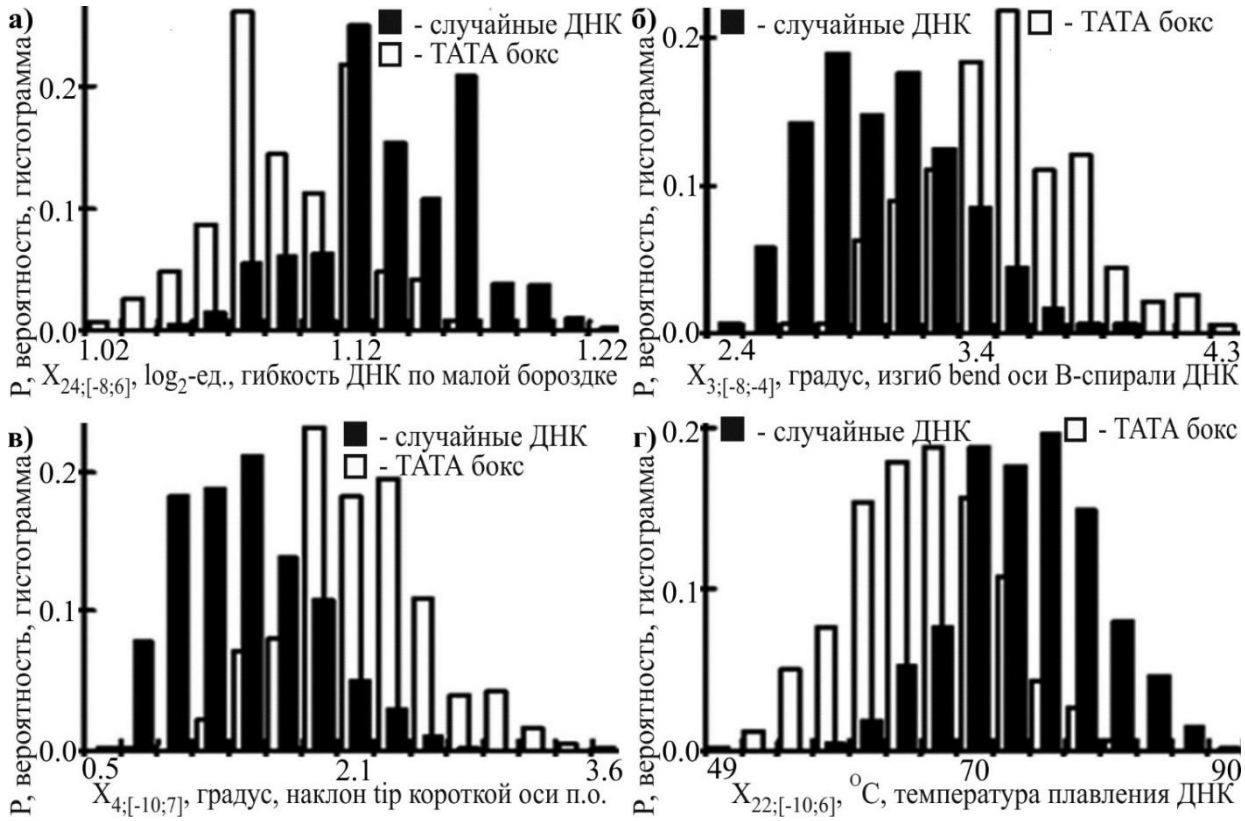

Рисунок 39 - Распределения свойств спирали ДНК для TATA-содержащих промоторов из базы данных EPD (Perier *et al*., 1999) и для случайных ДНК.

свободной ДНК до 90° в комплексе с TBP (Kim J. *et al*., 1993; Kim Y. *et al*., 1993; Juo *et al*., 1996) без ее частичной денатурации, что было впоследствии подтверждено экспериментально методом флюориметрии (Powell *et al*, 2002).

В рамках теории аддитивной полезности для принятия решений (Fishburn, 1970) все найденные отличия между TATA-боксами и случайными ДНК были приведены к общей шкале со значениями “+1” и “−1”, соответствующими среднеарифметическим оценкам для этих классов ДНК, а также со значением “0” в качестве порога между ними, и, затем, они были усреднены:

$$X(S) = \frac{\sum_{k=1}^{38} \Delta(X_k^0 \neq \emptyset) \dfrac{X_k^0(S) - (M_0(X_k^0(S_n^+) + M_0(X_k^0(S_n^-)/2)}{(M_0(X_k^0(S_n^+) - M_0(X_k^0(S_n^-)/2)}}{\sum_{k=1}^{38} \Delta(X_k^0 \neq \emptyset)}. \quad (37)$$

Результат формулы (37) можно видеть в последней строке Таблицы 16. На Рисунке 40 показано компьютерное распознавание TATA-боксов с

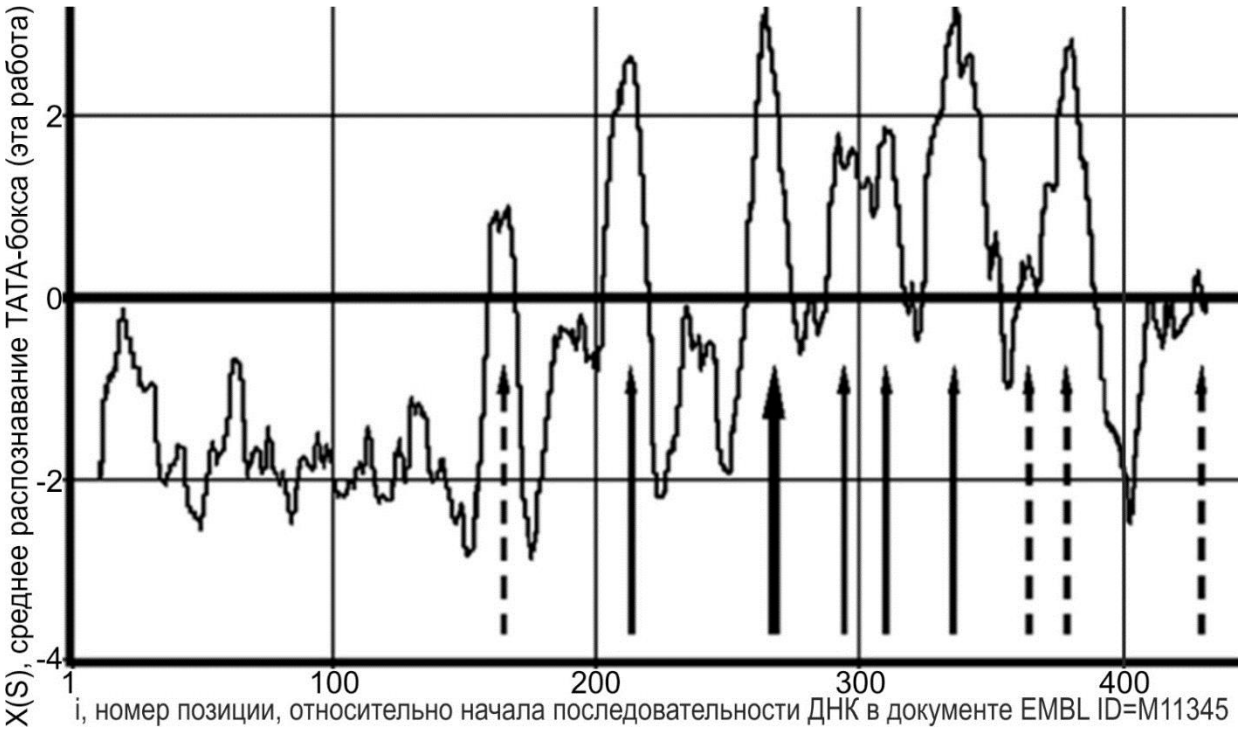

Рисунок 40 - Распознавание ТАТА-бокса путем усреднения (формула 37) значимых свойств его спирали ДНК (Таблица 16 и Рисунок 39) на примере промотора гена CYC1 дрожжей (изо-1-цитохром С, EMBL ID=M11345). Стрелки: жирная и тонкие – экспериментально доказанные ТАТА-боксы главного и минорных стартов транскрипции; пунктир – ошибки II рода (перепредсказание).

помощью формулы (37) на основе четырех свойств конформации В-формы их спиралей ДНК (ось у) в последовательности ДНК длиной 450 п.о. промотора гена CYC1 дрожжей для изо-1-цитохрома С (EMBL ID=M11345). Жирная и тонкие стрелки указывают на экспериментально доказанные ТАТА-боксы главного и минорных стартов транскрипции, соответственно; пунктирные стрелки – ошибки II рода, ТАТА-подобные боксы.

Для сравнения, единственным предшествующим “физико-химическим и конформационным” прогнозом ТАТА-боксов в этой последовательности ДНК было сообщение авторов работы (Karas *et al*, 1996), что все пять доказанных ТАТА-боксов (сплошные стрелки) оказались в числе 28 наибольших пиков

ширины малой бороздки спирали ДНК, выбор которой был эвристическим на основе известных TBP/TATA-комплексов (Kim J. *et al*., 1993; Kim Y. *et al*., 1993). Это означало, что созданная в диссертации компьютерная система bDNAvideo втрое увеличила точность “физико-химического и конформационного” распознавания TATA-боксов (Рисунок 40) относительно того уровня (Karas *et al*., 1996), который был достигнут в мире на момент создания этой системы.

Таким образом с помощью bDNAvideo (формулы 29 - 37) были исследованы TATA-содержащие промоторы генов эукариот. В результате были обнаружены биологически значимые особенности температуры плавления, кручения, раскрытия, гибкости, ширины большой и малой бороздок, наклона и изгиба оси их спирали ДНК. Они соответствовали 3D-структурам, моделям молекулярной динамики додекамеров и TBP/TATA-комплекса, а также достоверно распознавали потенциальные TATA-боксы в последовательностях ДНК промоторов генов эукариот (Рисунок 40).

2.2 Количественные характеристики спирали ДНК сайтов связывания транскрипционных факторов эукариот

2.2.1 Суперклассы транскрипционных факторов (введение)

Сайты связывания на геномной ДНК для каждого из тысяч транскрипционных факторов искали чаще всего в опытах, получивших наиболее общее название “футпринтинг” (от “footprinting”, англ. яз., “отпечаток”), модификации (Galas, Schmitz, 1978) секвенирования ДНК по Максаму-Гильберту (Maxam, Gilbert, 1977). Пример результата этого метода для транскрипционного фактора EN показан на Рисунке 41. На этом рисунке дорожки № 1- 5 соответствуют концентрациям 0 пг, 40 пг, 120 пг, 225 пг, 400 пг белка EN при 20 μг/мл ДНКазы I, которая в присутствии Mg^{2+} разрезает обе нити ДНК одна против другой равновероятно по их длине. Можно видеть пять

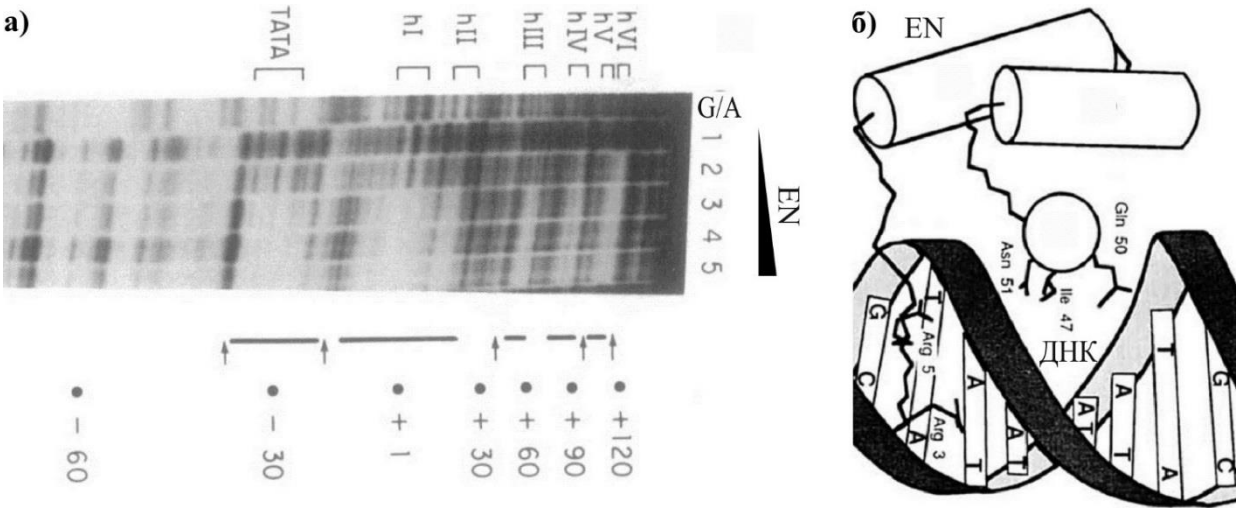

Рисунок 41 – Пример (а) пяти сайтов связывания транскрипционного фактора EN (**en**grailed - англ. яз.) в позициях -30, +1, +60, +90 и +120 промотора гена hsp 70 дрозофилы (Ohkuma *et al*., 1990), которые были идентифицированы методом футпринтинга (от “footprinting”, англ. яз.), и (б) ДНК/EN-комплекс (Kissinger *et al*., 1990). Дорожки: G/A – реакция “G>A” (Mirzabekov, Melnikova, 1974); 1, 2, 3, 4, 5 – концентрации 0 пг, 40 пг, 120 пг, 225 пг и 400 пг EN при 20 μг/мл ДНКазы I; ↑ – границы сайтов связывания EN на геномной ДНК. Рисунок автора на основе иллюстраций из статей (Ohkuma *et al*., 1990) и (Kissinger *et al*., 1990).

сайтов связывания транскрипционного фактора EN, где интенсивность полос убывает с ростом концентрации EN вследствие экранирования ДНК в комплексе с EN (Рисунок 41б) от разрезания ДНКазой I. Жирные и тонкие полосы дорожки “G/A” соответствуют нуклеотидам G и A определенной нити ДНК (Mirzabekov, Melnikova, 1974).

Пример выборки сайтов связывания транскрипционного фактора EN можно видеть в Таблице 16. Экспериментальные сайты связывания транскрипционных факторов (Рисунок 41) были документированы во многих базах данных, например, в TRRD (Kolchanov *et al*., 2002). На основе множественного выравнивания (Lawrence *et al*., 1993) таких сайтов были построены позиционно-весовые матрицы и консенсусы (Hawley, McClure, 1983) сайтов связывания транскрипционных факторов, документированные,

Таблица 16 - Пример 12 сайтов связывания транскрипционного фактора EN, документированных в базе данных SAMPLE (Ponomarenko M. *et al*., 1999b)

GenBank	Последовательность ДНК сайта связывания EN (заглавным шрифт)
V00213	cgaaaagagcgccggagTATAAATAGAGGCGCttcgtcgacggagcgtca
M11072	cgaaaagagcgccggagTATAAATAGAGGCGCttcgtcgacggagcgtga
V00219	cgaaaagagcgcagcagTATAAATAGAGGCGCttcgtctacggagcgaca
X05427	acattcgttcgatgGCAACGGATTGGATAACAGGCgcgcgctttgtttta
M29285	gcaaataaataaattaaatgTCAATTAAATatcaatcaatttcgtcagct
M29285	gctgtttttcaaggcACATTTAACTGGTTAATTGAaggcctcaaaaataa
X01765	ggtgtcccgtccgtacttaaCCAATTAGCCacgctcggccgaaaccgcaa
X04727	tcaaatcatctaagcaatcGAGCAATTAAATtataatttacaatgtgtcg
X04727	tagctagagaaagcccctggTCAATTAGCTaaatatcgtactaagcagcc
X02996	cgggtgttcctgaaggggggcTATAAAAgggggtgggggcgcgttcgtcc
X03000	caggggtccccgccgggggggTATAAAAgggggcggacctctgttcgtcc
J01917	cgggtgttcctgaaggggggcTATAAAAgggggtgggggcgcgttcgtcc

например, в базе данных TRANSFAC (Heinemeyer *et al*., 1999), с целью распознавания этих сайтов в произвольных ДНК. В рамках настоящей диссертации bDNAvideo использовали для анализа сайтов связывания транскрипционных факторов без выравнивания их последовательностей, чтобы минимизировать вклад символьного представления в биологически значимые количественные характеристики регуляторных ДНК.

Существенно, что к началу выполнения этой части диссертационной работы тысячи транскрипционных факторов оказались сгруппированными по сходству их 3D-структур и молекулярных механизмов их связывания с ДНК всего лишь в четыре суперкласса (Рисунок 42): основный, Zn-координированный, гомеодомен и β-слой (Вингендер, 1997). Однако, не было найдено какого-либо сходства между сайтами связывания транскрипционных факторов в рамках каждого из этих суперклассов ни по консенсусам, ни по позиционно-весовым матрицам, ни по каким-то иным особенностям нуклеотидного контекста ДНК. Поэтому в процессе выполнения диссертационной работы с использованием описанной в предыдущем разделе системы bDNAvideo (Пономаренко М. и др., 1997в) были проанализированы выборки сайтов связывания для 42 транскрипционных факторов,

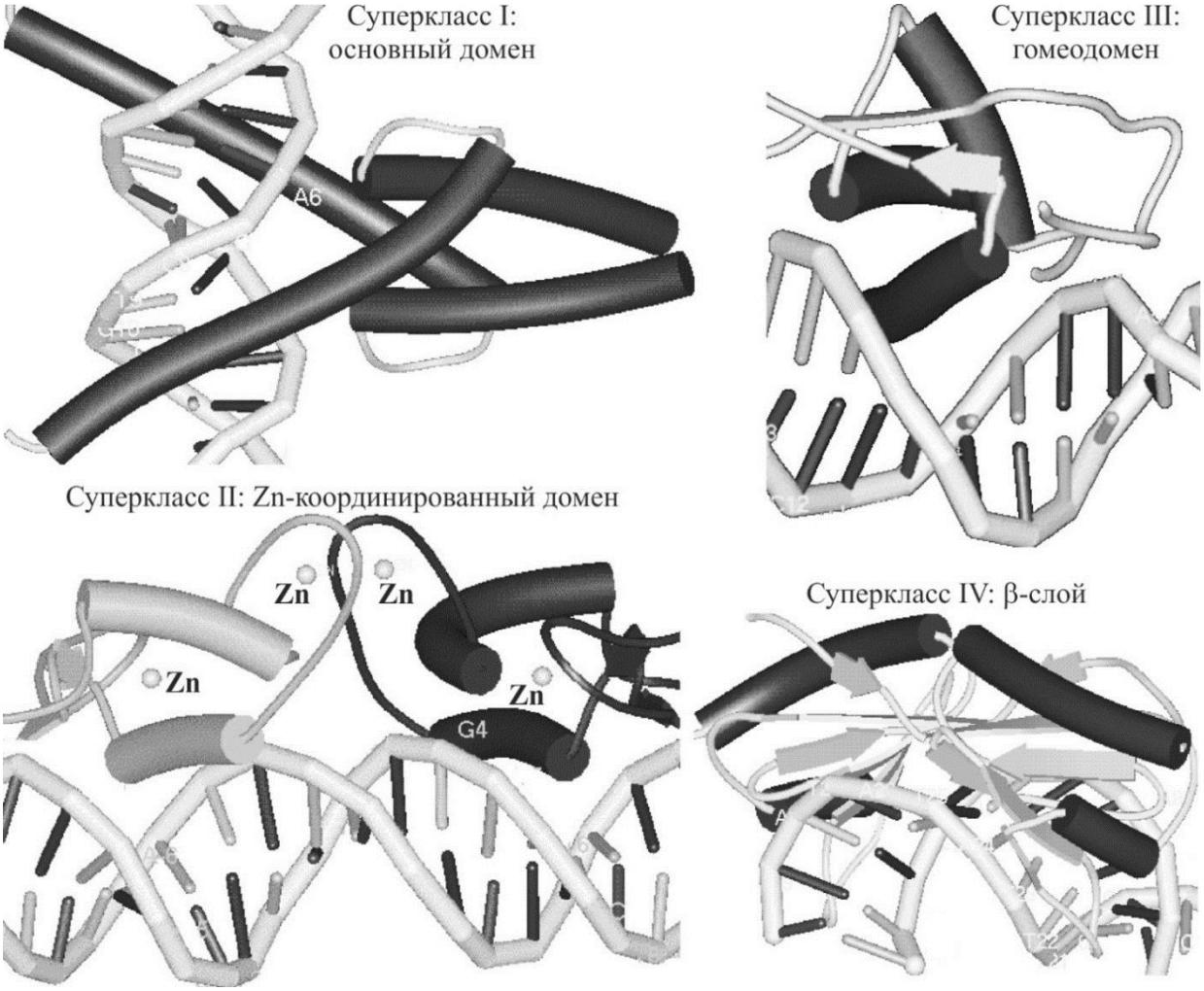

Рисунок 42 - Суперклассы транскрипционных факторов (Вингендер, 1997).

представлявших все четыре суперкласса, в сравнении с 500 случайными ДНК такой же длины.

Рассмотрим применение системы bDNAvideo на примере исследования сайтов связывания транскрипционного фактора EN (Таблица 16).

2.2.2 Компьютерный анализ конформационных и физико-химических свойств спирали ДНК на примере сайтов связывания транскрипционного фактора EN

Результат bDNAvideo (Пономаренко М. и др., 1997в) в случае сравнения 12 сайтов связывания транскрипционного фактора EN (Таблица 16) с 500 случайными последовательностями равновероятных независимых

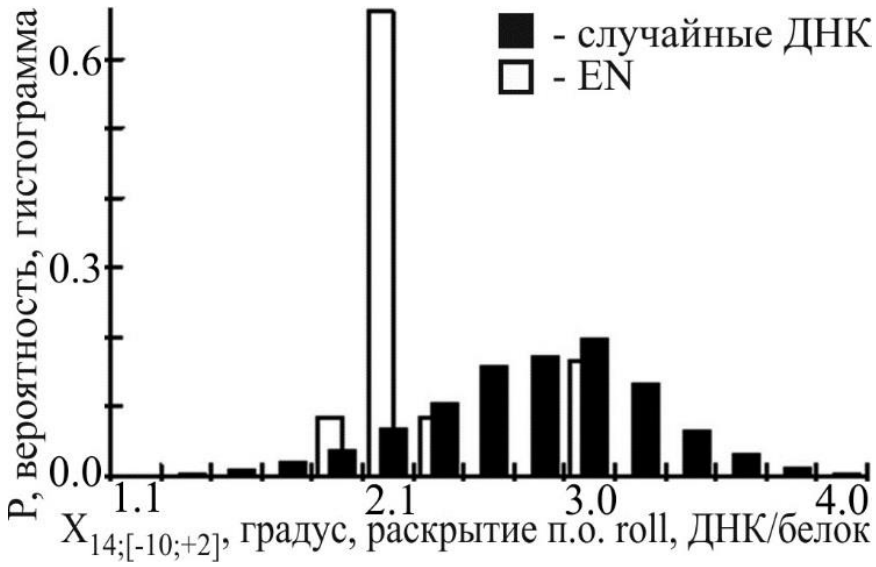

Рисунок 43 – Результат bDNAvideo для сравнения сайтов связывания EN (□) со случайными ДНК (■): высшую оценку $U(X_{14,-10,2})=0.989$ получило среднее значение "угла *roll* в комплексе белок/ДНК" (Suzuki *et al*., 1996) на участке [−10; 2] от центра сайта связывания EN, установленного методом футпринтинга.

нуклеотидов ДНК такой же длины можно видеть на Рисунке 43 и в Таблице 17. Самую большую полезность $U(X_{14,-10,2})=0.989$ получил достоверно низкий среднеарифметический угол раскрытия roll пары соседних оснований по их короткой оси в комплексе белок/ДНК (Suzuki *et al*., 1996) (Рисунок 43) между позициями −10, и 2 относительно центра сайта связывания EN, установленного методом футпринтинга. Он был 2.26 ± 0.20^{O} для сайтов связывания EN достоверно ($\chi^2=172.11$, $\alpha<0.0005$) меньше, чем в случайных ДНК, 2.72 ± 0.04^{O}, в согласии с 3D-структурой (Рисунок 41б) комплекса ДНК/EN (Kissinger *et al*, 1990).

Всего было выявлено 10 конформационных и физико-химических свойств спирали ДНК, средние оценки которых в районе ±20 п.о. от центра сайта связывания EN, установленного методом футпринтинга, которые достоверно дискриминировали эти сайты от случайных ДНК (Таблица 17). В последней строке Таблицы 17 можно видеть параметры метода для

Таблица 17 – Результат сравнения сайтов связывания EN и случайных ДНК.

Среднеарифметическое значение конформационного свойства спирали ДНК, X_{kab} (формула 38)				U, фор-мула	среднее ± ст.ош.средн.	
k	Название свойства	Ед/	[a; b]	(31)	EN сайт	случайные
14	Раскрытие длинной оси, roll	градус	[−10; 2]	0.989	2.26±0.20	2.72±0.04
22	Температура плавления, T_M	^OC	[−10; 3]	0.886	64.10±3.33	73.43±0.59
24	Гибкость большой бороздки	ln-ед.	[−9; 4]	0.885	1.07±0.01	1.05±0.002
4	Наклон по короткой оси, tip	градус	[−13; 5]	0.809	1.85±0.24	1.34±0.04
3	Изгиб оси спирали, bend	градус	[−8; 4]	0.710	3.41±0.14	−3.03±0.03
15	Кручение спирали, twist	градус	[−19; 0]	0.684	34.29±0.13	34.12±0.02
30	Перекрест пары, propeller	градус	[−8; 5]	0.661	-13.80±0.33	-12.52±0.09
17	Сдвиг по длинной оси, slide	ангстр.	[−1; 20]	0.657	−0.03±0.02	-0.06±0.004
35	Дисбаланс бороздок	ln-ед.	[−9; 5]	0.484	1.03±0.09	1.10±0.01
38	Свободная энергия Гиббса	kcal/mol	[−10; 0]	0.440	−1.42±0.15	−1.61±0.02
Среднее отличие		(ф-ла 37)	[−19; 20]		0.75±0.60	−0.74±0.09

компьютерного распознавания сайтов связывания транскрипционного фактора EN, основанный на использовании формулы (37), которая была описана в предыдущем разделе.

На Рисунке 44 представлен документ базы знаний FEATURES (Ponomarenko M. *et al*., 1999b), созданной для документирования результатов системы bDNAvideo (Таблица 17). Двухбуквенные коды полей документов этой базы знаний соответствуют общепринятой нотации EMBL Data Library (Rice *et al*., 1993), уникальное поле “C-CODE” содержит код программы на языке программирования “Си” для распознавания сайтов EN.

2.2.3 Компьютерный анализ спирали ДНК сайтов связывания транскрипционных факторов, представлявших все суперклассы

Для исследования с помощью системы bDNAvideo было собрано 1819 сайтов связывания 42 транскрипционных факторов, которые представляли все четыре их суперкласса (Вингендер, 1997) и которые были документированы в базе данных SAMPLE, созданной в рамках настоящей диссертации.

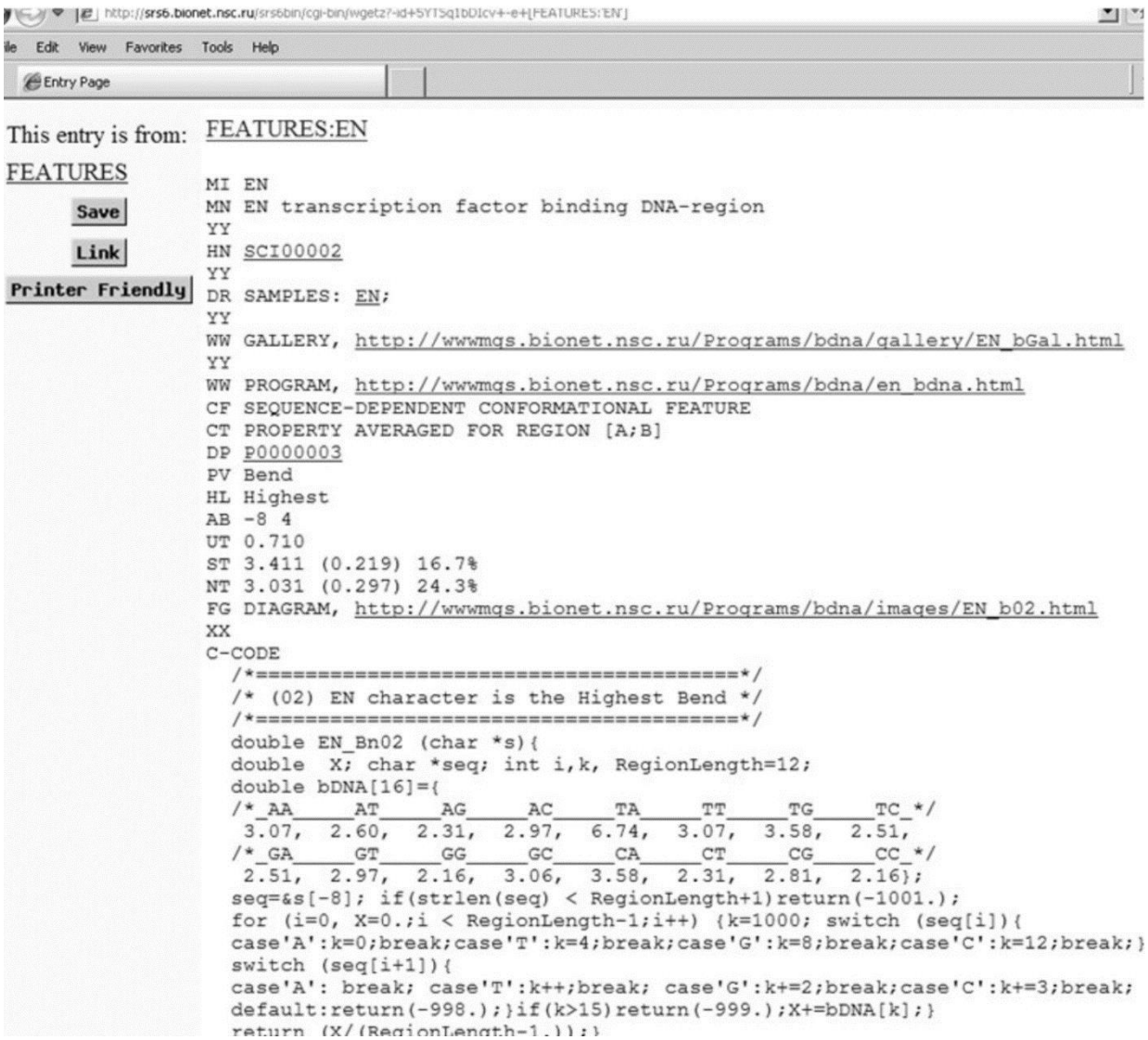

Рисунок 44 - База знаний FEATURES (Ponomarenko M. *et al*., 1999b), результатов системы bDNAvideo (Таблица 17). Фрагмент документа этой базы данных, в котором был документирован результат сравнения сайтов связывания транскрипционного фактора EN со случайными последовательностями такой же длины, состоящих из равновероятных независимых нуклеотидов ДНК (Таблица 16). Поле “C-CODE” – автоматически сгенерированный исходный код программы на языке программирования “Си” (стандарт ANSI; сокращение от “**A**merican **N**ational **S**tandards **I**nstitute”, англ. яз.) для дискриминации сайтов связывания EN от случайных ДНК. Двухбуквенные коды других полей документа были использованы по аналогии с нотацией EMBL Data Library (Rice *et al*., 1993).

```
ID   YY10017; DNA
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Rodentia;
OC   Sciurognathi; Muridae; Murinae; Mus.
DR   TRANSFAC; R01833; MOUSE$RPL30_04; 3.2;
DR   EMBL; K02928; MMRPL30; ; join(420..539)
FT   {0,0} [56;64] direct; EXP
SQ   cctctgtcgg cctagaagag ctttgcattg tgggagctcc ttcctttctc gctccccggc
     catcttggcg gctggtgttg gtgagtgagc tctgcggggt aaacgattag gcggctcggg
//
.....................................................................
ID   YY10020; DNA
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Rodentia;
OC   Sciurognathi; Muridae; Murinae; Mus.
DR   MEDLINE; 7969151; ; ;
CC   element c of THE MOUSE Hoxb-4 gene (+143 to +169) - binding site
CC   of YY1 /Gutman A. et al. Mol Cel Biol 1994. 14. 8143-8154./
FT   {0,0} [48;73] direct; EXP
SQ   gagtagggtc cgggtgagca gatttcctta tccgggaatc gcaggccggg tggccattgg
     ctcggaggat cacgtgggcc tctaactttg ttcacttgac agtaagtagg agggctttcg
//
```

Рисунок 45 - Примеры сайтов связывания транскрипционного фактора YY1 (“Инь-Янь 1” - “**Y**in-**Y**ang **1**”, англ. яз.) из базы данных SAMPLE (Ponomarenko M. *et al*., 1999b), созданной в рамках настоящей диссертации.

На Рисунке 45 показан пример сайтов связывания транскрипционного фактора YY1 (“**Y**in-**Y**ang **1**” – “Инь-Янь 1”, англ. яз.), которые были задокументированы в этой базе данных.

Таблица 18 характеризует 42 выборки последовательностей ДНК, которые были фазированы по центрам сайтов связывания 42 транскрипционных факторов (Ponomarenko M. *et al*., 1999b). В нижней части Таблицы 18 можно также видеть, что все 1819 анализируемых сайтов связывания транскрипционных факторов характеризуются достоверно равными частотами встречаемости нуклеотидов ($\alpha<0.05$, критерий χ^2).

Это достоверное сходство обосновало поиск биологически значимых свойств спирали ДНК сайтов связывания транскрипционных факторов путем сравнения их последовательностей ДНК со случайными ДНК.

Результат bDNAvideo на этих данных показан в Таблице 19. Всего было найдено 848 достоверных различий между свойствами спирали ДНК для сравнения 1819 сайтов связывания 42 транскрипционных факторов и со случайными ДНК (Таблица 19).

Таблица 18 – Исходные данные для системы bDNAvideo: 1819 последовательностей ДНК длиной 120 п.о., ±60 п.о. от центра сайта связывания транскрипционного фактора (ТФ), взятых из базы данных SAMPLE (Ponomarenko M. *et al*., 1999b) и сгруппированных в 42 выборки, № 1 - 42.

Сайт связывания			Количество N(ξ) нуклеотида ξ					Частота p(ξ) нуклеотида ξ			
№	ТФ	Коли-чество	N(A)	N(T)	N(C)	N(G)	Всего	p(A)	p(T)	p(C)	p(G)
1	AP-1	74	1960	2017	2518	2385	8880	0.22	0.23	0.28	0.27
2	ATF	28	785	799	973	803	3360	0.23	0.24	0.29	0.24
3	C/EBP	108	3639	3387	3077	2857	12960	0.28	0.26	0.24	0.22
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
8	CP-1	51	1551	1290	1582	1697	6120	0.25	0.21	0.26	0.28
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
13	EN	12	382	361	381	316	1440	0.27	0.25	0.26	0.22
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
17	GATA	76	2424	2228	2210	2258	9120	0.27	0.24	0.24	0.25
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
20	HNF1	42	1534	1420	1020	1066	5040	0.30	0.28	0.20	0.21
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
23	IRF-1	6	215	159	183	163	720	0.30	0.22	0.25	0.23
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
26	NF-1	101	3138	3063	2891	3028	12120	0.26	0.25	0.24	0.25
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
29	NF-κB	36	1013	966	1113	1228	4320	0.23	0.22	0.26	0.28
30	OCT	62	1894	1948	1802	1796	7440	0.25	0.26	0.24	0.24
31	PR	20	661	670	479	590	2400	0.28	0.28	0.20	0.25
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
35	Sp-1	176	3672	3636	7052	6760	21120	0.17	0.17	0.33	0.32
...	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
39	TFIID	500	16525	13992	14195	15288	60000	0.28	0.23	0.24	0.25
40	TTF-1	7	257	190	200	193	840	0.31	0.23	0.24	0.23
41	USF	20	542	517	719	622	2400	0.23	0.22	0.30	0.26
42	YY1	27	707	774	845	914	3240	0.22	0.24	0.26	0.28
Средняя частота нуклеотида ξ, $M_0[p(\xi)]$								0.25	0.24	0.26	0.25
Ошибка среднего (1% доверительный интервал)								±0.01	±0.01	±0.01	±0.01
t-тест Стьюдента (1% доверительный интервал)								±0.08	±0.06	±0.07	±0.08
Нормальность распределения частот *p(ξ)* нуклеотида ξ			Число степеней свободы, ν:					7	5	8	7
			Мера сходства, χ^2:					1.62	0.53	2.37	2.01
			Уровень значимости, α:					< 0.05	< 0.05	< 0.05	< 0.05

В 516 из 848 случаев (61%) среднеарифметические значения свойств сайтов связывания были больше, чем для случайных ДНК (символ “+” в Таблице 19), в 332 случаях (39%) – меньше (символ “−” в Таблице 19). Это отклонение от равновероятных исходов, 50% выше и 50% ниже, чем у случайной ДНК, было достоверным ($\alpha<10^{-9}$, биномиальное распределение). Оно означает достоверное сходство некоторых деталей в молекулярных механизмах связывания различных транскрипционных факторов с ДНК.

Прежде всего, в Таблице 20 эта закономерность была детализирована на случай отдельных свойств спирали ДНК. Самым достоверным ($\alpha<10^{-8}$) общим физико-химическим свойством сайтов связывания транскрипционных факторов была частота контакта с октамером гистонов (Satchwell *et al*., 1986). Это согласуется с результатами экспериментов, что нуклеосома является очень важной частью транскрипционных машин (например, (Felsenfeld, 1992)). Кроме того, самая высокая значимость предрасположенности сайтов связывания транскрипционных факторов к нуклеосомной ДНК согласуется с экспериментальными данными (Carrillo Oesterreich *et al*., 2011) о синхронизации перестройки хроматина, транскрипции и сплайсинга как единого процесса.

Самым важным контекстно-зависимым свойством конформации спирали ДНК для сайтов связывания 42 транскрипционных факторов было значимо большее среднеарифметическое значение кручения twist B-спирали ДНК (Karas *et al*., 1996) по сравнению со случайной ДНК.

В Таблице 21 можно видеть подробное описание результатов bDNAvideo для кручения twist, самого важного из свойств конформации спиралей ДНК сайтов связывания транскрипционных факторов. Оно было ассоциировано с двумя суперклассами: “Zn-координируемый” - достоверно слабое кручение спирали ДНК ($\alpha<0.05$); “гомеодомен” - достоверно сильное кручение ($\alpha<0.025$).

Таблица 19 – Анализ сайтов связывания 42 транскрипционных факторов

№ k	Фактор транскрипции (суперкласс: Рисунок 42) / Свойство В-ДНК	основный домен															Zn-координируемый												гомеодомен									β-слой					
		AP-1	c-Fos	c-Jun	NF-E2	CRE-BP1	ATF	CREB	CEBP	NF-IL6	MyoD	E2F	USF	NF-1	RFX1	CP1	ER	GR	PR	RAR	RXR	T3R	COUP	GATA	Sp1	GAGA	YY1	GAL4	OCT	HNF1	TTF-1	EN	HNF3	HSF	c-Myb	Ets	IRF1	NF-κB	MEF-2	SRF	E2	TFIID	TCF-1
1	Угол кручения [#]	+	+	+	+		+		+	+	+	+		+	+	+	+	+	+			+		+	−	+	+		+	+			+			+	+	+				−	+
2	Шаг спирали							+	−	−			+	+		−		−			+	+		−	+				−	−	−		−				−	+	−	+		−	+
3	Угол изгиба оси	−			−			−	+			−	−	−		+	−			−	−	−	−	+	−	−		−	+	+		+	+	−			−	−	+	+		+	−
4	Наклон п.о. tip	−	−	−	−	−	−	−			−		−	−	−	−	−	−	−	−	−	−	−	−	−	−	−			+	−	+	+			−	−	−	+	+		+	
5	Наклон inclination							−		+		−	+		−	−	−										+		+	+			−	−				+				−	
6	Ширина бол. бороз.				+			+	−	−	+		+	+		+		−			+	+		−	+		+		−	−			−				−	+	−	+		−	
7	Глубина бол. бороз.	−			−			−	+		−	−	−	−	−	−					−	−			−	−	−		+	+			+				−	−	+			+	
8	Ширина мал. бороз.	−			−			−			−	−	−	−		−	−				−	−	−	−	−	−	−			+			+				−	−	+			+	−
9	Глубина мал. бороз.	+										+	−	+		+	+						+	+	+	+	+	+	+	−				+	+		+	+				−	+
10	Угол открытия roll[&]				+			−			+	−		+		+	+	+	+	+	+	+	+	+	+	−			−	+	+			+		+	+	+	+	+		+	+
11	Угол кручения[&]	−			−				−		−	+	−	−		−	−	−					−		−				+								−	−					−
12	Угол открытия tilt[&]							−	−	−	+		+			+	+	−		+	+	+	+	+	+	+			−	−			−	+	+		+	+					+
13	Сдвиг длинной оси[&]	+	+	+	+			+		+	+	−	+	+	+	+	+			+	+	+	+	+	+		+	−	+	−	+		−				−	+	−	+		−	
14	Угол открытия roll[@]	+	+	+	+	+		+	−		+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		−	−	+	[−]	−	+			+	+	−			−	+
15	Угол кручения [@]	−			−			−	+			+	−		−	+	−	−		−		−	−		−				+	+		+	+				+	−	+		−	+	
16	Угол открытия tilt [@]							−	+			+	−			+	+	+	+	+	+	+	+	+		+			−	+	+		+	+			+	+	+				+
17	Сдвиг длинной оси [@]			+			+	+				+	+	+		−		−					−	−	+	+	+		−	−		+			+						+	−	
18	Угол кручения[%]							+	+	+		+		+		+											+		+	+			+				−		+	+			
19	Угол Эйлера, wedge							−				+	−		+	+			+					+	−	+		+	+		+		+	+	+		+		+				+
20	Угол Эйлера, direction									+		+		+	+	+				−	−				+	+	+		+								−	+				−	+
21	Персистентная длина					+		+	−	−	+		+	+			+			+	+	+	+	−	+		+		−	−	−		−		+		−	+	−		+	−	+
22	Т плавления, $T_М$					+	+	+	−	−	+	+	+	+		−	+				+	+		−	+	+	+		−	−		−	−				−	+	−	−	+	−	
23	Контакт в нуклеосоме	+	+	+			+		+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	−				+	+	+		+	+		+	+	+	+			+	+
24	Изгиб бол. бороздки							−	+	+	−	+	−	−			+	+						+	−				+	+		+	+	+			+	−	+	+		+	
25	Изгиб мал. Бороздки	+						+	−	−	+		+	+	+	−		−			+	+		−	+				−	−			−				−	+	−			−	−
26	Угол кручения, twist[$]	+	+	+				+	+	+			−			+	+	+	+			−		+	−				+	+			+	−			+	−	+	+		−	
27	Угол открытия, tilt[$]								+	+		+				−		−		−	−	−	−		−	−	+		+	+			+	−			−	−	+	+	+		−
28	Угол открытия, roll[$]	−							−	−			+	−		−	+	−		+	+	+		−	+	−			−	−	−		−				−	+		+	+		+
29	Сдвиг длинной оси[$]	+	+	+	+			+		+	+	−	+	+	+	+	+	+		+	+	+		+	+		+		+	−	+		+	−			−	+	−	+		−	+
30	Угол пропеллера	+	+	+	+	+	+	+	−		+	−	+			−	+	+		+			+	−		−			−	−		−	+	−			−	−	−		+	−	−
31	Размер мал.бороз., size					+		+	−		+	+	+	+		+					+	+		−	+				−	−			−				−	+	−			−	
32	Размер мал.бороз., dist			+			+	+		+			+							−	−	−	−	−	−	−			+	+				−			−	−	−		+		−
33	Размер бол.бороз., Size																							−	−																		
34	Размер бол.бороз., Dist							+						+	+		−								−				+								−	−		+		+	−
35	Дисбаланс бороздок				+			+	−	−	+	−	+				+	−			+	+		−	+	+			−	−		−	−	+	+		−	+	−		+	−	+
36	Энтальпия, ΔH	+						−	+	+		−	−			+		+	+					+	−				+	+	+		+				+	−	+	+		+	
37	Энтропия, ΔS	+						−	+	+		−	−				+	+	+					+	−	+			+	+	+		+				+	−	+	+		+	
38	Своб. энергия Гиббса							−	+	+		−	−	−		+	+	+	+					+	−				+	+	+	+	+				+		+	+		+	

“+”/“–” – достоверно >/<, чем случайно; рамка (□) – клетка для Рисунка 43); Данные [#)](Karas *et al*., 1996); [&)]свободная ДНК (Suzuki *et al*., 1996); [@)]ДНК/белок (Suzuki *et al*., 1996); [*)](Kabsh *et al*., 1982); [$)](Gorin *et al*., 1995); [%)](Shpigelman *et al*., 1993);

Таблица 20 – Конформационные и физико-химические свойства спирали ДНК, которые достоверно часто выявлялись системой bDNAvideo в качестве значимых различий между сайтами связывания транскрипционных факторов и случайными последовательностями равновероятных независимых нуклеотидов такой же длины

№	k	Свойство спирали ДНК (ссылки для альтернатив)	N	N_+	N_-	значимость α
I	23	Частота контакта с октамером гистонов	31	30	1	$<10^{-8}$
II	1	Кручение twist, *in silico* (Karas *et al*., 1996)	28	26	2	$<10^{-4}$
	11	Кручение twist, свобод. ДНК (Suzuki *et al*., 1996)	16	2	14	<0.0025
	18	Кручение twist (Shpigelman *et al*., 1993)	13	12	1	<0.0025
	26	Кручение twist (Gorin *et al*., 1995)	23	17	6	<0.025
III	4	Наклона п.о. tip	33	6	27	<0.00025
IV	10	Раскрытие roll, ДНК (Suzuki *et al*., 1996)	27	23	4	<0.00025
	14	Раскрытие roll, ДНК/белок (Suzuki *et al*., 1996)	35	27	8	<0.001
V	16	Раскрытие tilt, ДНК/белок (Suzuki *et al*., 1996)	23	20	3	<0.00025
	12	Раскрытие tilt, ДНК (Suzuki *et al*., 1996)	23	16	7	<0.05
VI	29	Сдвиг длинной оси п.о. (Gorin *et al*., 1995)	31	25	6	<0.0005
	13	Сдвиг длинной оси п.о. (Suzuki *et al*., 1996)	30	23	7	<0.005
VII	8	Ширина малой бороздки	23	4	19	<0.0025
	9	Глубина малой бороздки	20	17	3	<0.0025
	7	Глубина большой бороздки	22	6	16	<0.05
VIII	19	Угол Эйлера wedge	18	15	3	<0.005
IX	37	Энтропия	21	16	5	<0.025
	38	Свободная энергия Гиббса	21	16	5	<0.025
	36	Энтальпия	20	15	5	<0.025
	24	Жесткость изгиба по большой бороздке	21	15	6	<0.05

N, N_+ и N_- – число транскрипционных факторов, средние свойства спирали ДНК для сайтов связывания которых значимы, выше и ниже случайных, соответственно; значимость была оценена по биномиальному распределению.

Другими достоверными конформационными свойствами спирали ДНК для сайтов связывания транскрипционных факторов были угловые и линейные деформации спирали ДНК, которые, по-видимому, могут соответствовать механизму межмолекулярного распознавания “ключ/замок” для комплексов ДНК/белок (Fischer, 1966).

Таблица 21 – Корреляции между “суперклассом” транскрипционного фактора и достоверными отличиями сайтов их связывания от случайных ДНК по оценкам среднего “угла кручения twist ДНК в комплексе с белками” (Suzuki *et al*., 1996)

Транскрип-ционный фактор		район [a; b]	$U(X_{15ab})$ (47)	сред. ± станд.ош.среднего		значимость	
				сайт EN	случайная ДНК	χ^2	α
Осно-вный домен	NFE2	-18; 8	0.756	33.91±0.14	**34.12±0.02**	187.8	0.0005
	USF	-13; 8	0.695	33.91±0.07	**34.11±0.02**	32.0	0.05
	RF-X	-9; 16	0.586	34.00±0.11	**34.11±0.02**	123.9	0.0005
	CREB	-11; 6	0.462	33.98±0.08	**34.10±0.02**	34.7	0.05
	AP-1	-10; 13	0.453	30.00±0.05	**34.12±0.02**	231.2	0.0005
	C/EBP	-9; 3	0.462	**34.28±0.06**	34.10±0.03	63.4	0.005
	CP-1	-14; 12	0.504	**34.16±0.05**	34.11±0.02	73.0	0.005
	E2F	-4; 14	0.765	**34.45±0.18**	34.10±0.02	92.2	0.005
β-слой	NF-kB	-16; 0	0.452	33.89±0.08	**34.13±0.02**	116.1	0.0005
	E2	-7; 5	0.311	33.90±0.15	**34.11±0.03**	34.4	0.05
	TFIID	-4; 13	0.605	**34.29±0.02**	34.11±0.02	377.1	0.0005
	MEF-2	-10; 2	0.904	**34.44±0.24**	34.11±0.03	81.0	0.005
Zn-коор-дини-руе-мый	T3R	-14; 10	0.736	33.89±0.07	**34.11±0.02**	54.6	0.005
	Sp-1	-14; 11	0.642	33.90±0.03	**34.11±0.02**	396.0	0.0005
	COUP	-11; 5	0.550	33.92±0.14	**34.09±0.02**	53.4	0.005
	ER	-12; 14	0.536	33.92±0.06	**34.10±0.02**	43.9	0.05
	RAR	-14; 12	0.448	33.92±0.12	**34.11±0.02**	28.8	0.05
	GR	-9; 13	0.439	34.03±0.05	**34.11±0.02**	68.0	0.005
Оценка среднеарифметического для сайтов значимо **выше**							0.05
гомео-домен	IRF-1	-14; 12	0.611	**34.17±0.23**	34.11±0.02	31.0	0.05
	OCT	-9; 4	0.621	**34.29±0.07**	34.10±0.03	154.6	0.005
	EN	-19; 0	0.684	**34.29±0.13**	34.12±0.02	26.7	0.05
	HNF1	-6; 10	0.725	**34.46±0.10**	34.12±0.02	95.8	0.005
	HNF3	-10; 6	0.748	**34.43±0.18**	34.11±0.02	138.2	0.0005
Оценка среднеарифметического для сайтов значимо **выше**							0.025

В качестве достоверных физико-химических свойств спирали ДНК этих сайтов связывания были также энтропия, энтальпия и свободная энергия Гиббса как меры стабильности комплекса спирали ДНК с белком.

Таблица 22 – Значимые различия между суперклассами транскрипционных факторов по свойствам спирали ДНК, значимым для сайтов связывания этих транскрипционных факторов (значимость α по точному критерию Фишера)

Свойство		Среднее для сайтов связывания транскрипционных факторов			
k	Имя	$	меньше, чем у случайных	больше, чем у случайных	α
15	Кручения	II	COUP, T3R, Sp1, ER, RAR, GR		0.05
		III		OCT, IRF-1, EN, HNF1, HNF3	
3	Изгиб оси	II	T3R, Sp1, GAGA, RXR, GAL4, COUP, ER, RAR	GATA	0.05
		III	HSF, IRF-1	HNF3, OCT, EN, HNF1	
21	Персистентная длина	II	GATA	YY1, COUP, RAR, T3R, RXR, ER, Sp1	0.025
		III	HNF3, OCT, HNF1, IRF-1, TTF-1	c-Myb	
22	T плавления	II	GATA	ER, YY1, T3R, Sp1, RXR, GAGA	0.01
		III	OCT, HNF1, HNF3, EN, MEF-2, IRF-1		
24	Изгиб большой бороздки	I	Sp1, MyoD, NF-E2, USF, CP-1, NF-1, AP-1, CREB, E2F, RF-X	C\EBP	0.05
		III	IRF-1	OCT, HNF1, HNF3	
25	Изгиб малой бороздки	I	C\EBP, CP-1, NF-IL6	AP-1, RF-X, NF-1, CREB, USF, MyoD	0.05
		III	OCT, HNF3, HNF1, IRF-1		
28	Раскрытие п.о. roll	I	CP-1, C/EBP, NF-1, NF-IL6, AP-1	USF	0.025
		IV		SRF, E2, TCF-1, NF-kB	
		III	OCT, HNF3, TTF-1, HNF1, IRF-1		0.01
30	Пропеллера	I	E2F, C/EBP, CP-1	MyoD, c-Fos, AP-1, c-Jun, ATF, USF, NF-E2, CRE-BP1, CREB	0.05
		III	IRF-1, HSF, EN, OCT, HNF1	HNF3	
31	Размер мал.бор оз. size	I	C/EBP	NF-1, E2F, CP-1, MyoD, CREB, USF, CRE-BP1	0.025
		III	HNF3, HNF1, OCT, IRF-1		
32	Размер мал.бор оз. dist	I		c-Jun, CREB, ATF, NF-IL6, USF	0.01
		II	GATA, GAGA, Sp1, T3R, COUP, RXR, RAR		
38	Свобод. энергия Гиббса	I	E2F, CREB, USF, NF-1	NF-IL6, C/EBP, CP-1	0.05
		III		EN, TTF-1, OCT, HNF3, IRF-1, HNF1	

$) Суперклассы: основный (I), Zn-коорд. (II), гомеодомен (III), β-слой (IV).

В Таблице 22 показаны достоверные различия между суперклассами транскрипционных факторов по выявленным свойствам спирали ДНК сайтов

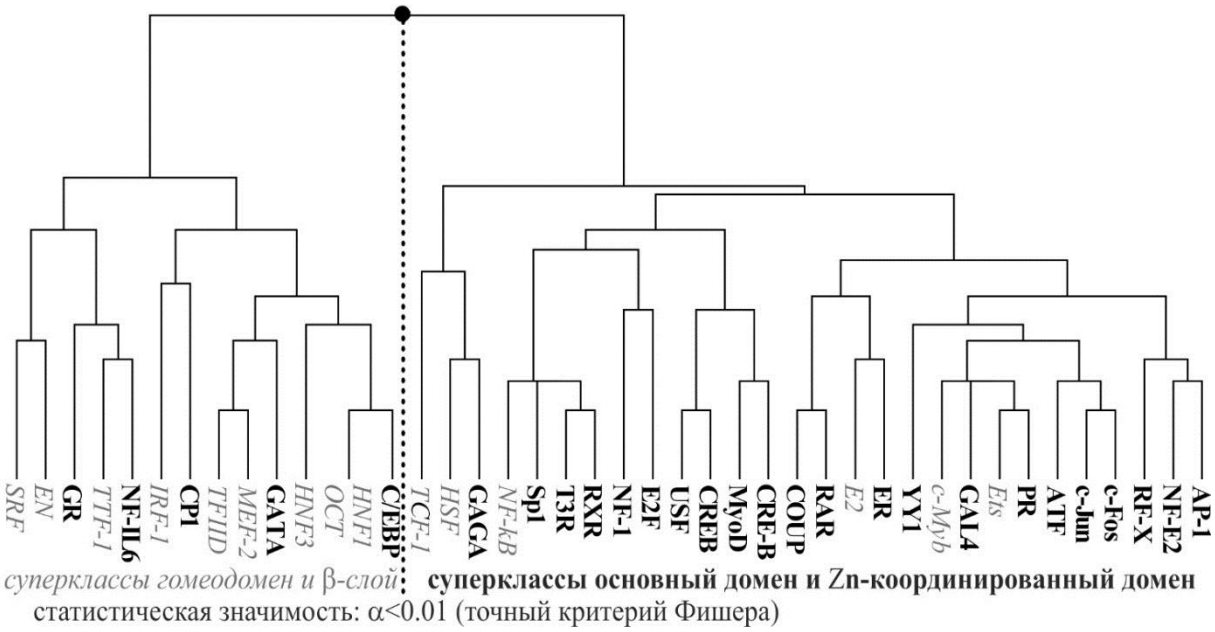

Рисунок 46 – Граф сходства между суперклассами транскрипционных факторов по физико-химическим и конформационным свойствам спирали ДНК сайтов их связывания, полученный с помощью стандартного пакета Statistica (Statsoft™, Tulsa, USA, параметры: мера сходства Евклида и метод кластер-анализа UPGA).

их связывания. Например, суперкласс “гомеодомен” характеризуется большим кручением twist и большим изгибом оси bend спирали ДНК при меньшей персистентной длине и меньшей температуре плавления ДНК сайтов связывания транскрипционных факторов в сравнении с суперклассом “Zn-координируемый”.

В качестве обобщения частных закономерностей, представленных в Таблицах 20 - 22, с помощью стандартного статистического пакета Statistica (Statsoft™, Tulsa, USA) был проведен кластерный анализ всех выявленных с помощью bDNAvideo свойств спиралей ДНК для 1819 сайтов связывания 42 транскрипционных факторов (Таблица 19).

Полученный результат представлен на Рисунке 46 в виде бинарного графа сходства между транскрипционными факторами по конформационным и физико-химическим характеристикам сайтов их связывания. Можно видеть, что суперклассы “Zn-координированный” и “основный” транскрипционных

Таблица 23 – Оценка устойчивости графа сходства транскрипционных факторов по значимым физико-химическим и конформационным свойствам спирали ДНК сайтов их связывания к варьированию мер сходства и методов кластер-анализа в рамках стандартного статистического пакета Statistica (Statsoft™, Tulsa, USA)

STATISTICA (Statsoft™, Tulsa, USA) / Мера сходства / Метод кластеризации	мера Евклида		% отличий	
	N[#]	α	N[#]	α
Невзвешенное попарное усреднение (UPGA)	Рисунок 46		0	<0.01
Взвешенное попарное усреднение (WPGA)	0	<0.01	0	<0.01
Невзвешенное попарное центрирование (UPGC)	2	<0.01	0	<0.01
Взвешенное попарное центрирование (WPGC)	0	<0.01	0	<0.01
Метод Уорда (“Ward's method”)	0	<0.01	5	<0.01
Метод полной связи (“Complete Linkage”)	0	<0.01	0	<0.01

[#)]N – число отличий от приведенной на Рисунке 46 корневой дихотомии “●”

факторов с характерным локальным избытком электростатического заряда (правый подграф) достоверно ($\alpha<0.01$, критерий Фишера) отличаются от суперклассов “β-слой” и “гомеодомен” транскрипционных факторов без локального избытка электростатического заряда (левый подграф).

Таблица 23 характеризует устойчивость графа сходства суперклассов транскрипционных факторов по физико-химическим и конформационным свойствам сайтов их связывания при варьировании мер сходства и методов кластер-анализа пакета Statistica (Statsoft™, Tulsa, USA). Как можно видеть, лишь в двух из одиннадцати испытаниях были отклонения от представленного на Рисунке 46 графа сходства транскрипционных факторов, которые, однако, не влияли на достоверность ($\alpha<0.01$) их кластеризации на белки с локальным избытком электростатического заряда (80%) и белки без локального избытка электростатического заряда (64%) на основе учета значимых конформационных и физико-химических характеристик спирали ДНК этих сайтов связывания.

Представленные в настоящей главе результаты позволяют сделать следующий вывод:

- На основе теории аддитивной полезности для принятия решений и нечетких множеств создана компьютерная система bDNAvideo для выявления контекстно-зависимых конформационных и физико-химических характеристик спирали ДНК, достоверно дискриминирующих сайты связывания транскрипционных факторов от случайных последовательностей. С использованием этой системы впервые получена достоверная кластеризация транскрипционных факторов на две группы, первая из которых включает преимущественно основные и Zn-координируемые белки с локальным избытком электростатического заряда, вторая - белки с β-слоем и с гомеодоменом без локального избытка электростатического заряда.

ЗАКЛЮЧЕНИЕ ПО ГЛАВЕ 2

Представленные в настоящей главе диссертации результаты компьютерного анализа количественных характеристик регуляторных сайтов в составе геномных ДНК были опубликованы в 1997, 1998 и 1999 годах прошлого века (Пономаренко М. и др., 1997в; Колчанов и др., 1999; Ponomarenko M. *et al.*, 1997b; Kolchanov *et al.*, 1998, 1999; Ponomarenko J. *et al.*, 1999a,b). За прошедшие с тех пор почти два десятилетия на их основе другими авторами были созданы новые компьютерные методы анализа регуляторных районов геномов.

Прежде всего, в отделе системной биологии ИЦиГ СО РАН, где выполнялась настоящая диссертационная работа, была создана компьютерная система SITECON (Oshchepkov *et al.*, 2004), основанная на результатах этой главы и на множественном символьном выравнивании экспериментально доказанных сайтов связывания транскрипционных факторов (Lawrence *et al.*, 1993). Достигнутая при этом точность предсказаний регуляторных сайтов в

геномах эукариот заложила основы компьютерно-экспериментальной идентификации сайтов для связывания транскрипционных факторов (Ощепков и др., 2005, 2009; Ощепков, 2010; Игнатьева и др., 2007, 2009; Брызгалов и др., 2008; Кузнецова и др., 2008; Ершов и др., 2009; Левицкий и др. 2010, 2011; Kolchanov *et al*., 2007; Furman *et al*., 2009; Omelina *et al*., 2011).

Кроме того, с помощью системы bDNAvideo (Ponomarenko M. *et al*., 1999b) были созданы программы распознавания нуклеосомной ДНК по ее последовательности (Levitsky *et al*., 1999). С помощью этих программ был, например, экспериментально идентифицирован минимальный промотор гена RFP2 супрессора опухолей на основе опыта с делециями в районе 13q14.3 генома человека (Skoblov *et al*., 2006).

В свою очередь, независимый анализ геномных карт посадки нуклеосом выявил (Tillo, Hughes, 2009) улучшение их прогноза при учете конформационных углов *propeller* и *slide* спирали ДНК. Учет температуры плавления ДНК в дополнение к символьным закономерностям улучшил прогнозы сайтов связывания регуляторных белков по нуклеотидным последовательностям геномной ДНК (Fu *et al*., 2009).

Предсказанная (Пономаренко М., 1997в) в настоящей диссертации конкуренция TBP с октамером гистонов за TATA-бокс (Рисунок 35) согласуется с недавно открытыми синхронизацией перестройки хроматина, транскрипции и сплайсинга (Bieberstein *et al*., 2012), а также с импульсами транскрипции (Hornung *et al*., 2012), амплитуда которых растет с ростом TBP/TATA-сродства, тогда как их длительность убывает с ростом предрасположенности кор-промотора гена к нуклеосомной ДНК.

Наконец, совпадения нуклеотидов были обобщены с помощью критерия χ^2 в меру сходства позиций в промоторах генов эукариот по контекстно-зависимым количественным величинам конформационных и физико-химических свойств спирали ДНК в локальных окрестностях этих позиций в системе FeatureScan (Deyneko *et al*., 2006), которая выявляет на этой основе конформационно сходные районы промоторов генов. Результаты настоящей

диссертации применялись при создании компьютерных систем BiDaS (Paraskevopoulou et al., 2013), CRoSSeD (Meysman et al., 2011), DISCOVER (Fu et al., 2009), BioBayesNet (Nikolajewa et al., 2007), ProMapper (Pudimat et al., 2005) для распознавания сайтов связывания транскрипционных факторов в геномной ДНК. Общепринятым строением промотора стала мозаика конформационно-контрастных участков ДНК (Левицкий, 2001; Goni *et al*., 2007), например, чередование гибкий/жесткий район спирали ДНК промотора (Perez *et al*., 2008). Дополнения консенсуса и позиционно-весовой матрицы нейронными (Beiko, Charlebois, 2005) и байесовскими сетями (Gunewardena *et al*., 2006), а также оценками корреляции позиций ДНК (Levitsky *et al*., 2007) в качестве учета кооперативных взаимодействий нуклеотидов в функциональных сайтах в составе геномной ДНК расширили представления о регуляции экспрессии генов.

ГЛАВА 3 КОМПЬЮТЕРНАЯ СИСТЕМА ACTIVITY: КОРРЕЛЯЦИЯ МЕЖДУ СРОДСТВОМ ТАТА-СВЯЗЫВАЮЩЕГО БЕЛКА К ТАТА-БОКСУ И КОЛИЧЕСТВЕННЫМИ ХАРАКТЕРИСТИКАМИ ДНК

Настоящая глава описывает созданную в диссертационной работе компьютерную систему Activity для выявления количественных характеристик регуляторных сайтов в составе геномных ДНК, коррелирующих с количественными характеристиками биологической активности ДНК (Пономаренко М. и др., 1996, 1997а,б,в, 1998, 1999б, 2006, 2008; Колчанов и др., 1998, 1999; Савинкова и др., 2007, 2009; Пономаренко П. и др., 2008, 2009, 2010; Втюрина и др., 2012; Ponomarenko M. *et al*., 1997a, 1999a, 2013a,e,g; Ponomarenko J. *et al*., 2000a,b, 2001a,b, 2002a,b,c; Suslov *et al*., 2010a,b; Kirpota *et al*., 2011; Drachkova *et al*., 2011; Savinkova *et al*., 2013), на примере ее применения для анализа связывания между ТАТА-связывающим белком и ТАТА-боксом.

Аналогично предыдущей главе диссертации, к началу работ по этой ее главе в конце 90-ых годов XX века регуляцию экспрессии генов чаще всего анализировали методами консенсуса (Hawley, McClure, 1983) и позиционно-весовой матрицы (Mulligan *et al*., 1984), соответствующих линейно-аддитивному приближению независимых нуклеотидов сайта (Berg, von Hippel, 1987). Для их оптимизации на “обучающих” экспериментальных данных применяли перцептрон (Stormo *et al*., 1982), линейную регрессию (Schneider *et al*., 1986) и алгоритмы распознавания образов в системах MATRIX SEARCH (Chen *et al*., 1995), MatInspector (Quandt *et al*., 1995) и TESS (Schug, Overton, 1997). Оказалось, что эти подходы давали достоверно разные математические модели регуляторных сайтов, прогнозы которых были, тем не менее, достоверно неразличимыми (Barrick *et al*., 1994) и не всегда коррелировали с экспериментальными данными (Roulet *et al*., 1998). Более того, было обнаружено отсутствие корреляций между разными экспериментальными

данными, например, как о влиянии определенного регуляторного белка на один и тот же ген в разных клеточных линиях (Hyde-DeRuyscher *et al*., 1995), так и об одновременно измеренных разных количественных характеристиках экспрессии одного и того же гена (Javahery *et al*., 1994). Эти несоответствия между теорией и опытом, а также между разными опытами объясняли различием кооперативных взаимодействий нуклеотидов сайта при различных условиях. Их изучению на примере связывания TBP с TATA-боксом посвящена эта глава.

3.1 Создание компьютерной системы Activity на основе системы bDNAvideo

Представленная в предыдущей главе диссертации система bDNAvideo (Ponomarenko M. *et al*., 1997b) выявляла контекстно-зависимые количественные характеристики спирали ДНК для сайтов связывания регуляторных белков на основе теории аддитивной полезности для принятия решений (Fishburn, 1970). Она учла кооперативные взаимодействия групп нуклеотидов в задаче распознавания регуляторных сайтов в составе геномной ДНК. Поэтому, чтобы получить компьютерную систему Activity (Ponomarenko M. *et al*., 1997a) для задачи поиска контекстно-зависимых количественных характеристик регуляторной ДНК, учитывающих кооперативный вклад нуклеотидов в величины биологической активности, было естественным заменить в системе bDNAvideo (Ponomarenko M. *et al*., 1997b) критерии дискриминантного анализа (Рисунок 30) на критерии корреляционного анализа (Рисунок 47). Соответственно, во “входных данных” были заменены случайные ДНК $\{S^-\}$ (Рисунок 30) на экспериментально измеренные количественные величины $\{F(S^+)\}$ биологической активности, характеризующие последовательности $\{S^+\}$ сайтов в составе геномной ДНК (Рисунок 47), имевшиеся в bDNAvideo.

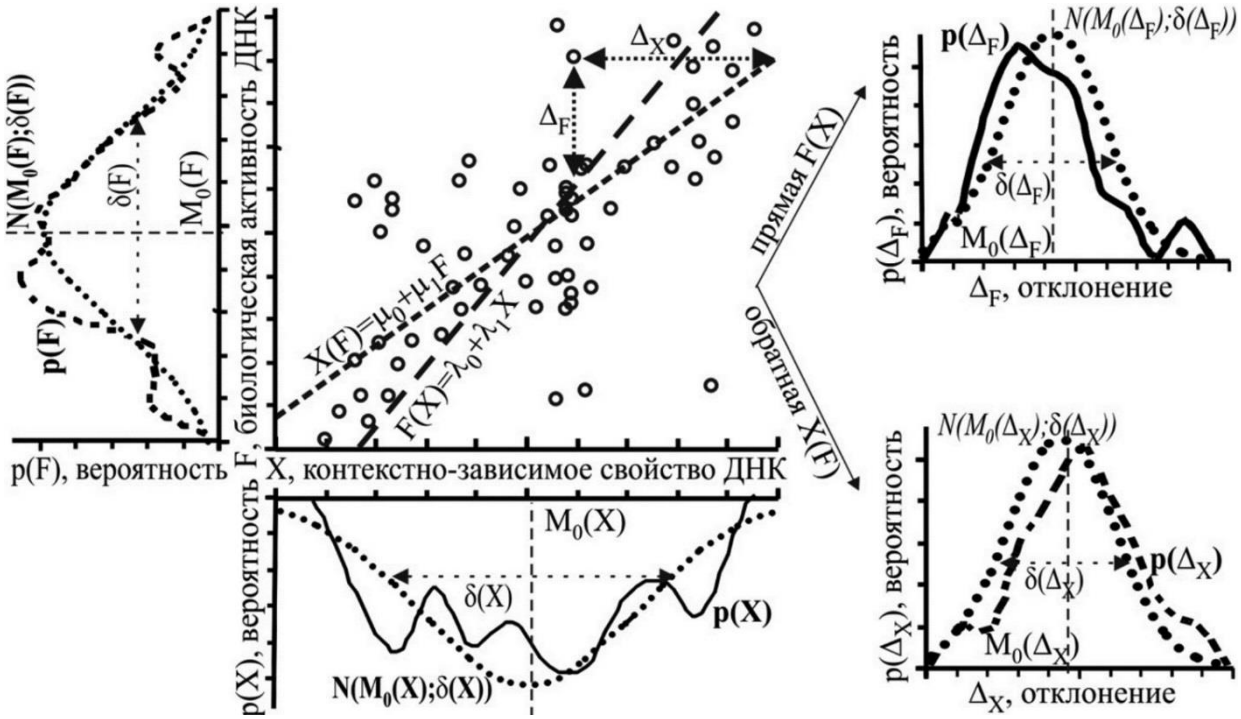

Рисунок 47 - Частные оценки полезности контекстно-зависимого свойства ДНК, X, для линейной регрессии биологической активности F на основе учета степени соответствия между выборочными распределениями p(X(S)), p(F), $p(\Delta_X)$, $p(\Delta_F)$ и нормальными распределениями $N[M_0(X(S)); \delta(X(S))]$, $N[M_0(F); \delta(F)]$, $N[M_0(\Delta_X); \delta(\Delta_X)]$ и $N[M_0(\Delta_F); \delta(\Delta_F)]$ в компьютерной системе Activity. Обозначения: M_0 и δ - среднее и стандартное отклонение; Δ – отклонение анализируемых данных от их линейных регрессий: прямой $F(X)=\lambda_0+\lambda_1 X$ и сопряженной $X(F)=\mu_0+\mu_1 F$.

Наконец, среднеарифметические $X_{k[ab]}$ конформационные и физико-химические свойства спирали ДНК (формула 29) были дополнены содержанием олигонуклеотидов $[\xi_1 \ldots \xi_m]_f$ длины m в сайте длины L, взвешенных позиционно по эвристическому правилу: “чем выше вес f(i), тем больше вклад $s_i \ldots s_{(i+m-1)} \in \xi_1 \ldots \xi_m$ в количественную величину целевой активности F” (здесь: $1 \leq m \leq 4 << L$):

$$[\xi_1 \ldots \xi_m]_f(S = \{s_1 \ldots s_i \ldots s_L\}) = \sum_{\substack{i=1 \\ s_i \ldots s_{i+m-1} \in \xi_1 \ldots \xi_m}}^{L-m+1} f(i). \qquad (38)$$

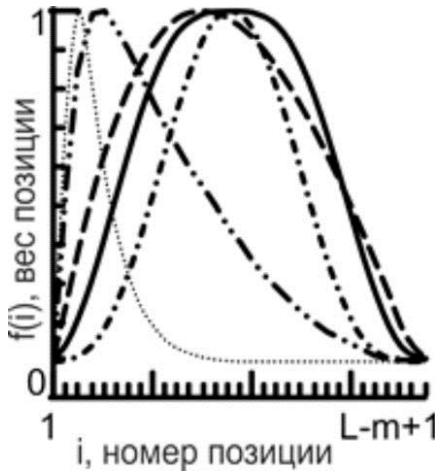

Рисунок 48 - Примеры весовых функций f(i) коротких олигонуклеотидов дины m в локальной окрестности сайта ДНК длины L (здесь: m<<L) для формулы (38).

Всего система Activity (Ponomarenko M. *et al*., 1997a) анализирует 360 унимодальных весовых функций f(i), в том числе: 90 функций с одним максимумом $f(i_{\#})=1$ внутри сайта и двумя минимумами на его концах, f(1) = f(L-m+1) = 0.1; 90 функций с аналогичными минимумом внутри и максимумами на концах сайта; 90 функций с одним максимумом на 5'-конце сайта и одним минимумом на его 3'-конце; 90 функций с минимумом на 5'-конце и максимумом на 3'-конце сайта (здесь: $1 < i_{\#} < L-m+1$). Они различаются формой монотонных переходов между минимальными и максимальными весами (Рисунок 48).

Поэтому в сравнении с "материнской" системой bDNAvideo (Ponomarenko M. *et al*., 1997b), "дочерняя" система Activity (Ponomarenko M. *et al*., 1997a) дополнительно генерирует и проверяет $360 \times 15^4 = 18225000 \approx 10^7$ вариантов взвешенного содержания $[\xi_1 \ldots \xi_m]_f$ олигонуклеотидов в сайте. Необходимость анализа столь большого числа дополнительных контекстно-зависимых количественных характеристик регуляторных сайтов в составе геномной ДНК была вызвана отсутствием сведений о вкладе количественных характеристик контекста ДНК в количественные характеристики регуляции экспрессии генов.

Система Activity (Ponomarenko M. *et al*, 1997a) анализирует любую контекстно-зависимую количественную характеристику $X \in \{[\xi_1 \ldots \xi_m]_f, X_{k;[a;b]}\}$ независимо от других, следуя принципу беспристрастности искусственного интеллекта (Нильсон, 1995). Для иллюстративного примера на Рисунке 47 показана проверка критериев применимости корреляционного анализа к "входным данным" путем оценки достоверности α соответствия ("≈") между выборочными и нормальными распределениями: $p(X(S)) \approx N[M_0(X(S)); \delta(X(S))]$, $p(F) \approx N[M_0(F); \delta(F)]$, $p(\Delta_X) \approx N[M_0(\Delta_X); \delta(\Delta_X)]$ и $p(\Delta_F) \approx N[M_0(\Delta_F); \delta(\Delta_F)]$ (здесь: $\Delta_X = X(S) - X(F(S))$, $\Delta_F = F - F(X(S))$, $X(F(S))$ и $F(X(S))$ – простые регрессии Пирсона "X по F(S)" и "F по X(S)"). Кроме того, она проверяет два критерия независимости выборочных распределений знака отклонений, $p(sing(\Delta_X))$ и $p(sing(\Delta_F))$, данных от их простых регрессий Пирсона. Наконец, Activity оценивает пять типов корреляций между X(S) и F: линейная, две ранговые Спирмена и Кендалла, а также две бинарные Фишера и χ^2.

Всего Activity оценивает 11 критериев: 5 корреляций и 6 условий их применимости к "входным данным". Каждый критерий анализируется с использованием метода bootstrap (Efron *et al*., 1996) многократной проверки каждой статистической гипотезы на семи подвыборках "входным данных": (i) на всех данных, (ii) на 50% наибольший X(S), (iii) на 50% наименьших X(S), (iv) на 50% ближайших X(S) к среднеарифметическому $M_0(X(S))$, (v) на 50% наибольший F, (vi) на 50% наименьших F и (vii) на 50% ближайших F к среднеарифметическому $M_0(F)$. Аналогично "материнской" системе bDNAvideo (Ponomarenko M. *et al*., 1997b), "дочерняя" Activity (Ponomarenko M. *et al*., 1997a) перемасштабирует качественные оценки α достоверности в количественные оценки полезности (формула 30) и усредняет их (формула 31) в итоговую оценку U(X), обладающую свойствами (формулы 32 - 33). Согласно неравенству Бонферрони и биномиальному распределению, верхняя оценка (формула 35) вероятности случайного выбора Activity контекстно-

зависимой характеристики X с позитивной оценкой ее интегральной полезности U(X; F) для построения линейной регрессии биологической активности F из количества 10^7 рассматриваемых характеристик была $p(U(X) > 0) < 10^{-20}$, как это было рассмотрено выше в разделе 2.1.3 настоящей диссертационной работы на примере системы bDNAvideo (формулы 34 и 35).

3.2 Сродство ТАТА-связывающего белка к однонитевым олигонуклеотидам ДНК

Согласно флюориметрическим измерениям в растворе *in vitro* (Powell *et al*., 2002), комплекс "TBP/TATA" был стабилизирован увеличением изгиба оси спирали ДНК от 19^O до 90^O в деформированном и частично денатурированном гетеродуплексе ДНК, как это предсказывали методы молекулярной динамики (Flatters, Lavery, 1998). Поэтому настоящий раздел описывает анализ данных *in vitro* о сродстве TBP к однонитевым олигоДНК, онДНК (Соколенко и др., 1996). Результатом (Пономаренко и др., 1997а,б) этого анализа были достоверные корреляции между сродством TBP/онДНК и содержанием динуклеотидов TV и WR (номенклатура, IUPAC-IUB, 1971) в онДНК. На их основе с помощью пакета Statistica (Statsoft™, Tulsa, USA) была выведена оценка количественной величины сродства TBP/онДНК по последовательности ДНК длиной 15 п.о. Анализ с ее помощью 776 промоторов эукариот из базы данных EPD, вып. 45 (Perier *et al*., 1999) обнаружил единственный пик сродства TBP/онДНК, который соответствовал оптимальному положению ТАТА-бокса в позиции -30 относительно старта транскрипции, при достоверно низком таком сродстве вокруг этого пика. Сотня из миллиона случайных ДНК длины 35 п.о., выбранная по критерию наибольшего превышения сродства TBP/онДНК в центре над флангами, оказался достоверно неотличим от природных ТАТА-боксов такой же длины.

3.2.1 Анализ сродства ТАТА-связывающего белка к однонитевой ДНК

С помощью описанной выше в разделе 3.1 компьютерной системы Activity (Ponomarenko M. *et al*., 1997a) были проанализированы 19 однонитевых олигоДНК, онДНК, длиной 15 нт, в том числе модельные ТАТА-несодержащие ДНК, природные и мутантные варианты ТАТА-содержащих районов промоторов генов гистона H1 и β-актина человека (Соколенко и др., 1996). Каждая олигоДНК была охарактеризована измеренной *in vitro* величиной ($-\ln[K_D]$) сродства ТВР дрожжей к этой олигоДНК, варьирующей в диапазоне значений от 11.78 до 24.23 ln-ед. Эти модельные олигоДНК (Таблица 24) были выбраны так, чтобы они представляли неспецифическое сродство ТВР/ДНК и специфическое сродство ТВР/ТАТА-бокс.

Эти данные были разделены на две части: 8 олигоДНК № 1 – 8 были представительной обучающей выборкой для ее анализа с помощью системы Activity (Ponomarenko M. *et al*., 1997a), остальные 11 олигоДНК № 9 – 19 были независимым контролем для оценки достоверности результатов этого анализа. Вследствие короткой длины 15 нт олигоДНК система Activity анализировала содержание динуклеотидов, m=2 (формула 38). Всего было проанализировано $\approx 15^2 \times 360 = 7290000$ вариантов $[\xi_1\xi_2]_f$ всех $\approx 15^2 = 225$ возможных динуклеотидов в 15-символьном коде (номенклатура, IUPAC-IUB, 1971), взвешенных с помощью 360 весовых функций f(i).

Наибольшая оценка U=0.406 (формула 31) была у содержания динуклеотида WR (IUPAC-IUB, 1971) с максимумом его веса на концах олигоДНК, как это показано пунктиром на Рисунке 49, $[WR]_{(- -)}$. Все четыре AA, TA, AG, TG динуклеотида WR входят в состав общепринятого TATA(a/t)A(a/t)g консенсуса ТАТА-бокса. На независимом контроле величины $[WR]_{(- -)}$ достоверно коррелировали со сродством ТВР/онДНК (Таблица 24: r=0.69, $\alpha<0.025$).

Вторым по убыванию U=0.351 было содержание динуклеотида TV (IUPAC-IUB, 1971) с наибольшим весом в центре олигоДНК, показанным

Таблица 24 – СродствоТВР к нити ДНК длины 15 нт (Соколенко и др, 1996)

№	нить ДНК, онДНК	сродство ТВР/онДНК, $-\ln[K_D]$	$[WR]_{(--)}$		$[TV]_{\{\_\}}$		Этап
1	gcgcccTATActacc	24.08	2.27		2.15		о
2	ggtagTATAgggcgc	24.08	2.12		2.34		б
3	gcgcccTgTActacc	21.50	2.27		2.15		у
4	ggtagTAcAgggcgc	21.50	1.45		1.37		ч
5	cgcccaaacccTATA	20.72	2.90		0.38		е
6	TATAgggtttgggcg	20.72	1.93		1.25		н
7	gttttttttttcgcg	16.60	0.00		0.48		и
8	cccccccccccccc	11.78	0.00		0.00		е
9	cgccc**TATAAAA**ccc	24.23		3.61		1.92	
10	gggtttTATAgggcg	24.23		2.16		1.87	
11	gcgcccTATgctacc	21.17		2.27		2.15	к
12	ggtagcATAgggcgc	21.17		1.72		1.38	о
13	cggctTATAtAAgcc	21.06		3.80		2.63	н
14	ggctTATAtAAgccg	21.06		3.33		2.71	т
15	TATAAAAcccagcgg	17.73		2.64		0.54	р
16	ccgctgggtttTATA	17.73		2.21		1.21	о
17	ggtggcgcacgcctg	16.27		1.14		0.51	л
18	agagttcaagacgat	15.76		2.37		0.95	ь
19	catggcggcggggcg	11.87		0.14		0.41	
	Коэффициент линейной корреляции, r=		0.82	0.69	0.84	0.76	
	Статистическая значимость, α<		0.025	0.025	0.01	0.001	

непрерывной линией на Рисунке 49, $[TV]_{\{\_\}}$. Два TA и TG из трех динуклеотидов TV (кроме TC) входят в общепринятый TATA(a/t)A(a/t)g консенсус TATA-бокса. Величины $[TV]_{\{\_\}}$ достоверно коррелировали с величинами $-\ln[K_D]$ сродства ТВР/онДНК на независимых контрольных данных (Таблица 24: r=0.76, α<0.01).

Существенно, что корреляция между $[WR]_{(--)}$ и $[TV]_{\{\_\}}$ оказалась недостоверной (r=0.04, α>0.25, данные не показаны). Поэтому с помощью стандартного пакета Statistica (Statsoft™, Tulsa, USA) была построена линейная регрессия количественных величин сродства ТВР/онДНК по содержанию динуклеотидов TV и WR в нити ДНК длины 15 нт:

$$-ln[K_D(s_1 \ldots s_{15})] = 14.53 + 2.53[TV]_{(-)}(s_1 \ldots s_{15}) + \\ +0.87[WR]_{(--)}(s_1 \ldots s_{15}). \quad (39)$$

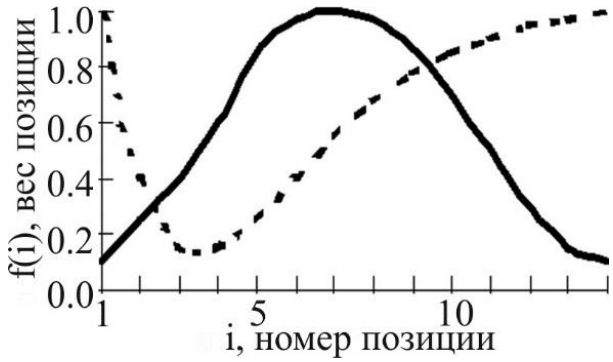

Рисунок 49 - Примеры весов f(i) динуклеотидов в позициях i нити ДНК длиной 15 нт для оценки их линейно-аддитивного вклада в сродство TBP/ДНК-, $-\ln[K_D]$.

На независимом контроле прогнозы формулы (39) достоверно (Рисунок 50: r=0.79, $\alpha<0.05$, ●) коррелируют с измерениями *in vitro* сродства TBP/онДНК, $-\ln[K_D]$, (Соколенко и др., 1996). Это означает, что формула (39) объясняет 62% вариансы экспериментальных данных (Соколенко и др., 1996), то есть соответствует точности этих измерений методом “задержки в геле”, для которой является общепринятой оценка ±25% от диапазона измеренных величин.

Таким образом, использование системы Activity (Ponomarenko M. *et al*., 1997a) позволило выявить количественные характеристики последовательности ДНК TATA-бокса, величины которых достоверно линейно коррелируют с величинами количественных характеристик связывания TBP с TATA-боксом.

3.2.2 Верификация результатов системы Activity для сродства TATA-связывающего белка к нитям ДНК

Для проверки результатов (Рисунок 50) анализа Activity измерений *in vitro* сродства TBP к синтетическим нитям ДНК длиной 15 п.о., $-\ln[K_D]$, были проанализированы 776 фрагментов ДНК [-250; +100] относительно старта транскрипции промоторов генов эукариот из базы данных EPD, вып. 45 (Perier

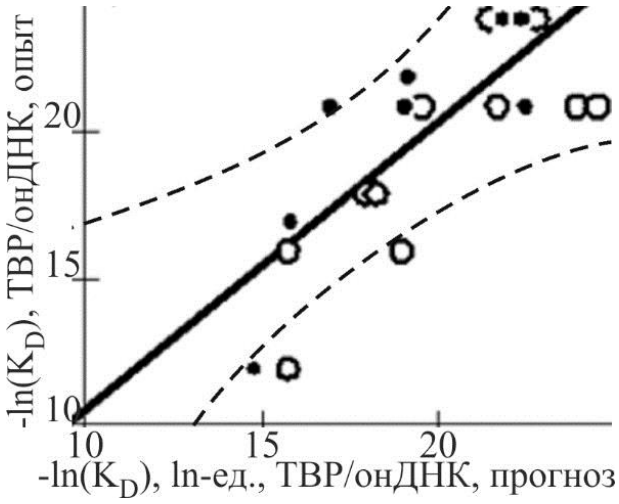

Рисунок 50 – Корреляция между прогнозом (формула 39) и измерением *in vitro* (Соколенко и др., 1996) сродства TBP/онДНК; ○ – обучение, ● – контроль, r=0.79, $\alpha < 0.05$; пунктир – границы 95% доверительных интервалов для контроля.

et al., 1999). Для каждого участка длиной 15 нт с центром в позиции i по формуле (39) оценивалась величина сродства TBP к онДНК, которая усреднялась по всем промоторам. Полученные результаты показаны на Рисунке 51 жирной линией, которая имеет единственный пик в позиции -30, соответствующей общепринятой оптимальной локализации TATA-бокса. Тонкой линией на этом рисунке показан результат аналогичных расчетов для 500 случайных последовательностей независимых равновероятных нуклеотидов A, T, G и C: сродство TBP к случайным ДНК было больше такового для природных ДНК везде кроме пика в позиции -30.

В целях интерпретации этого явления был сгенерирован миллион последовательностей S^+ из 35 случайных равновероятных независимых нуклеотидов A, T, G и C. Из этого миллиона вариантов с помощью формулы (39) была выбрана сотня вариантов с наибольшим превышением Δ оценок сродства TBP/онДНК на участке [11; 25] в центре сгенерированной S^+, над наибольшей из его оценок на флангах [1; 15] и [21; 35] нити S^+ и на комплементарной ей нити S^-:

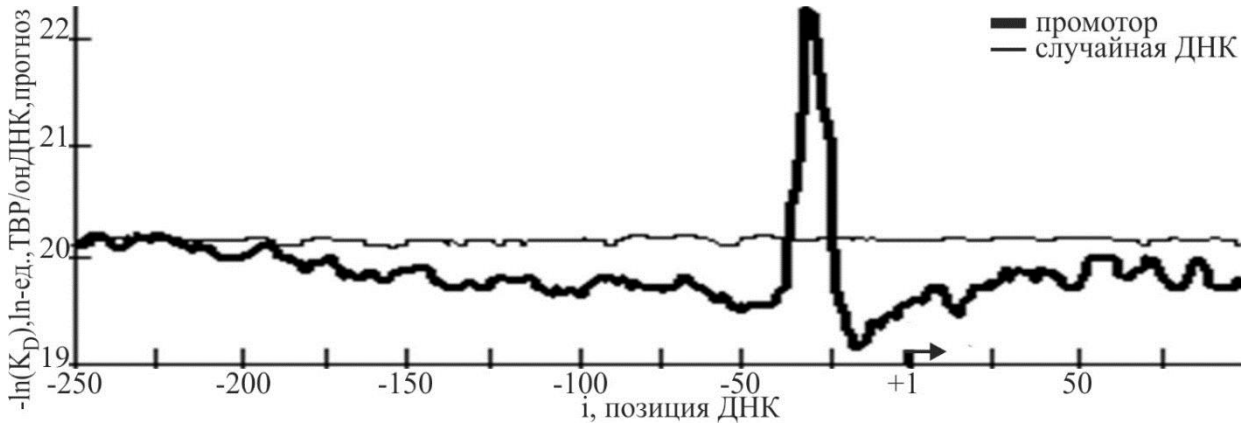

Рисунок 51 - Прогноз сродства ТВР/онДНК (формула 39) к 776 природным промоторам негомологичных генов из базы данных EPD, вып. 45 (Perier *et al*., 1999) (жирная линия) и к 500 последовательностям случайных независимых равновероятных нуклеотидов (тонкая линия).

$$\Delta = -ln\left[K_D\left(S^{+}_{[11;25]}\right)\right] - \\ - \underset{x\in\{[1;15];[21;35]\}}{\mathrm{MAX}}\left\{-ln[K_D(S^{+}_{x})]; -ln[K_D(S^{-}_{x})]; -ln\left[K_D\left(S^{-}_{[11;25]}\right)\right]\right\}. \quad (40)$$

Как можно видеть, формула (40) соответствует жирной кривой на Рисунке 51. Отобранные по ней 100 модельных ДНК охарактеризованы в Таблицах 25 и 26, а также на Рисунках 46 и 47.

Прежде всего, в Таблице 25 представлены 10 лучших из 100 отобранных случайных ДНК с наибольшим сродством ТВР/онДНК в центре, которое в наибольшей степени превышало его оценки ко всем пяти остальным участкам этих ДНК (формула 40). Видно, что варианты № 7 и № 10 содержат по одному “TATA”; № 2, 6 и 9 - по два перекрывающихся; варианты № 1, 3, 4 и 5 - по два тандемных; вариант №8 не содержит “TATA”. Всего 80 из 100 случайных ДНК, отобранных по формуле (40), содержали “TATA”, 20 остальных - нет (данные не показаны).

В двух последних строках Таблицы 25 представлен консенсус (номенклатура IUPAC-IUB, 1971) при уровне значимости $\alpha < 0.01$ (биномиальное распределение) для 100 отобранных по формуле (40) случайных ДНК в сравнении с общепринятым консенсусом природных TATA-боксов (Bucher, 1990). Между этими двумя консенсусами наблюдается 13 из

Таблица 25 – Результат отбора (формула 40) 100 из 10^9 случайных ДНК длины 35 нт на превышение сродства TBP к центру нити ДНК над концами

№	-ln[$K_D(S^+_{[11;25]})$]	10 лучших из 100 модельных ДНК(формула 40)
1	26.74	ggaagcgccgctca**TATATATA**tggcccgagccac
2	26.66	cagaaaaaccgtct**TATATAAAA**gaaggcccttca
3	26.00	cgatggccgccca**TATATATA**caggcagcccggtg
4	26.00	cgacgcgccccca**TATATATA**tgggggcaatggtg
5	25.83	gggcccgaagtgct**TATATATAA**gcccggcgacgt
6	25.73	cgccgggccggcct**TATATA**tGagggcgttttcac
7	25.54	cagaagccgcgtcgttc**TATAgAA**ggccgcggggg
8	25.14	gaacccgcgcccc**TAT**ct**TA**gaaaggcggcgctcg
9	24.79	aagccgaacgggc**TATATA**ctgagcgccgggggat
10	22.73	tccccggcggctttg**TATAA**cAgcagccccgcgtg
Консенсус	100 отобранных (Bucher, 1990)	---c-gssssccCTTTWWWWAAAGSSSssssc—g ----------STWTAWADRSSSSSS--------

15 возможных совпадений, что является достоверным сходством между ними ($\alpha < 0.005$, биномиальное распределение).

В свою очередь, участки [1;10], [11; 25] и [26;35] из 100 модельных (формула 40) ДНК сравнили с участками [-46; -37], [-36; -22] и [-21; -12] из 776 промоторов негомологичных генов эукариот из базы данных EPD, вып. 45 (Perier *et al*., 1999) по частотам нуклеотидов в них. Полученные результаты представлены на Рисунке 52. На этом рисунке можно видеть, что частоты нуклеотидов в модельных и в природных TATA-боксах коррелируют достоверно ($r=0.91$, $\alpha<0.01$).

Кроме того, с помощью формулы (28) общепринятого критерия TATA-бокса (Bucher, 1990) сравнили результаты распознавания TATA-боксов в 100 модельных ДНК (формула 40) и в 776 природных промоторов эукариот из базы данных EPD, вып. 45 (Perier *et al*., 1999). Полученные результаты представлены в Таблице 26 и на Рисунке 53. На этом рисунке можно видеть достоверную ($r=0.98$, $\alpha<0.01$) корреляцию между оценками частот распознавания TATA-бокса (формулы 23) с помощью общепринятого для него критерия (Bucher, 1990) в 100 случайных ДНК, отобранных с помощью

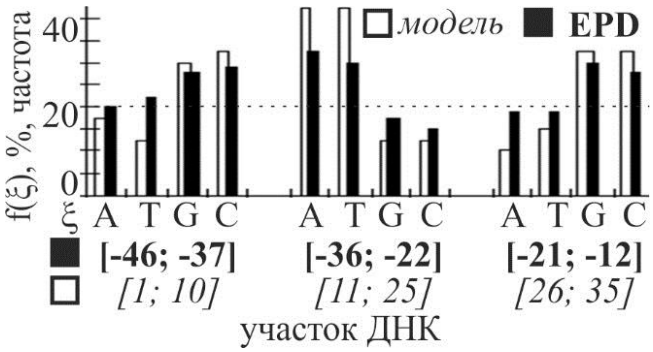

Рисунок 52 - Достоверная (r=0.91, α<0.01) корреляция между частотами встречаемости нуклеотидов в 100 отобранных по формуле (40) случайных ДНК (□, *курсив*) и в 776 промоторах негомологичных генов эукариот (■, **жирный шрифт**) из базы данных EPD, вып. 45 (Perier *et al*., 1999).

формулы (40), и в 776 природных промоторов, взятых из базы данных EPD, вып. 45 (Perier *et al*., 1999).

Наконец, в Таблице 26 можно видеть, что общепринятый критерий TATA-бокса (Bucher, 1990) распознал (формулы 27-28) его в 73% природных промоторах эукариот (Perier *et al*., 1999) и в 67% модельных ДНК (формула 40). Это различие результатов распознавания TATA-бокса было недостоверным как по критерию Фишера (α>0.12), так и по биномиальному распределению (α>0.07).

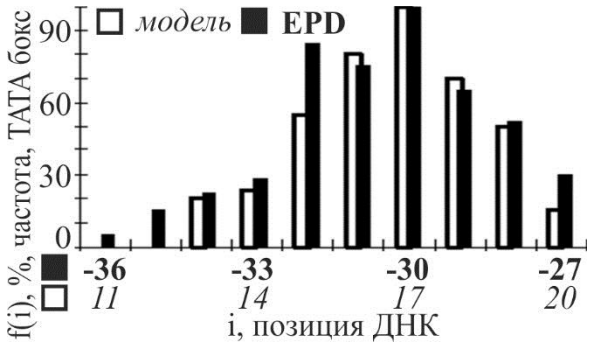

Рисунок 53 - Корреляции (r=0.98, α<0.01) частот P(i) распознавания TATA-бокса (формула 23) в позициях 776 промоторов из базы данных EPD, № 45 (Perier *et al*., 1999) (■, **жирный шрифт**), и в 100 модельных (формула 40) ДНК (□, *курсив*).

Таблица 26 – Распознавание ТАТА-бокса по критерию (Bucher, 1990) в 100 модельных ДНК (формула 40) и в 776 промоторах из EPD (Perier *et al*., 1999)

Набор последовательностей ДНК	ТАТА-бокс		критерий Фишера	биномиальное распределение
	число	%		
100 случайных ДНК, формула (40)	67	67		
776 промоторов из базы данных EPD (Perier *et al*., 1999)	567	73	$\alpha>0.12$	$\alpha>0.07$

Таким образом, ни точный критерий Фишера, ни биномиальное распределение, ни корреляция Пирсона не смогли различить результаты распознавания ТАТА-бокса с помощью общепринятого для него (формулы 27-28) критерия (Bucher, 1990) в 776 природных промоторах из базы данных EPD, вып. 45 (Perier *et al*., 1999) и в 100 отобранных с помощью формулы (40) случайных ДНК. Это означает, что контекстно-зависимые количественные характеристики сайтов в составе геномных ДНК, которые были выявлены с использованием компьютерной системы Activity (Ponomarenko M. *et al*., 1997a), созданной в рамках настоящей диссертации, адекватно учитывают связывание TBP с нитями ДНК ТАТА-бокса.

3.3 Сродство ТАТА-связывающего белка к двунитевым олигонуклеотидам ДНК

Согласно опыту (Coleman, Pugh, 1995), связыванию TBP с ТАТА-боксом предшествует скольжение TBP вдоль B-формы спирали ДНК в силу неспецифического сродства между ними, 10^{-5} M (Hahn *et al*., 1989). Для учета этого этапа TBP/ТАТА-связывания в настоящем разделе с помощью системы Activity (Ponomarenko M. *et al*., 1997a) были выявлены контекстно-зависимые количественные характеристики двунитевой ДНК ТАТА-бокса, коррелирующие со сродством TBP дрожжей к двунитевым олигоДНК, измеренным методом “задержки в геле” (Savinkova *et al*., 1998). На их основе с помощью стандартного пакета Statistica (Statsoft™, Tulsa, USA) была

Таблица 27 – Сродство TBP к олигоДНК длины 15 п.о. (Savinkova *et al*,1998)

№	прямая нить ДНК	комплементарная нить ДНК	K_D,$\times10^{-10}$ M	$-\ln[K_D]$
1	cgcccTATAAAAccc	gggTTTTATAgggcg	0.060	23.54
2	cggcTTATATAAgcc	ggcTTATATAAgccg	0.421	21.59
3	gcgcccTATAcTAcc	ggTAgTATAgggcgc	0.056	23.61
4	gcgcccTgTAcTAcc	ggTAgTAcAgggcgc	0.540	21.34
5	gcgcccTATgcTAcc	ggTAgcATAgggcgc	0.781	20.97
6	cgcccAAAcccTATA	TATAgggTTTgggcg	1.390	20.39
7	TATAAAAcccAgcgg	ccgcTgggTTTTATA	66.240	16.53
8	ggTggcTcAcgccTg	cAggcgTgAgccAcc	90.310	16.22
9	AgAgTTcAAgAccTg	cAggTcTTgAAcTcT	95.900	16.16
10	cccccccccccccc	gggggggggggggg	8895.000	11.63

выведена формула для прогноза величин сродства TBP/днДНК по последовательности ДНК и показана достоверная корреляция между прогнозом этой формулы и независимым опытом.

В Таблице 27 можно видеть величины (K_D) от 6×10^{-12} M до 9×10^{-7} M сродства *in vitro* TBP дрожжей к 10 двунитевым олигоДНК длины 15 п.о., днДНК (Savinkova *et al*., 1998). Результат их обработки с помощью созданной в рамках настоящей диссертации системы Activity (Ponomarenko M. *et al*., 1997a) представлен на Рисунке 54. Среди среднеарифметических контекстно-зависимых характеристик B-формы спирали ДНК TATA-бокса наивысшую оценку полезности $U(X_{8;[6;9]})$ =0.37 получила (формула 29) ширина малой бороздки между позициями 6 и 9 в центре олигоДНК. Она достоверно ($r=0.95$, $\alpha<10^{-4}$) коррелирует со сродством TBP к днДНК (Рисунок 54а). Этот результат согласуется с известными 3D-структурами комплекса TBP/ДНК (Kim J. *et al*., 1993; Kim Y. *et al*., 1993; Jou *et al*., 1996), согласно которым TBP связывает ДНК путем ввода боковых групп фенилаланина в малую бороздку спирали ДНК на краях консенсуса TATA-бокса.

Среди содержаний динуклеотидов (формула 38, m=2) наивысшую оценку $U([TA]_f)=0.26$ получило содержание динуклеотида TA в 3'-половине олигоДНК (Рисунок 54б,в). Величины $[TA]_f$ достоверно ($r=0.80$, $\alpha<10^{-3}$) коррелируют с величинами $-\ln(K_D)$ сродства TBP/ДНК, Рисунок 54б.

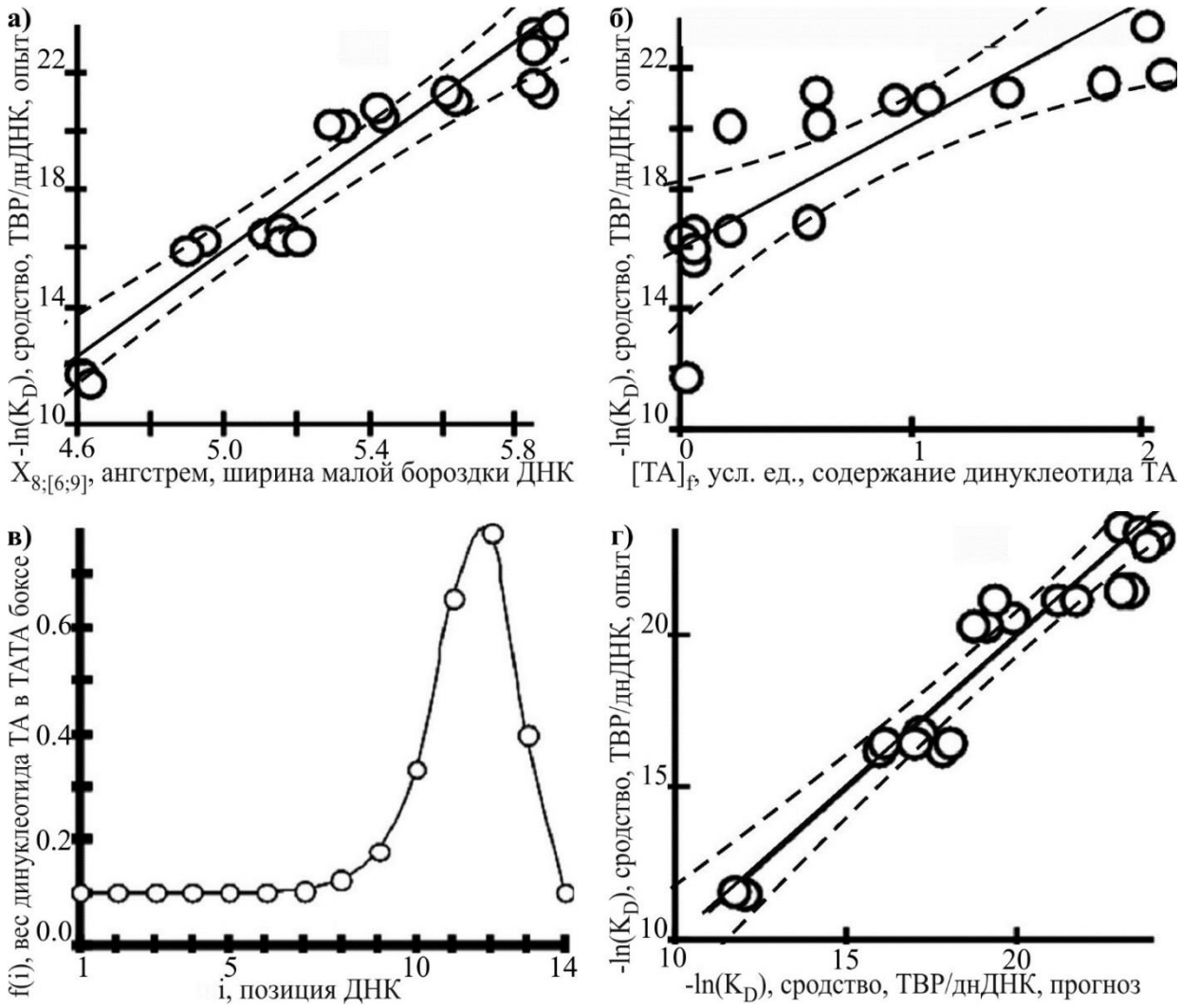

Рисунок 54 – Результат Activity (Ponomarenko M. *et al*., 1997a) для 10 двунитевых олигоДНК с измеренными *in vitro* величинами, -ln[K_D], сродства TBP дрожжей к ним в опыте (Savinkova *et al*., 1998). Достоверные корреляции между сродством TBP/днДНК (ось y) и (а) средней шириной малой бороздки спирали ДНК в центре олигоДНК; (б) с содержанием динуклеотидов TA с наибольшим весом в 3'-половине олигоДНК (в); (г) с регрессией (формула 41) сродства TBP/днДНК по средней ширине малой бороздки ДНК и взвешенному содержанию TA. Пунктир – границы 95%-доверительных интервалов.

На основе этих двух корреляций с помощью стандартного пакета Statistica (Statsoft™, Tulsa, USA) была построена линейная регрессия ($r=0.96$, $\alpha<10^{-4}$) сродства TBP/днДНК:

$$-ln[K_D(s_1 \dots s_{15})] = -35.13 + \\ +10.21X_{8[6;9]}(s_1 \dots s_{15}) - 0.72[TA]_f(s_1 \dots s_{15}). \quad (41)$$

Отрицательный коэффициент -0.72 в формуле (41) соответствует негативному вкладу динуклеотидов ТА вблизи 3’-конца олигоДНК длины 15 п.о. в величину сродства TBP/днДНК, вопреки позитивной корреляции этих величин (Рисунок 54б). Прежде всего, консенсусу TATA-бокса соответствуют динуклеотиды ТА в центре олигоДНК, а не в его 3’-половине. Это негативное влияние смещенных в 3’-направлении динуклеотидов ТА соответствует выводам авторов работы (Jou *et al*., 1996) по анализу известных 3D-структур TBP/TATA-комплексов, что самая большая ширина 6.4Å малой бороздки спирали ДНК (Karas *et al*., 1996) у динуклеотида ТА превышает в 3’-части TATA-бокса оптимум ширины этой бороздки для ввода боковой группы фенилаланина.

Прогноз формулы (41) проверили с помощью контрольных данных независимого опыта (Wiley *et al*., 1992) по измерению величин TBP/ДНК-сродства. Полученный результат показан на Рисунке 55. Как можно видеть на

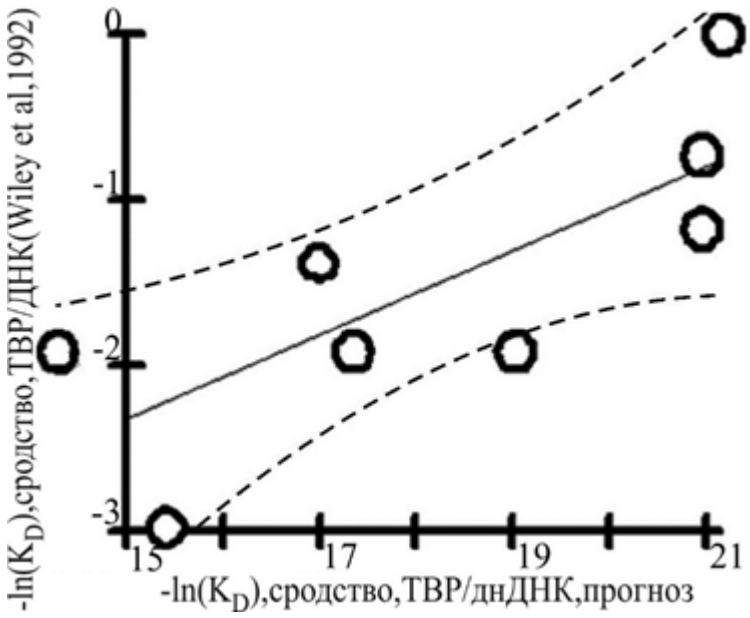

Рисунок 55 – Проверка формулы (41) на независимых данных (Wiley *et al*., 1992). Пунктир – границы 95%-доверительных интервалов.

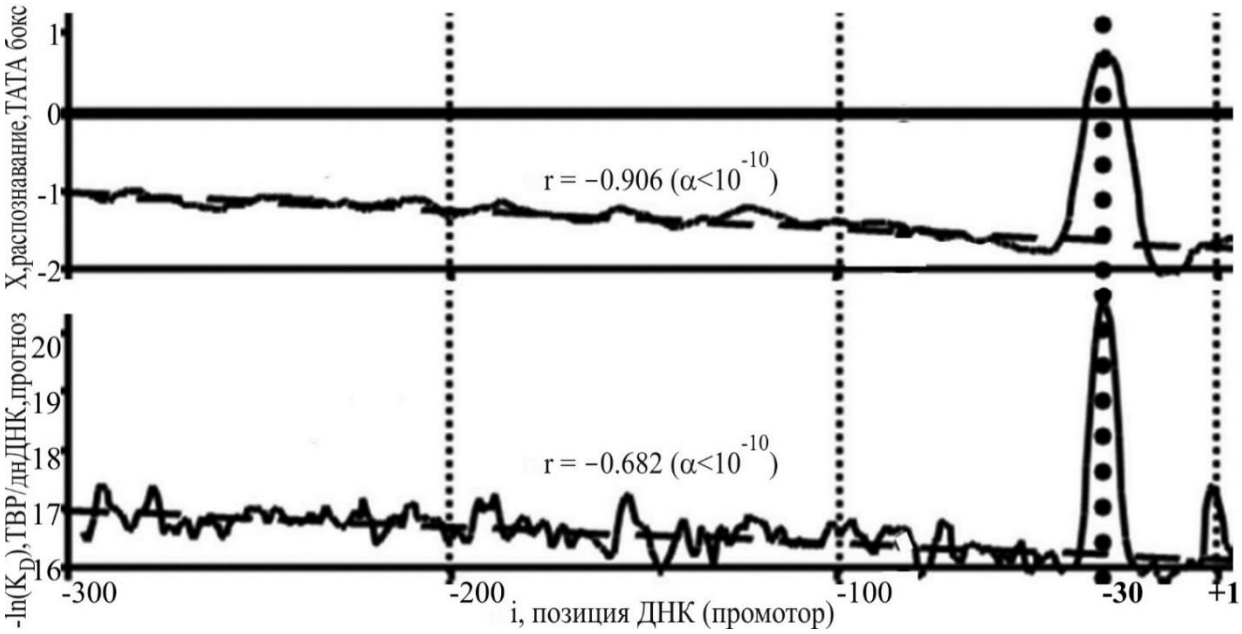

Рисунок 56 – Прогноз систем bDNAvideo (сверху) и Activity (снизу) для 500 негомологичных промоторов эукариот. Рисунок автора на основе марериалов его статьи (Kolchanov *et al*., 1999).

этом рисунке, отложенные по оси x предсказания формулы (41) достоверно (r= 0.76, $\alpha<0.05$) коррелируют с экспериментальными данными, ось y.

В качестве заключения сравним (Рисунок 56) результаты двух созданных в рамках настоящей диссертации компьютерных систем: bDNAvideo (сверху) и Activity (снизу), усредненные по 500 участкам ДНК длиной 300 п.о. перед главными стартами транскрипции негомологичных генов эукариот (Kolchanov *et al*., 1999). Видно, что оба предсказания имеют максимальные пики в оптимальной для ТАТА-бокса позиции -30 относительно старта транскрипции. При этом они оба имеют достоверные негативные тренды, r= −0.91 ($\alpha<10^{-10}$) и r= −0.68 ($\alpha<10^{-10}$), в районе [-300; -30] непосредственно перед этой позицией, которые соответствуют скольжению TBP вдоль ДНК (Coleman, Pugh, 1995) в направлении к этой оптимальной позиции ТАТА-бокса. Оценка протяженности скольжения TBP от места его случайной посадки на ДНК до его остановки на ТАТА-боксе, 300 п.о. совпала с оценкой 240 ± 105 п.о. недавнего независимого экспериментально-компьютерного исследования (Marklund *et al*., 2013).

Таким образом, количественные характеристики последовательности ДНК ТАТА-бокса, выявленные обеими созданными в диссертации системами Activity (Ponomarenko M. *et al*., 1997a) и bDNAvideo (Ponomarenko M. *et al*., 1997b) на основе независимых экспериментальных данных, взаимно соответствуют друг другу в адекватном учете скольжения TBP вдоль ДНК до его остановки на ТАТА-боксе при образовании TBP/TATA-комплекса, как указано в главе 1, Рисунок 5.

3.4 Эмпирическое уравнение связывания TBP с ТАТА-боксом

В предыдущих разделах диссертации были выведены формулы (39) и (41) для оценки сродства TBP/онДНК и TBP/днДНК, соответственно, по последовательности ДНК длины 15 п.о. на основе анализа системой Activity (Ponomarenko M. *et al*., 1997a) данных *in vitro* о сродстве TBP дрожжей к одно- (Соколенко и др., 1996) и двунитевым (Savinkova *et al*., 1998) олигоДНК. В настоящем разделе на основе новых измерений *in vitro* сродства TBP человека к двунитевым олигоДНК длины 26 п.о., содержащим ТАТА-боксы генов человека (Савинкова и др., 2007), обе эти формулы были объединены эмпирическим уравнением трехшагового TBP/TATA-связывания (Пономаренко П. и др, 2008), что нашло подтверждение в опыте *in vitro* (Delgadillo *et al*., 2009).

В Таблице 28 можно видеть величины от 18.02 до 21.54 ln-ед. сродства *in vitro* ($-\ln(K_{D;TBP/TATA})$) рекомбинантного TBP человека к 10 синтетическим олигоДНК длиной 26 п.о., идентичных фрагментам ТАТА-содержащих промоторов генов протеиназы-3 (GenBank: AF015446), АТФ-связывающей кассеты 1A1 (AF287262), одорант-связывающего белка (J251025), эластазы-2 (AY596461), циклооксигеназы (AF044206), фактора Phox2A (AJ320270), α2-глобина (Z84721) и фактор роста гепатоцитов А (AY246560) человека, кристаллина αB крысы (U04320) и цитокина мыши (X70058), которые были

Таблица 28 – ОлигоДНК длиной 26 п.о., идентичные природным ТАТА-боксам генов человека, крысы и мыши, а также величины ($-\ln(K_D)$) сродства ТВР человека к этим олигоДНК, измеренные *in vitro* (Савинкова и др., 2007), в сравнении с критерием Бухера для ТАТА-бокса (Bucher, 1990), PWM (формула 28), оценками $-\ln(K_{D;онДНК})$ и $-\ln(K_{D;днДНК})$ сродства ТВР к одно- и двунитевым олигоДНК (формулы (39) и (41)), соответственно.

измерения *in vitro* (Савинкова и др., 2007)			оценки *in silico* (формулы)			прог-ноз
№	последовательность олигоДНК (ТАТА-бокс – ЗАГЛАВНЫЙ шрифт)	сродство	PWM (28)	$-\ln(K_{D;онДНК})$ (39)	$-\ln(K_{D;днДНК})$ (41)	(43)
1	agactgcATATATAAggggcaggctg	21.54	-2.48	24.47	22.29	20.59
2	cggctgccTATAAAAgaggaggcaga	21.34	-1.79	22.26	21.00	19.95
3	tggcattgggcTATAAgaggagcttg	19.34	-4.12	22.31	17.77	18.86
4	agccgaatcTATAAAAggaactagtc	20.25	-1.95	22.16	20.61	19.82
5	gccaggggcTATAAAgaacatctcg	18.97	-4.15	21.58	19.26	19.05
6	agcacagggcTATAAgaggagccggg	18.42	-3.86	22.44	19.20	19.25
7	gttttcagtcTTATAAAAAggaaggg	18.93	-2.74	23.18	20.29	19.83
8	cctgcgcgTTAAAAggcgcgcgggcc	18.24	-5.41	19.20	20.44	18.65
9	cgccccaagcATAAAccctggcgcgc	18.02	-8.66	21.24	18.78	18.18
10	ctcatcgcAATAAAAAgcagctcaga	18.24	-4.10	21.44	19.39	19.06
	Коэффициент корреляции, r		0.72	0.64	0.69	0.840
	(значимость, α)		(0.025)	(0.05)	(0.05)	(10^{-3})

Закономерность № 1	Закономерность № 2	r	α >
Критерий ТАТА-бокса (23)	ТВР/онДНК-сродство (39)	0.554	0.09
Критерий ТАТА-бокса (23)	ТВР/днДНК-сродство (41)	0.547	0.10
ТВР/онДНК-сродство (39)	ТВР/днДНК-сродство (41)	0.360	0.30

измерены экспериментально в равновесных условиях метода “задержки в геле”(Савинкова и др., 2007).

В этой таблице показаны также три контекстно-зависимые количественные характеристики этих олигоДНК, достоверно коррелирующие с этими экспериментальными данными: максимум общепринятого критерия Бухера (Bucher, 1990) для ТАТА-бокса (формулы 27-28) в случае обеих нитей ДНК, локализованный в позиции i олигоДНК (PWM_i, $r=0.72$, $\alpha<0.025$; $8 \le i \le 18$); среднее сродство ТВР к каждой из нитей ДНК (формулы 39) в позиции i максимума критерия Бухера, ($-\ln(K_{D;онДНК;[i;i]}$, $r=0.64$, $\alpha<0.05$);

среднее сродство TBP к обеим нитям олигоДНК (формула 41) ($-\ln(K_{D;\text{днДНК};[8;18]}$, r=0.69, α<0.05), вычисленные как:

$$-ln\left[K_{D,\xi;[a;b]}\right] = \frac{1}{2(b-a+1)}\sum_{i=a}^{b}\left(-ln\left[K_{D,\xi}\left(S^{+}_{[i-7;i+7]}\right)\right] - ln\left[K_{D,\xi}\left(S^{-}_{[i-7;i+7]}\right)\right]\right), \quad (42)$$

здесь: ξϵ{онДНК; днДНК}; $S^{+}[\zeta;\eta]$ и $S^{-}[\zeta;\eta]$ – две нити участка [ζ;η] олигоДНК.

Тем не менее, прогнозы этих формул не коррелируют между собой (Таблица 28, внизу), хотя каждый из них по отдельности значимо коррелирует с экспериментальными данными из работы (Савинкова и др., 2007).

Это противоречие было преодолено эмпирическим уравнением связывания TBP с TATA-боксом за три последовательных шага (Пономаренко П. и др., 2008):

$$-ln\left[K_{D;TATA}\right] = 10.90 - 0.23ln\left[K_{D;\text{д};[8;18]}\right] + \quad (43)$$
$$+0.15PWM_{i''} - 0.20ln\left[K_{D;\text{о};[i'';i'']}\right].$$

Результаты применения формулы (43) для 10 олигоДНК с природными TATA-боксами (Савинкова и др., 2007) показаны в Таблице 28 и на Рисунке 57. Как можно видеть, прогноз (формула 43) и измерения сродства TBP/TATA, $-\ln[K_{D;TATA}]$, коррелируют достоверно: r=0.84, α<0.0025.

С помощью стандартного статистического пакета Statistica (Statsoft™, USA) было также установлено, что общепринятый критерий (Bucher, 1990) TATA-бокса объясняет 33% вариансы данных опыта (Савинкова и др., 2007), поскольку он не учитывает ни скольжение TBP вдоль ДНК (Coleman, Pugh, 1995), ни стабилизацию TBP/TATA-комплекса за счет изомеризации ДНК (Kim J *et al*., 1993; Kim Y *et al*., 1993; Powell *et al*., 2002). В разделах 3.2 и 3.3 настоящей диссертации были выведены формулы (39) и (41) для эмпирического учета этих стадий связывания TBP/TATA-бокс, объяснившие 38% вариансы данных (Савинкова и др., 2007)

Для верификации формулы (43) с ее помощью были предсказаны величины сродства TBP к 27 вариантам TATA-бокса генов человека (Таблица 29). Эти величины были независимо измерены методом “задержки в

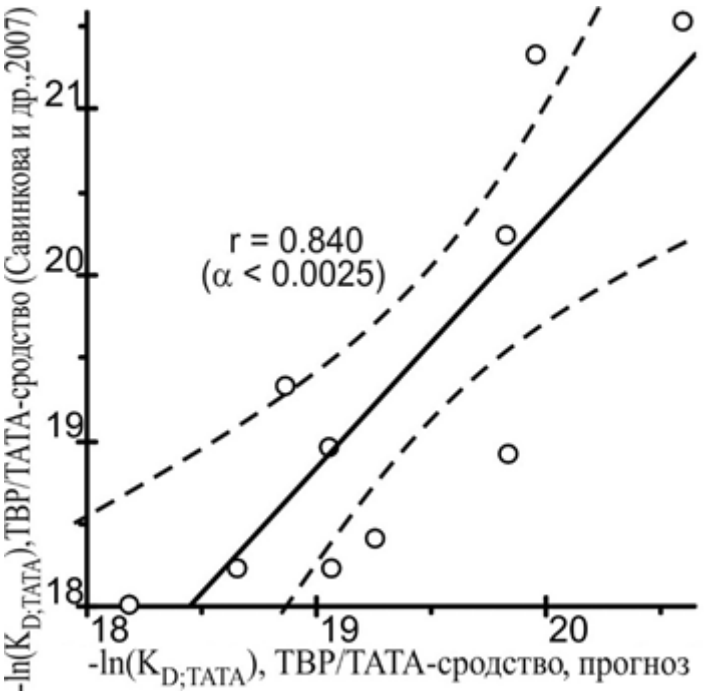

Рисунок 57 – Достоверная корреляция (r=0.84, α<0.0025) между предсказанным (Пономаренко П. и др., 2008) по формуле (43) и экспериментальным сродством TBP человека к олигоДНК с TATA-боксами генов человека (Савинкова и др., 2007). Пунктир – границы 95%-доверительных интервалов.

геле" в равновесных условиях *in vitro* опыта (Savinkova *et al*., 2013), а также, дополнительно, для 17 из этих 27 вариантов, в неравновесных условиях *in vitro* опыта (Drachkova *et al*., 2014).

Сравнение результатов этих опытов с прогнозом формулы (43) показано в Таблице 29 и на Рисунке 58. Все корреляции между предсказанными и экспериментальными оценками сродства TBP/TATA-бокс были достоверными как в случае равновесных (r=0.82, $\alpha<10^{-7}$, (Savinkova *et al*., 2013)), так и в случае неравновесных (r=0.84, $\alpha<10^{-6}$, (Drachkova *et al*., 2014)) условий *in vitro* метода "задержки в геле". Таким образом, в рамках диссертации были впервые предсказаны величины сродства TBP человека к природным TATA-боксам промоторов генов человека, подтвержденные с использованием независимых экспериментальных измерений (Savinkova *et al*., 2013; Drachkova *et al*., 2014).

Таблица – 29 Проверка формулы (43) для предсказания величин TBP/TATA-сродства, $-\ln[K_D]$, в эксперименте *in vitro* (Savinkova *et al*., 2013)

последовательность, 26 п.о., ТАТА-бокс (ЗАГЛАВНЫЙ шрифт)	$-\ln[K_D]$		Литература
	прогноз	опыт	
cagggctgggCATAAAAgtcagggca	18.47	16.94	Fei *et al*, 1988
cagggctgggCgTAAAAgtcagggca	17.87	16.13	Takihara *et al*, 1986
cagggctgggCAaAAAAgtcagggca	17.01	15.97	Fei *et al*, 1988
cagggctgggCAcAAAAgtcagggca	17.49	16.01	Cai *et al*, 1989
cagggctgggCATgAAAgtcagggca	17.23	14.76	Antonarakis et al, 1984
cagggctgggCATAgAAgtcagggca	17.76	16.26	Orkin *et al*.,1983
cagggctgggCATAcAAgtcagggca	17.94	15.02	Poncz *et al*., 1982
cagggctgggCATAAtAgtcagggca	18.11	16.56	Badens *et al*, 1999
acaggaccagCATAAAAggcagggca	18.94	16.89	Frischknecht, Dutly, 2005
acaggaccagCgTAAAAggcagggca	18.33	15.94	
ctgccacacccacattattagaaaat	17.72	15.70	De Gobbi et al., 2006
ctgccacaccCACATTATCagaaaat	18.28	16.00	
cgcggcgctcTATATAAgtgggcagt	20.11	19.15	Watanabe *et al*, 1996
cgcggcgctcTATAgAAgtgggcagt	19.08	15.71	
atggggtgagTATAAATActtcttgg	19.85	20.14	Clark *et al*., 2003
atggggtgagTATAAATAcctcttgg	20.06	20.25	
tttcaggcagTATAAAggcaaaccac	19.87	17.89	Pitarque *et al*, 2001
tttcaggcagTAgAAAggcaaaccac	18.38	16.34	
aggtctggccTATAAAgtagtcgcgg	19.21	17.03	Niemann *et al*, 2007
aggtctggccTgTAAAgtagtcgcgg	18.04	15.59	
acagctcagcTTGTACTTTggtacaa	18.24	14.49	Reijnen et al., 1992
acagctcagcTTcTACTTTggtacaa	17.75	14.51	
ttttgaaagcCATAAAAAcagcgagg	18.67	17.36	Wu *et al*, 2010
ttttgaaagctATAAAAAcagcgagg	19.85	18.78	
catctatttcTATATAgcctgcaccc	19.68	19.39	Boldt *et al*., 2006
catctatttcTAcATAgcctgcaccc	18.57	16.66	
gccggcctttatagcgcgcggggca	18.91	16.45	Arnaud *et al*., 2000
gccggcccTTTATAGTgcgcggggca	19.43	17.47	
Линейная корреляция, r (значимость)	0.82 ($\alpha<10^{-7}$)		Savinkova *et al.*, 2013

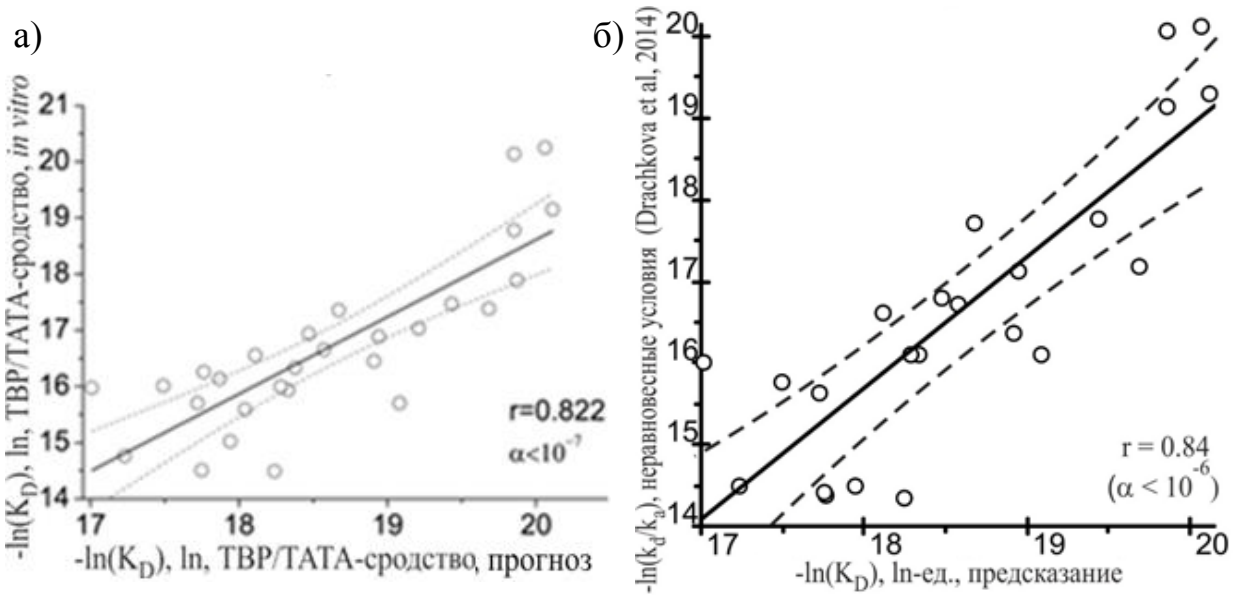

Рисунок 58 – Достоверные корреляции между предсказанным (формула 43) и независимо измеренным в (а) равновесных (Savinkova *et al*., 2013) и в (б) неравновесных (Drachkova *et al*., 2014) условиях *in vitro* методом “задержки в геле” сродством TBP к олигоДНК, представлявшим TATA-боксы генов человека. Пунктир – границы 95% доверительных интервалов. Рисунок автора на основе его статей (Savinkova *et al*., 2013) и (Drachkova *et al*., 2014).

Представленные в настоящей главе результаты позволяют сделать следующие выводы:

- На основе теории аддитивной полезности для принятия решений и нечетких множеств создана компьютерная система Activity для:
  - анализа выборок нуклеотидных последовательностей сайтов в составе геномной ДНК с известными величинами специфической биологической активности и выявления контекстных, а также контекстно-зависимых конформационных и физико-химических характеристик B-формы ДНК, достоверно коррелирующих с анализируемой активностью сайтов ДНК;
  - построения регрессионных уравнений для предсказания величин специфической биологической активности по произвольной последовательности сайта в составе геномной ДНК на основе

выявленных контекстных, а также контекстно-зависимых конформационных и физико-химических характеристик B-формы ДНК, коррелирующих с этой активностью.

- Выявлены достоверные корреляции равновесной константы диссоциации K_D TATA-связывающего белка (TBP) к олигоДНК длиной 15 нт с содержанием динуклеотида WR на флангах и динуклеотида TV в центральной части однонитевой ДНК, а также с содержанием динуклеотида TA в 3'-половине и шириной малой бороздки в центре дуплексов ДНК. На основе этих корреляций были впервые предсказаны величины равновесной константы диссоциации K_D комплекса TBP/ДНК, которые были подтверждены независимыми экспериментами.

ЗАКЛЮЧЕНИЕ ПО ГЛАВЕ 3

Представленные в этой главе настоящей диссертации результаты компьютерного анализа комплексов TBP с олигоДНК были опубликованы в конце 90-х годов XX века (Пономаренко М. и др., 1997а,б; Ponomarenko M. *et al*., 1999a), и нашли свое практическое применение через десять лет в рамках эмпирического регрессионного уравнения трехшагового связывания TBP с TATA-боксом для прогноза абсолютных и относительных величин сродства между ними (Пономаренко П. и др., 2008), а также в планировании экспериментов *in vitro* по верификация этих прогнозов (Savinkova *et al*., 2013; Drachkova *et al*., 2014). На основе использования этих данных настоящей диссертации другими авторами был получен ряд важных результатов. Прежде всего, был предложен молекулярный механизм трехшагового связывания белка TBP с TATA-боксом промоторов генов эукариот: (1) неспецифическое связывание и скольжение TBP вдоль ДНК; (2) первичное распознавание TBP-белком TATA-бокса; (3) стабилизация комплекса "TBP/TATA-бокс" путем изомеризации двойной спирали ДНК (Пономаренко П. и др., 2008). Годом позже этот механизм был подтвержден независимым экспериментом *in vitro*

(Delgadillo *et al*., 2009). На основе эмпирической трехшаговой модели связывания TBP/TATA-бокс (Пономаренко П. и др., 2008) были достоверно предсказаны результаты 69 независимых опытов (Пономаренко П. и др., 2010), биохимические манифестации более 100 ассоциированных с патологиями человека вариантов TATA-боксов генов (Пономаренко П. и др., 2009), а также вариантов TATA-боксов, ассоциированных с селекционно-ценными признаками растений и животных (Суслов и др., 2010). Кроме того, публикация (Савинкова и др., 2009) обширной коллекции ассоциированных с патологиями человека вариантов TATA-боксов, стимулировала создание полногеномной карты локализации 17181 потенциальных TATA-боксов в промоторах генов референсного генома человека в качестве платформы для изучения их роли в патогенезе (Yang *et al*., 2011).

В свою очередь, Рассказов и соавт. (2013б) создали в системе Интернет сервис SNP_TATA_Comparator для практического применения прогнозов изменения сродства TBP к мутантным вариантам TATA-боксов промоторов генов человека, который вошел в программный комплекс SNP-MED для анализа влияния полиморфизма на функцию генов человека, связанных с социально значимыми заболеваниями (Подколодный и др., 2013). С его помощью были предсказаны кандидатные SNP-маркеры для широкого круга патологий человека, в том числе для гендер-зависимых аутоиммунных заболеваний (Ponomarenko *et al.*, 2016) и для заболеваний, ассоциируемых с избыточным весом (Arkova et al., 2015).

В свою очередь, в работе (Bazykin, Kondrashov, 2006) на дрожжах было открыто переключение генов-гомологов между классами TATA-содержащих и TATA-несодержащих промоторов в процессе их дивергенции от общего предка, биологическая природа которого осталась неясной вследствие слабой изученности таких генов у дрожжей. На основе использования результатов этой главы настоящей диссертации Миронова и соавт. (Mironova *et al*., 2010) показали аналогичное эволюционное переключение у растений в семействе генов *ARF* транскрипционных факторов ответа на ауксин. Среди

млекопитающих нашли корреляцию между переключением регулируемых каскадом Hh передачи сигналов детерминации/дифференцировки генов *Gli1* и *Gli3* факторов транскрипционного контроля морфогенеза и переключением от r- к K-стратегий: от достижения эволюционного успеха за счет многочисленного потомства у грызунов к ограничению плодовитости и эволюционному успеху за счет воспитания ограниченного количества более приспособленных потомков у приматов, включая человека (Ponomarenko M. *et al*., 2013c). Кроме того, с использованием формулы (43) была выявлена линейная корреляция между вариабельностью экспрессии генов путей передачи сигналов в мозге человека и сродством TATA-связывающего белка к промоторам этих генов (Пономаренко М. и др., 2014). Наконец, в качестве расширения области применения этого результата на случай высших приматов, в работе (Gunbin *et al.,* 2017) было показано что человек как биологический вид характеризуется систематическим расширением нормы реакции генов, которые экспрессируются в мозге, с каждым шагом дивергенции его предков от общего предка высших приматов, тогда как в случаях орангутана, гориллы и шимпанзе норма реакции этих генов сужалась, как только каждый из этих высших приматов выделялся в отдельный биологический вид.

В целом, дополнение общепринятого критерия (Bucher, 1990) TATA-бокса достоверными регрессиями константы равновесия K_D комплекса «TBP/ДНК» в этой главе диссертации, дало импульс компьютерно-экспериментальному изучению TATA-боксов. Полученные результаты расширили представления о молекулярных механизмах инициации транскрипции генов эукариот, о патогенезе наследственных заболеваний человека и об эволюции геномов. Это увеличило достоверность прогноза биохимического проявления мутаций в промоторах генов человека до уровня, достаточного для постгеномной предиктивно-превентивной персонифицированной медицины. Компьютерный анализ количественных характеристик биологической активности сайтов в составе геномных ДНК

стал новым научным направлением биоинформатики и математической биологии, которое возникло накануне начала данной диссертационной работы и получило существенное развитие как непосредственно в ее рамках, так и в работах автора вне ее на основе использования результатов этой диссертации (Deplancke et al., 2016).

ГЛАВА 4 КОМПЬЮТЕРНАЯ СИСТЕМА ACTIVITY: ОЦЕНКА ВЛИЯНИЯ КОНТЕКСТА НА ЭФФЕКТИВНОСТЬ МУТАГЕНЕЗА ГЕНОМНОЙ ДНК

В настоящей главе на примере трех типов горячих точек мутагенеза представлены оценки границ применимости компьютерной системы Activity (Ponomarenko M *et al*., 1997a), создание которой было описано в предыдущей главе диссертации на примере связывания между TATA-связывающим белком (TBP) и TATA-боксом. Оценка границ применимости компьютерных систем является общепринятым обязательным разделом статей многих ведущих международных научных журналов по биоинформатике и математической биологии. Она заключается в оценке результатов нового вычислительного метода, предлагаемого для решения определенной задачи, для широкого круга других содержательных задач. Это может быть некоторым эквивалентом общепринятому требованию усреднять экспериментальные значения, измеренные на нескольких независимо приготовленных образцах, для уменьшения влияния неконтролируемых условий на результаты опытов. В заключении к этой главе можно также найти ряд дополнительных примеров применения компьютерной системы Activity (Ponomarenko M. *et al*., 1997a) к анализу связывания регуляторных районов геномов фагов и бактерий с их активаторами и репрессорами, а также к анализу пре-мРНК и микроРНК эукариот.

4.1 Количественные характеристики ДНК, коррелирующие с частотами повреждений гуанина лазерным ультрафиолетовым излучением с длиной волны 193 нм

Ультрафиолетовый (UV-) лазер используется в терапии рака (Thogersen *et al*., 2007), микрохирургии (Schastak *et al*., 2007), биолюминесценции в ходе генной терапии (Maloney *et al*., 2006) и UV-футпринтинге (Engelhorn *et al*.,

1995). Однако, в многочисленных экспериментах было установлено, что индуцированное лазером UV-облучение способно повреждать молекулы ДНК (Gurzadyan *et al*., 1993; Melvin *et al*., 1998; Angelov *et al*., 2005; Втюрина и др., 2011).

Самым частым UV-повреждением ДНК при длинах волн $\nu^{-1} \leq 290$ нм является образование 7,8-дигидро-8-оксогуанина, 2,2-диамино-оксозалона и других продуктов окисления гуанин-катиона (G^+) как предмутационным повреждениям ДНК, приводящим к заменам нуклеотидов (Kirpota *et al*., 2011) в отличие от повреждений ДНК другими мутагенами: например, ультразвуком (Grokhovsky *et al*., 2011). Эти наиболее частые UV-повреждения G общепринято ассоциировать с достаточностью энергии $h\nu$ поглощенного фотона (h – постоянная Планка) для выхода электрона (e^-) из ДНК с образованием "дырки" в ДНК (Gurzadyan *et al*., 1993; Melvin *et al*., 1998; Angelov *et al*., 2005). Блуждание этой "дырки" вдоль ДНК чаще всего завершается появлением катиона G^+ из-за самого низкого порога ионизации G в сравнении с A, T, и C (Saito *et al*., 1995) и немедленной атакой этого G^+ свободными радикал-анионами. Было показано, что частота UV-повреждений ДНК по гуанинам в отдельной нуклеотидной последовательности может варьировать в широких пределах, что указывает на ее зависимость от нуклеотидного контекста ДНК вокруг гуанинов (Gurzadyan *et al*., 1993; Melvin *et al*., 1998; Angelov *et al*., 2005). Однако эта зависимость все еще остается неизвестной. Вопрос о систематическом компьютерном анализе характеристик нуклеотидного контекста ДНК, влияющих на частоту UV-повреждения гуанинов, до сих пор даже не ставился.

В настоящем разделе был впервые осуществлен (Втюрина и др., 2012) систематический *in silico* анализ контекстных характеристик ДНК, влияющих на частоту повреждений гуанинов при UV-облучении лазером. Для UV-лазера с длиной волны 193 нм были ранее (Втюрина и др., 2011) измерены частоты $F(s_i)$ повреждений нуклеотидов $s_i \in \{A, T, G, C\}$ в каждой позиции i от 146 до 461 плазмиды pGEM7(f+) *Escherichia coli* и установлено достоверно частое

повреждение всех 43 гуанинов, $\{s_i=G_n\}_{1\le n\le 43}$, этой ДНК. Частота F(G) UV-повреждения G (Втюрина и др., 2011) варьировала от 0 до 1.59 ln-ед., как это можно видеть в Таблице 30.

Эти частоты $F(G_n)$ UV-повреждений и $S_{\pm 10}(G_n)=\{s_{i-10}\ldots s_i=G_n\ldots s_{i+j}\ldots s_{i+10}\}$ окрестности с флангами ±10 п.о. были проанализированы общепринятыми консенсусом, матрицей ω_{sj} встречаемости нуклеотидов s в позициях j относительно G_n и описанной выше системой Activity (Ponomarenko *et al.*, 1997a). В результате были обнаружены достоверные корреляции частот F(G) UV-повреждений G с оценкой частоты встречаемости нуклеотидов вокруг UV-повреждений G, а также с консенсусом редких нуклеотидов в локальном окружении G, с содержанием тетрануклеотида YNVW (номенклатура, IUPAC-IUB, 1971) перед G и с оценкой средней частоты контакта гистон-подобных белков (Satchwell et al., 1986) с локальным окружением G. Простые регрессии Пирсона на основе этих четырех корреляций были суммированы в формулу линейно-аддитивного прогноза неизвестных частот $F(S_{\pm 10}(G))$ UV-повреждений G по их известной окрестности $S_{\pm 10}(G)$. С помощью этой формулы была впервые продемонстрирована возможность достоверных таких прогнозов UV-повреждений ДНК на примере независимых опытов с геном MIP-1α мыши (Melvin *et al.*, 1998) и с синтетическими олигоДНК (Angelov *et al.*, 2005).

Идея формулы для линейно-аддитивного прогноза неизвестных частот F(G) по известной локальной окрестности G в ДНК состояла в учете особенностей $X[S_{\pm 10}(G)]$ окрестности G в ДНК, достоверно коррелирующих с измерениями частот F(G) UV- облучении G для их прогноза:

$$F(G)=\sum_{k=1}^{K}\lambda_k X_k[S_{\pm 10}(G)]. \qquad (44)$$

Для применения этой эмпирической формулы требовалось найти контекстно-зависимые количественные оценки $X[S_{\pm 10}(G)]$, коррелирующие с экспериментальными данными F(G).

Таблица 30 – Участки ДНК длины 21 нт вокруг гуанинов (G), частоты F(**G**) предмутационного повреждения (Втюрина и др., 2011) и результат их анализа

Данные опыта (Втюрина и др., 2011)			Результаты анализа *in silico*								прогноз
±10 п.о. окрестности вокруг G		F(G_i)	консенсус		PWM (46)		[YNVW]$_\omega$		$P_{23;-7;+2}$		
№	Последовательность ДНК	ln	поиск	тест	поиск	тест	поиск	тест	поиск	тест	
1	cagaacattt**G**ataccaaacc	1.59	2		−2.02			1.40		14.41	1.44
2	ttaaaccctg**G**gaaccgcaag	1.57	0		−1.95		2.06		12.45		1.61
3	tttaaaccct**G**ggaaccgcaa	1.55	2		−2.76			1.55		13.48	1.28
4	cgcaaggttg**G**gcaaataaag	1.53		2		−2.69	1.20		13.48		1.21
5	ccccttcctt**G**gtatggaaaa	1.43	1		−2.95			1.16		14.15	1.27
6	agaacactaa**G**agctcagatc	1.43	1		−2.74		0.51		12.63		1.07
7	ggcaaataaa**G**gctaatcata	1.39	4		−2.78			0.93		13.96	1.04
8	cttgttttaa**G**aacagtttgt	1.38	2		−3.45		1.13		14.48		1.13
9	ataagttttt**G**cagaataatg	1.33	3		−2.88			0.30		14.90	1.01
10	accaactcag**G**aaaccacttg	1.29	0		−2.59		1.06		13.26		1.33
11	aaccaactca**G**gaaaccactt	1.28	2		−2.56			0.73		13.81	1.16
12	gctcagatca**G**aacatttgat	1.19		3		−3.21	1.42		12.69		1.05
13	ccgcaaggtt**G**ggcaaataaa	1.15	3		−3.37			1.02		14.16	1.03
14	ttatattatg**G**tttacataag	1.13	4		−2.87		1.34		10.16		0.87
15	aagaacatag**G**aaaatagaac	1.12	1		−2.35			0.10		11.71	0.99
16	agtttttgca**G**aataatgttc	1.10	3		−3.85		0.32		14.70		0.82
17	ggaaccgcaa**G**gttgggcaaa	1.10	0		−2.78			0.24		10.69	0.95
18	cctgggaacc**G**caaggttggg	1.02	5		−4.34		0.10		9.60		0.22
19	agaacagttt**G**taaccataaa	0.95		3		−2.89		0.50		14.16	1.00
20	gcagaataat**G**ttctatcagt	0.93	4		−3.48		1.23		11.83		0.84
21	tccagccact**G**cccttcctt	0.86	3		−3.52			1.32		11.40	0.89
22	cataccataa**G**ttttgcaga	0.85	2		−3.01		0.39		11.58		0.86
23	tcacatcctt**G**tttaagaac	0.85	2		−3.08			1.16		12.90	1.10
24	aacactaaga**G**ctcagatcag	0.82	5		−3.41		0.91		11.85		0.72
25	taaaccctgg**G**aaccgcaagg	0.81	2		−2.87			1.29		11.56	1.08
26	gacttggata**G**attccaaaag	0.80	3		−3.25		0.60		11.85		0.81
27	tatcagtcca**G**ccactgcccc	0.78	3		−3.98			0.52		12.16	0.68
28	cataagactt**G**gatagattcc	0.77		3		−3.12	0.20		14.15		0.90
29	gaaaccactt**G**tctcacatcc	0.73	6		−3.89			0.44		13.14	0.54
30	gcaaataaag**G**ctaatcataa	0.72	4		−3.25		1.26		12.31		0.92
31	caaaaaatca**G**cactctttta	0.71	4		−2.91			1.20		13.64	1.06
32	atttacataa**G**acttggatag	0.71	3		−3.19		0.40		11.21		0.74
33	tttaagaaca**G**tttgtaacca	0.69	4		−3.63			1.02		14.15	0.91
34	taagagctca**G**atcagaacat	0.67	0		−2.62		0.32		12.29		1.10
35	gattccaaaa**G**aacataggaa	0.67		4		−3.48		0.63		14.78	0.90
36	ataagacttg**G**atagattcca	0.64	4		−3.31		0.30		13.26		0.76
37	gcaaggttgg**G**caaataaagg	0.51	2		−3.25			0.20		11.83	0.79
38	tgttctatca**G**tccagccact	0.51	6		−3.52		1.01		11.36		0.62
39	gtttacataa**G**catttacata	0.50	3		−2.86			0.40		10.45	0.75
40	tttatattat**G**gtttacataa	0.43	4		−3.19		1.13		9.79		0.74
41	taggaaaata**G**aacactaaga	0.41	3		−2.81			0.30		13.21	0.92
42	gaaccgcaag**G**ttgggcaaat	0.37		4		−3.81	0.29		10.69		0.50
43	aaagaacata**G**gaaaatagaa	0.00	4		−2.82			0.10		11.71	0.71
	Линейная корреляция, r=		−0.41	−0.91	0.36	0.86	0.46	0.48	0.45	0.47	0.68
	Значимость, α<		0.025	0.025	0.05	0.05	0.05	0.025	0.05	0.05	10^{-6}

Прежде всего, в качестве $X[S_{\pm 10}(G)]$ был построен консенсус UV-повреждений G. С этой целью все $\{G_n\}_{1\le n\le 43}$ были разбиты по убыванию $F(G_n)$ на 6 равновеликих групп ($6^2<43<7^2$) и от каждой группы в контроль был выбран G_n с ближайшей частотой $F(G_n)$ к их внутригрупповому среднему $M_0(F)$, Таблица 30: № 4, 12, 19, 28, 35, 42. Остальные 37 проб G_n были взвешены величинами $F(G_n)$ в ln-ед. согласно принципу Больцмана и затем, с использованием этого взвешивания, для каждого s в каждой позиции j окрестности $S_{\pm 10}(G_n)$ были оценены частоты f_{sj}, их среднее и стандартная ошибка, $M_0(f_{sj})\pm\delta(f_{sj})$. На этой основе с помощью *t*-критерия Стьюдента для каждой позиции консенсуса был найден достоверно самый частый нуклеотид, aaccttg$G_n$gtc-cgca. Однако, число совпадений $S_{\pm 10}(G_n)$ с этим консенсусом, $N_>(G_n)$, не коррелировало с частотой предмутационных повреждений $F(G_n)$ ни на поисковых данных, ни на контроле (данные не показаны). Это означало отсутствие в эволюции *E. coli* отбора на повышенную UV-индуцируемую изменчивость ДНК по гуанину (G), что хорошо согласуется с существующими представлениями. Поэтому для UV-повреждений G_n был построен консенсус достоверно самых редких нуклеотидов, ttaaagc(G_n)tcg-actgc, который мы назвали “антиконсенсусом” как противоположность по отношению к традиционно используемому понятию “консенсус”. Число совпадений с этим антиконсенсусом $N_<(G_n)$ достоверно ($r=-0.91$, $\alpha<0.025$) коррелирует с F(G) на независимом контроле. Эта негативная корреляция означает, что происходит давление отбора на ближайших соседей гуанина (G) в геномной ДНК *E. coli* против UV-повреждений: чем лучше контекст окружения гуанина в ДНК совпадает с консенсусом ttaaagc(G_n)tcg-actgc, тем ниже частота UV-повреждения этого гуанина.

В свою очередь на основе оценок частот f_{sj} нуклеотидов s в позициях j вокруг G была построена позиционно-частотная матрица ω_{sj} и соответствующая ей оценка $PWM(S_{\pm 10})$ сходства контекста произвольного G в ДНК с этим наиболее частым ω_{sj} контекстом UV-повреждений G:

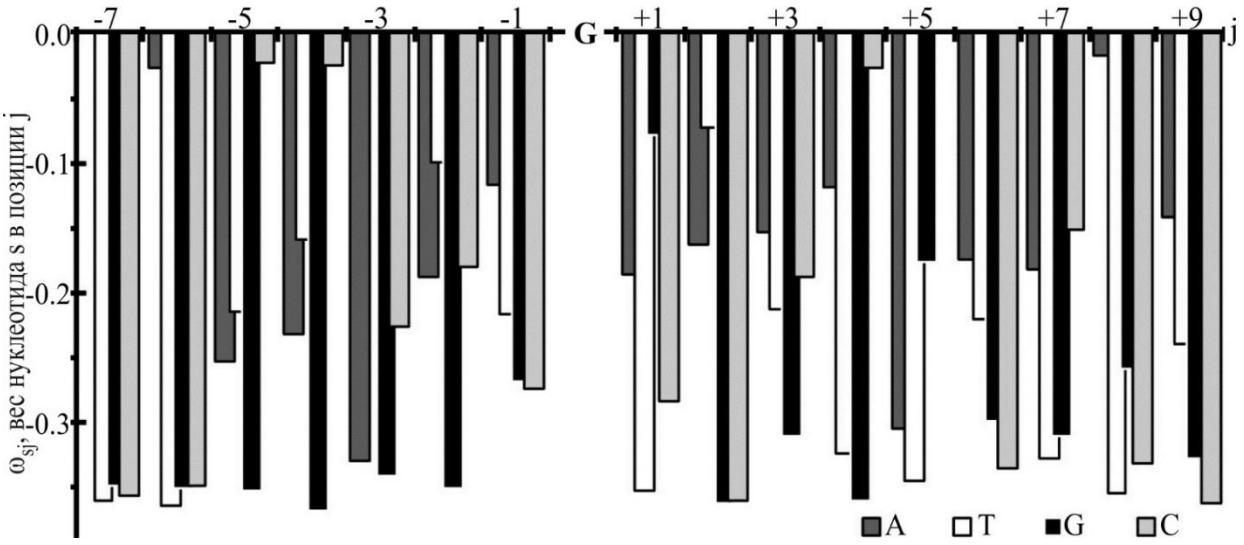

Рисунок 59 - Графическое представление позиционно-частотной матрицы ω_{sj}. Из-за отсутствия достоверно частых (редких) нуклеотидов в позициях $j \in \{-10, -9, -8, 10\}$, в них установлено $\omega_{sj} \equiv 0$ (не показано), что обосновывает эвристический выбор ±10 п.о. окрестности вокруг G.

$$\begin{cases} \omega_{sj} = \dfrac{f_{sj}}{\max_{\xi}(f_{\xi j}) + \max_{\zeta}(f_{s\zeta})} \ln\left(\dfrac{{f_{sj}}^2}{\max_{\xi}(f_{\xi j})\max_{\zeta}(f_{s\zeta})}\right). \\ \mathrm{PWM}(S_{\pm 10}(G)) = \sum_{j=-10}^{10} \omega_{s_j j}. \end{cases} \quad (45)$$

Графическое представление ω_{sj} показано на Рисунке 59. В силу отсутствия достоверно частых или редких нуклеотидов в позициях $j \in \{-10, -9, -8, 10\}$, в них было установлено $\omega_{sj} \equiv 0$. Это обосновывает эвристический выбор границ ±10 п.о. окрестности $S_{\pm 10}(G_n)$ UV-повреждений G_n. Величины $\mathrm{PWM}(S_{\pm 10}(G_n))$ приведены в Таблице 30. Они достоверно коррелируют с $F(G_n)$ на независимом контроле: $r=0.86$ ($\alpha > 0.05$). Эта позитивная корреляция соответствует влиянию окрестности UV-повреждений G на их частоту F(G) в согласии с выводами опытов (Gurzadyan *et al.*, 1993; Melvin *et al.*, 1998; Angelov *et al.*, 2005). Таким образом, было впервые установлено (Втюрина и др., 2012), что чем лучше локальное окружение гуанина в ДНК соответствует позиционно-частотной матрице на Рисунке 59, тем больше частота повреждения этих гуанинов.

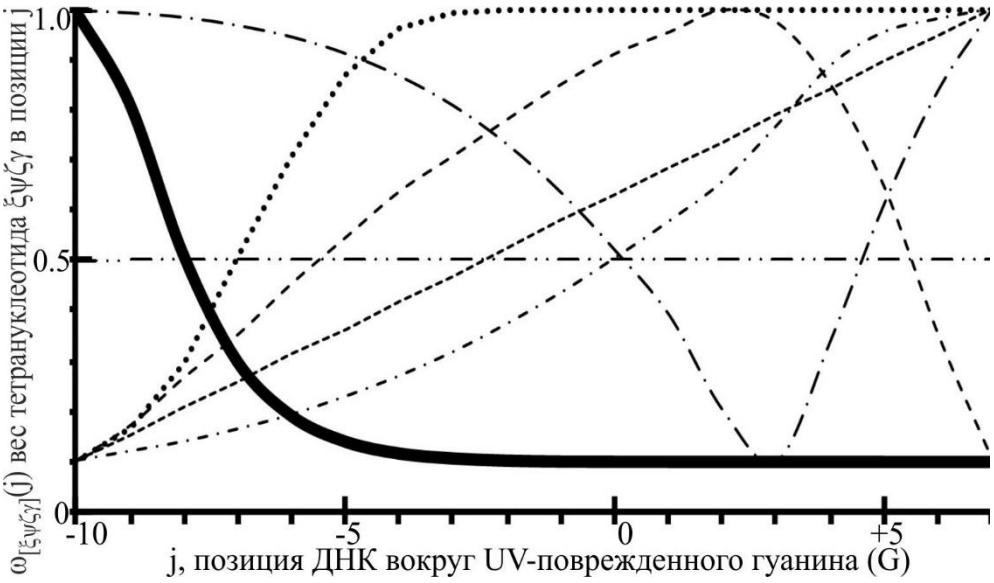

Рисунок 60 - Примеры 7 из 360 вариантов весов ω(j) позиций j вокруг G, j=0, используемых (формула 38) в качестве оценки линейно-аддитивного вклада тетрануклеотида $z_jz_{j+1}z_{j+2}z_{j+3}$ в позиции j в частоту F(G) UV-повреждений G на основе правила "чем выше $0.1\leq\omega(i)\leq1$, тем выше вклад $z_jz_{j+1}z_{j+2}z_{j+3}$ в F(G)". Жирной линией показан вес тетрануклеотида YNVW, найденного компьютерной системой Activity (Ponomarenko M. *et al*., 1997a)

Затем с помощью описанной в предыдущей главе системы Activity (Ponomarenko *et al*., 1997a) была проанализирована локальная окрестность $S_{\pm10}(G_n)$ UV-повреждений G_n с целью поиска достоверной корреляции контекста с частотой $F(G_n)$. Для этого были использованы все окрестности $S_{\pm10}(G_{2k})$ с четными номерами проб $\{G_{2k}\}_{1\leq2k\leq43}$ в порядке убывания измеренных *in vitro* частот $F(G_n)$ UV-повреждений G (Таблица 30: всего 21 проба), тогда как все оставшиеся 22 пробы $\{G_{2k-1}\}_{1\leq2k-1\leq43}$ с нечетными номерами были использованы для независимого контроля результатов этого анализа. В качестве $X[S_{\pm10}(G)]$ было содержание тетрануклеотидов $z_1z_2z_3z_4$, взвешенных с помощью 360 весовых функций $0\leq\omega\leq1$ (формула 38) по правилу: "чем выше $0.1\leq\omega(i)\leq1$, тем выше вклад $z_jz_{j+1}z_{j+2}z_{j+3}$ в F(G)" (Рисунок 60). Всего проверено $360\times15^4\approx10^7$ вариантов $[z_1z_2z_3z_4]_\omega$, (формула 38). Наибольшая $U(F;[YNVW]_S)=0.25$ указала тетрануклеотид YNVW

(номенклатура, IUPAC-IUB, 1971) с S-образным весом его локализаций в окрестности $S_{\pm 10}(G_n)$ UV-повреждений G_n (Рисунок 60: жирная линия). Величины $[YNVW]_S$ для всех 43 проб представлены в Таблице 30, где они достоверно коррелируют с частотой $F(G_n)$ на независимом контроле ($r=0.48$, $\alpha<0.025$).

В свою очередь, с помощью компьютерной системы Activity (Ponomarenko *et al.*, 1997a) в качестве $X[S_{\pm 10}(G)]$ аналогично были проанализированы оценки $P_{k;[a;b]}$ средних величин для 38 физико-химических и конформационных свойств B-спирали ДНК из базы данных PROPERTY (Колчанов и др., 1997) на участке между позициями a и b окрестности UV-повреждений G в ДНК (формула 36). Всего было проанализировано $38\times(20\times19)/2=7220$ вариантов $P_{k;[a;b]}$.

Наибольшая оценка $U(F; P_{23;[-7;2]})=0.15$ указала участок [-7; 2] вокруг UV-повреждения G для усреднения свойства №23 спирали ДНК из (Колчанов и др., 1997) частота контакта динуклеотида с гистон-подобным белком. Численные значения $P_{23;[-7;2]}$ показаны в Таблице 30. Они достоверно коррелируют с частотами F на контроле: $r=0.45$ ($\alpha<0.05$). Хотя для бактерий не известна нуклеосомная упаковка геномной ДНК в хроматин, корреляция частот UV-повреждений и частот контакта с гистон-подобным белком согласуется с открытием редких нуклеосома-подобных шарообразных структур в качестве уровня организации геномной ДНК бактерий (Griffith, 1976). Эти нуклеосома-подобные шарообразные структуры оказались результатом связывания белков семейства HU с геномной ДНК как это было обнаружено методами иммуноэлектронной микроскопии с антителами к этим белкам для *Escherichia coli* (Киселева и др., 1986), стрептомицетов (Киселева и др., 1988а), хлоропластов (Киселева и др., 1988б) и митохондрий (Salganik *et al.*, 1990, 1991). Линии *E. coli*, дефектные по генам *HU*, были охарактеризованы аномально высокой чувствительностью к ультрафиолетовому излучению (Li, Waters, 1998). Методами сравнительной

геномики установили степень эволюционного родства (гомологии) между генами семейства HU у бактерий и генами гистонов у эукариот (Wong et al., 2003). Все это вместе взятое позволило высказать гипотезу, что в ходе эволюции бактериальных геномов мог отобраться механизм защиты наиболее повреждаемых UV-излучением участков геномной ДНК в случае их связывания с гистон-подобными белками, который и привел в итоге к наследованию эукариотами "изобретенного" бактериями хроматина. Поэтому, чем выше была частота UV-повреждений гуанинов в опыте (Втюрина и др., 2011), тем выше оказалась предрасположенность локального окружения этих гуанинов к контакту с гистон-подобными белками как это можно видеть в Таблице 30: колонка "$P_{23;-7;+2}$".

Наконец, для конструирования прогностической формулы (46) оказалось ключевым, что достоверно коррелирующие с частотой F(G) UV-повреждений G особенности ДНК не коррелировали между собой (данные не показаны). Поэтому после суммирования простых регрессий и их нормирования искомая формула приобрела следующий окончательный вид:

$$\begin{aligned} \mathrm{F}(S_{\pm 10}(G)) = 0.69 + 0.22[YNVW]_S(S_{\pm 10}(G)) + \\ +0.07P_{23;[-7;2]}(S_{\pm 10}(G)) + 0.19PWM(S_{\pm 10}(G)) - 0.07\mathrm{N}_<(S_{\pm 10}(G)). \end{aligned} \quad (46)$$

Оценки частот $F(S_{\pm 10}(G_n))$ по формуле (46) достоверно коррелируют с их измерениями (Втюрина и др., 2011) $F(G_n)$: r=0.68, $\alpha<10^{-6}$ (Таблица 30, Рисунок 61). Формула (46) была выведена на основе экспериментальных данных о бактериальной ДНК (Втюрина и др., 2011). Поэтому с помощью данных независимого опыта (Melvin at el., 1998) с геном MIP-1α мыши (**M**acrophage **I**nflammatory **P**rotein 1α, GenBank AC=X12531: 201-301) была установлена достоверность этой формулы в случае эукариот (Рисунок 62a: r=0.82 и $\alpha<0.005$). При этом позитивный вклад $P_{23;[-7;2]}$ в формулу (46) соответствует также экспериментально известной нуклеосомной защите эукариотической ДНК от UV-излучения (Brown *et al*., 1993).

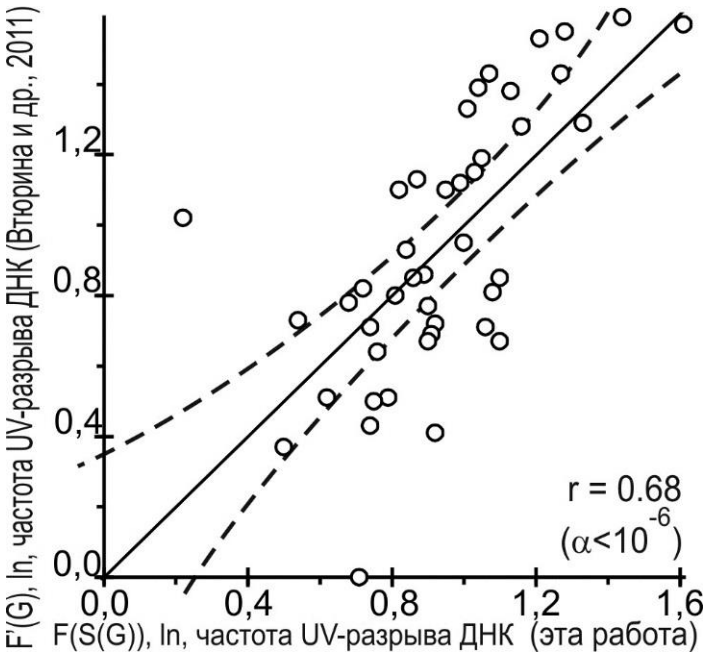

Рисунок 61- Достоверная ($r=0.68$, $\alpha<10^{-6}$) корреляция между предсказанными (формула 46) и экспериментальными величинами $F(G_n)$ для данных из статьи (Втюрина и др., 2011). Пунктир - границы 95%-доверительных интервалов.

Аналогично, на независимых данных опыта по UV-повреждению не имеющих природных аналогов олигоДНК (Angelov *et al.*, 2005) была установлена достоверность ($r=0.47$, $\alpha<0.05$) формулы (46) для синтетических олигоДНК (Рисунок 62б). Это означает, что частота F(G) UV-повреждения G в ДНК является инвариантной для любых источников молекул ДНК.

В свою очередь, важно отметить, что все выявленные контекстные особенности локального окружения гуанинов вносят независимые линейно-аддитивные вклады в величины частот UV-повреждений этих гуанинов с равными минимальными уровнями значимости $\alpha<0.05$, в то время как их суммарный вклад (формула 46) имеет тысячекратно больший уровень значимости, $\alpha<10^{-6}$. Это указывает на высокую сложность многостадийного многовариантного кооперативного молекулярного механизма повреждения гуанинов в ДНК при воздействии UV-излучения.

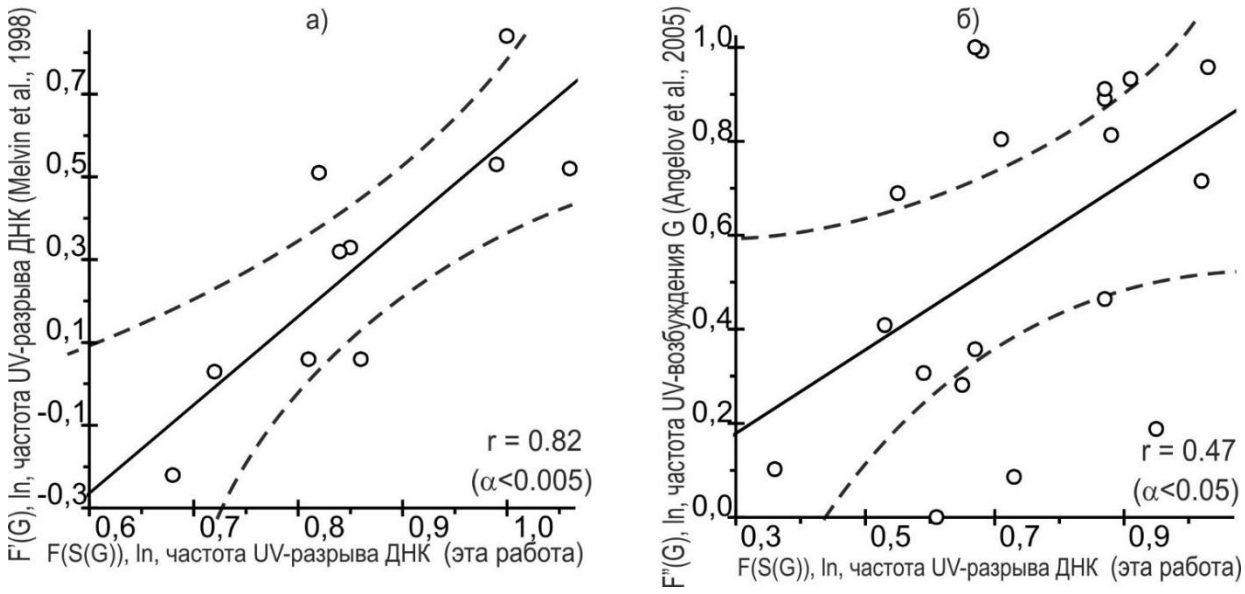

Рисунок 62 - Корреляции прогнозов (46) для независимых опытов: а) частота F'(G) UV-разрывов ДНК в G лазером 193 нм (Melvin *et al*., 1998), r=0.82, $\alpha<0.005$; б) частота F''(G) UV-возбуждения G первым фотоном двухфотонной ионизации лазером 366 нм (Angelov *et al*., 2005), r=0.47, $\alpha<0.05$. Пунктир – границы 95%-доверительных интервалов.

Кроме того, в главе 2 настоящей диссертации была обнаружена предрасположенность сайтов связывания транскрипционных факторов к нуклеосомной ДНК, препятствующей UV-повреждениям ДНК (Li, Waters, 1998). Это согласуется с экспериментально установленным ранее мозаичным строением промоторов из районов постоянной низкой алгоритмической сложности и низкими частотами мутаций (Chuzhanova *et al*., 2000, 2003a,b), насыщенных палиндромами, повторами и трактами, которыми являются сайты связывания транскрипционных факторов (Orlov *et al*., 2006; Abnizova *et al*., 2007), и с горячих точек мутагенеза: замен, вставок, делеций, дупликаций и транслокаций (Costa *et al*., 2005), - на границах между такими районами.

Наконец, в двух оставшихся разделах настоящей главы можно найти выявленные автором независимые свидетельства о том, что локальный нуклеотидный контекст геномной ДНК влияет не только на возникновение предмутационных повреждений в ней, но на и на эффективность их репарации.

4.2 Количественные характеристики локальных окрестностей 8-оксогуанина, коррелирующие с константой Михаэлиса и каталитической константой фермента 8-оксогуанин-ДНК гликозилаза человека

Важную роль в мутагенезе, канцерогенезе и старении человека играет 7,8-дигидро-8-оксогуанин (oxoG), продукт окисления G, предмутационное повреждение ДНК для замены G:C→T:A за два цикла репликации: G:C→oxoG:A→T:A (Beckman, Ames, 1998). Ферменты 8-оксогуанин-ДНК гликозилаза, OGG1 у человека и Fpg у бактерий, находят oxoG в ДНК (Banerjee *et al*., 2005) для первого шага репарации oxoG→G (Rosenquist *et al*., 1997). Однако между OGG1 и Fpg нет ни гомологии, ни сходства их 3D-структур (Bruner *et al*., 2000). Поэтому выявление взаимосвязей между кинетическими, конформационными и контекстными деталями активности OGG1 человека представляется особенно интересным (Kuznetsov *et al*., 2007). Было изучено 21 олигоДНК длиной 21 п.о., состоящих из пар нитей $S(X)/S(C)=\{s^0_{-10}...s^0_{-1}Xs^0_1...s^0_{10}/s^\#_{-10}...s^\#_{-1}Cs^\#_1...s^\#_{10}\}$, в которых было до 6 некомплементарных пар $s^0_i:s^\#_i$ и для которых были *in vitro* измерены (Kirpota *et al*., 2011) величины от 8 нМ до 400 нМ константы Михаэлиса K_M и от 0.4 мин$^{-1}$ до 2.9 мин$^{-1}$ каталитической константы k_{CAT} для фермента OGG1 (Таблица 31).

Для учета частичных нарушений комплементарности п.о., $\{s^0_{-10}\ldots X\ldots s^0_{+10}/s^\#_{-10}\ldots C\ldots s^\#_{+10}\}$, при прогнозе K_M и k_{CAT} эвристически была предложена аппроксимация в рамках приближения “лимитирующей стадии” соответствующих взаимодействий ДНК/OGG1, $k_{CAT}\{S\}$ и $K_M\{S\}$ с G вместо oxoG в силу отсутствия сведения о влиянии oxoG на свойства спирали ДНК:

$$k_{CAT}\{s^0_{-10}\ldots X\ldots s^0_{+10}/s^\#_{-10}\ldots C\ldots e^\#_{+10}\} = \\ = \mathrm{MIN}(k_{CAT}\{s^0_{-10}\ldots G\ldots s^0_{+10}\};\ k_{CAT}\{s^\#_{-10}\ldots C\ldots s^\#_{+10}\}); \qquad (47)$$

$$K_M\{s^0_{-10}\ldots X\ldots s^0_{+10}/s^\#_{-10}\ldots C\ldots e^\#_{+10}\} = \\ = \mathrm{MAX}(K_M\{s^0_{-10}\ldots G\ldots e^0_{+10}\};K_M\{s^\#_{-10}\ldots C\ldots s^\#_{+10}\}); \qquad (48)$$

Таблица 31 – Величины константы Михаэлиса K_M и каталитической константы k_{CAT} фермента *8*-оксогуанин-ДНК-гликозилаза OGG1 человека для олигоДНК длиной 21 п.о. с числом $N_{\#}$ нарушений комплементарности п.о. (Kirpota *et al*., 2011).

олигоДНК S(X)/S(C)	код пробы	Последовательность олигоДНК[#], 21 п.о. -10----+----0----+----+-12	$N_{\#}$	K_M, нМ	k_{CAT}, мин$^{-1}$
ODN1 (≡ODN13)	1 1c	5'-ctctcccttcXctcctttcctct-3' 3'-gagagggaagCgaggaaaggaga-5'	0	11±3	1.0±0.1
ODN2	2 2c	5'-ctctcccttcXctcctctcctct-3 3'-gagagggaagCgaggagaggaga-5'	0	23±4	1.4±0.1
ODN3	3 3c	5'-ctctcccctcXctcctctcctct-3' 3'-gagaggggagCgaggagaggaga-5'	0	21±5	1.1±0.1
ODN4	4 4c	5'-ctctcccctcXctcctttcctct-3' 3'-gagaggggagCgaggaaaggaga-5'	0	20±4	0.9±0.1
ODN5	5 5c	5'-ctctcctttcXctcctttcctct-3' 3'-gagaggaaagCgaggaaaggaga-5'	0	16±4	1.1±0.1
ODN6	6 6c	5'-aaaaaaaaacXcgcccgcccgcg-3' 3'-ttttttttgCgcgggcgggcgc-5'	0	26±5	1.2±0.1
ODN7	7 7c	5'-ttttttttcXctttttttttt-3' 3'-aaaaaaaaagCgaaaaaaaaaaa-5'	0	11±3	1.3±0.1
ODN8	8 8c	5'-tttttttttXttttttttttt-3' 3'-aaaaaaaaaaCaaaaaaaaaaaa-5'	0	29±6	0.96±0.07
ODN9	9 9c	5'-ccgcccgcgcXcaaaaaaaaaaa-3' 3'-ggcgggcgcgCgttttttttttt-5'	0	64±8	2.9±0.2
ODN10	10 10c	5'-ttttttgggXgggtttttttt-3' 3'-aaaaaaacccCcccaaaaaaaaa-5'	0	66±12	1.0±0.1
ODN11	11 11c	5'-gagcgagcgcXcgcgagcgagcg-3' 3'-ctcgctcgcgCgcgctcgctcgc-5'	0	78±8	2.4±0.1
ODN12	12 12c	5'-ctctcccttcXatcctttcctct-3' 3'-gagagggaagCtaggaaaggaga-5'	0	41±11	1.2±0.1
ODN13	1 12c	5'-ctctcccttcXCtcctttcctct-3' 3'-gagagggaagCTaggaaaggaga-5'	1	22±2	0.94±0.10
ODN14	7 8c	5'-tttttttttCXCttttttttttt-3'◊ 3'-aaaaaaaaaACAaaaaaaaaaaa-5'	2	390±20	1.3±0.4
ODN15	8 10c	5'-ttttttttTTTXTTTttttttttt-3' 3'-aaaaaaaCCCCCCCaaaaaaaaa-5'	6	400±50	0.40±0.02
ODN16	7 10c	5'-ttttttttTTCXCTTttttttttt-3' 3'-aaaaaaaCCCCCCCaaaaaaaaa-5'	6	310±50	0.48±0.07
ODN17	8 7c	5'-tttttttttTXTttttttttttt-3' 3'-aaaaaaaaaGCGaaaaaaaaaaa-5'	2	70±10	1.7±0.05
ODN18	1 7c	5'-ctctcccttcXctcctTtcctct-3' 3'-gagagggaagCgaggaGaggaga-5'	1	16±7	0.80±0.10
ODN19	1 3c	5'-ctctcccTtcXctcctTtcctct-3' 3'-gagagggGagCgaggaGaggaga-5'	2	59±14	1.9±0.10
ODN20	1 4c	5'-ctctcccTtcXctcctttcctct-3' 3'-gagagggGagCgaggaaaggaga-5'	1	22±6	1.4±0.10
ODN21	1 5c	5'-ctctccCttcXctcctttcctct-3' 3'-gagaggAaagCgaggaaaggaga-5'	1	8.0±4.0	1.4±0.10

#) две *крайние п.о. (курсив*) проигнорированы (Kirpota *et al*., 2011).

Чтобы максимизировать разнообразие последовательностей ДНК для *in silico* анализа на начальных стадиях комплементарных олигоДНК ODN1-ODN12, были выбраны 12 нитей из 6 олигоДНК ODN1-ODN6, представлявших все варианты трактов A_n, T_n, $(C/G)_n$ и перестановок динуклеотидов CT и AG вокруг oxoG (Таблица 31). Соответственно все 12 нитей из 6 олигоДНК ODN7-ODN12 были независимым контролем результатов этого анализа. Аналогично, на завершающей стадии анализировались все нити частично некомплементарных ODN13–ODN15, независимым контролем для результатов анализа которых были все нити ODN16–ODN21.

С использованием компьютерной системы Activity (Ponomarenko M. *et al*., 1997a) во всех 20×(20-1)/2=190 возможных участках [a; b] окружения oxoG в олигоДНК длиной 21 п.о. оценили среднеарифметические значения $P_{n;[a;b]}$ для всех 38 конформационных и физико-химических свойств ДНК (формула 36; $1 \le n \le 38$) из базы данных PROPERTY (Колчанов и др., 1998), всего 190×38=7220 вариантов $P_{n;[a;b]}$. Для каждого из них оценили степень их коррелированности $U(K_M; P_{n;[a;b]})$ с константой Михаэлиса K_M и $U(k_{CAT}; P_{n;[a;b]})$ с каталитической константой k_{CAT}. Наибольшая $U(k_{CAT};P_{11;[-6;6]})=0.389$ указала на линейно-аддитивный вклад среднего кручения спирали ДНК ±6 п.о. вокруг oxoG в величину каталитической константы k_{CAT}. Следующая $U(k_{CAT};P_{11;[-8;8]})=0.065$ указала то же свойство B-ДНК в более широкой окрестности oxoG. Величины $P_{11;[-6;6]}$ показаны на Рисунке 63: они достоверно коррелируют с величинами k_{CAT} (а) на поисковых данных $r=0.86$, $\alpha<0.0005$, и (б) на контроле ($r=0.84$, $\alpha<0.001$). Достоверная ($r = 0.87$, $\alpha < 0.00025$) простая регрессия Пирсона для величин k_{CAT} имела следующий вид:

$$k_{CAT}\{s_{-10}\ldots s_{+10}\} = 0.88(P_{11;[-6;+6]}\{s_{-10}\ldots s_{+10}\} - 33.76). \quad (49)$$

В случаях нарушения комплементарности (Таблица 31, ODN13-ODN21), прогноз *in silico* (формулы 36, 47 и 49) также оказался достоверным (Рисунок 63г: $r = 0.667$, $\alpha < 0.05$). Ошибка $\Delta k_{CAT} = k_{CAT} - k_{CAT}(s^0_{-10}\ldots X\ldots s^0_{+10}/s^{\#}_{-}$

$_{10}$…C…$e^{\#}_{+10}$) этого прогноза варьировала около 0 мин$^{-1}$ от −0.47 мин$^{-1}$ до 0.92 мин$^{-1}$ и, более того, не коррелировала с количеством $N_{\#}$ нарушений комплементарности (r = –0.02, $\alpha > 0.925$, данные не показаны). Это означает, что эвристическая нелинейная аппроксимация “лимитирующей стадии” (формула 47) исчерпывает линейно-аддитивные вклады нуклеотидов олигоДНК с нарушением комплементарности в величину константы k_{CAT}, скорости вырезания oxoG из ДНК ферментом OGG1 в экспериментальных данных (Kirpota *et al*., 2011).

В случае константы Михаэлиса K_M, наибольшая $U(K_M;P_{38;[-10;10]})$= 0.083 (формулы 36, 38) указала на оценку средней свободной энергии Гиббса $\Delta G°$ олигоДНК, которая в отсутствие нарушений комплементарности достоверно коррелирует с K_M на проанализированных (Рисунок 64а: r = –0.83, $\alpha < 0.001$) и на контрольных (Рисунок 64б: r = –0.81, $\alpha < 0.0025$) данных. При этом, достоверная (r = –0.46, $\alpha < 0.025$) простая регрессия K_M на всех ODN1-ODN12 имеет вид:

$$K_M\{s_{-10}...s_{+10}\} = -42.68(P_{38;[-10;+10]}\{s_{-10}...s_{+10}\} + 0.86). \quad (50)$$

В отличие от случая k_{CAT}, ошибка аппроксимации “лимитирующей стадии” (формула 50) здесь, $\Delta K_M = K_M - K_M(s^0_{-10}…X…s^0_{+10}/s^{\#}_{-10}…C…e^{\#}_{+10})$, варьировала от –27 nM до 373 nM и достоверно коррелировала с числом $N_{\#}$ нарушений комплементарности (r = 0.814, $\alpha < 10^{-5}$, данные не показаны).

Поэтому для *in silico* прогноза неизвестных значений K_M по известным последовательностям частично некомплементарных нитей олигоДНК вокруг oxoG был дополнительно учтен линейно-аддитивный вклад числа нарушений комплементарности, $N_{\neq}$:

$$K_M\{s^0_{-10}...X...s^0_{+10}/s^{\#}_{-10}...C...s^{\#}_{+10}\} = 72.7\times N_{\neq} + \\ + \mathrm{MAX}(K_M\{s^0_{-10}...G...s^0_{+10}\}; K_M\{s^{\#}_{-10}...C...s^{\#}_{+10}\}). \quad (51)$$

Прогноз (формула 51) величин K_M для всех 21 олигоДНК показан на Рисунке. 64в,г. Они достоверно коррелируют с измерениями этой величины в

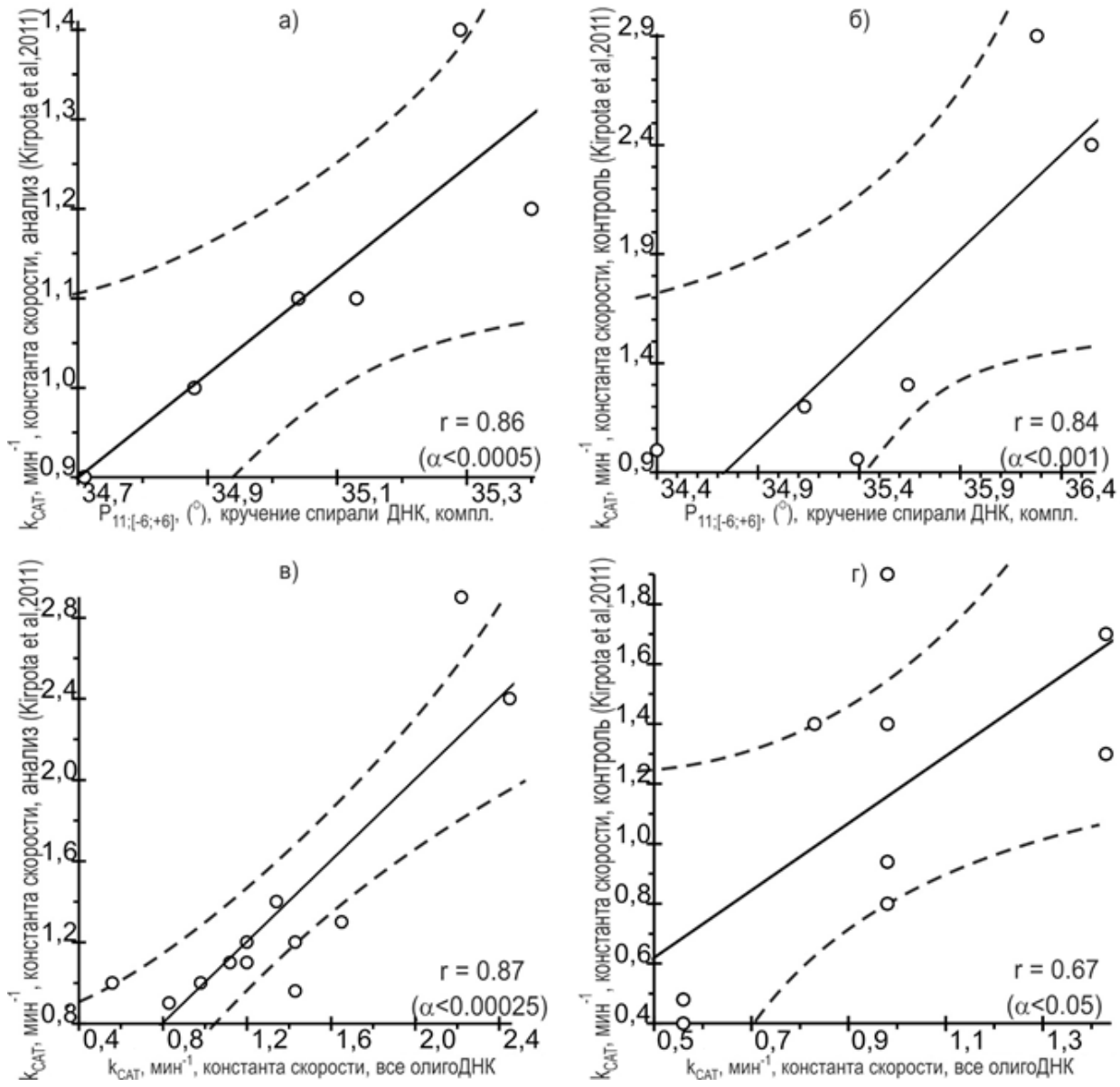

Рисунок 63 - Корреляции между измерениями k_{CAT} в опыте (Kirpota *et al*., 2011) и кручением спирали ДНК ±6 п.о. вокруг oxoG без нарушения комплементарности: а) анализ, r=0.86, α<0.0005; б) контроль, r=0.84, α<0.001, - и для всех олигоДНК: в) анализ, r = 0.87, α<0.00025; г) контроль, r=0.67, α<0.05. Пунктир – границы 95%-доверительных интервалов.

опыте (Kirpota *et al*., 2011) на проанализированных (r = 0.85, α < 0.005) и на контрольных (r = 0.86, α < 0.0005) олигоДНК.

Для независимой проверки аппроксимации “лимитирующей стадии” с ее помощью была оценена удельная свободная энергия Гиббса для олигоДНК:

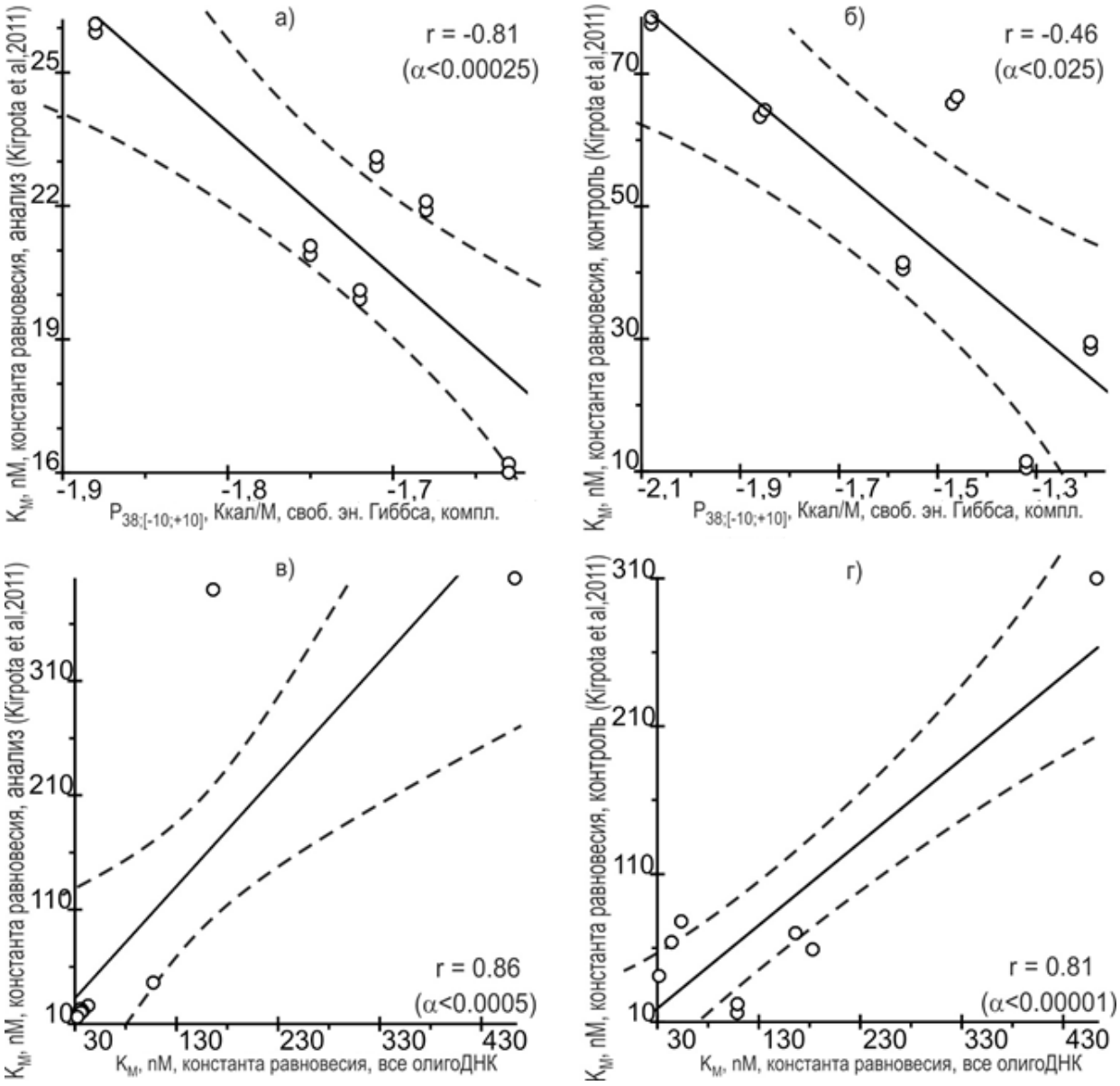

Рисунок 64 - Достоверные корреляции между измерениями K_M (Kirpota *et al*., 2011) и средней свободной энергией Гиббса спирали ДНК всего олигоДНК без нарушения комплементарности – а) анализ, $r = –0.81$, $\alpha < 0.00025$; б) контроль, $r = –0.46$, $\alpha < 0.025$, - а также прогноз для всех олигоДНК: в) анализ, $r = 0.86$, $\alpha < 0.0005$; г) контроль, $r = 0.81$, $\alpha < 0.00001$. Пунктир – границы 95%-доверительных интервалов.

$$\delta G\{s^0_{-10}\ldots \mathrm{X}\ldots s^0_{+10}/s^{\#}_{-10}\ldots \mathrm{C}\ldots s^{\#}_{+10}\} = \\ = \mathrm{MAX}(P_{38;[-10;+10]}\{s^0_{-10}\ldots \mathrm{G}\ldots s^0_{+10}\};\ P_{38;[-10;+10]}\{s^{\#}_{-10}\ldots \mathrm{C}\ldots s^{\#}_{+10}\}). \quad (52)$$

Прогноз (формула 52) величин $\delta G°$ для всех олигоДНК представлен в Таблице 32. Они достоверно коррелируют ($r = 0.81$, $\alpha < 0.00001$) с величинами свободной энергии Гиббса олигоДНК из опыта (Kirpota *et al*., 2011), но не коррелируют с другими энергетическими параметрами олигоДНК из этого опыта: ни с энтальпией ($\Delta H°$; $r = 0.03$, $\alpha > 0.9$), ни с энтропией ($\Delta S°$; $r = 0.27$, $\alpha > 0.2$). Ранее вне рамок настоящей диссертации автором была установлена достоверная корреляция между частотами амбер-мутаций, индуцируемых 2-аминопурином, и температурой T_M плавления ДНК (Ponomarenko M. *et al*., 1997a). Прогноз формулы (52) также достоверно коррелирует ($r = –0.88$, $\alpha < 10^{–6}$) с этой величиной T_M из опыта (Kirpota *et al*., 2011). Все это в целом свидетельствует в пользу адекватности эвристически предложенной нелинейной аппроксимации "лимитирующей стадии" (47, 48, 51 и 52) условиям опыта (Kirpota *et al*., 2011). Наконец, прогноз δG достоверно коррелирует с величинами k_{CAT} из опыта (Kirpota *et al*., 2011), но не с отклонениями (Δk_{CAT}) прогноза (формулы 47 и 49) этих величин (Таблица 31: соответственно, $r = –0.52$, $\alpha < 0.025$ и $r = –0.28$, $\alpha > 0.2$). Это означает, что линейно-аддитивный вклад свободной энергии Гиббса $\delta G°$ в каталитическую константу k_{CAT} исчерпывается вкладом внутренней энергии кручения B-спирали олигоДНК вблизи oxoG, который был учтен в формулах (36, 47 и 49) и который входит в состав эмпирических оценок свободной энергии Гиббса.

Константа Михаэлиса, K_M, и каталитическая константа, k_{CAT}, традиционно используют в качестве количественных характеристик биологической активности ферментов. Однако оказалось, что они по-разному характеризовали работу фермента OGG1 в случае частичного нарушения комплементарности ДНК вокруг oxoG (Kirpota *et al*., 2011). Выявление с помощью системы Activity физико-химической обусловленности этих отличий (формулы 36, 38) существенно уточнило детали (формулы 50 - 52) взаимодействия OGG1 с геномной ДНК.

В случае константы Михаэлиса K_M, межмолекулярного распознавания OGG1/oxoG программа Activity (Ponomarenko M. *et al*., 1997a) с помощью формул (36, 38) выявила наиболее обоснованной негативную корреляцию этой константы со свободной энергией Гиббса, ΔG (Рисунок 64а). Этот результат соответствует результату независимого опыта (Banerjee *et al*., 2005), согласно которому фермент OGG1 распознает oxoG в ДНК по отклонению oxoG-содержащей спирали ДНК от канонической спирали Уотсон-Крика: чем больше отклонение, тем ниже свободная энергия Гиббса ΔG. Существенно, что независимая контрольная проверка удельной свободной энергии Гиббса δG олигоДНК (Таблица 32: $r = 0.81$, $\alpha < 10^{-5}$) обосновала допустимость использования аппроксимации "лимитирующей стадии" опыта (Kirpota *et al*., 2011) по изучению отклонений структуры ДНК от канонической спирали Уотсон-Крика вследствие частичного нарушения комплементарности нитей ДНК, например, в случае ошибок репликации.

Для каталитической константы k_{CAT} фермента OGG1 (Kirpota *et al*., 2011) самым важным оказалось кручение спирали ДНК вокруг oxoG, что согласуется с данными независимого опыта (Bruner *et al*., 2000) о раскручивании спирали ДНК ферментом OGG1. Поэтому, чем больше кручение спирали ДНК, тем больше вклад этой биологической функции OGG1 в величину константы k_{CAT} скорости OGG1 вырезания oxoG из ДНК (Kirpota *et al*., 2011). Кручение спирали ДНК было наиболее частым биологически значимым свойством конформации спирали ДНК сайтов связывания многих регуляторных белков (Nussinov, 1984; MacLeod, 1994), включая исследованные в главе 2 диссертации 42 транскрипционных фактора (Пономаренко и др., 1997в).

Таким образом, полученные в настоящем разделе диссертации результаты компьютерного анализа влияния контекста ДНК на количественные величины констант Михаэлиса K_M и каталитической константы k_{CAT} фермента 8-оксогуанин-ДНК гликозилаза OGG1 человека

Таблица 32 – Проверка аппроксимации “лимитирующей стадии” на основе сравнения удельной свободной энергии Гиббса δG (52) с независимыми измерениями энергетических констант и температуры плавления (T_M) в опыте (Kirpota *et al*., 2011)

олигоДНК S(X)/S(C)	Прогноз (52), δ*G*°, ККалМ/п.о.	опыт *in vitro* (Kirpota *et al*., 2011)					Δk_{CAT}, min^{-1}
		Δ*G*°, Ккал/М	энтальпия Δ*H*, Ккал/М	энтропия Δ*S*°, кал/(М×К)	T_M, °C	k_{CAT}, min^{-1}	
ODN1	−1.68	−19.6	–138.47	383	64.0	1.00	0.02
ODN2	−1.71	−18.4	–113.59	307	66.7	1.40	0.06
ODN3	−1.75	−20.7	–129.05	349	69.3	1.10	–0.10
ODN4	−1.72	−21.2	–139.77	382	67.9	0.90	0.07
ODN5	−1.63	−19.4	–139.68	388	63.1	1.10	–0.02
ODN6	−1.88	−20.7	–121.12	324	71.8	1.20	–0.23
ODN7	−1.36	−14.6	–146.86	426	50.5	1.30	–0.35
ODN8	−1.23	−11.6	–108.02	311	46.0	0.96	–0.47
ODN9	−1.89	−21.1	–137.78	376	68.2	2.90	0.68
ODN10	−1.50	−16.7	–136.68	387	56.7	1.00	0.44
ODN11	−2.12	−19.9	–91.50	231	81.3	2.40	–0.05
ODN12	−1.61	−17.9	–123.44	340	62.5	1.20	0.00
ODN13	−1.61	−16.0	–129.44	366	56.0	0.94	–0.04
ODN14	−1.23	−7.6	–106.92	320	34.2	1.30	–0.13
ODN15	−1.23	−3.7	–102.34	318	23.1	0.40	–0.16
ODN16	−1.36	−2.2	–101.60	320	18.9	0.48	–0.08
ODN17	−1.23	−8.7	–137.07	414	37.5	1.70	0.27
ODN18	−1.68	−16.8	–120.95	336	59.9	0.80	–0.18
ODN19	−1.68	−14.6	–109.83	307	55.2	1.90	0.92
ODN20	−1.68	−18.6	–140.74	394	61.0	1.40	0.57
ODN21	−1.63	−13.8	–94.70	261	55.3	1.40	0.42
Корреляция, r		0.81	0.03	0.27	−0.88	–0.52	–0.28
Значимость, α		$<10^{-5}$	>0.9	>0.2	$<10^{-6}$	<0.025	>0.2

адекватно согласуются с самыми общими знаниями о связывании регуляторных белков с ДНК.

4.3 Количественные характеристики нуклеотидного контекста, значимые для сродства белка RecA к нитям ДНК

В результате вырезания предмутационного повреждения из геномной ДНК *E. coli* (например, oxoG в случае действия 8-оксогуанин-ДНК гликозилазы Fpg), вызванного мутагеном (например, UV-излучением), на месте этого повреждения в соответствующей нити ДНК возникает брешь со свободными 3’- и 5’-концами неповрежденных участков этой нити ДНК, которые немедленно связывает белок RecA, запуская экспрессию генов SOS-системы для репарации этой бреши (Qin *et al*., 2015). Комплексом RecA с однонитевой ДНК (онДНК) является RecA-филамент, виток спирали которого составляют 6 мономеров RecA и 18 нуклеотидов ДНК (Cox, 1993; West, 1994). Открытие чувствительности RecA к контексту онДНК (Mazin, Kowalczykowski, 1996) означало стократное различие уровней защиты локальных районов генома *E. coli* от их предмутационных повреждений. Поэтому возникла необходимость найти контекстные особенности ДНК, биологически значимые для сродства RecA к онДНК.

С помощью описанной в предыдущей главе диссертации системы Activity (Ponomarenko *et al*., 1997a) были проанализированы измеренные *in vitro* величины сродства RecA/онДНК (Mazin, Kowalczykowski, 1996) и было установлено (Пономаренко М. и др., 1998), что сродство RecA/онДНК убывает с ростом содержания в онДНК тринуклеотидов DRV (номенклатура, IUPAC-IUB, 1971). Эти тринуклеотиды: AAA, AAC, AGA, AGC, GAA, GAC, GGA, GGC, TAA, TGA, AAG, AGG, TAG, TGG, GAG, GGG, TAC, TGC, - значимо коррелируют ($\alpha<0.025$) с кодонами аминокислот, которые достоверно часто встречаются на поверхности белковых глобул, но достоверно редко – в ядрах этих глобул. Этот результат соответствовал общепринятым механизмам молекулярной эволюции белков путем блочно-модульных перестановок доменов, имеющих консервативные ядра глобул.

Таблица 33 – Сродство онДНК/RecA (Mazin, Kowalczykowski, 1996)

№	Проба	Последовательность ДНК, $\mathbf{S_n}$	Сродство, Φ_n, ln-ед.	Activity
1	dC	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	0.54	
2	#40	ACCACCACACACGCGCACACCACCACACACGC	0.48	
3	htr#3	TTCACAAACGAATGGATCCTCATTAAAGCCAG	0.34	
4	#39	GCGTGTGTGGTGGTGTGCGCGTGTGTGGTGGT	0.33	
5	dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	0.09	АНА-
6	IDENT	CCATCCGCAAAAATGACCTCTTATCAAAAGGA	0.00	ЛИЗ
7	htr#4	CATGGAGCAGGTCGCGGATTTCGACACAATTT	-0.02	
8	#7	GGCGGGCGGCGCGGCCGGGCGGCGGGCGCGCG	-1.99	
9	htr#2	AATTCTTCGAAGCTAGCCCTCAGGCCTAGGCA	-2.42	
10	dA	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	-5.01	
11	A>T	CCTTCCGCTTTTTTGTCCTCTTTTCTTTTGGT	1.20	
12	G>C	CCATCCCCAAAAATCACCTCTTATCAAAACCA	0.03	
13	G>T	CCATCCTCAAAAATTACCTCTTATCAAAATTA	-0.40	КОНТ-
14	C>G	GGATGGGGAAAAATGAGGTGTTATGAAAAGGA	-1.00	РОЛЬ
15	C>T	TTATTTGTAAAAATGATTTTTTATTAAAAGGA	-1.20	
16	C>A	AAATAAGAAAAAATGAAATATTATAAAAAGGA	-3.40	

Измеренные величины сродства RecA/онДНК (Mazin, Kowalczykowski, 1996) можно видеть в Таблице 33. Для 16 нитей ДНК длиной 32 нт, S, величины сродства RecA/онДНК, Ф(S), варьировали от -5.01 до 1.20 натуральных логарифмических единиц (ln-ед.). Поскольку в филаменте один RecA связывает 3 нт онДНК (Mazin, Kowalczykowski, 1996), то с использованием системы Activity (Ponomarenko m. *et al.*, 1997a) было проанализировано взвешенное содержание в онДНК тринуклеотидов $[z_1z_2z_3]_f$ (формула 38).

На Рисунке 65 изображены примеры функции f(i) для наибольшего вклада в сродство RecA/онДНК для $z_1z_2z_3$ на 5’-конце онДНК (линия) и в позиции 19 (пунктир), через один виток филамента длиной 18 нт от 5’-конца онДНК (Mazin, Kowalczykowski, 1996). С помощью 360 таких f(i) было оценено $360\times15^3\approx10^6$ вариантов $[z_1z_2z_3]_f$.

Компьютерная система Activity (формулы 36, 38) обрабатывала 10 онДНК из верхней части Таблицы 33, остальные 6 онДНК были использованы для независимого контроля результатов этого анализа.

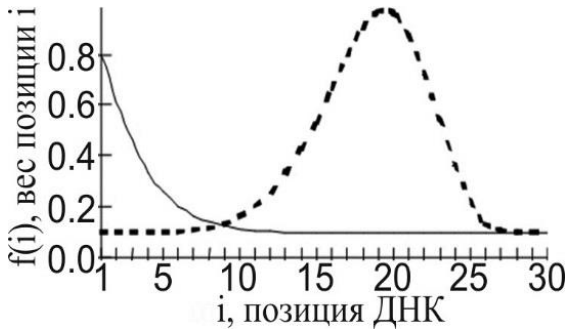

Рисунок 65 - Примеры весов f(i) как модели наибольшего вклада в сродство RecA/онДНК в случае $z_1z_2z_3$ на 5’-конце онДНК (линия) и в позиции 19 через виток RecA-филамента вокруг ДНК длиной 18 нт (пунктир)

При этом каждый $[z_1z_2z_3]_f$ из $\approx 10^6$ вариантов был исследован независимо от остальных, в результате чего он получил оценку $U([z_1z_2z_3]_f)$ полезности для прогноза *in silico* неизвестной величины Ф(S) сродства RecA/онДНК по известной последовательности онДНК, S. Всего пять из $\approx 10^6$ $[z_1z_2z_3]_f$ получили (Таблица 34) позитивные оценки полезности (формула 31).

Наибольшей $U([DRV]_f) = 0.27$ оказалась полезность тринуклеотидов DRV={AAA, AGA, TAA, TGA, GAA, GGA, AAG, AGG, TAG, TGG, GAG, GGG, AAC, AGC, TAC, TGC, GAC, GGC} (номенклатура, IUPAC-IUB, 1971), взвешенных f(i) с максимумом на 5’-конце онДНК (Рисунок 65: линия). Это означает, что сродство RecA/онДНК определяется локальным содержанием тринуклеотидов DRV вблизи 5’-конца неповрежденной части нити ДНК на фланге бреши в этой нити, которая возникла после вырезания из нее предмутационного повреждения как молекулярного события эксцизионной репарации этого повреждения геномной ДНК (Cox, 1993; Qin *et al.*, 2015).

Таблица 34 – Значимые тринуклеотиды $[z_1z_2z_3]_f$ для сродства RecA/онДНК

$z_1z_2z_3$	F(i), Рисунок 65	$U([z_1z_2z_3]_f)$	связь с лучшим (номенклатура IUPAC-IUB)
DRV	линия	0.270	Лучший
RVD	линия	0.229	Циклическая перестановка в лучшем $z_1z_2z_3$
VDR	линия	0.170	Циклическая перестановка в лучшем $z_1z_2z_3$
RRV	пунктир	0.191	Частный случай лучшего: DRV=RRV+TRV
RRM	пунктир	0.180	Частный случай лучшего: DRV=RRM+*TRV*+*RRG*

Две следующие по величине оценки $U([z_1z_2z_3]_f)$ оказались циклическими перестановками лучшего DRV с тем же весом f(i) (Таблица 34: RVD и VDR), две последние - у частных случаев DRV (Таблица 34: RRV и RRM), взвешенных $f^{\#}(i)$ с максимумом в позиции i=19 (Рисунок 65: пунктир) на удалении 1 виток спирали филамента от 5'-конца онДНК (Mazin, Kowalczykowski, 1996). Таким образом, позитивно оцененные $[z_1z_2z_3]_f$ указали на DRV вблизи 5'-конца разрыва нити ДНК в качестве наилучшей основы для предсказания сродства этой нити ДНК к RecA:

$$\Phi(S) = 0.54 - 1.03[DRV]_f. \quad (53)$$

Отрицательный коэффициент "−1.03" свидетельствует, что неспецифическое сродство RecA/онДНК является наибольшим в согласии с общепринятым мнением о равной защите всех районов геномной ДНК *E. coli* от их предмутационных повреждений, уровень которой может контекстно-зависимо уменьшаться с ростом локального содержания тринуклеотидов DRV вокруг таких повреждений ДНК.

Прогнозы *in silico* сродства Ф(S) RecA/онДНК (формула 53) показаны на Рисунке 66: а) анализ и б) контроль (Таблица 33). Их достоверность (r=0.81, $\alpha<0.05$) на независимом контроле означает, что формула (53) адекватно учитывает влияние нуклеотидного контекста на уровень репарационной защиты геномной ДНК *E. coli* от ее предмутационных повреждений.

Поскольку большая часть генома *E. coli* кодирует белки, то тринуклеотид DRV был сопоставлен с кодонами генетического кода с помощью точного критерия Фишера. Полученные результаты можно видеть в Таблице 35. Триплет DRV соответствует кодонам аргинина, глицина, лизина, серина, цистеина, тирозина, триптофана, аспарагина, аспарагиновой и глютаминовой кислот (Таблица 35). Были найдены три достоверных корреляции (Таблица 35) между кодоном DRV и общеизвестными свойствами аминокислот (Karlin *et al.*, 1989; Cohen *et al.*, 1991), две из которых были позитивными - избыток электростатического заряда ($\alpha<0.05$) и предпочтение

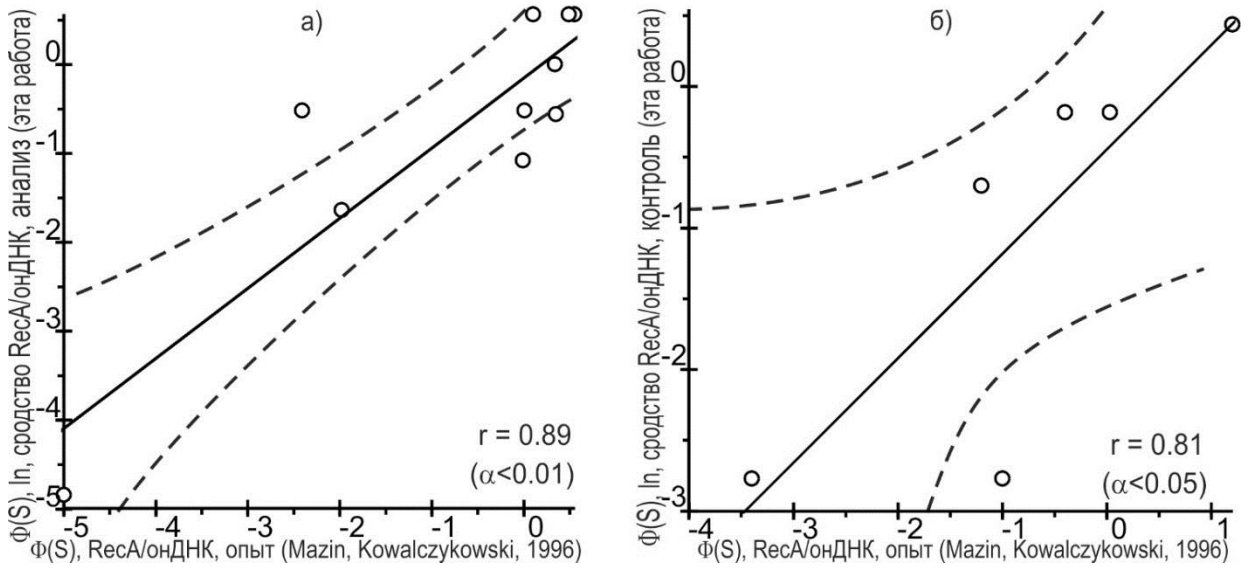

Рисунок 66 - Корреляция прогноза сродства RecA-филамента к ДНК с его измерением (Mazin, Kowalczykowski, 1996) на анализируемых данных (а) и на контроле (б). Пунктир – границы 95%-доверительных интервалов.

случайному клубку ($\alpha<0.025$), - и одна негативная корреляция с предпочтением к ядру глобулы ($\alpha<0.0025$).

Таким образом, наибольшее сродство RecA/онДНК и высшую меру защиты от предмутационных повреждений имеют DRV-бедные участки кодирования ядер глобулярных доменов белков (формула 53), тогда как низшую меру такой защиты имеют DRV-богатые районы кодирования неструктурных перемычек “случайный клубок” на поверхности белковых молекул, соединяющих их глобулярные домены. Этот результат настоящей диссертационной работы согласуется с независимыми экспериментальными данными (Vlasic et al., 2011) о том, что линии *E. coli* с мутацией recA730, увеличивающей сродство RecA/онДНК в сравнении с нормой, имеют повышенную устойчивость к мутагенному воздействию UV-излучения. Таким образом, открытие (Mazin, Kowalczykowski, 1996) чувствительности RecA к контексту онДНК, которая может обуславливать разную меру защиты различных районов геномной ДНК *E. coli* от ее предмутационных повреждений, соответствует общепринятым представлениям молекулярной биологии и генетики.

Таблица 35 – Проекция тринуклеотида DRV на генетический код *E. coli*

Аминокислота		Генетический код	Количество кодонов		
Название	Код	(DRV – подчеркнуты)	всех	DRV	не DRV
Аланин	A	GCG GCA GCT GCC	4		4
Аргинин	R	AGG AGA CGG CGA CGT CGC	6	2	4
Аспарагин	N	AAT AAC	2	1	1
Аспарагиновая кислота	D	GAT GAC	2	1	1
Цистеин	C	TGT TGC	2	1	4
Глютамин	Q	CAG CAA	2		2
Глютаминовая кислота	E	GAG GAA	2	2	0
Глицин	G	GGG GGA GGT GGC	4	3	1
Гистидин	H	CAT CAC	2		2
Изолейцин	I	ATA ATT ATC	3		3
Лейцин	L	TTG TTA CTG CTA CTT CTC	6		6
Лизин	K	AAG AAA	2	2	
Метионин	M	ATG	1		1
Фенилаланин	F	TTT TTC	2		2
Пролин	P	CCG CCA CCT CCC	4		4
Серин	S	AGT AGC TCG TCA TCT TCC	6	1	5
Треонин	T	ACG ACA ACT ACC	4		4
Валин	V	GTG GTA GTT GTC	4		4
Триптофан	W	TGG	1	1	0
Тирозин	Y	TAT TAC	2	1	1
СТОП-кодон		TAA, TAG, TGA	3	3	

а.к.о.		кодон	18 DRV		46 не DRV		критерий Фишера	
№	свойство а.к.о.	список а.к.о.	есть	нет	есть	нет	DRV	α
1	случайный клубок	YDNEKGAV	10	8	12	34	есть	0.025
2	заряженные	DEKRH	7	11	7	39	есть	0.05
3	ядро глобулы	LIMFV	0	18	16	30	нет	0.0025

а.к.о. – аминокислотный остаток.

В целом, представленные примеры применения системы Activity (Ponomarenko M. *et al.*, 1997a) к анализу количественных характеристик предмутационных повреждений геномной ДНК при воздействии мутагенов и репарации этих повреждений свидетельствуют о том, что анализ регуляторных районов геномов находится в границах адекватного применения

и надежной работы этой компьютерной системы, которая была создана в рамках выполнения настоящей диссертационной работы.

Представленные в этой главе результаты позволяют сделать следующие выводы:

- Выявлены контекстные характеристики ДНК плазмиды pGEM7(f+) *Escherichia coli*, достоверно коррелирующие с частотой повреждений ДНК по гуанинам под действием ультрафиолетового излучения лазера с длиной волны 193 нм. На этой основе впервые получено регрессионное уравнение для предсказания величин частоты таких повреждений гуанина в ДНК. Это уравнение подтверждено независимым экспериментом с дуплексами ДНК, идентичными фрагментам гена *MIP-1α* мыши.
- Выявлены достоверные корреляции: (а) между каталитической константой k_{CAT} 8-оксогуанин-ДНК-гликозилазы OGG1 человека и углом кручения B-формы ДНК в окрестности 8-оксогуанина (oxoG), а также (б) между константой Михаэлиса K_M этого фермента и изменением свободной энергии Гиббса при образовании гетеродуплекса ДНК в окрестности этого oxoG. На этой основе впервые выведены регрессионные уравнения для оценки величин этих констант при частичном нарушении комплементарности ДНК вокруг oxoG, которые были подтверждены независимыми экспериментальными данными.
- Показано, что сродство RecA к однонитевой ДНК достоверно убывает с ростом встречаемости в нити ДНК тринуклеотидов DRV в 15-буквенном коде IUPAC, которые достоверно соответствуют кодонам заряженных аминокислотных остатков.

ЗАКЛЮЧЕНИЕ ПО ГЛАВЕ 4

Репликация, транскрипция, сплайсинг, трансляция и другие молекулярно-генетические процессы регулируются сайтами со специфической активностью, которые функционируют в результате их

взаимодействия с соответствующими белками или РНК-белковыми комплексами (Neidle, 1994). В настоящее время известны тысячи конкретных вариантов таких сайтов, для которых определены их последовательности и локализация в ДНК или РНК (Колчанов, 1997; Бухер, 1997; Игнатьева и др., 1997; Кель О. и др., 1997; Кель А. и др., 1997). Компьютерный анализ сайтов до последнего времени был направлен на построение методов их распознавания в произвольных нуклеотидных последовательностях (см. обзор, Gelfand, 1995). Хотя в этой области достигнуты грандиозные успехи (Gelfand, 1995), становится все более ясно, что только распознавания сайтов недостаточно для понимания функциональной организации геномных ДНК. Это обусловлено тем, что любой сайт характеризуется, помимо его расположения в геноме, также и количественной величиной его биологической активности. Экспериментальные данные свидетельствуют о том, что сайты одного и того же типа, расположенные в различных участках ДНК, могут различаться по величине активности на несколько порядков (Berg, von Hippel, 1987).

Изучение особенностей функциональных сайтов, влияющих на величины их активности, становится все более важной проблемой молекулярной биологии, в первую очередь потому, что различия в уровнях активности сайтов создают основу для дифференциальной активности генов и их координированного функционирования в организмах про- и эукариот. Исследование этих особенностей является важным для конструирования молекулярно-генетических систем с заданными уровнями экспрессии целевых генов. Еще одним обстоятельством, привлекающим к этой проблеме большое внимание, является возможность мутаций в сайтах, которые зачастую приводят к потере их активности или, напротив, вызывают ее резкое усиление (Berg, von Hippel, 1987) и, тем самым, ведут к возникновению патологий (Подколодная, Степаненко, 1997).

С использованием компьютерной системы Activity (Ponomarenko *et al.*, 1997a) было исследовано более 70 выборок функциональных сайтов, для

каждой из которых был построен метод количественного прогноза активности, показывающий достоверное согласие с экспериментальными данными. Примеры таких выборок представлены в Таблице 36. Количество последовательностей в отдельной выборке варьирует от 7 до 50 при их общем числе более 1500. Они представляют природные сайты, их варианты при искусственном мутагенезе, а также искусственно синтезированные олигонуклеотиды и их гетеродуплексы. Количественными характеристиками специфической активности сайтов являются такие величины как кинетические и термодинамические константы комплекса “сайт/белок”; время жизни таких комплексов; количество специфического продукта гена (например, пре-мРНК для сайтов, регулирующих транскрипцию генов эукариот, или зрелой мРНК для сайтов 3’-концевого процессинга и сплайсинга, а также белка для сайтов инициации трансляции). Кроме того, экспериментально определяют контекстно-зависимые конформационные характеристики сайтов, зависящие от их нуклеотидного контекста, например, угол изгиба ДНК, а также частоты мутаций, вызванных определенным мутагеном. Например, наибольшее и наименьшее значения силы промоторов *E. coli* в условиях *in vivo* эксперимента (Jonsson *et al*., 1993) различались в 1000 раз; в случае таких величин равновесной константы диссоциации комплекса между активатором CRP в ответ на цикло-АМФ и сайтами его связывания в геноме *E. coli* – в 600 раз (Gartenberg, Crothers, 1988); в случае сродства Cro-репрессора к оператору OR1 в геноме фага λ - в 100 раз (Kim *et al*., 1987).

Для сайтов в составе геномных ДНК, РНК и их синтетических аналогов, документированных в созданной в рамках настоящей диссертации базе данных Activity (Ponomarenko J *et al*., 2001a), были выявлены значимые статистические, физико-химические и конформационные характеристики и на этой основе были оптимизированы методы предсказания биологической активности этих сайтов. Примеры выявленных характеристик и методов предсказания активности даны в Таблице 36. Рассмотрим некоторые из них.

Таблица 36 – Примеры анализа сайтов ДНК и РНК с помощью системы Activity (Ponomarenko M. *et al.*, 1997a)

сайт				контекстно-зависимая особенность			значимость			Литера-тура
название	позиция 1	n	активность, F	X_k	район	свойство, формула прогноза	U	r	α	
горячие точки 2АР-мутагенеза	точка мутации	26	количество замен С→Т	X_1	-1, 2	температура плавления, $T_{плав}$	0,20	0,76	10^{-5}	Ponoma-renko M. *et al.*, 1997a
				$F=-8.56+0.159X_1$						
промотор *E. coli*	старт транскрипции	27	сила промотора	X_1	-12; 14	Содержание [ASM]	0.59	0.82	10^{-64}	
				X_2	-4; 14	угол Direction	0.50			
				$F=0.3+0.6X_1+0.0008X_2$						
сайт связывания белка CRP (*E. coli*)	центр повтора в сайте	10	сродство CRP/ДНК	X_1	-15; 14	шаг rise спирали ДНК	0,15	0.87	0.0025	Колчанов и др., 1998
				X_2	-17; 12	ширина малой бороздки ДНК	0.06			
				$F=190-66.8\times X_1+7.5\times X_2$						
оператор OR1 фага λ для связывания репрессора Cro	первый нуклеотид сайта	7	сродство Cro/ДНК	X_1	1; 16	ширина малой бороздки ДНК	0,55	0.99	0.00005	
				X_2	6; 19	раскрытие roll	0,44			
				X_3	6, 19	шаг rise спирали ДНК	0,41			
				$F=-72+4X_1+X_2+13X_3$						
сайт 3'процес-синга пре-мРНК	точка разрезания	16	выход мРНК	X_1	-27; 25	содержание [VUKK]	0.24	0,88	0.001	
				$F=-5.98+3X_1$						
микроРНК человека	последний нуклеотид зрелой микроРНК	28	сродство микроРНК к Ago2 и Ago3	X_1	-22; -1	Содержание [RHHK]	0.36	0.66	0.00025	Омельян-чук и др., 2011
				X_2	-22; -1	Содержание [DRYD]	0.36			
				$F_{AGO2}=4.97 + 0.52X_1 + 1.35X_2$						
				$F_{AGO3}=6.11 – 0.52X_1 + 1.35X_2$						
микроРНК арабидопсиса	первый нуклеотид микроРНК	27	содержание микроРНК в органах	X_1	1; 20	Содержание [WRHW]	0.48	0.58	0.0025	Пономаренко М. и др., 2006
				X_2	1; 20	Содержание [DRYD]	0.47			
				$F=-0.78+1.31X_1+0.76X_2$						
элемент ответа генов растений на ауксин	центр сайта	50	коэффициент индукции экспрессии	X_1	-16; 16	Содержание [SNW]	0.24	0.71	10^{-9}	Mironova et al., 2013
				X_2	-12; 12	Сдвиг slide п.о.	0.47			
				$F= 3.33 -0.26X_1 – 1.39X_2$						

n – объем выборки; X_k – значимая контекстно-зависимая количественная характеристика ДНК сайта; $F=F_0+\Sigma_{k=1,K} F_k\times X_k$ – линейно-аддитивное приближение; цАМФ – циклоАМФ.

Предложенный подход оказался применим к анализу широкого круга экспериментальных данных, которые имеют наиболее общий формат “нуклеотидная последовательность и величина ее специфической активности”. Например, с помощью предложенного в главе 3 диссертации подхода было показано, что частота индукции мутаций C→T под действием 2-аминопурина (Coulondre *et al*., 1978) зависит от температуры плавления ДНК вокруг точки мутации (Рисунок 67а). Это согласуется с существующими представлениями (Mhaskar, Goodman, 1984), согласно которым первичное повреждение ДНК в случае 2-аминопурина не зависит от контекста ДНК, тогда как частота ошибок репарационной системы определяется его стэкинг взаимодействием с соседними нуклеотидами в нити ДНК и температурой плавления ДНК вокруг точки мутации.

Система Activity (Ponomarenko M. *et al*., 1997) выявила содержание тринуклеотида ASM (номенклатура, IUPAC-IUB, 1971) на участке [-12; 14] и средний угол Direction спирали ДНК на участке [-4; 14] в качестве значимых для силы 27 промоторов *E. coli* (Jonson *et al*., 1993). Это соответствует выводу об этих экспериментальных данных, сделанному их авторами о том, что нуклеотиды на участках [-12; -8] и [4; 14] вносят наибольший линейно-аддитивный вклад в величины этой биологической активности геномной ДНК *E. coli* (Jonson *et al*., 1993). При этом итоговый прогноз значимо коррелирует с результатом эксперимента (Рисунок 67б).

В случае анализа экспериментальных данных (Gartenberg, Crothers, 1988) о величинах сродства белка CRP, система Activity (Ponomarenko M. *et al*., 1997) выявила в качестве наиболее значимых контекстно-зависимых конформационных характеристик спирали ДНК сайтов для связывания этого активатора генов *E. coli* в ответ на цикло-АМФ ширину малой бороздки и шаг rise спирали ДНК (Таблица 36). На их основе были достоверно (Рисунок 67в: $r = 0.87$, $\alpha < 0.0025$) предсказаны количественные величины сродства регуляторного белка CRP к сайтам его связывания в составе геномной ДНК *Escherichia coli*.

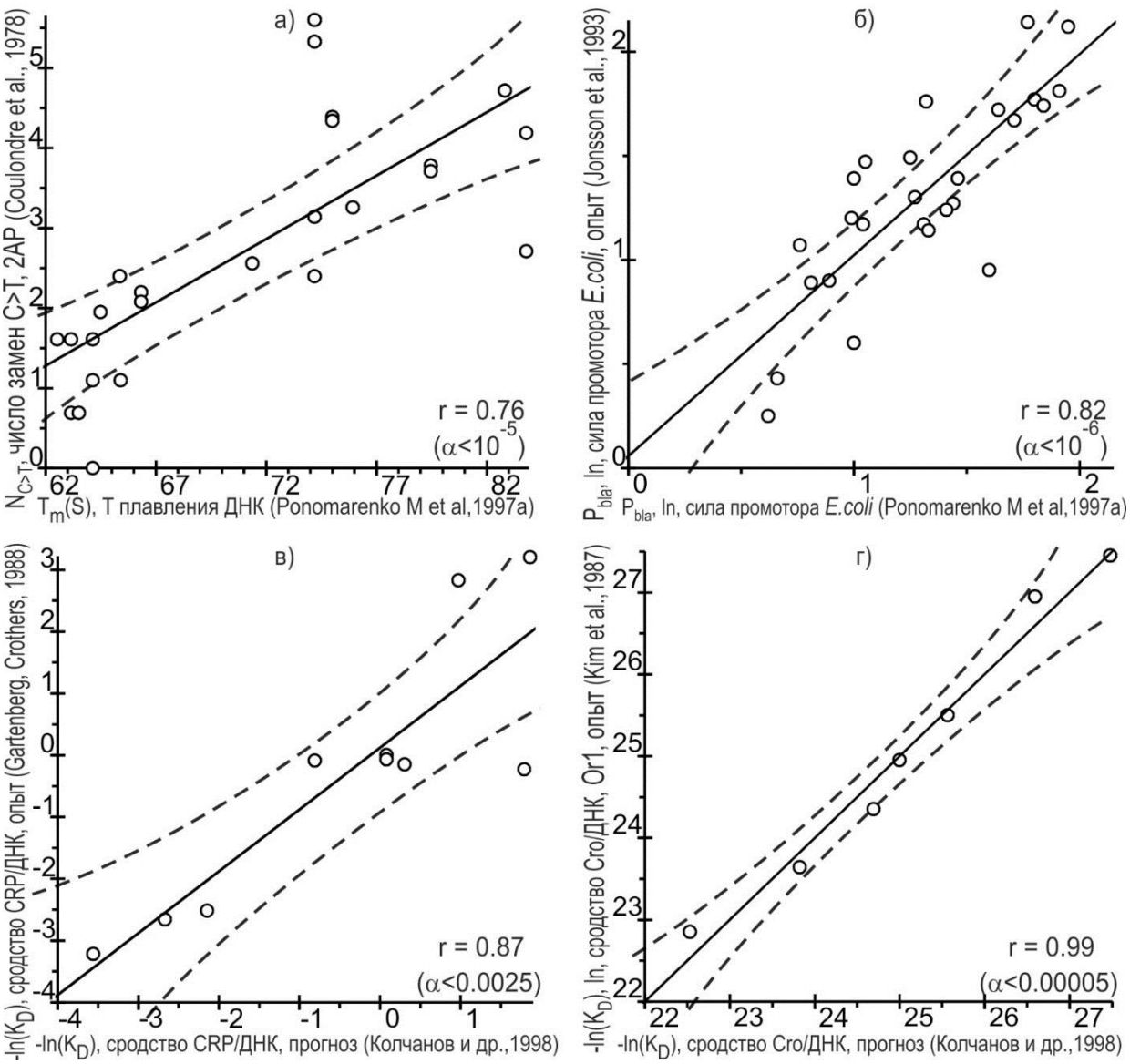

Рисунок 67 - Достоверные корреляции между прогнозами системы Activity (Ponomarenko M. *et al*., 1997a), ось x, и экспериментальными данными, ось y, (а) число замен C → T под действием 2-аминопурина (Coulondre *et al*., 1978), (б) силы промоторов *E. coli* (Jonson *et al*., 1993), (в) сродства белка CRP к сайтам его связывания в геноме *E. coli* (Gartenberg, Crothers, 1988) и (г) сродства Cro-репрессора к оператору OR1 в геноме фага λ (Kim *et al*., 1987). Пунктир – границы 95%-доверительных интервалов. Рисунок автора на основе из его статей (Ponomarenko M. *et al*., 1997a; Колчанов и др., 1998).

Аналогично, с использованием системы Activity (Ponomarenko M. *et al*., 1997a) были даны достоверные (Рисунок 67г: $r = 0.99$, $\alpha < 0.00005$)

предсказания величин сродства репрессора Cro к нормальному и природным мутантным вариантам оператора OR1 фага λ (Kim *et al*., 1987) на основе выявленных этой системой значимых контекстно-зависимых оценок ширины малой бороздки, угла roll раскрытия п.о. и шага rise спирали ДНК этого сайта-переключателя (Таблица 36). В свою очередь, единственной характеристикой контекста первичного транскрипта РНК, достоверно (Рисунок 68а: r=0.88, $\alpha<10^{-4}$) коррелировавшей с интенсивностью выхода пре-мРНК вируса SV40 в опыте (McDevitt *et al*., 1986) оказалось содержание тетрануклеотида VUKK справа от точки 3'-процессинга пре-мРНК. Тетрануклеотид VUKK является G/U-богатым, поскольку V={A, C, **G**} и K={**U**, **G**} (номенклатура, IUPAC-IUB, 1971), что соответствует общеизвестной локализации точки 3'-концевого отрезания пре-мРНК от первичного транскрипта посередине между облигатным консенсусом AATAAA и G/U-богатым районом (McDevitt *et al*., 1986).

Использование системы Activity (Ponomarenko M. *et al*., 1997a) позволило предсказать относительные содержания микроРНК в арабидопсисе (Пономаренко М. и др., 2006, 2008), а также величины сродства микроРНК человека к белкам Ago2 и Ago3 семейства *Argonaute* (Омельянчук Н. и др., 2011; Ponomarenko M. *et al*., 2013), которые достоверно коррелируют (Рисунок 68б,в,г) с данными экспериментов (Axtell, Bartel, 2005) и (Azuma-Mukai *et al*., 2008), соответственно.

Наконец, в статье (Mironova *et al*., 2013) применили (Рисунки 69а,б,в) систему Activity (Ponomarenko M. *et al*., 1997a) к данным сканирующего мутагенеза ауксин-чувствительного района в промоторе гена *IAA4/5* гороха (Ballas *et al*., 1995) и достоверно (Рисунок 69г: r=0.62, $\alpha<0.025$) предсказала влияние замен нуклеотидов в опыте (Ulmasov *et al*., 1997) по созданию репортерного конструкта P3(x4), индикатора ауксина в клетках растений.

В целом, созданная в рамках настоящей диссертации система Activity (Ponomarenko M *et al*., 1997a) анализирует данные в условиях определенного

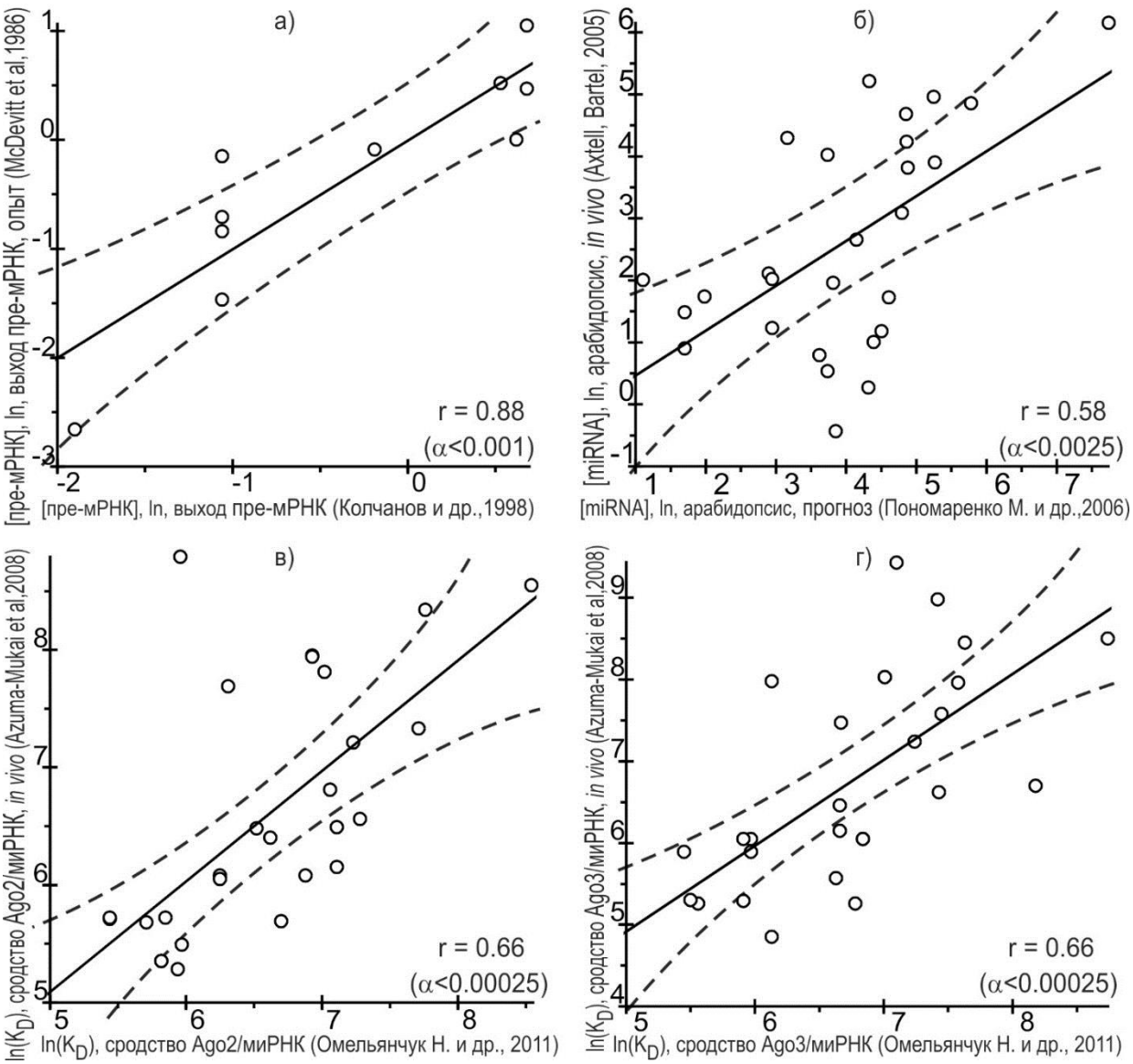

Рисунок 68 - Достоверные корреляции между прогнозами системы Activity (Ponomarenko M. *et al*., 1997a), ось x, и экспериментальными данными, ось y, (а) выход пре-мРНК вируса SV40 (McDevitt *et al*., 1986), (б) содержания миРНК в арабидопсисе (Axtell, Bartel, 2005) и (в, г) сродства миРНК человека к белкам Ago2 и Ago3 (Azuma-Mukai *et al*., 2008). Пунктир – границы 95%-доверительных интервалов. Рисунок автора на основе иллюстраций из его статей (Пономаренко М. и др., 2006; Колчанов и др., 1998, 2013; Омельянчук и др., 2011).

эксперимента, для которых она выявляет биологически значимые количественные характеристики контекста.

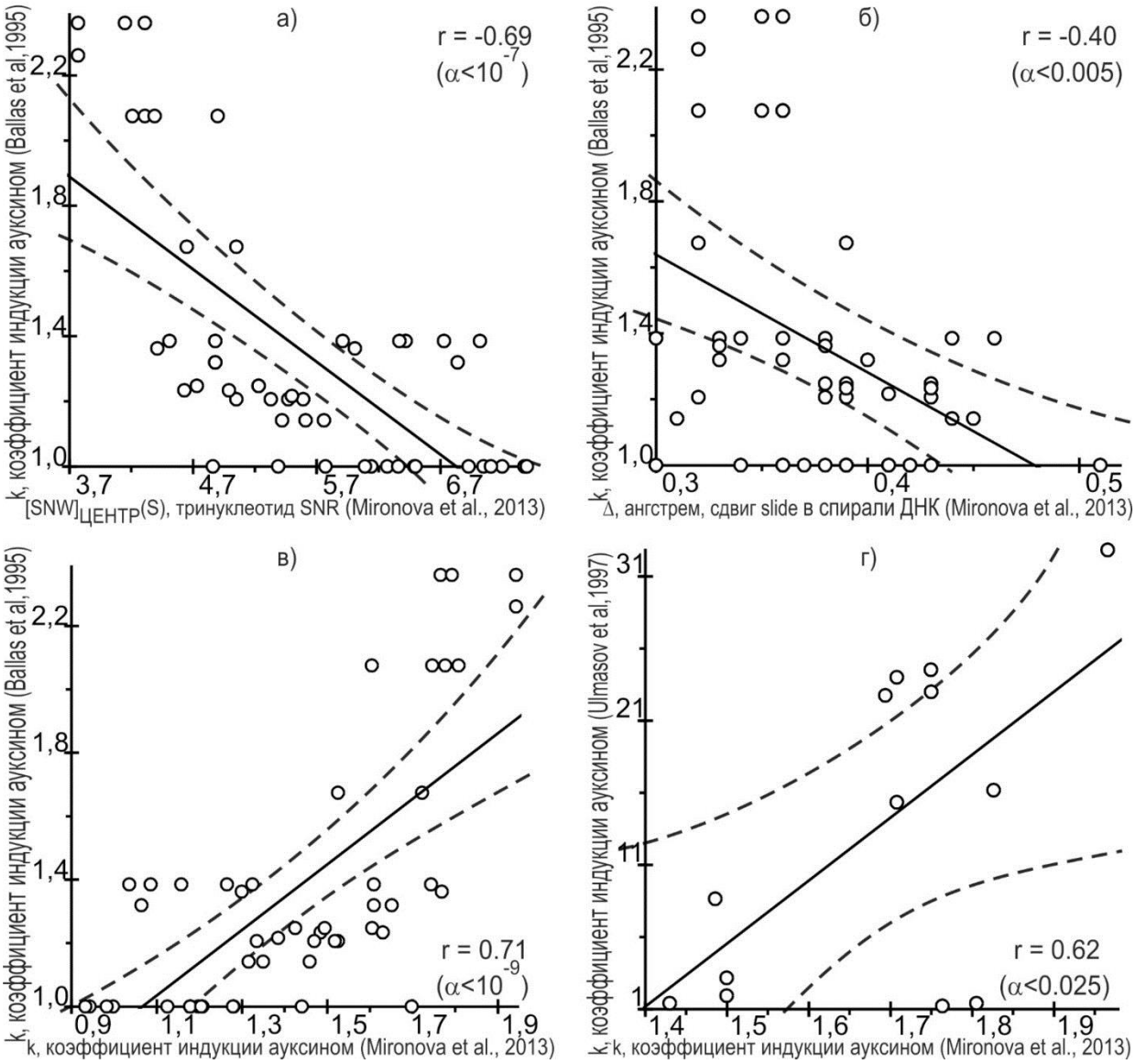

Рисунок 69 - Достоверные корреляции между прогнозами системы Activity (Ponomarenko M. *et al*., 1997a), ось x, и экспериментальными данными, ось y, об индукции генов растений ауксином (а, б, в) по сканирующему мутагенезу промотора гена IAA4/5 гороха (Ballas *et al*., 1995) и по созданию репортерного конструкта P3(x4) для индикации ауксина в клетках растений (Ulmasov *et al*., 1997). Пунктир – границы 95%-доверительных интервалов.

Представленные примеры результатов системы Activity (Ponomarenko M *et al*., 1997a) демонстрируют ее применимость к широкому кругу экспериментальных величин биологической активности сайтов в составе геномных ДНК и РНК для изучения связывания белков про- и эукариот в условиях опытов *in vitro, ex vivo* и *in vivo*.

ГЛАВА 5 КОНТЕКСТНО-ЗАВИСИМЫЕ КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ ДНК, КОРРЕЛИРУЮЩИЕ С АКТИВНОСТЬЮ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ

Настоящая глава, на примере сайтов связывания трех транскрипционных факторов MEF-2, USF и YY1, описывает применение представленной в предыдущих главах диссертации компьютерной системы Activity (Ponomarenko M *et al*., 1997a) с целью преодоления ряда проблем традиционных методов консенсуса (Hawley, McClure, 1983) и позиционно-весовых матриц (Mulligan *et al*., 1984) для компьютерного распознавания таких сайтов в составе геномных ДНК на основе множественного символьного выравнивания их экспериментально доказанных вариантов.

Сначала будет изучена проблема неоднородности выборок экспериментально доказанных сайтов связывания транскрипционных факторов вследствие включения в эти выборки сайтов из разных таксонов, которые, соответственно, связывают разные таксон-специфичные варианты одного и того же транскрипционного фактора в разных таксон-специфичных транскрипционных машинах при разных таксон-специфичных внутриклеточных условиях в ответ на разные таксон-специфичные молекулярные сигналы.

Затем будет рассмотрена проблема неоднородности этих выборок вследствие различий между вариантами определенного сайта по условиям экспериментов, в которых была доказана функциональность каждого из них в соответствующих клеточных линиях.

В заключение будет исследована одна из возможностей преодоления этих проблем за счет эмпирического учета существенных условий экспериментов в компьютерном прогнозе результатов определенного опыта на основе использования контекстно-зависимых количественных характеристик ДНК, выявленных с помощью данных других опытов.

5.1 Количественные характеристики ДНК сайта связывания транскрипционного фактора MEF-2

Одной из проблем неоднородности выборок последовательностей ДНК экспериментально доказанных сайтов связывания транскрипционных факторов является использование сайтов из разных таксонов. ДНК-связывающие домены ортологов определенного регуляторного белка могут иметь таксономические различия в структуре и функции их ДНК-связывающих центров. В настоящем разделе диссертации эта проблема рассматривается на примере трех мутантных форм транскрипционного фактора MEF-2C мыши, из которых форма INS-1GG может отличать сайты MEF-2 от A/T-богатых участков ДНК (например, от сайтов связывания транскрипционного фактора SRF с консенсусом $CC(A/T)_6GG$ (Pellegrini *et al*., 1995)), что является нормой, в отличие от дефектных форм M1DEL и R3K, не обладающих этой способностью (Meierhans *et al*., 1997).

Транскрипционные факторы семейства MEF-2 (**M**yocyte **E**nhancer **F**actor-2) участвуют в регуляции дифференцировки мышечных и нервных клеток в эмбриональном развитии, содержат MADS-бокс и, в форме гомодимера, связывают сайт с консенсусом $CTA(A/T)_8TAG$ (Meierhans *et al*., 1997; Meierhans, Allemann, 1998). В опытах *in vitro* (Meierhans *et al*., 1997; Meierhans, Allemann, 1998) измерили (Таблица 37) величины $\Delta\Delta G$ от -0.14 Ккал/М до 2.77 Ккал/М изменения свободной энергии Гиббса комплексов ДНК с тремя мутантными вариантами INS-1GG, M1DEL и R3K фактора транскрипции MEF-2C мыши, охарактеризованными выше.

Сначала был исследован вариант INS-1GG: вставка двух глицинов, GG, перед N-концевым метионином, - неотличимый от нормы по способности распознавать сайты MEF-2 (Meierhans *et al*., 1997; Meierhans, Allemann, 1998). В этом случае с помощью представленной в предыдущих главах диссертации компьютерной системы Activity (Ponomarenko M. *et al*., 1997a) были проанализированы шесть “обучающих” проб MEF-Site, MEF-D1, MEF-1T,

Таблица 37 – Измеренные (Meierhans, Allemann, 1998) изменения свободной энергии Гиббса, $-\Delta\Delta G$ (Ккал/М), комплексов ДНК с мутантами факторами транскрипции MEF-2C мыши, их разделение на данные для анализа (*курсив*) и для независимого контроля (подчеркнуты).

Проба	MEF-2 сайт, ДНК, позиция 1 - Заглавная	Варианты фактора транскрипции MEF-2C		
		INS-1GG (норма)	M1DEL	R3K
MEF(-2A,-1T)	ctgctaaTaatagag	-0.47	*0.01*	
MEF-D4	gctgctaTaaatgag	-0.16	0.01	*-0.08*
MEF-Site	ctgctatAaatagag	*0.00*	*0.00*	*0.00*
MEF(-1T)	ctgctatTaatagag	-0.40	-0.04	*-0.10*
MEF(-3T)	ctgctttAaatagag	-0.05	*-0.04*	
MEF-D(-1,1)	gctgctaTatagagt	-0.07	-0.08	*-0.11*
MEF-D1	gctgctaTaatagag	*-0.41*	*-0.09*	
MEF-1T	ctgctatAtatagag	*0.14*	-0.10	*0.01*
MEF-D(-4)4(SRF)	gctgcatAaatgagt	*-1.72*	*-0.66*	*-0.42*
MEF-A-tract	ctgcaaaAaaaagag	*-1.34*	-1.13	
MEF(-1G,1C)	ctgctatGcatagag	-1.68	-1.34	*-1.33*
NO-Ebox	aggcagcAggtggtg		*-2.77*	

MEF-D(-4)4(SRF), MEF-A-tract из Таблицы 37, в числе которых были наибольшее и наименьшее значения $\Delta\Delta G$ в качестве учета диапазона значений этой величины, уникальный поли-А тракт, и оба целевых сайта MEF-2 и SRF. Шесть оставшихся проб MEF(-2A, -1T), MEF-D4, MEF(-1T), MEF-D(-1, 1), MEF(-3T), MEF(-1G, 1C) были независимым контролем для результатов этого анализа (Таблица 37).

Всего с помощью Activity (Ponomarenko M *et al*., 1997a) было проанализировано (15-1)×(15-2)×38=3458 вариантов оценок $P_{k,[a;b]}$ средних значений для 38 свойств спирали ДНК из базы данных PROPERTY (Колчанов и др., 1998). Наибольшая оценка $U(\Delta\Delta G; P_{9,[-5;4]})=0.48$ указала на ширину малой бороздки В-формы спирали ДНК (Karas *et al*., 1996), коррелирующую с $\Delta\Delta G$ на контроле (Рисунок 70б: $r=0.68$, $\alpha<0.025$). Это согласуется с отнесением белка MEF-2C к суперклассу “β-слой в контакте с малой бороздкой спирали ДНК” (Вингендер, 1997), Рисунок 42.

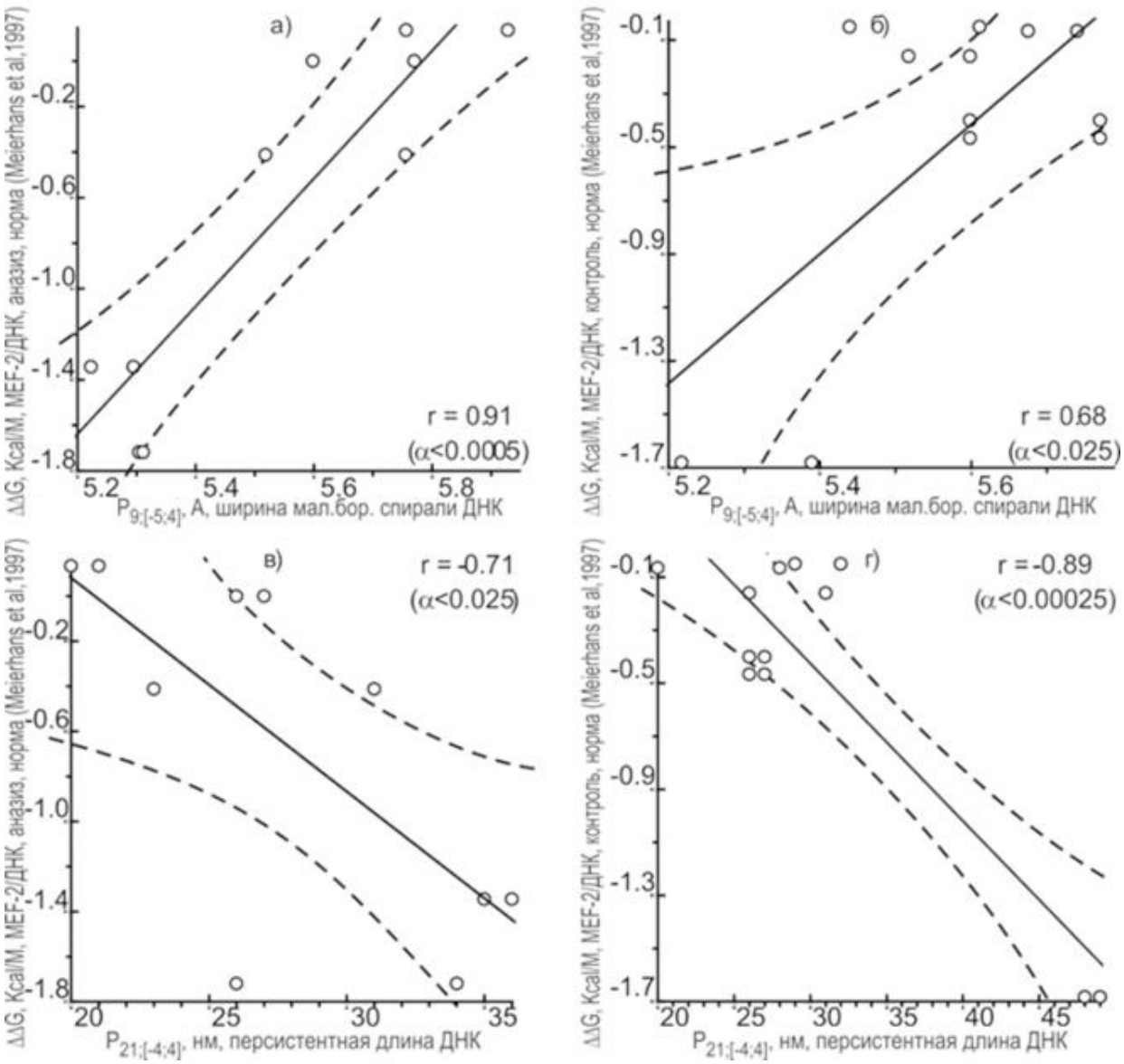

Рисунок 70 - Достоверные корреляции свободной энергии Гиббса, -ΔΔG, комплекса ДНК с мутантной формой INS-1GG фактора транскрипции MEF-2 мыши, отличающим сайты MEF-2 от сайтов связывания других факторов транскрипции (например, SRF) в АТ-богатой ДНК, что является нормой (а, б) с шириной малой бороздки спирали ДНК; (в, г) с персистентной длиной ДНК. В Таблице 37 *курсив* выделяет данные для анализа, подчеркивание – контроль. Пунктир – границы 95%-доверительных интервалов.

Поскольку $P_{9,[-5;4]}$ объясняла менее 50% вариансы измерений опыта (Meierhans *et al*., 1997; Meierhans, Allemann, 1998), то была также рассмотрена следующая $U(\Delta\Delta G; P_{21,[-4;4]})=0.39$ по убыванию, указавшая на среднюю

персистентную длину спирали ДНК (Hogan, Austin, 1987) сайта MEF-2 и достоверно коррелирующая с ΔΔG на контроле (Рисунок 70г: r=−0.89, $\alpha<0.00025$).

На этой основе с помощью пакета Statistica (Statsoft™, Tulsa, USA) на "обучающих" пробах была построена множественная регрессия величин ΔΔG по последовательностям ДНК, S:

$$-\Delta\Delta G_{\text{ДНК;норма}}(S) = -17.43 + 2.98P_{8;[-5;5]}(S) + 0.01P_{21;[-4;3]}(S). \quad (54)$$

Оценки формулы (54) коррелируют с измерениями ΔΔG (Meierhans *et al*., 1997; Meierhans, Allemann, 1998) на независимых контрольных пробах (рис.71б: r=0.61, $\alpha<0.05$).

Затем был исследован вариант M1DEL - делеция N-концевого метионина, - не отличающий сайты MEF-2 от сайтов SRF (Meierhans *et al*., 1997; Meierhans, Allemann, 1998). Для него был проведен аналогичный анализ с помощью Activity (Ponomarenko M. *et al*., 1997a), начиная с формирования наборов проб для анализа и для контроля, заканчивая выводом формулы для компьютерного прогноза неизвестных ΔΔG. При этом наибольшая оценка U(ΔΔG; $P_{21,[-4;3]}$)=0.50 была у средней персистентной длины (Hogan, Austin, 1987) района [-4; 3] в центре MEF-2 сайта, которая достоверно негативно коррелировала с величинами ΔΔG на независимом контроле (Рисунок 72б, r=−0.88, $\alpha<0.00025$). Как можно видеть на Рисунке 72а, она исчерпывающе (r=−0.97, $\alpha<0.000001$) объяснила вариансу оценок ΔΔG для варианта M1DEL из (Meierhans *et al*., 1997; Meierhans, Allemann, 1998).

На основе указанной линейной корреляции была построена простая регрессия Пирсона:

$$-\Delta\Delta G_{\text{ДНК}/M1DEL}(S) = 1.25 - 0.05P_{21;[-4;3]}(S). \quad (55)$$

Наконец был исследован третий вариант R3K: вставка GG перед N-концевым метионином и замена аргинина на лизин в третьей позиции белка, который также не отличал сайты MEF-2 от сайтов SRF (Meierhans *et al*., 1997;

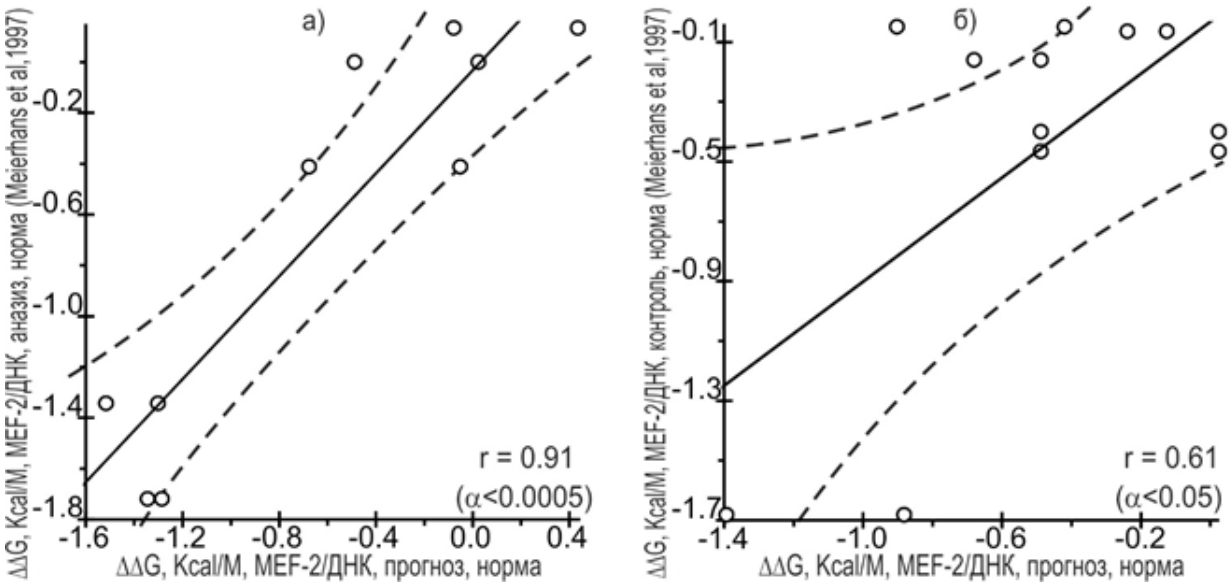

Рисунок 71 - Достоверные корреляции между предсказанными (формула 54) и экспериментально измеренными (Meierhans *et al*., 1997; Meierhans, Allemann, 1998) величинами ΔΔG свободной энергией Гиббса комплекса ДНК с мутантным вариантом INS-1GG (норма) фактора транскрипции MEF-2 мыши, способным отличать сайты MEF-2 от сайтов связывания других факторов транскрипции (например, от SRF) в АТ-богатой ДНК: а) анализ (Таблица 37, *курсив*); б) контроль (Таблица 37, подчеркнут). Пунктир – границы 95%-доверительных интервалов.

Meierhans, Allemann, 1998). Для него были проанализированы все пробы (Таблица 37). Наибольшая оценка $U(\Delta\Delta G; P_{21,[-4;3]})=0.45$ снова указала на среднюю персистентную длину (Hogan, Austin, 1987) района [-4; 5] в центре сайта связывания транскрипционного фактора MEF-2, достоверно коррелирующую со всеми измерениями ΔΔG (Рисунок 72а: $r=-0.97$, $\alpha<0.000001$).

На Рисунке 73 показано сравнение измерений –ΔΔG в случае мутанта R3K (вертикальная ось) с прогнозами этих величин, горизонтальная ось, (а) по формуле (55) для дефектного варианта M1DEL, не способного различать сайты MEF-2 от сайтов SRF ($r=0.95$, $\alpha<0.000001$) и (б) по формуле (54) для варианта INS-1GG, способного различить эти сайты ($r=0.70$, $\alpha<0.01$).

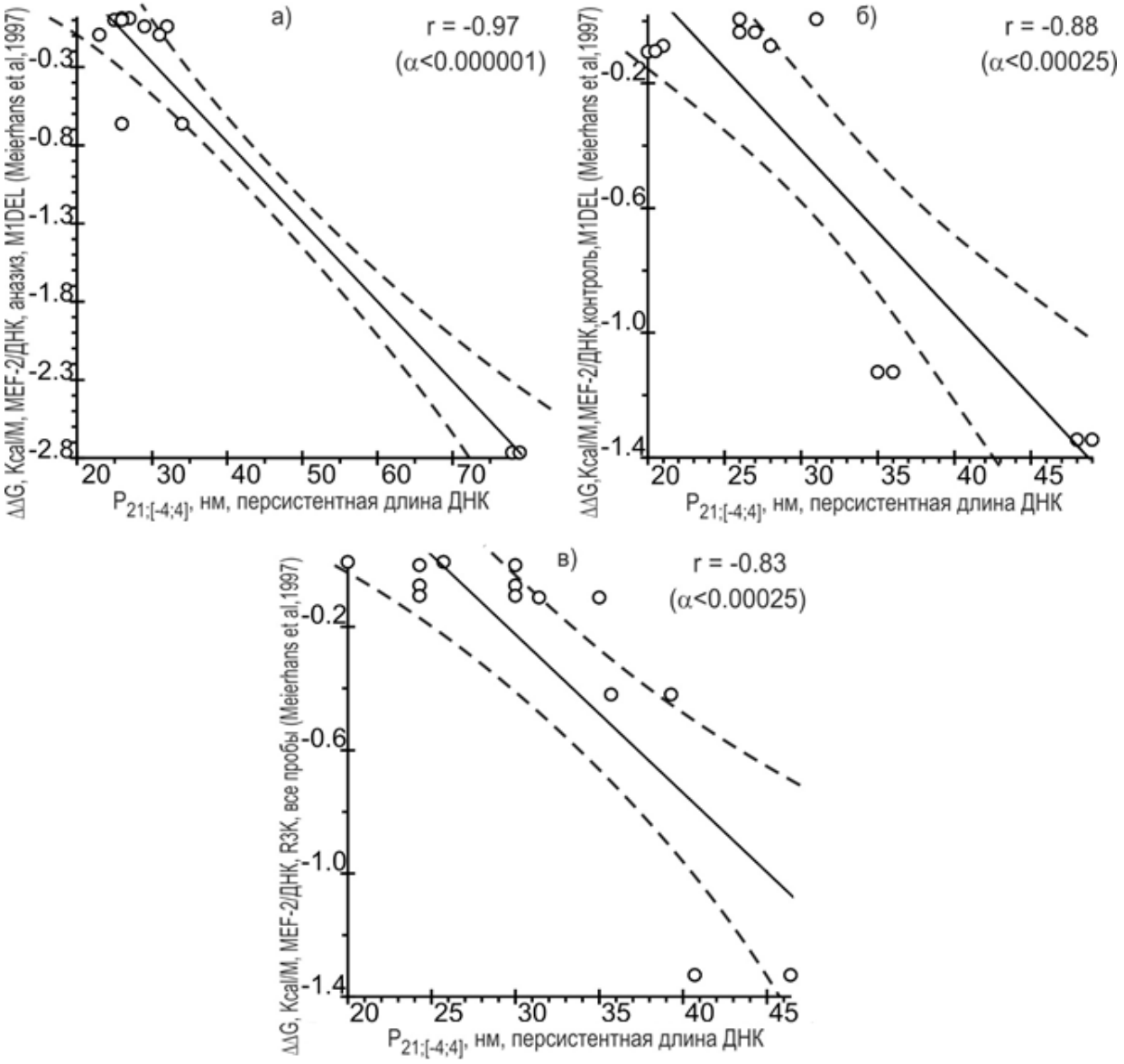

Рисунок 72 - Достоверные корреляции между персистентной длиной ДНК сайта MEF-2 и изменением ΔΔG свободной энергии Гиббса комплексов ДНК с двумя неспособными отличить их от других A/T-богатых сайтов мутантными формами транскрипционного фактора MEF-2: мутант M1DEL на анализируемых (а: r=-0.91, α<0.000001, *курсив* в Таблице 37) и контрольных (б: r=-0.88, α<0.00025; подчеркнуты в Таблице 37) данных, а также мутант R3K, на независимых контрольных данных (в: r=-0.83, α<0.00025). Пунктир – границы 95%-доверительных интервалов.

Видно, что основанная на данных об одном дефектном варианте M1DEL формула (55) исчерпывающе объясняет вариансу измерений –ΔΔG для

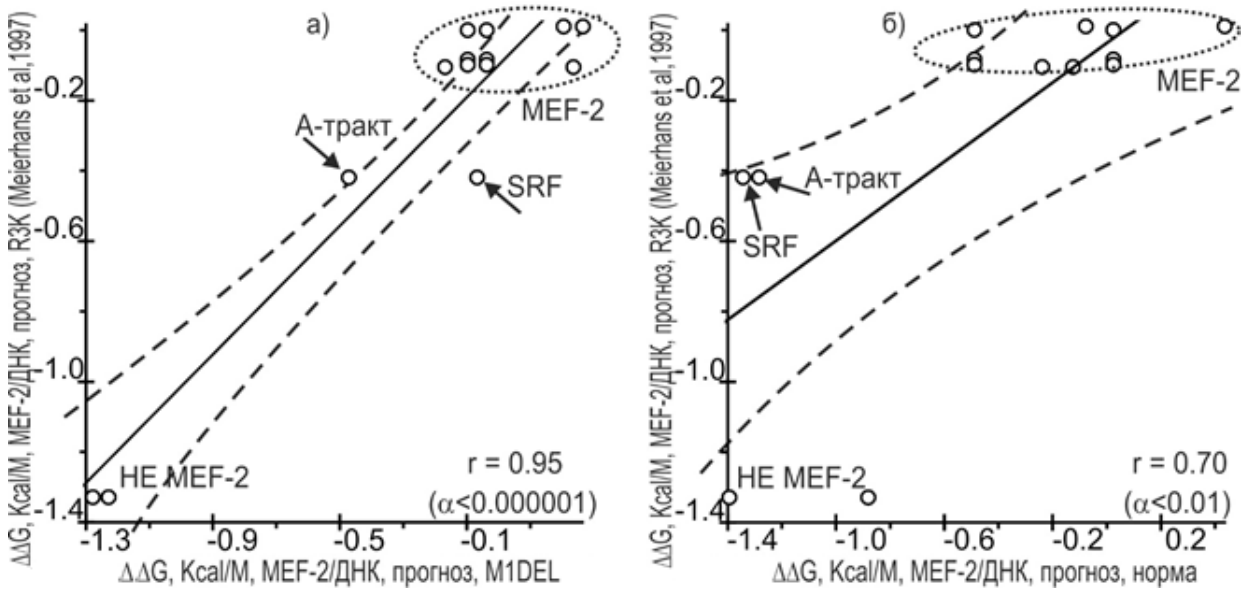

Рисунок 73 - Линейные корреляции (а) с уровнем значимости ($\alpha < 0.000001$) между предсказанными оценками изменения $\Delta\Delta G$ свободной энергии Гиббса комплексов ДНК с мутантами M1DEL и R3K, не способными отличать сайты MEF-2 от сайтов SRF в силу утраты этими мутантами чувствительности к ширине малой бороздки ДНК и (б) значимая ($\alpha < 0.01$) между такими оценками для мутанта R3K, а также для мутанта INS-1GG с нормой, способной отличать сайты MEF-2 от сайтов SRF и от A/T-богатых сайтов ДНК. Пунктир – границы 95%-доверительных интервалов.

другого дефектного варианта R3K. Это означает, что дефектные варианты R3K и M1DEL обладают общей биологически значимой закономерностью: они оба являются не чувствительными к ширине малой бороздки B-формы спирали ДНК.

Таким образом, с помощью системы Activity (Ponomarenko M. *et al*., 1997a) было установлено, что делеция N-концевого метионина (вариант M1DEL) и замена аргинин→лизин в позиции 3 (вариант R3K) в транскрипционном факторе MEF-2C мыши ведут к утрате этим белком способности отличать сайты MEF-2 от AT-богатых сайтов связывания других белков (в частности, от сайтов SRF) вследствие утраты чувствительности к ширине малой бороздки B-формы спирали ДНК.

Этот вывод согласуется с экспериментально установленным фактом (Pellegrini *et al*., 1995), что комплекс ДНК/MEF-2 стабилизируют три водородные связи между атомами трех тиминов в ДНК и введенными в малую бороздку ДНК атомами N-концевого метионина и аргинина в позиции 3 транскрипционного фактора MEF-2.

Выявленные с помощью компьютерной системы Activity (Ponomarenko m *et al*., 1997a) различия между тремя вариантами транскрипционного фактора MEF-2C мыши, способными и не способными отличать сайты MEF-2 от других A/T-богатых сайтов, свидетельствуют о возможности уточнения традиционных консенсусов и позиционно-весовых матриц, построенных по таксономически-неоднородным выборкам экспериментально доказанных сайтов связывания регуляторных белков, путем учета таксономически значимых различий этих сайтов при их компьютерном распознавании.

В главе 2 настоящей диссертации на примере различий между TATA-боксами промоторов генов у дрожжей, беспозвоночных, позвоночных и *E. coli* была продемонстрирована способность другой созданной в рамках настоящей диссертации компьютерной системы bDNAvideo (Ponomarenko M *et al*., 1997b) выявлять значимые контекстно-зависимые различия конформационных и физико-химических свойств спирали ДНК сайтов связывания регуляторных белков у разных групп организмов, стоящих на разных ступенях эволюции, которые адекватно отражали измерения локального контекста TATA-боксов в процессе эволюционного усложнения молекулярных транскрипционных машин.

5.2 Количественные характеристики ДНК сайта связывания транскрипционного фактора USF

В качестве одного из путей преодоления неоднородности выборок экспериментально доказанных сайтов связывания регуляторных белков вследствие различий между вариантами определенного сайта по условиям

опытов, в которых доказывали функциональность каждого из этих его вариантов в соответствующих клеточных линиях (см. обзоры, Gold *et al*., 1997; Werstuck, Green, 1998; Ponomarenko J *et al*., 2000a, 2002c), в начале 1990-х появились принципиально новые подходы *in vitro* к экспериментальному поиску олигоДНК, способных связывать заданные белки: SELEX (Systematic Evolution of Ligands by EXponential enrichment - Ellington, Szostak, 1990; Tuerk, Gold, 1990), SAAB (Selected And Amplified Binding site imprint assay - Blackwell, Weintraub, 1990), CASTing (Cyclical Amplification and Selection of Targets – Wright *et al*., 1991), REPSA (Restriction Endonuclease Protection Selection and Amplification – Hardenbol *et al*., 1997) и ряд других (например, (Kinzler, Vogelstein, 1989; Pollock, Treisman, 1990)).

Однако, для более чем 50 регуляторных белков *E. coli* было установлено (Robison *et al*., 1998), что общепринятые методы консенсуса (Hawley, McClure, 1983) и позиционно-весовых матриц (Mulligan *et al*., 1984) на основе результатов селекции *in vitro* оказываются неадекватными для распознавания сайтов связывания регуляторных белков по ее последовательности из-за достоверного различия частот нуклеотидов в соответствующих позициях природных и селектированных *in vitro* сайтов связывания определенного белка (Shulzaberger, Schneider, 1998).

В этой связи возникла необходимость найти адекватный подход к учету этих новых экспериментальных данных наряду с общепринятыми природными регуляторными сайтами в составе геномной ДНК, поскольку игнорирование этих новых экспериментальных данных, очевидно, нарушало целостность существующих биологических представлений, на основе которых создаются методы распознавания сайтов связывания ДНК с белками (Roberts, Ja, 1999). Поэтому предложенная в предыдущих главах диссертации система Activity (Ponomarenko M *et al*., 1997a) была апробирована на примере экспериментальных данных SELEX-протокола (Bendall, Molloy, 1994) для сайтов связывания транскрипционного фактора USF (**U**pstream **S**timulatory **F**actor).

Таблица 38 – Сродство $\ln(n/n_{max})$ транскрипционного фактора USF человека к аффинным ему олигоДНК 5’-gctggatcctN_{25}tctagatcgagctcg-3’ (Bendall, Molloy, 1994) и результат их анализа

Сродство USF/ДНК (Bendall, Molloy, 1994)				Activity		Про-гноз
Проба	Последовательность ДНК#, N_{25}	n_S,%	$n(n_S/n_{max})$	$P_{1,[10;18]}$	$P_{9,[10;13]}$	
A50	ACCACGTGACTACAGTGGGTGTGAA	100	0.00	34.61	8.96	-0.28
A54	ACCACGTGTTTAAGTTGTACCTGAG	23.	-1.47	34.58	8.95	-0.09
A16	ACCACGTGAGATATCGTGATTTCGG	17	-1.77	35.86	8.98	-1.48
A4	ACCACGTGTGAGCTTCGGTTAGCGA	10	-2.30	35.65	8.99	-1.49
A42	ACCACATGACGGACAGGTTGTGATA	7.8	-2.55	36.37	8.99	-2.00
A9	ACCACGTGTAGCGGAGCCACGAAGA	6.4	-2.75	36.20	9.03	-2.53
A52	ACCACTTGTGAATTGGCGTTATGTG	5.3	-2.94	36.60	9.01	-2.48
A45.3	ACCACGTTACTATGGAGTCAAGTCG	3.9	-3.24	36.75	8.99	-2.26
A30	ACCACGTGATGTGGGTGTACAGGAT	66	-0.42	34.50	8.96	-0.20
A56	ACCACGTGCTTTGTTACGGTTAGTA	22.	-1.51	34.85	8.98	-0.77
A60	ACCACGTGAGGGTTCGGAGTTAAGA	15.	-1.90	35.41	8.98	-1.16
A2	ACCACATGGTACAAAGAAGCAAAGT	8.9	-2.42	37.04	8.98	-2.30
A51	ACCACGGGGTAAAGCGGAACTTCTA	4.2	-3.17	36.40	9.00	-2.18
A20	ACCACGCGGCCAGCGCACCCTCTCG	3.8	-3.27	36.21	9.03	-2.54
Линейная корреляция	анализ: A50-A45.3		r=	-0.89	−0.76	0.87
	контроль A56-A20			−0.83	−0.87	0.85
Значимость	анализ: A50-A45.3		α	0.005	0.05	0.005
	контроль A56-A20			0.05	0.005	0.005

Фактор USF является повсеместным (ubiquitous, англ. яз.) ядерным белком, связывающим E-бокс, консенсус CACGTG, промоторов генов эукариот наряду с другими транскрипционными факторами. В опыте (Bendall, Molloy, 1994) на основе олигоДНК 5’-gctggatcctN_{25}tctagatcgagctcg-3’ с помощью SELEX-протокола (Tuerk, Gold, 1990) были отобраны *in vitro* 14 олигоДНК длиной 25 п.о., N_{25}, на сродство к фактору USF человека и охарактеризованы относительной величиной n_S от 100% до 3.8% этого сродства (Таблица 38). Фрагменты S участка N_{25} восьми проб A50, A54, A16, A4, A42, A9, A52, A45.3 и соответствующие им оценки $\Psi(S)=\ln(n_S/n_{max})$ были отобраны для анализа с помощью компьютерной системы Activity (Ponomarenko M. *et al*., 1997a), остальные 6 проб A30, A56, A60, A2, A51, A20 составили независимый контроль результата этого анализа.

Поскольку общепринятые подходы на основе частот нуклеотидов были признаны неадекватными (Shulzaberger, Schneider, 1998) для применения селектированных *in vitro* олигоДНК к распознаванию сайтов связывания регуляторных белков в природных ДНК, то количественные характеристики B-формы двойной спирали ДНК, достоверно коррелирующие с измеренными *in vitro* величинами сродства между транскрипционным фактором USF и высокоаффинным к нему олигоДНК, были выявлены с помощью Activity (Ponomarenko M. *et al*., 1997a). Для этого оценивались среднеарифметические $P_{k,[a;b]}$ значения 38 свойств динуклеотидных шагов спирали ДНК из базы данных PROPERTY (Колчанов и др., 1998) на всех (25-1)×(25-2)/2=276 возможных участках [a;b] длины от 2 п.о. до 25 п.о. последовательности $S=\{s_1...s_{25}\}$, всего 276×38=10488 вариантов $P_{k,[a;b]}$. Наибольшая оценка $U(\Psi; P_{1,[10;18]})=0.23$ указала на среднее кручение спирали ДНК (Karas *et al*., 1996) участка [10; 18] анализируемых олигоДНК. Величины $P_{1,[10;18]}$ для всех проб даны в Таблице 38. Они достоверно коррелируют с измерениями Ψ сродства на контроле: $r= -0.82$, $\alpha<0.05$ (Рисунок 74б). Эта негативная корреляция согласуется с данными рентгеноструктурного анализа комплекса USF/ДНК (Ferre-D'Amare *et al*., 1994), где кручение спирали ДНК было 33^O, которое было меньше кручения идеальной спирали ДНК Уотсон-Крика, 36^O.

Следующая по убыванию оценка $U(\Psi; P_{9,[10;13]})=0.22$ указала глубину малой бороздки спирали ДНК (Karas *et al*., 1996) на участке [10; 13] в центре фрагмента N_{25} USF-связывающих олигоДНК с шаблоном "5'-gctggatcctN_{25}tctagatcgagctcg-3'" в эксперименте (Bendall, Molloy, 1994). Количественные величины $P_{9,[10;13]}$ (Таблица 38) достоверно коррелируют с измерениями сродства Ψ на контроле: $r=-0.87$, $\alpha<0.005$ (Рисунок 74г). Эта негативная корреляция также соответствует выводу авторов рентгеноструктурного анализа комплекса ДНК/USF (Ferre-D'Amare *et al*., 1994), что глубина малой бороздки ДНК была меньше в сравнении с таковой в случае идеальной двойной спирали ДНК Уотсон-Крика.

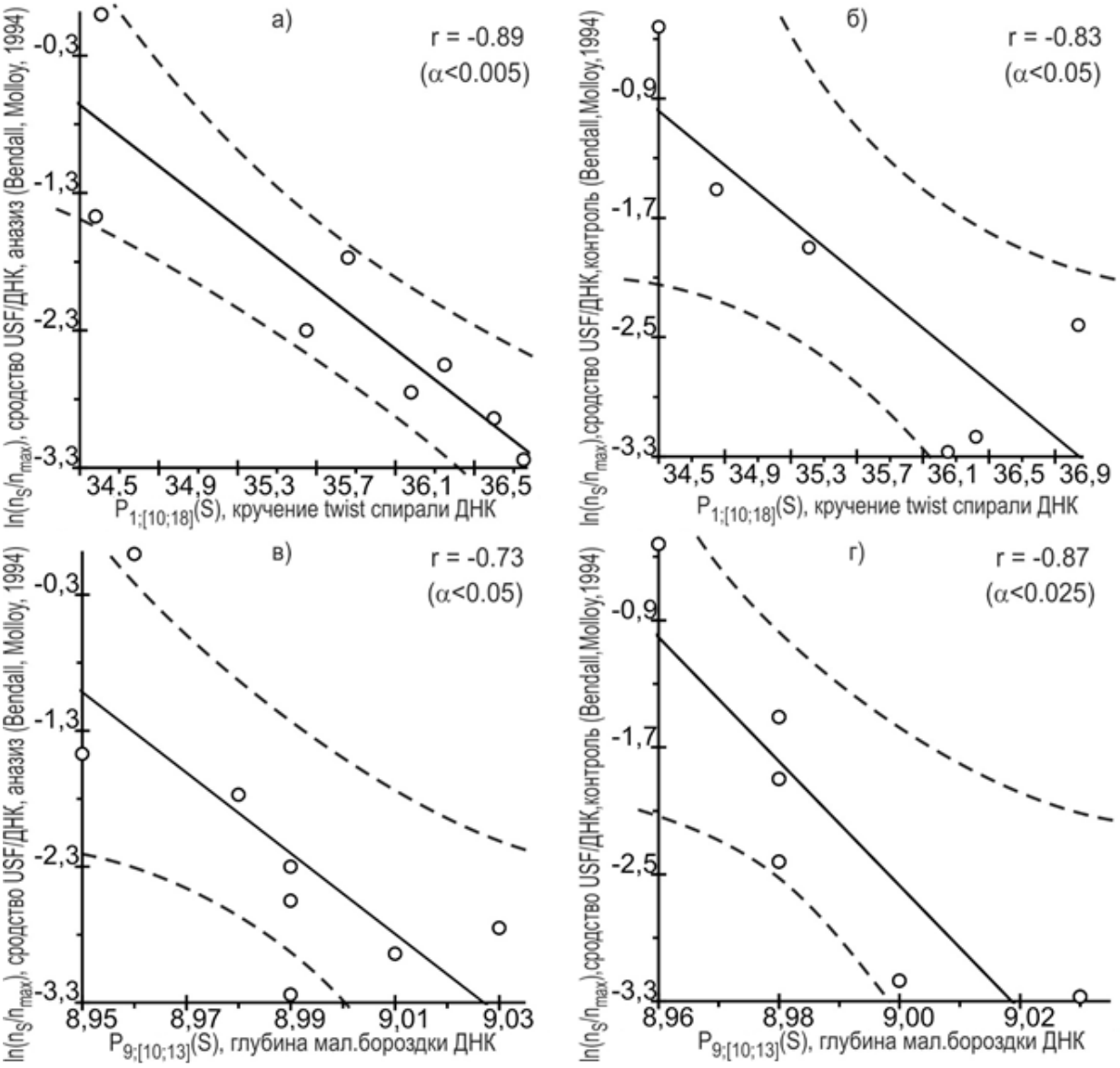

Рисунок 74 - Значимые свойства спирали ДНК селектированных *in vitro* USF-аффинных олигоДНК 5’-gctggatcctN_{25}tctagatcgagctcg-3’ (Bendall, Molloy, 1994): кручение района [10; 18] (а) анализ, проб A50, A54, A16, A4, A42, A9, A52, A45.3 из Таблицы 38 ($r=-0.89$, $\alpha<0.005$) и (б) контроль, пробы A30, A56, A60, A2, A51 и A20 ($r=-0.83$, $\alpha<0.05$); глубина малой бороздки спирали ДНК района [10; 13] (в) анализ, $r=-0.73$, $\alpha<0.05$) и (г) контроль, $r=-0.87$, $\alpha<0.005$. Пунктир – границы 95%-доверительных интервалов.

На этой основе с помощью стандартного пакета Statistica (Statsoft™, Tulsa, USA) для проанализированных проб была оптимизирована линейная регрессия (Рисунок 75а: $r=0.87$, $\alpha<0.005$):

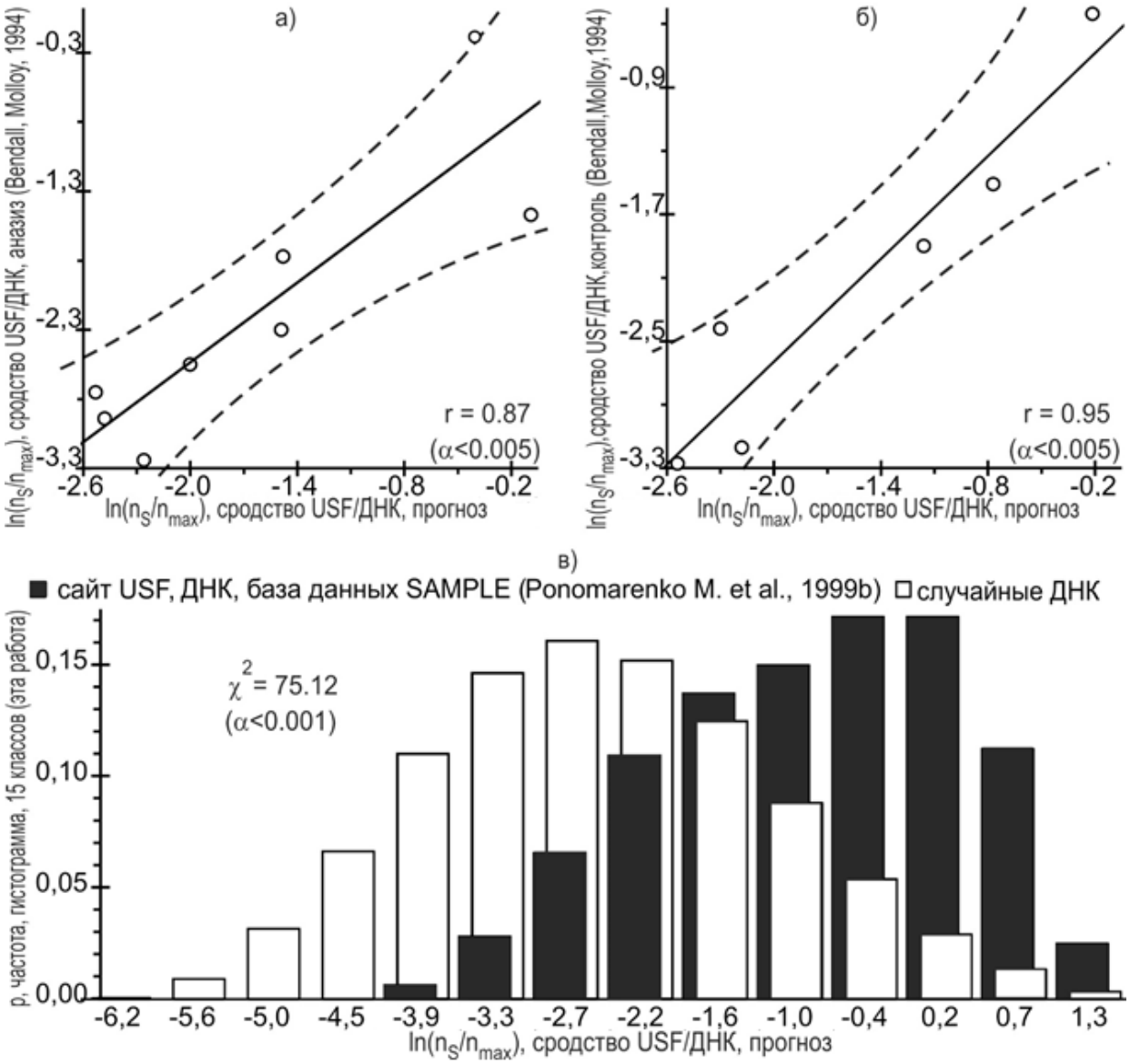

Рисунок 75 - Прогноз (формула 56) сродства USF/ДНК для (а) обработанных системой Activity (Ponomarenko m *et al*., 1997a) данных (пробы A50, A54, A16, A4, A42, A9, A52, A45.3), r=0.87 ($\alpha<0.005$) и (б) для контрольных проб A30, A56, A60, A2, A51, A20 (r=0.95, $\alpha<0.005$), отобранных *in vitro* на USF-аффинность олигоДНК (Bendall, Molloy, 1994); (в) достоверное ($\chi^2=75.12$, $\alpha<10^{-3}$) отличие природных сайтов USF (■, база данных, (Ponomarenko M. *et al*., 1999b)) и случайных ДНК (□). Пунктир – границы 95%-доверительных интервалов.

$$\Psi(S) = 170 - 0.7P_{1;[10;18]}(S) - 16.3P_{9;[10;13]}(S). \qquad (56)$$

Прогноз формулы (56) можно видеть в Таблице 38. Он достоверно коррелирует с измерениями сродства USF/ДНК (Bendall, Molloy, 1994) на контроле, r=0.95, $\alpha<0.005$ (Рисунок 75б). Наконец, с помощью формулы (56) был сделан прогноз величин Ψ(S) для 20 экспериментально доказанных сайтов связывания транскрипционного фактора USF из базы данных SAMPLE (Ponomarenko M. *et al*., 1999b) и для 1000 ДНК такой же длины из независимых равновероятных случайных нуклеотидов.

На Рисунке 75в дано достоверное ($\chi^2=75.12$, $\alpha<0.001$) различие между гистограммами природных сайтов USF (■) и случайными ДНК (□), построенными на 15 классах равновеликих интервалов оценок сродства USF/ДНК (формула 56).

Это достоверное отличие природных сайтов USF от случайных ДНК по оценкам (формула 56) сродства USF/ДНК на основе USF-аффинных олигоДНК, селектированных *in vitro* (Bendall, Molloy, 1994), свидетельствует об адекватности результатов системы Activity (Ponomarenko m *et al*., 1997a) для селектированных *in vitro* олигоДНК на аффинность к заданным регуляторным белкам-мишеням.

Выявленные таким путем значимые контекстно-зависимые конформационные и физико-химические свойства спирали ДНК сайтов связывания этих регуляторных белков могут дополнить консенсус и/или позиционно-весовой матрицу, которые являются традиционными методами распознавания регуляторных сайтов на основе учета частот встречаемости нуклеотидов в позициях этих сайтов.

В разделе 3.4 настоящей диссертации на примере позиционно-частотной матрицы (Bucher, 1990) было продемонстрировано дополнение общепринятого критерия TATA-бокса оценками сродства TBP к одно- и двунитевым ДНК. Это усовершенствование втрое увеличило долю объясненной вариансы экспериментально измеренных величин TBP/TATA-сродства и дало достоверный прогноз их изменений при минимальных

заменах одного нуклеотида в промоторах генов человека (Savinkova *et al*., 2013; Drachkova *et al*., 2014).

Таким образом, в рамках настоящей диссертационной работы удалось впервые обнаружить такие контекстно-зависимые характеристики сайтов связывания транскрипционных факторов, которые могли бы улучшить прогностические свойства позиционно-весовых матриц сайтов связывания транскрипционных факторов. В недавней работе (Ponomarenko, Ponomarenko, 2015) вне рамок настоящей диссертации ее автору удалось впервые показать статистическую достоверность этого улучшения.

5.3 Количественные характеристики ДНК сайта связывания транскрипционного фактора YY1

Любопытной биологической особенностью транскрипционного фактора **YY1** является его способность быть одновременно как активатором экспрессии генов с YY1-зависимыми инициаторными Inr-элементами стартов транскрипции, так и репрессором транскрипции этих генов при наличии других сайтов связывания YY1 в их промоторах (Hyde-DeRuyscher *et al*, 1995), что было отражено в названии этого регуляторного белка ("**Y**in-**Y**ang **1**" англ. яз., или "Инь-Янь 1" русс. яз.).

В эксперименте (Hyde-DeRuyscher *et al*., 1995) в клетках линии HeLa, трансфецированных плазмидой pTiLUC, измеряли *ex vivo* величины φ биолюминесценции, репортерной активности гена люциферазы, LUC, регулируемого промотором-химерой, сконструированным из TATA-бокса позднего промотора аденовируса, Inr-элемента из гена терминальной деоксинуклеотидил трансферазы и синтетического YY1-аффинного олигоДНК между позициями −88 и −66 от старта транскрипции. Всего было 13 вариантов олигоДНК, амплифицированных/селектированных *in vitro* и охарактеризованных величинами φ от 1±0.05% до 97±25%, что

Таблица 39 – Уровни YY1-зависимой люминесценции (активность репортерного гена LUC в плазмиде pTiLUC, промотор которого содержит сайт связывания транскрипционного фактора YY1, репрессора для этого гена) при 13 *in vitro* селектированных YY1-аффинных олигоДНК длиной 23 п.о. между позициями -88 и -66 от старта транскрипции (Hyde-DeRuyscher *et al*., 1995).

Экспериментальные данные (Hyde-DeRuyscher *et al*., 1995)				кручение ДНК
in vitro YY1-аффинный олигоДНК		φ, LUC, *ex vivo*, HeLa		район [10; 21],
проба	Последовательность ДНК	%	ln-единицы	формула (57)
11	GACGCCATTTTAAGTCCTAACGA	12±3	-2.12	34.63
32	TCGTTAAATCCGCCATTTGCGTC	15±4	-1.90	34.36
61	TCGTCCATTTTGTTCCTCCCGTC	13±6	-2.04	34.25
15	GACGCCATTATCCTCCATTACGA	1	-4.60	33.72
43	TCGTCCATTTGTAATATGTCGTC	3±1	-3.50	34.38
44	TCGTATGTCCGCCATGTTGCGTC	14±5	-2.0	34.19
55	GACGGCGCCATTTTGTGTTACGA	16±3	-1.83	34.22
28	TCGTCCATTTTTGTCATGTCGTC	8±2	-2.53	34.30
72	GACGCGTCCATTTTGTTGTACGA	10±4	-2.30	34.22
79	GTCGTCCATATTGTAATGGCGTC	7±1	-2.66	34.10
80	TCGTCGGCCATCTTGTCTGCGTC	1	-4.60	33.91
91	GACGCATCCATCTTGACTTACGA	6±1	-2.81	33.98
14	TCGTTTAGTTAATACTTCGCGTC	97±25	-0.03	34.79
			Коэффициент линейной корреляции, r	0.80
			Значимость, α	<0.001

соответствовало диапазону величин от -4.60 до -0.03 натуральных логарифмических единиц (ln), Таблица 39.

Все эти данные были проанализированы компьютерной системой Activity (Ponomarenko M. *et al*., 1997a), как это было описано в главах 3 и 4, а также в предыдущих разделах этой главы. В результате набольшая оценка U(LUC; $P_{14,[10;21]}$)=0.27 указала на среднее кручение спирали участка [10; 21] олигоДНК в комплексах ДНК/белок (Suzuki *et al*., 1996), свойство №14 из базы данных PROPERTY (Колчанов и др., 1998). Величины $P_{14,[10;21]}$ для всех проб показаны в Таблице 39. Они достоверно (r=0.80, α<0.001) коррелируют с экспериментальными величинами φ репортерной активности гена LUC на всех

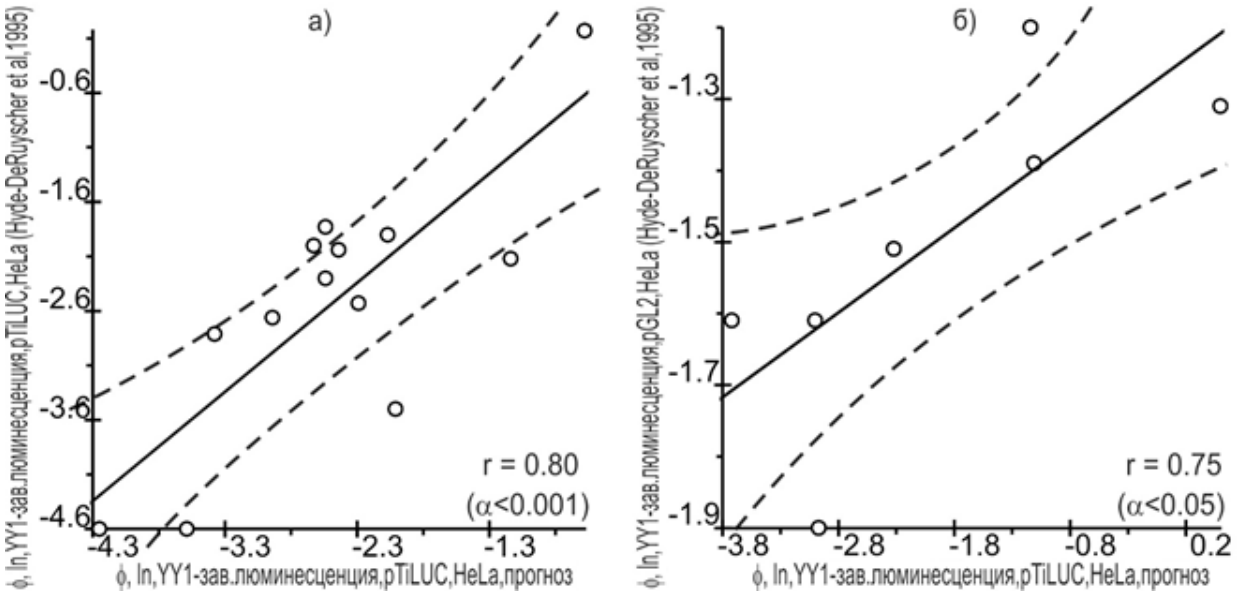

Рисунок 76 – Достоверные корреляции между предсказанными (формула 57) и измеренными величинами репортерной LUC активности *ex vivo* в опыте (Hyde-DeRuyscher *et al*., 1995) в случае трансфекции клеток HeLa (а) плазмидой pTiLUC (13 проб из Таблицы 39, проанализированных с использованием Activity (Ponomarenko M *et al*., 1997a)), и (б) плазмидой pGL2 (8 таких проб, которые были независимыми контрольными данными для системы Activity). Пунктир – границы 95%-доверительных интервалов.

“обучающих” олигоДНК. Это означает уменьшение репрессорного воздействия транскрипционного фактора YY1 на экспрессию гена люциферазы с увеличением кручения спирали ДНК в комплексе YY1/ДНК. Этот вывод соответствует данным рентгеноструктурного анализа комплекса YY1/ДНК (Houbaviy *et al*., 1996), где виток спирали ДНК 12 п.о. имел кручение 30^{O}, меньше кручения 36^{O} идеальной спирали ДНК Уотсона-Крика.

На основе этой достоверной корреляции была построена простая регрессия:

$$\varphi_{i=9}(S) = -119.46 + 3.42P_{14;[10;21]}(S). \quad (57)$$

Оценки (формула 57) величин φ(S) для всех олигоДНК достоверно коррелируют с их измерениями (Hyde-DeRuyscher *et al*., 1995), Рисунок 76а.

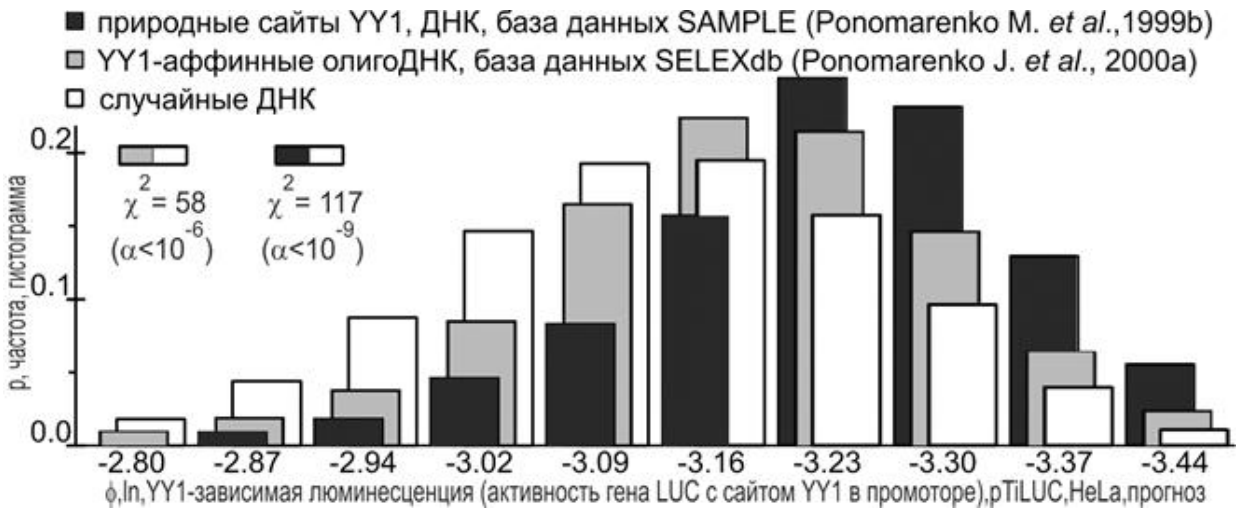

Рисунок 77 - Сравнение прогноза (формула 57) величин φ репортерной LUC-активности в клетках HeLa для (■) экспериментально установленных сайтов связывания транскрипционного фактора YY1 из базы данных SAMPLE (Ponomarenko M *et al*., 1999b), (■) селектированных *in vitro* YY1-аффинных олигоДНК из базы данных SELEXdb (Ponomarenko J. *et al*., 2000a) и (□) случайных ДНК.

На Рисунке 76б показано сравнение прогноза формулы (57) в случае четырех вариантов замен в олигоДНК №11 и четырех вариантах замен в олигоДНК №15, встроенных в другой вектор pGL2 с тем же конструктом. Этот прогноз достоверно ($r=0.75$, $\alpha<0.05$) коррелирует с измерениями *ex vivo* на этом независимом контроле.

Наконец, на Рисунке 77 представлена гистограмма сравнения прогноза (формула 57) величин φ репортерной LUC-активности в клетках HeLa человека для 27 экспериментально доказанных сайтов связывания транскрипционного фактора YY1 из базы данных SAMPLES (Ponomarenko M. *et al*., 1999b) и всех 106 селектированных *in vitro* YY1-аффинных олигоДНК из базы данных SELEXdb (Ponomarenko J. *et al*., 2000a), а также для 8000 последовательностей такой же длины из случайных равновероятных независимых нуклеотидов. Случайные ДНК были достоверно

дискриминированы от природных (χ^2=116.68, $\alpha<10^{-9}$) и от селектированных *in vitro* (χ^2=58.01, $\alpha<10^{-6}$) сайтов YY1. При этом селектированные *in vitro* YY1-аффинные олигоДНК (▪) заняли промежуточное положение между случайными ДНК (□) и природными сайтами YY1 (■).

Этот результат подтверждает вывод предыдущего раздела этой главы диссертации в случае USF: результаты системы Activity (Ponomarenko M *et al*., 1997a) на данных селекции *in vitro* олигоДНК на аффинность к заданному белку-мишени могут способствовать улучшению традиционного распознавания сайтов ДНК для этого белка на основе консенсуса и позиционно-весовых матриц.

Авторы опыта (Hyde-DeRuyscher *et al*., 1995) обнаружили также на примере олигоДНК № 32, 44, 72 и 91 из Таблицы 39, что измеренные *ex vivo* величины φ репортерной LUC-активности в клетках линий HeLa и PYS-2 не коррелируют между собой (Рисунок 78а), как это обсуждалось в главе 1 настоящей диссертации при обосновании ее темы на основе обзора литературы. Это указало на возможное различие молекулярных механизмов YY1-зависимой репрессии между клеточными линиями HeLa и PYS-1, например, вследствие различия между мозаиками транскрипционных машин, которые способны собираться на одном и том же промоторе в этих клеточных линиях.

В порядке дискуссии был эвристически смоделирован эмпирический учет возможности специфических различий в сборке транскрипционных машин на промоторе-химере из опыта *ex vivo* (Hyde-DeRuyscher *et al*., 1995) в терминах точечных (φ_i), наибольших ($\max(\varphi_i)$), наименьших ($\min(\varphi_i)$) и средних ($\mathrm{mean}(\varphi_i) \pm \sigma(\varphi_i)$) оценок для олигоДНК, а также их сумм и разностей.

В рассматриваемом случае (Рисунок 78а) в рамках использования формулы (58) для величин φ репортерной LUC-активности в HeLa была эмпирически найдена оценка $\varphi_{\#,}$ коррелирующая с величинами φ в линии клеток PYS-2 (Рисунок 78б: τ=1.00, $\alpha<0.05$, ранговая корреляция Кендалла):

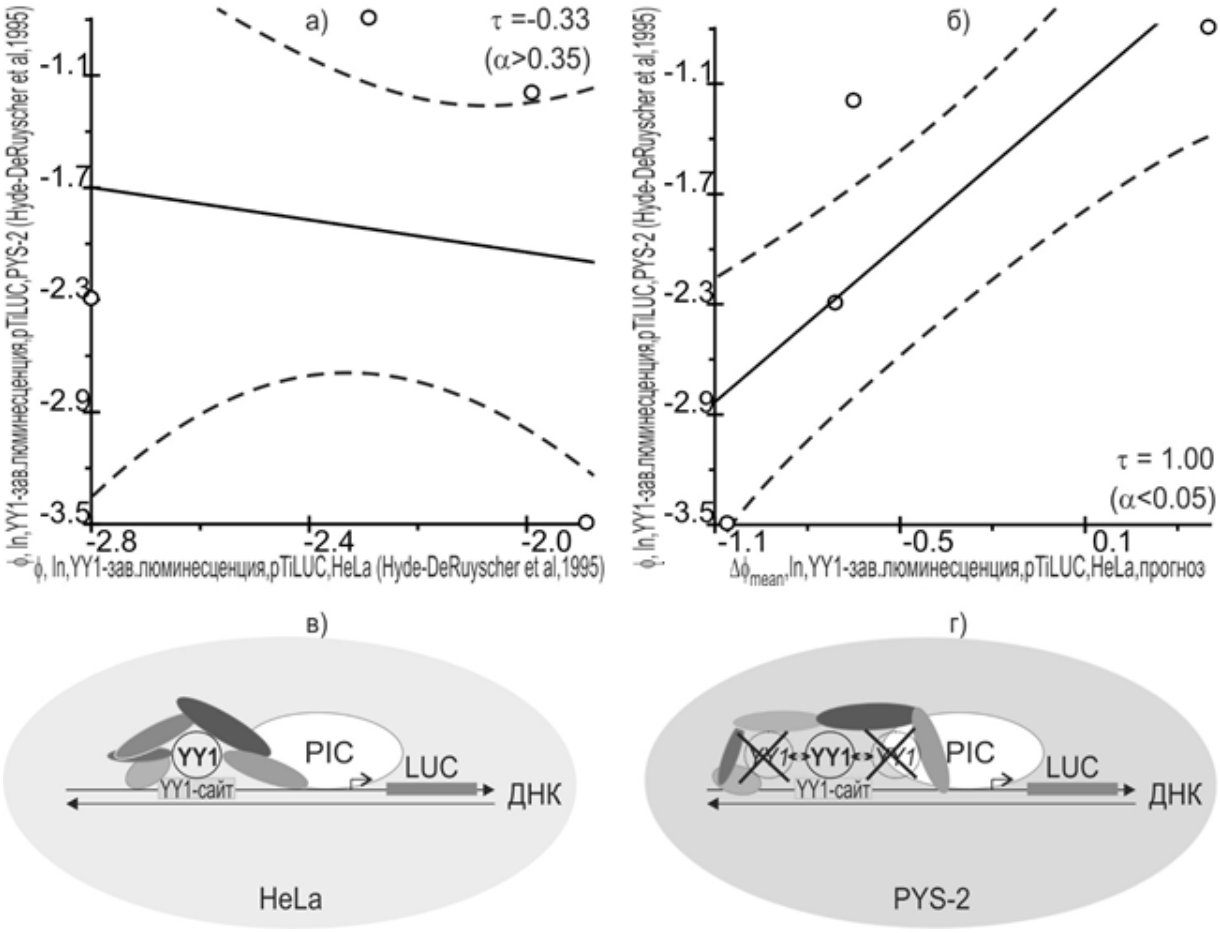

Рисунок 78 - Сравнение измерений *ex vivo* величин φ репортерной LUC активности вариантов № 32, 44, 72 и 91 плазмиды pTiLUC в клетках PYS-2 мыши (Hyde-DeRuyscher *et al*., 1995) с (а) такими измерениями в клетках HeLa человека и (б) с прогнозом *in silico* (формула 58) на основе случая HeLa с эмпирическим учетом возможного отличия (в, г) между механизмами YY1-зависимой репрессии в PYS-2 клетках: кластер из 7 потенциальных сайтов YY1, один из которых был репрессором, а остальные - конкурентные ингибиторы для него. Пунктир – границы 95%-доверительных интервалов.

$$\varphi_{\#}(S) = \varphi_{YY1-\text{сайт};i=6} - mean_{4\le i\le 10}(\varphi_i). \quad (58)$$

Формула (58) описывает гипотетический случай связывания транскрипционного фактора YY1 с соответствующим ему сайтом со стартом в позиции 6 олиго ДНК при его конкуренции с другими перекрывающимися с

ним вариантами связывания ДНК/белок между ними со стартами между позициями 4 и 10 вокруг сайта YY1, Рисунок 78г. Это означает, что учет возможного различия условий HeLa (Рисунок 78г) и PYS-2 (Рисунок 78в) обеспечил соответствие между результатами опытов.

Кроме того, авторы (Hyde-DeRuyscher *et al*., 1995) показали (Рисунок 79а), что измеренные *in vitro* величины $-\ln[K_D]$ сродства фактора транскрипции YY1 к 9 олигоДНК № 11, 14, 15, 32, 28, 43, 44, 79 и 91 не коррелировали с измеренными *ex vivo* величинами φ репортерной LUC-активности в клетках HeLa, что также отмечалось в главе 1 при обосновании темы диссертации.

В порядке продолжения дискуссии о поиске возможных соответствий между результатами разных опытов был эмпирически учтен случай сайта связывания YY1 с наибольшим сродством YY1/ДНК в границах олигоДНК, который подвержен конкурентному ингибированию потенциальными сайтами связывания YY1 во всех остальных позициях этой же олигоДНК, что можно оценить с помощью формулы:

$$\Delta\varphi_{95\%}(S) = -[min_{2\le i\le 9}(\varphi_i) - \\ -\{mean_{2\le i\le 9}(\varphi_i) - \tau_{\alpha<0.05;\nu=9-2}\sigma_{2\le i\le 9}(\varphi_i)\}]. \quad (59)$$

Прогнозы формулы (59), основанной на *t*-критерии Стьюдента, отложены по вертикали на Рисунке 79б. Как можно видеть, они достоверно (ранговая корреляции Кендалла $\tau=-0.39$, $\alpha<0.025$) коррелируют с величиной φ активности репортера LUC в HeLa (Hyde-DeRuyscher *et al*., 1995): чем выше сродство YY1/ДНК, тем ниже величина φ LUC-активности, репрессируемой YY1.

Существенно, что формула (59) была впоследствии использована в статье (Миронова и др., 2010) для классификации промоторов генов ARF арабидопсиса и риса на TATA-содержащие и TATA-несодержащие, подтвержденной независимыми измерениями экспрессии этих генов.

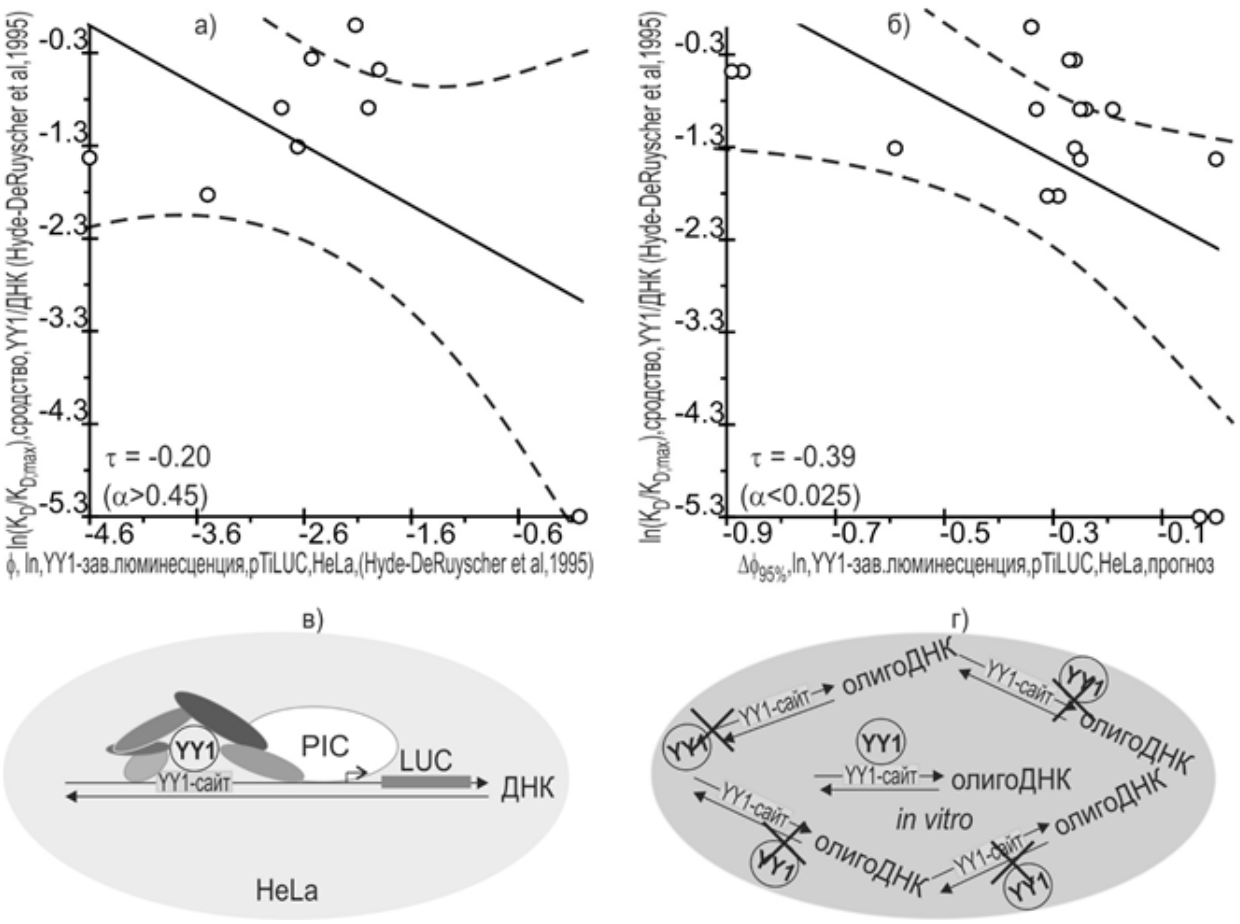

Рисунок 79 - Сравнение измерений *in vitro* сродства фактора транскрипции YY1 к олигоДНК № 11, 14, 15, 32, 28, 43, 44, 79 и 91 (Hyde-DeRuyscher *et al*., 1995) и (а) измерений *ex vivo* величин φ репортерной LUC-активности в клетках HeLa и (б) с прогнозом *in silico* (формула 59) величин сродства YY1 к олигоДНК на основе этих величин φ и эмпирических моделей (в, г) учета в качестве возможного различия между механизмами YY1-зависимой репрессии транскрипции репортерного гена *ex vivo* и связывания YY1 с олигоДНК *in vitro*: позицию олигоДНК с наибольшим сродством YY1/ДНК как сайт связывания YY1, который подвержен конкурентному ингибированию потенциальными сайтами связывания YY1 во всех остальных позициях этой олигоДНК *in vitro* (г). Пунктир – границы 95%-доверительных интервалов.

Для активации транскрипционным фактором YY1 транскрипции генов эукариот с их YY1-зависимыми инициаторными Inr-элементами в опыте (Javahery *et al*., 1994) было установлено отсутствие корреляции между транскрипционной активностью этих генов и сродством YY1/ДНК (Таблица 40, Рисунок 80а: $r=-0.11$, $\alpha>0.8$), как обсуждалось в главе 1 в обосновании темы настоящей диссертации. Поэтому в завершение дискуссии о восстановлении соответствий между разными одновременно измеренными количественными характеристиками транскрипционной активности были найдены эмпирические прогнозы на основе формулы (57), достоверно коррелирующие с каждым из этих измерений по-отдельности. В случае уровней φ_{INR} YY1-зависимой активации транскрипции (Javahery *et al*., 1994) оказалось, что они достоверно негативно коррелируют с прогнозом уровней YY1-зависимой репрессии по участку плазмиды pSVPyTK длиной 12 п.о. сразу перед стартом транскрипции, т.е. с сайтом YY1 (Таблица 40, Рисунок 80б: $r=-0.85$, $\alpha<0.005$).

При этом YY1/Inr-сродство достоверно (Таблица 40, Рисунок 80в: $r=-0.76$, $\alpha<0.05$) коррелирует с его прогнозом на основе известного различия консенсусов YCANAT и YCAT для Inr-элемента и сайта связывания YY1, соответственно. Позитивные φ_{TSS-1} и φ_{TSS+1} и негативный "$-\varphi_{TSS}$" вклады совпадений и несовпадений между ними, характеризуют (Рисунок 80г) смещенные долгоживущие неактивные и несмещенный короткоживущий активный преинициаторные комплексы:

$$\varphi_{YCA-AT}(S) = \varphi_{TSS-1} + \varphi_{TSS+1} - \varphi_{TSS}. \qquad (60)$$

Таким образом, в отличие от экспериментальных величин, несопоставимых друг с другом при различии условий измерений, количественные характеристики регуляторных сайтов в составе геномных ДНК сопоставимы между собой при эмпирическом учете этих условий. Это важно для экспериментально-компьютерных исследований регуляции экспрессии генов при планировании условий предстоящих экспериментов с

Таблица 40 – Измеренные *in vitro* (Javahery *et al*., 1994) в экстракте ядер HeLa сродство фактора транскрипции YY1 к инициаторному Inr-элементу (A – старт транскрипции) и YY1- активация транскрипции с плазмиды pSVPyTK не коррелируют между собой по-отдельности, но достоверно коррелируют с прогнозами *in silico* на основе (формулы 57 и 60) независимых измерений *ex vivo* уровней YY1-репрессии LUC-репортера (Hyde-DeRuyscher *et al*., 1995).

Экспериментальные данные (Javahery *et al*., 1994)				φ_{LUC}, прогноз	
YY1-зависимый инициатор, Inr		измерения *in vitro*		φ_{TSS}, (ф-ла 57)	ф-ла (60)
Проба	последовательность[&]	φ_{INR}	YY1/Inr, ранг		
TdT	gctcggccctcAttctggagac	0.00	1	-2.85	-4.49
p5+1	gctcggtctccAttttgaagcg	-1.27	3	-3.13	-4.87
p5-60	gctcgggcgacAttttgcgaca	-0.84	2	-2.60	-4.34
p5-60m3	gctcgggtgatAttttgcgaca	-1.43	0	-2.79	-4.53
p5+1m2	gctcggtctccAgttggaagcg	-0.76	0	-3.13	-4.06
p5+1m1	gctcggtctaaAttttgaagcg	-3.91	0	-1.39	-3.13
E1	gctcggcggccAtcttgaagcg	-3.00	4	-2.32	-5.21
Линейная корреляции, r		-0.11		-0.85[#]	-0.76[$]
Значимость, α		>0.80		<0.025[#]	<0.05[$]

учетом предыдущих опытов. Благодаря возможности сопоставления количественных характеристик регуляторных сайтов в составе геномных ДНК при различных экспериментальных условиях возникли новые исследовательские возможности. Рассмотрим их на практическом примере.

В опыте (Comings *et al*., 1996) были обнаружены варианты WT, M1 и M2 интрона 6 гена триптофан 2,3-оксигеназы, TDO2, человека, ассоциированные с предрасположенностью к синдрому Туретта, алкоголизму, зависимости от лекарств, гиперактивности, дефициту внимания и к другим поведенческим отклонениям (Таблица 41).

На Рисунке 81a представлены данные опыта “задержки в геле” комплексов олигоДНК, содержащих эти варианты, с экстрактом ядер из клеток печени крысы (Vasiliev *et al*., 1999). Стрелкой на этом рисунке указан комплекс, мутационные изменения которого были проанализированы на основе использования формулы (57).

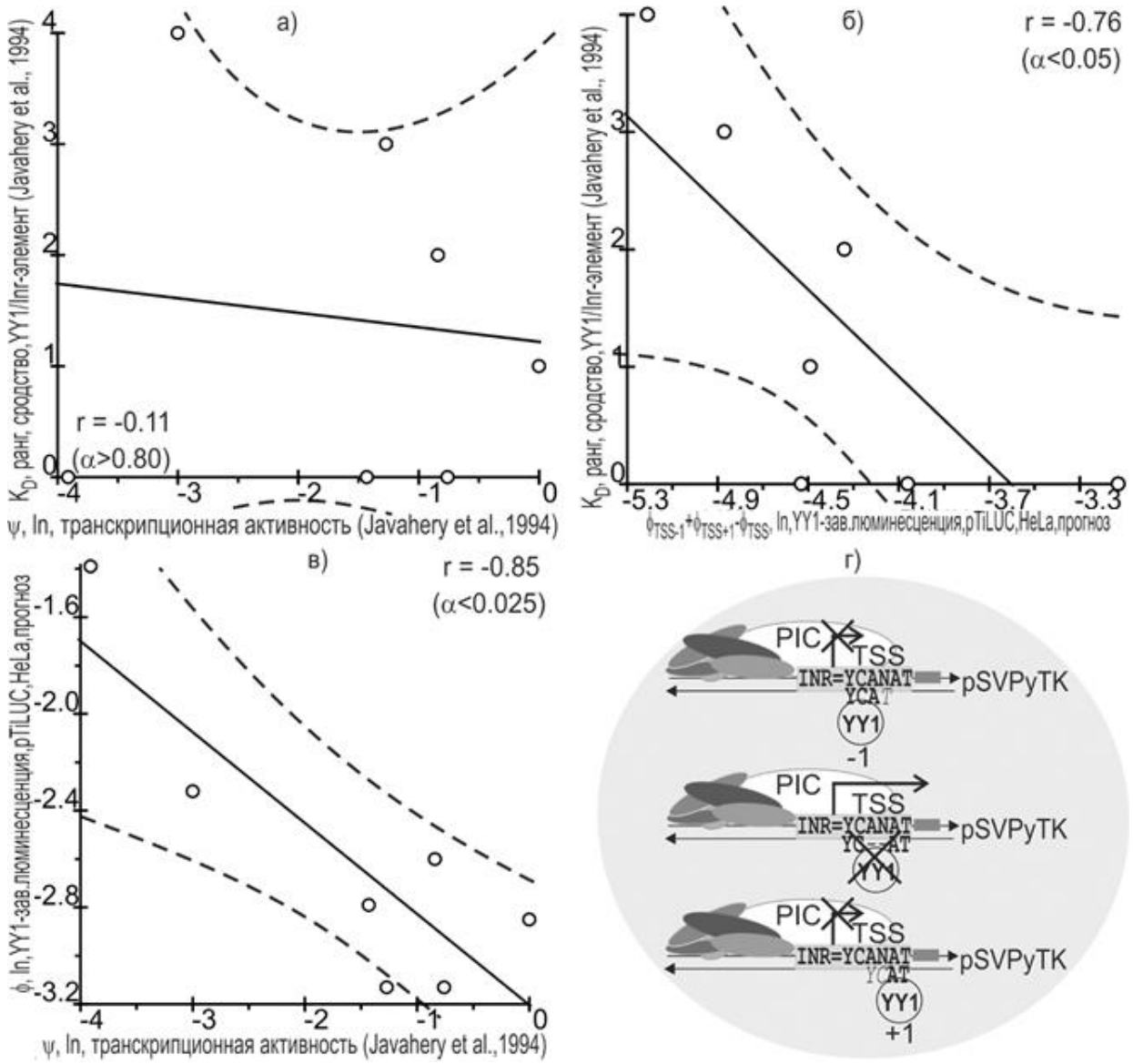

Рисунок 80- Измеренные *in vitro* (Javahery *et al*., 1994) в экстракте ядер *HeLa* величины сродства фактора транскрипции YY1 к инициаторному Inr-элементу и YY1- активация транскрипции с плазмиды pSVPyTK не коррелируют между собой (а) и достоверно коррелируют с прогнозами на основе YY1-репрессии LUC-репортера (Hyde-DeRuyscher *et al*., 1995): б) формула (57) и в) формула (60) в рамках эмпирической модели (г). Пунктир – границы 95% доверительных интервалов.

В Таблице 41 приведены величины сродства олигоДНК к неизвестному белку из экстракта ядер клеток печени крысы (Vasiliev *et al*., 1999), образующего комплекс, указанный на Рисунке 81а, "→". В последней колонке

Таблица 41 – Ассоциированный с поведенческими отклонениями (Comings *et al*., 1996) полиморфизм интрона 6 гена TDO2 человека и результат его анализа в настоящей работе

Экспериментальные данные (Vasiliev *et al*., 1999)				прог-
полиморфизмы гена TDO2 (Comings *et al*., 1996)		комплекс №3		ноз
вариант	последовательность	%	ln	(61)
WT	tgccaaataa**tg g ca g at**aagaatagggag	100	0.0	1.6
M1	tgccaaataa**tg A ca g at**aagaatagggag	5	-3.0	2.6
M2	tgccaaataa**tg g ca T at**aagaatagggag	50	-0.7	1.9
Коэффициент линейной корреляции, r (значимость, α)		-0.997 (<0.05)		

даны эмпирические оценки сродства транскрипционного фактора YY1 к олигоДНК (Рисунок 81г), несущим варианты WT, M1 и M2 интрона 6 гена *TDO2* человека:

$$\varphi_*(S) = max_{-3\le i\le 3}(\varphi_i) - min_{-3\le i\le 3}(\varphi_i). \qquad (61)$$

Как можно видеть на Рисунке 81б, прогноз формулы (61) достоверно ($r=-0.99$, $\alpha<0.05$) коррелирует с результатами экспериментальных измерений. Эта достоверная связь между транскрипционным фактором YY1 и анализируемым SNP интрона 6 гена TDO2 человека была независимо подтверждена *in silico* с помощью метода среднего распознавания (формула 37) на основе олигонуклеотидных позиционно-частотных матриц (Ponomarenko M. *et al.*, 1999b), как это показано на Рисунке 81в. На этой основе для контрольного эксперимента был выбран транскрипционный фактор YY1 (Vasiliev *et al*., 1999) из трех потенциальных факторов-кандидатов YY1, SRF и NF-IL6 (Рисунок 81в), найденных с помощью программы TESS (Schug, Overton, 1997) из более 400 транскрипционных факторов, описанных в базе данных TRANSFAC (Heinemeyer *et al*., 1999). На Рисунке 81д представлен результат этого контрольного эксперимента (Vasiliev *et al*., 1999) с использованием анти-YY1 антител (дорожка 3), связывание которых с неизвестным белком из указанного на Рисунке 81а комплекса, "→", однозначно идентифицировало его: комплекс олигоДНК с транскрипционным фактором YY1 экстракта ядер клеток печени крысы (Рисунок 81д: дорожка 2).

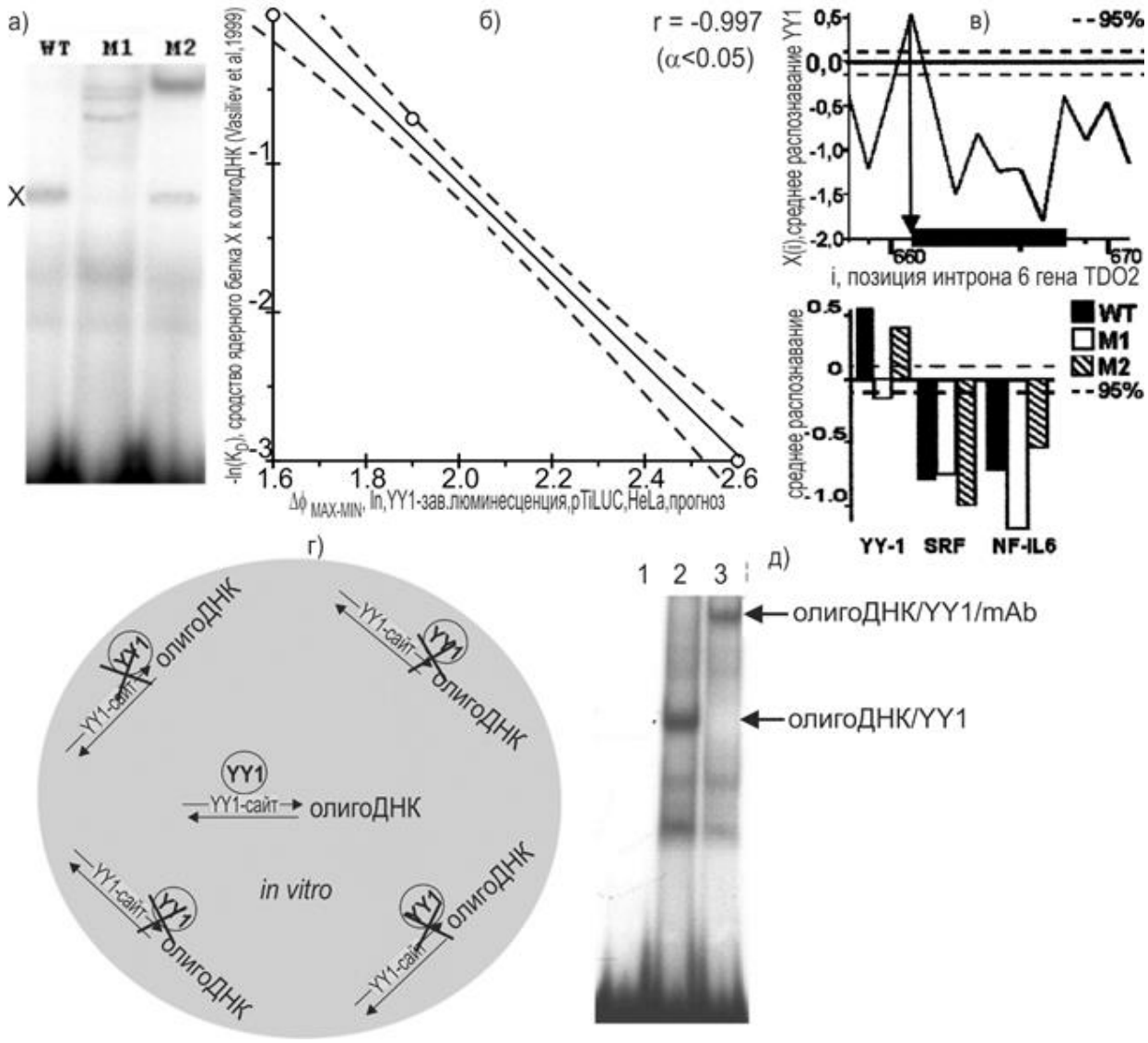

Рисунок 81 Анализ вариантов WT, M1 и M2 интрона 6 гена TDO2, связанных с поведенческими отклонениями человека (Comings *et al*., 1996): (а) данные “задержки в геле” (Vasiliev *et al*., 1999), здесь: X – комплекс олигоДНК с неизвестным белком X из экстракта ядер печени крысы; (б) прогноз (формула 61) YY1-зависимой люминесценции (Hyde-DeRuyscher *et al*., 1995); (в) независимое подтверждение *in silico* прогноза формулы (61) с помощью метода среднего распознавания (формула 37); (г) модель сродства YY1/олигоДНК для эмпирической формулы (61); (д) подтверждение *in vitro* прогноза в опыте с антителами против YY1 (дорожка №3, mAb – **m**onoclonal **A**nti**b**ody). Пунктир – границы 95%-доверительных интервалов.

Таким образом, с помощью компьютерной системы Activity (Ponomarenko m *et al*., 1997a) на основе оригинальных экспериментальных данных (Vasiliev *et al*., 1999) о биохимическом проявлении (Рисунок 81а) вариантов WT, M1 и M2 интрона 6 гена TDO2, ассоциированных с поведенческими отклонениями человека (Comings *et al*., 1996), был впервые предсказан (Рисунок 81б) транскрипционный фактор, сайт связывания которого в составе геномной ДНК был патогенно изменен в минорных вариантах M1 и M2 полиморфизма этого гена, что подтвердил контрольный опыт на основе этого прогноза (Рисунки 81в,г).

Представленные в настоящей главе результаты позволяют сделать следующие выводы:

- С помощью системы Activity впервые построены регрессионные уравнения для предсказания величин сродства регуляторных белков к сайтам их связывания в составе геномной ДНК:
  - Cro-репрессора к оператору OR1 фага λ на основе оценок ширины малой бороздки, угла раскрытия пар оснований по малой оси и шага B-формы ДНК;
  - активатора CRP к промоторам генов *Escherichia coli* на основе оценок ширины малой бороздки и шага B-формы ДНК;
  - транскрипционного фактора USF к сайтам его связывания в промоторах генов человека на основе оценок угла кручения и глубины малой бороздки B-формы ДНК;
  - транскрипционного фактора MEF2 к сайтам его связывания в промоторах генов мыши на основе оценок персистентной длины и ширины малой бороздки B-формы ДНК.
- Впервые построено регрессионное уравнение, которое достоверно предсказывает величину подавления транскрипционной активности генов человека транскрипционным фактором YY1 на основе оценки угла кручения B-формы ДНК сайтов связывания этого регуляторного белка. С использованием этого уравнения было впервые предсказано, что мутации

663G>A и 666G>T, ассоциированные с комплексом поведенческих расстройств человека и локализованные в интроне 6 гена TDO2, затрагивают сайт связывания транскрипционного фактора YY1 и нарушают его активность за счет изменения угла кручения B-формы ДНК этого сайта. Спланированный на этой основе эксперимент с использованием антител против транскрипционного фактора YY1 подтвердил результаты предсказания.

ЗАКЛЮЧЕНИЕ ПО ГЛАВЕ 5

Представленные в настоящей главе диссертации результаты компьютерного анализа количественных характеристик регуляторных сайтов в составе геномных ДНК были опубликованы на рубеже XX и XXI веков (Ponomarenko M. *et al*., 1999a; Vasiliev *et al*., 1999; Васильев и др., 2000; Ponomarenko J *et al*., 2001a). В дальнейшем на основе их использования совместно с другими авторами в сети Интернет были созданы новые компьютерные системы и был получен ряд важных биологических результатов. Прежде всего, был предложен (Пономаренко М и др., 1999a; Ponomarenko M *et al*., 1999b) набор алфавитов из независимых равновероятных олигонуклеотидов ДНК длиной от 2 п.о. до 12 п.о в 15-символьном коде (IUPAC-IUB, 1971). Усреднение результатов распознавания регуляторных сайтов в составе геномной ДНК по $N > 10$ произвольным таким алфавитам уменьшило ошибку II рода, которая оказалась асимптотически убывающей пропорционально $N^{-1/2}$ (Пономаренко М и др., 1999a) вследствие центральной предельной теоремы.

В свою очередь, на этой основе был предложен эмпирический учет условий опыта по “задержке в геле” олигонуклеотидов ДНК, представляющих ассоциированные с заболеваниями человека варианты ДНК, и белков из экстракта ядер соответствующей линии клеток человека, компьютерная система rSNP_Guide (Ponomarenko J *et al*., 2001b, 2002a,b, 2003, 2005) для

прогноза транскрипционного фактора, сайт связывания которого в составе геномной ДНК изменяется в случае SNP. С ее помощью в интроне 2 гена K-ras мыши было предсказано (Ponomarenko J *et al*., 2002a) появление сайта связывания транскрипционного фактора GATA при SNP, ассоциированном с раком легких, что было подтверждено в опыте (Тимофеева и др, 2002). Также, при использовании системы rSNP_Guide были получены компьютерно-экспериментальные данные, которые являются основанием для проведения дальнейшего стандартного медико-генетического исследования связи SNPs rs75996864, rs76241113, rs78037487, rs80112297 и rs80313086 гена APC с раком толстого кишечника человека (Рассказов и др., 2013a).

В свою очередь, применение эмпирического учета различия белков Ago2 и Ago3 семейства *Argonaute* человека по величинам их сродства к зрелым микроРНК (миРНК) позволило в рамках так называемого приближенного решения некорректно поставленной задачи оценить содержание 96 зрелых миРНК в клеточной линии HEK293T почки эмбриона человека в норме и после обработки актиномицином D (ингибитор транскрипции), а также содержание 318 зрелых миРНК в мозге человека (Ponomarenko M. *et al*., 2013g).

Наконец, был предложен эмпирический учет (Mironova *et al*., 2013) искусственного мутагенеза кластеров сайтов ответа на ауксин, объяснивший репрессию базальной транскрипции ауксин-зависимых генов без необходимости гипотез о каких-либо неизвестных регуляторных сайтах в составе геномной ДНК вне рамок общепринятых представлений об индукции генов растений ауксином. На этой основе было выведено (Ponomarenko P. and Ponomarenko M., 2015) эмпирическое уравнение импульсов транскрипции, амплитуда, длительность и частота которых соответствуют сродству промотора к TBP, к октамеру гистонов и к специфическому транскрипционному фактору-активатору, и с его помощью было предсказано содержание 70 мРНК генов первичного ответа на ауксин в транскриптоме арабидопсиса через 1 час после обработки растения ауксина.

В целом, эмпирический учет условий экспериментов в рамках регрессионных уравнений для биоинформатических прогнозов стал новым перспективным направлением исследований, которое возникло в настоящей диссертации.

ЗАКЛЮЧЕНИЕ

В настоящей диссертационной работе был предложен новый подход к компьютерному анализу экспериментальных данных о влиянии нуклеотидных последовательностей на величины специфической биологической активности сайтов в составе геномной ДНК с использованием теории аддитивной полезности для принятия решений и нечетких множеств. Он использует для анализа более миллиона различных вариантов контекстных и контекстно-зависимых конформационных и физико-химических характеристик B-формы двойной спирали ДНК, единообразно оценивает влияние каждой из них на экспериментально наблюдаемое соответствие между первичной структурой и исследуемой биологической активностью сайтов в составе геномной ДНК и выявляет на основе этих оценок ограниченный набор таких характеристик, которые статистически достоверно коррелируют с экспериментальными величинами этой активности. В качестве результата, этот подход выводит эмпирические регрессионные уравнения для оценки величин биологической активности сайтов в составе геномной ДНК по нуклеотидным последовательностям в рамках линейно-аддитивного приближения, которое оценивает базовые уровни этой активности на основе облигатных символьных характеристик консенсуса и позиционно-весовой матрицы и модуляцию активности сайтов в составе геномной ДНК вблизи этих уровней – на основе выявленных контекстных и контекстно-зависимых конформационных и физико-химических характеристик B-формы двойной спирали ДНК.

В диссертационной работе автор реализовал предложенный подход посредством создания двух компьютерных систем bDNAvideo (Ponomarenko M et al., 1997и) и Activity (Ponomarenko M et al., 1997a) для автоматического анализа данных об экспериментально измеренных величинах специфической биологической активности сайтов в составе геномной ДНК для выявления контекстных и контекстно-зависимых конформационных и физико-химических характеристик сайтов в составе геномных ДНК, величины

которых достоверно коррелируют с величинами этой активности. Эти системы содержат в качестве своей основы базы данных PROPERTY (Колчанов и др., 1998) о физико-химических и конформационных свойствах динуклеотидных шагов B-формы спирали ДНК (раздел 2.1.6), SAMPLE (Ponomarenko M et al., 1999b) по сайтам связывания транскрипционных факторов на B-форме спирали ДНК (раздел 2.2.1) и Activity (Ponomarenko J et al., 2001a) по экспериментально измеренным количественным величинам специфических биологических активностей сайтов в составе геномных ДНК (раздел 4.4), а также вспомогательные базы данных SELEXdb (Ponomarenko J. et al., 2000a) по синтетическим рандомизированным олигонуклеотидам ДНК, амплифицированным/селектированным *in vitro* на повышенное сродство к белкам-мишеням, и SYSTEM (Ponomarenko J. et al., 2002c) по условиям экспериментов с сайтами в составе геномной ДНК, документированных в остальных базах данных этих систем (раздел 1.1). Также эти системы содержат в себе соответствующие активные приложения, автоматически анализирующие информацию из вышеуказанных основных баз данных (входные данные) и находят в качестве результата (выходные данные) ограниченные наборы контекстно-зависимых характеристик сайтов в составе геномной ДНК, которые достоверно коррелируют с количественными величинами ее исследованной биологической активности (разделы 2.1.3 и 3.1). В свою очередь, эти результаты документируются в базах знаний FEATURES (Ponomarenko M et al., 1999b) для системы bDNAvideo и KNOWLEDGE (Ponomarenko M. et al., 1997a) для системы Activity в форме кода на языке программирования “Си” для вычислительных процедур оценки количественных величин заданной специфической биологической активности сайтов в составе геномных ДНК (выходные данные) по ее нуклеотидной последовательности (входные данные), как это показано на Рисунке 44 в разделе 2.2.3. Наконец, обе компьютерные системы bDNAvideo (Ponomarenko M. et al., 1997и) и Activity (Ponomarenko M. et al., 1997a) содержат общую для них базу знаний CROSS_TEST (Ponomarenko J et al., 2002c) по результатам

перекрестных тестов закономерностей, выявленных в результате анализа данных одних опытов и подтвержденных данными независимых опытов (раздел 1.1).

В рамках диссертационной работы с использованием созданной в ней системы bDNAvideo впервые была получена достоверная ($p < 0.01$) кластеризация транскрипционных факторов по сходству сайтов в составе геномной ДНК для их связывания на две группы, в первой из которых доминировали Zn-координируемые и основные белки с локальным избытком электростатического заряда, во второй - белки с гомеодоменом или с β-слоем без локального избытка электростатического заряда (раздел 2.2.3). В частности, с помощью bDNAvideo были обнаружены достоверные контекстно-зависимые конформационные и физико-химические отличия TATA-содержащих промоторов про- и эукариот от случайных последовательностей нуклеотидов ДНК. В результате было показано, что районы геномной ДНК вокруг TATA-боксов, которые соответствовали их значимым контекстно-зависимым количественным характеристикам, достоверно укорачивались в ряду организмов "*Escherichia coli*, дрожжи, беспозвоночные, позвоночные", стоящих на разных ступенях эволюции (раздел 2.1.4). На примере TATA-боксов было также показано, что учет характеристик геномной ДНК, выявленных системой bDNAvideo, втрое увеличивает точность "физико-химического и конформационного" распознавания TATA-боксов в сравнении с уровнем, достигнутым в мире на момент ее создания (раздел 2.1.6).

Кроме того, в диссертационной работе с использованием созданной компьютерной системы Activity были впервые построены эмпирические регрессионные уравнения, достоверно предсказывающие количественные величины сродства таких регуляторных белков, как Cro-репрессор и активатор CRP (раздел 4.4), транскрипционных факторов USF и MEF2 к сайтам их связывания в составе геномной ДНК (разделы 5.1 и 5.2), а также для оценки частот повреждений гуанинов в ДНК под действием лазерного

ультрафиолетового излучения с длиной волны 193 нм, подтвержденный данными независимых экспериментов (раздел 4.1). Также, для фермента 8-оксогуанин-ДНК гликозилаза (OGG1) человека были впервые выведены эмпирические регрессионные уравнения для оценки констант Михаэлиса, K_M, и каталитической константы, k_{CAT}, фермента 8-оксогуанин-ДНК гликозилазы OGG1 человека, подтвержденные на независимых данных (раздел 4.2). Аналогично, для белка RecA *Escherichia coli* было выведено эмпирическое уравнение для оценки сродства его филамента к нити ДНК, согласно которому наибольшее такое сродство имеют участки кодирования глобулярных ядер белков в геномных ДНК бактерий, наименьшее – участки бактериальных геномов, которые кодируют заряженные аминокислотные остатки и элементы “случайный клубок” вторичной структуры белков (раздел 4.3). В свою очередь, с использованием регрессионного уравнения для оценки уровня репрессирующего воздействия транскрипционного фактора YY1 на гены человека было предсказано, что связанные с психическими расстройствами человека замены 663G>A и 666G>T в интроне 6 гена *TDO2* повреждают сайт связывания этого транскрипционного фактора (раздел 5.3). Это предсказание было подтверждено экспериментом с антителами к транскрипционному фактору YY1 (раздел 5.3).

Наконец, выведенные уравнения для оценок сродства TATA-связывающего белка (TBP) к нити ДНК и к двунитевым гетеродуплексам были обобщены вместе с общепринятым критерием Бухера для TATA-бокса (Bucher, 1990) в эмпирическое уравнение трехшагового связывания TBP с ДНК, которое впервые достоверно ($p<10^{-6}$) предсказало величины константы K_D диссоциации их комплекса для независимых опытов в равновесных и в неравновесных условиях *in vitro*. В целом, полученные в диссертационной работе новые научные результаты нашли свое практическое применение при создании современных компьютерных систем, в том числе: SITECON (Россия), BiDaS (Греция), CRoSSeD (Бельгия), DISCOVER (США), а также FeatureScan, BioBayesNet и ProMapper (все: Германия). В свою очередь,

выведенные в рамках диссертации эмпирические регрессионные уравнения были использованы в планировании опытов для предиктивно-превентивной персонифицированной медицины, которые были направлены, в том числе, на предсказание кандидатных SNP-маркеров для ожирения человека и связанных с ним сопутствующих заболеваний (Arkova *et al.*, 2015), а также гендер-зависимых аутоиммунных заболеваний человека (Ponomarenko *et al.*, 2016). Все это вместе взятое свидетельствует, что результаты настоящей диссертации представляют новое направление в решении задач таких актуальных наук о жизни как предиктивно-превентивная персонифицированная медицина, глубокая функциональная аннотация геномов, исследования влияния генетической изменчивости на экспрессию генов, конструирования искусственных молекулярно-генетических систем с заданными свойствами, создания новых подходов к маркер-ориентированной селекции животных и растений.

ВЫВОДЫ

1. На основе теории аддитивной полезности для принятия решений и нечетких множеств создана компьютерная система Activity для:
   - анализа выборок нуклеотидных последовательностей сайтов в составе геномной ДНК с известными величинами специфической биологической активности и выявления контекстных, а также контекстно-зависимых конформационных и физико-химических характеристик B-формы ДНК, достоверно коррелирующих с анализируемой активностью сайтов ДНК;
   - построения регрессионных уравнений для предсказания величин специфической биологической активности по произвольной последовательности сайта в составе геномной ДНК на основе выявленных контекстных, а также контекстно-зависимых конформационных и физико-химических характеристик B-формы ДНК, коррелирующих с этой активностью.
2. С помощью системы Activity впервые построены регрессионные уравнения для предсказания величин сродства регуляторных белков к сайтам их связывания в составе геномной ДНК:
   - Cro-репрессора к оператору OR1 фага λ на основе оценок ширины малой бороздки, угла раскрытия пар оснований по малой оси и шага B-формы ДНК;
   - активатора CRP к промоторам генов *Escherichia coli* на основе оценок ширины малой бороздки и шага B-формы ДНК;
   - транскрипционного фактора USF к сайтам его связывания в промоторах генов человека на основе оценок угла кручения и глубины малой бороздки B-формы ДНК;
   - транскрипционного фактора MEF2 к сайтам его связывания в промоторах генов мыши на основе оценок персистентной длины и ширины малой бороздки B-формы ДНК.
3. Впервые построено регрессионное уравнение, которое достоверно предсказывает величину подавления транскрипционной активности генов человека транскрипционным фактором YY1 на основе оценки угла кручения B-формы ДНК сайтов связывания этого регуляторного белка. С

использованием этого уравнения было впервые предсказано, что мутации 663G>A и 666G>T, ассоциированные с комплексом поведенческих расстройств человека и локализованные в интроне 6 гена TDO2, затрагивают сайт связывания транскрипционного фактора YY1 и нарушают его активность за счет изменения угла кручения B-формы ДНК этого сайта. Спланированный на этой основе эксперимент с использованием антител против транскрипционного фактора YY1 подтвердил результаты предсказания.

4. Выявлены достоверные корреляции равновесной константы диссоциации K_D TATA-связывающего белка (TBP) к олигоДНК длиной 15 нт с содержанием динуклеотида WR на флангах и динуклеотида TV в центральной части однонитевой ДНК, а также с содержанием динуклеотида TA в 3'-половине и шириной малой бороздки в центре дуплексов ДНК. На основе этих корреляций были впервые предсказаны величины равновесной константы диссоциации K_D комплекса TBP/ДНК, которые были подтверждены независимыми экспериментами.

5. Выявлены контекстные характеристики ДНК плазмиды pGEM7(f+) *Escherichia coli*, достоверно коррелирующие с частотой повреждений ДНК по гуанинам под действием ультрафиолетового излучения лазера с длиной волны 193 нм. На этой основе впервые получено регрессионное уравнение для предсказания величин частоты таких повреждений гуанина в ДНК. Это уравнение подтверждено независимым экспериментом с дуплексами ДНК, идентичными фрагментам гена *MIP-1α* мыши.

6. Выявлены достоверные корреляции: (а) между каталитической константой k_{CAT} 8-оксогуанин-ДНК гликозилазы OGG1 человека и углом кручения B-формы ДНК в окрестности 8-оксогуанина (oxoG), а также (б) между константой Михаэлиса K_M этого фермента и изменением свободной энергии Гиббса при образовании гетеродуплекса ДНК в окрестности этого oxoG. На этой основе впервые выведены регрессионные уравнения для оценки величин этих констант при частичном нарушении комплементарности ДНК вокруг oxoG, которые были подтверждены независимыми экспериментальными данными.

7. Показано, что сродство RecA к однонитевой ДНК достоверно убывает с ростом встречаемости в нити ДНК тринуклеотидов DRV в 15-буквенном коде IUPAC, которые достоверно соответствуют кодонам заряженных аминокислотных остатков.

8. На основе теории аддитивной полезности для принятия решений и нечетких множеств создана компьютерная система bDNAvideo для выявления контекстно-зависимых конформационных и физико-химических характеристик спирали ДНК, достоверно дискриминирующих сайты связывания транскрипционных факторов от случайных последовательностей. С использованием этой системы впервые получена достоверная кластеризация транскрипционных факторов на две группы, первая из которых включает преимущественно основные и Zn-координируемые белки с локальным избытком электростатического заряда, вторая - белки с β-слоем и с гомеодоменом без локального избытка электростатического заряда.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных журналах

1. **Пономаренко, М.П.** Контекстные сигналы и антисигналы сайта встраивания td-интрона. / М.П. Пономаренко, А.Э. Кель, А.Н. Колчанова, Н.А. Колчанов // ДАН. – **1996.** – Т. 348, № 3. - С. 424 - 427.

2. **Пономаренко, М.П**. Компьютерное моделирование последовательностей ТАТА-боксов промоторов эукариот. / М.П. Пономаренко, Л.К. Савинкова, А.Э. Кель, Н.А. Колчанов // ДАН. - **1997**. - Т. 355, № 4. - С. 557 - 561.

3. **Пономаренко, М.П**. Моделирование последовательностей ТАТА-боксов генов эукариот. / М.П. Пономаренко, Л.К. Савинкова, Ю.В. Пономаренко, А.Э. Кель, И.И. Титов, Н.А. Колчанов // Мол. биол. - **1997**. - Т. 31, № 4. - С. 726-732.

4. **Пономаренко, М.П**. Компьютерный анализ конформационных особенностей ДНК ТАТА-боксов промоторов эукариот. / М.П. Пономаренко, Ю.В. Пономаренко, А.Э. Кель, Н.А. Колчанов, Х. Карас, Э. Вингендер, Х. Скленар // Мол. биол. – **1997**. - Т. 31, № 4. - С. 733 - 740.

5. **Ponomarenko, M.P**. Generating programs for predicting the activity of functional sites. / M.P. Ponomarenko, A.N. Kolchanova, N.A. Kolchanov // J. Comput. Biol. - **1997**. - V. 4, N. 1. - P. 83 - 90.

6. Колчанов, Н.А. Функциональные сайты геномов про- и эукариот: компьютерное моделирование и предсказание активности. / Н.А. Колчанов, **М.П. Пономаренко**, Ю.В. Пономаренко, А.С. Фролов, Н.Л. Подколодный // Мол. биол. – 1998. - Т. 32, № 2. - С. 255 - 267.

7. **Пономаренко, М.П**. Предпочтительность RecA-филамента к последовательностям ДНК коррелирует с генетическим кодом. / М.П. Пономаренко, Ю.В. Пономаренко, И.И. Титов, Н.А. Колчанов, А.В. Мазин, С. Ковальчиковски // ДАН. – 1998. - Т. 363, № 1. - С. 122 - 125.

8. Levitsky, V.G. Nucleosomal DNA property database. / V.G. Levitsky, **M.P. Ponomarenko**, J.V. Ponomarenko, A.S. Frolov, N.A. Kolchanov // Bioinformatics. – **1999**. - V. 15, N. 7/8. - P. 582 - 592.

9. **Ponomarenko, M.P**. Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. / M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, O.A. Podkolodnaya, D.G. Vorobyev, N.A. Kolchanov, G.C. Overton // Bioinformatics. - **1999**. - V. 15, N. 7/8. - P. 631 - 643.

10. Ponomarenko, J.V. Conformational and physicochemical DNA features specific for transcription factor binding sites. / J.V. Ponomarenko, **M.P. Ponomarenko**, A.S. Frolov, D.G. Vorobyev, G.C. Overton, N.A. Kolchanov // Bioinformatics. – **1999**. - V. 15, N. 7/8. - P. 654 - 668.

11. Kolchanov, N.A. Integrated databases and computer systems for studying eukaryotic gene expression. / N.A. Kolchanov, **M.P. Ponomarenko**, A.S. Frolov, E.A. Ananko, F.A. Kolpakov, E.V. Ignatieva, O.A. Podkolodnaya, T.N. Goryachkovskaya, I.L. Stepanenko, T.I. Merkulova, V.N. Babenko, J.V. Ponomarenko, A.V. Kochetov, N.L. Podkolodny, D.G. Vorobyev, S.V. Lavrushev, D.A. Grigorovich, Yu.V. Kondrakhin, L. Milanesi, E. Wingender, V.V. Solovyev, G.C. Overton // Bioinformatics. – **1999**. - V. 15, N. 7/8. - P. 669 - 686.

12. **Ponomarenko, M.P**. Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. / M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, N.L. Podkolodny, L.K. Savinkova, N.A. Kolchanov, G.C. Overton // Bioinformatics. - **1999**. - V. 15, N. 7/8. - P. 687 - 703

13. Vasiliev, G.V. Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY1 transcription factor binding site. / G.V. Vasiliev, V.M. Merkulov, V.F. Kobzev, T.I. Merkulova, **M.P. Ponomarenko**, N.A. Kolchanov // FEBS Lett. – **1999**. - V. 462, N. 1/2. - P. 85 - 88.

14. Васильев, Г.В. Точковые мутации в районе 663-666 п.н. интрона 6 гена триптофаноксигеназы, связанные с рядом психических расстройств, разрушают сайт связывания фактора транскрипции YY1. / Г.В. Васильев, В.М. Меркулов, В.Ф. Кобзев, Т.И. Меркулова, **М.П. Пономаренко**, Ю.В. Пономаренко, О.А. Подколодная, Н.А. Колчанов // Мол. биол. – **2000**. - Т. 34, № 2. - С. 214 - 222.

15. Колпаков, Ф.А. Методы интеграции неоднородных молекулярно-генетических информационных ресурсов в электронной библиотеке GENEEXPRESS. / Ф.А. Колпаков, Н.Л. Подколодный, С.В. Лаврюшев, Д.А. Григорович, **М.П. Пономаренко**, Н.А. Колчанов // Программирование. – **2000**. - Т. 4, № 3. - С. 72 - 80.

16. Ponomarenko, J.V. SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. / J.V. Ponomarenko, G.V. Orlova, **M.P. Ponomarenko**, S.V. Lavryushev, A.S. Frolov, S.V. Zybova, N.A. Kolchanov // Nucleic Acids Res. - **2000**. - V. 28, N. 1. - P. 205 - 208.

17. Ponomarenko, J.V. ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another. / J.V. Ponomarenko, D.P. Furman, A.S. Frolov, N.L. Podkolodny, G.V. Orlova, **M.P. Ponomarenko**, N.A. Kolchanov, A. Sarai // Nucleic Acids Res. - **2001**. - V. 29, N. 1. - P. 284 - 287.

18. Ponomarenko, J.V. rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. / J.V. Ponomarenko, T.I. Merkulova, G.V. Vasiliev, Z.B. Levashova, G.V. Orlova, S.V. Lavryushev, O.N. Fokin, **M.P. Ponomarenko**, A.S. Frolov, A. Sarai // Nucleic Acids Res. - **2001**. - V. 29, N. 1. - P. 312 - 316.

19. Ponomarenko, J.V. SELEX_DB: a database on in vitro selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data. / J.V. Ponomarenko, G.V. Orlova, A.S. Frolov, M.S. Gelfand, **M.P. Ponomarenko** // Nucleic Acids Res. - **2002**. - V. 30, N. 1. - P. 195 - 199.

20. Ponomarenko, J.V. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. / J.V. Ponomarenko, G.V. Orlova, T.I. Merkulova, E.V. Gorshkova, O.N. Fokin, G.V. Vasiliev, A.S. Frolov, **M.P. Ponomarenko** // Hum. Mutat. – **2002**. - V. 20, N. 4. - P. 239 - 248.

21. Ponomarenko, J.V. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. / J.V. Ponomarenko, T.I. Merkulova, G.V. Orlova, O.N. Fokin, E.V. Gorshkova, A.S. Frolov, V.P. Valuev, **M.P. Ponomarenko** // Nucleic Acids Res. - **2003**. - V. 31, N. 1. - P. 118 - 121.

22. Савинкова, Л.К. Полиморфизмы TATA-боксов промоторов генов человека и ассоциированные с ними наследственные патологии. / Л.К. Савинкова, **М.П Пономаренко**, П.М. Пономаренко, И.А. Драчкова, М.В. Лысова, Т.В. Аршинова, Н.А. Колчанов // Биохимия. – **2009**. - Т. 74, № 2. - С. 149 - 163.

23. Suslov, V.V. SNPs in the HIV-1 TATA box and the AIDS pandemic. / V.V. Suslov, P.M. Ponomarenko, V.M. Efimov, L.K. Savinkova, **M.P. Ponomarenko**, N.A. Kolchanov // J. Bioinform. Comput. Biol. - **2010**. - V. 8, N. 3. - P. 607 - 625.

24. Kirpota, O.O. Thermodynamic and kinetic basis for recognition and repair of 8-oxoguanine in DNA by human 8-oxoguanine-DNA glycosylase. / O.O. Kirpota, A.V. Endutkin, **M.P. Ponomarenko**, P.M. Ponomarenko, D.O. Zharkov, G.A. Nevinsky // Nucleic Acids Res. - **2011**. - V. 39, N. 11. - P. 4836 - 4850.

25. Втюрина, Н.Н. Контекстные характеристики ДНК, значимые для ее повреждения ультрафиолетовым лазерным излучением с длиной волны 193 нм. / Н.Н. Втюрина, С.Л. Гроховский, А.Б. Васильев, И.И. Титов, П.М. Пономаренко, **М.П. Пономаренко**, С.Е. Пельтек, Ю.Д. Нечипуренко, Н.А. Колчанов // ДАН. – **2012**. - Т. 447, № 2. - С. 217 - 222.

26. Savinkova, L.K. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. / L.K. Savinkova, I.A. Drachkova, T.V. Arshinova, P.M. Ponomarenko, **M.P. Ponomarenko**, N.A. Kolchanov // PLoS ONE. - **2013**. - V. 8, N. 2. - P. e54626.

27. Drachkova, I. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. / I. Drachkova, L. Savinkova, T. Arshinova, **M. Ponomarenko**, S. Peltek, N.A. Kolchanov // Hum. Mutat. - **2014**. - V. 35, N. 5. – P. 601 – 608.

28. Ponomarenko, P.M. Sequence-based prediction of transcription upregulation by auxin in plants. / P.M. Ponomarenko, **M.P. Ponomarenko** // J. Bioinform. Comput. Biol. - **2015**. - V. 13, N. 1. - P. 1540009.

29. Arkova, O.V. Obesity-related known and candidate SNP markers can significantly change affinity of TATA-binding protein for human gene promoters. / O.V. Arkova, **M.P. Ponomarenko**, D.A. Rasskazov, I.A. Drachkova, T.V. Arshinova, P.M. Ponomarenko, L.K. Savinkova, N.A. Kolchanov // BMC Genomics. - **2015**. - V. 16, Suppl. 13. - P. S5.

30. **Ponomarenko, M.P**. Candidate SNP markers of gender-biased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. / M.P. Ponomarenko, O. Arkova, D. Rasskazov, P. Ponomarenko, L. Savinkova, N. Kolchanov // Front. Immunol. - **2016**. – V. 7. – P. 130.

Статьи в сборниках научных трудов

31. **Ponomarenko, M.P**. Search for DNA conformational features for functional sites. Investigation of the TATA box. In: Pac. Symp. Biocomput. / M.P. Ponomarenko, J.V. Ponomarenko, A.E. Kel, N.A. Kolchanov; Eds. R. Altman, A.K. Dunker, L. Hunter, T.E. Klein - Singapore: World Sci. - **1997**. - V. 2, P. 340 - 351.

32. **Пономаренко, М.П**. Компьютерное представление и автоматическая генерация знаний о функциональной активности молекул ДНК и РНК. В: Труды ИВМиМГ СО РАН. Сер. Мат. моделирование в геофизике / М.П. Пономаренко, Ю.В. Пономаренко, А.С. Фролов, С.В. Лаврюшов, Н.А. Колчанов, Н.Л. Подколодный, Г.Н. Ерохин - Новосибирск: Изд. ИВМиМГ СО РАН. - **1998**. - Вып. 5. - С. 181 - 198.

33. Kolchanov, N.A. GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences in eukaryotic genome. In: Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology. / N.A. Kolchanov, **M.P. Ponomarenko**, A.E. Kel, Yu.V. Kondrakhin, A.S. Frolov, F.A. Kolpakov, T.N. Goryachkovsky, O.V. Kel, E.A. Ananko, E.V. Ignatieva, O.A. Podkolodnaya, V.N. Babenko, I.L. Stepanenko, A.G. Romashchenko, T.I. Merkulova, D.G. Vorobiev, S.V. Lavryushev, Yu.V. Ponomarenko, A.V. Kochetov, G.B. Kolesov, V.V. Solovyev, L. Milanesi, N.L. Podkolodny, E. Wingender, T. Heinemeyer; Eds J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, C. Sensen - Palo Alto: AAAI Press. - **1998**. - P. 95 - 104.

34. **Ponomarenko, M.P**. Mean-recognition: a systematic approach increasing the accuracy of the functional site recognition for the genomic DNA annotation. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, 1998, Orlando, Florida. / M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, O.A. Podkolodnaya, D.V. Vorobyev, N.A. Kolchanov, G.C. Overton; Eds. N. Callaos, L. Holmes - Piscataway: IEEE Press, **1998.** V. 4, P. 224-230.

35. Ponomarenko, J.V. Diffusion additive to TATA-box recognition score detects the non-contextual TBP-binding site at -30 position of TATA-less promoters. In: Proceedings of WSES/MIUE/HNA International Conference: Mathematics and Computers in Biology and Chemistry (MCBC'2000), December 20-22, 2000, Montego Bay, Jamaica. / J.V. Ponomarenko, **M.P. Ponomarenko**, A.S. Frolov, I. Ivan Zvolsky; Ed. N. Mastorakis - NY: WSES Press Publ. - **2000.** - P. 901 - 906.

36. Ponomarenko, J.V. Annotation of potential transcription factor binding sites using rSNP_Guide with the models of experimentally characterized altered TF sites. In: EuroQSAR 2002: Designing drugs and crop protectants. / Ponomarenko J.V., Orlova G.V., Merkulova T.I., Gorshkova E.V., Valuev V.P., **Ponomarenko M.P.**; Eds. M. Ford, D. Livingstone, J. Dearden, H. Van de Waterbeemd - Oxford: Blackwell Publ. Ltd (UK), - **2003.** - P. 347-351.

Главы в монографиях

37. Ощепков, Д.Ю. Регуляторные последовательности ДНК: исследование компьютерными методами. В сб.: Системная компьютерная биология. / Д.Ю. Ощепков, Д.Н. Бугреев, Г.А. Невинский, Н.А. Колчанов, Е.В. Игнатьева, Н.В. Климова, Г.В. Васильев, Т.И. Меркулова, **М.П. Пономаренко**, Ю.Н. Воробьев, Д.Ю. Емельянов, В.Г. Левицкий, Е.А. Ананько, О.В. Вишневский, В.Ф. Кобзев, Т.В. Бусыгина; Отв. ред. Н.А. Колчанов, С.С. Гончаров, В.А. Лихошвай, В.А. Иванисенко. - Новосибирск: Изд. СО РАН, **2008**. - Интеграционные проекты СО РАН; Вып 14. - С. 38-125.

38. Кочетов, А.В. Трансляция. В сб.: Системная компьютерная биология. / А.В. Кочетов, О.В. Вишневский, О.А. Волкова, Н.В. Владимиров, В.А. Лихошвай, Ю.Г. Матушкин, Д.А. Григорович, Х. Нишида, А. Сараи, О.Г. Смирнова, С.С. Ибрагимова, **М.П. Пономаренко**, А.Ю. Пальянов, И.И. Титов, Н.А Колчанов; Отв. ред. Н.А. Колчанов, С.С. Гончаров, В.А. Лихошвай, В.А. Иванисенко.- Новосибирск: Изд. СО РАН, **2008**. - Интеграционные проекты СО РАН; Вып 14. - С. 184-228.

39. **Ponomarenko M**. Cladogenesis. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, K Gunbin, A. Doroshkov, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013.** - V. 2. – P. 21 - 24.

40. **Ponomarenko, M**. Degrees of Freedom. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, V. Babenko, A. Kochetov, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013**. - V. 2. – P. 290 - 292.

41. **Ponomarenko, M**. Heat Shock Proteins. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, I. Stepanenko, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013**. - V. 3. – P. 402 - 405.

42. **Ponomarenko, M**. Hogness Box. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, V. Mironova, K. Gunbin, L. Savinkova; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013**. - V. 3. – P. 491 - 494.

43. **Ponomarenko, M**. Initiation Factors. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, L. Savinkova, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013**. - V. 4. – P. 83 - 85.

44. **Ponomarenko, M**. Unique DNA. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, G. Orlova, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – **2013**. - V. 7. – P. 259 - 262.

Тезисы конференций

45. Ponomarenko, J.V., ACTIVITY: a database for activities of functional DNA/RNA sites. In: Proceedings of the First Conference on Bioinformatics of Genome Regulation and Structure: BGRS'98. / J.V. Ponomarenko, D.P. Furman, T.M. Mishchenko, L.V. Katokhina, V.P. Valuev, E.L. Peregoedova, A.S. Frolov, N.L. Podkolodny, **M.P. Ponomarenko**; Eds. N.A. Kolchanov, E. Borovskikh, G. Chirikova, D. Afonnikov, S. Lavryushev - Novosibirsk: IC&G Press, **1998**. - V. 1. - P. 62 - 65.

46. Savinkova, L.K. Quantitative computer-assisted analysis of the TATA-binding protein affinity for complementary duplexes of synthetic oligodeoxyribonucleotides. In: Proceedings of the First Conference on Bioinformatics of Genome Regulation and Structure: BGRS'98 / L.K. Savinkova, A.A. Sokolenko, V.A. Rau, V.F. Kobzev, **M.P. Ponomarenko**, J.V. Ponomarenko, N.A. Kolchanov; Eds. N.A. Kolchanov, E. Borovskikh, G. Chirikova, D. Afonnikov, S. Lavryushev - Novosibirsk: IC&G Press. - **1998**. - V. 1. - P. 165-169.

47. **Ponomarenko M.P.**, B-DNA-VIDEO: an active database for the significant B-DNA features of transcription factor binding sites. In: Proceedings of the First Conference on Bioinformatics of Genome Regulation and Structure: BGRS'98 / L.K. Savinkova, A.A. Sokolenko, V.A. Rau, V.F. Kobzev, M.P. Ponomarenko, J.V. Ponomarenko, N.A. Kolchanov; Eds. N.A. Kolchanov, E. Borovskikh, G. Chirikova, D. Afonnikov, S. Lavryushev - Novosibirsk: IC&G Press. - **1998**. - V. 1. - P. 158 - 162.

48. Ponomarenko, J.V. Sequence-dependent B-helix DNA features common for transcription factor superclasses. In: The First Cold Spring Harbor Workshop "Bridging the Gap between Sequences and Functions", September 7 - 9, 1999b, New York, USA" / J.V. Ponomarenko, **M.P. Ponomarenko**, O.A. Podkolodnaya, A.S. Frolov; Ed. M. Zhang - Cold Spring Harbor: CSHL Press. – **1999**. - P. 12.

49. Ponomarenko, J.V. Quantitative computer-assisted analysis of the TATA-binding protein affinity for complementary duplexes of synthetic oligodeoxyribonucleotides. In: Proceedings of The Second Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2000 / J.V. Ponomarenko, G.V. Orlova, **M.P. Ponomarenko**, S.V. Lavryushev, S.V. Zybova, A.S. Frolov; Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2000**. - V. 1. - P. 37 - 40.

50. **Ponomarenko M.P**. A database on DNA sequence/activity relationships: application to phylogenetic footprinting. In: Proceedings of The Fourth Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2004 / M.P. Ponomarenko, J.V. Ponomarenko; Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2004**. - V. 1. - P. 166-169.

51. Gunbin K.V. Why TATA-box hides at gli gene molecular evolution? In: Proceedings of The Fifth Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2000 / K.V. Gunbin, P.M. Ponomarenko, **M.P. Ponomarenko**, N.A. Kolchanov; Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2008**. - P. 95.

52. Suslov V.V. SNPs in the HIV-1 TATA box and the AIDS pandemic. In: Proceedings of The 7th Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2010 / V.V. Suslov, P.M. Ponomarenko, V.M. Efimov, **M.P. Ponomarenko**, L.K. Savinkova, N.A. Kolchanov; Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2010**. - P. 283.

53. Mironova V.V. Quantitative sequence / activity relationships of auxin response elements (AuxRE) in plant promoters. In: Proceedings of International Moscow Conference on Computational Molecular Biology, MCCMB'2011 / Mironova V.V., Ponomarenko P.M., Omelyanchuk N.A., **Ponomarenko M.P.**; Ed. M. Gelfand - Moscow: MSU Press - **2011.** - P. 286-287.

54. Vtyurina N.N. Contextual DNA features significant for the DNA damage by the 193 nm ultraviolet laser beam. In: Proceedings of The 8th Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2012 / N.N. Vtyurina, S.L. Grokhovsky, A.B. Vasiliev, I.I. Titov, P.M. Ponomarenko, **M.P. Ponomarenko**, S.E., Peltek Yu.D. Nechipurenko, N.A. Kolchanov; Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2012**. - P. 328.

55. Ponomarenko P.M. An empirical equilibrium equation of a gene response to auxin in plants allows to predict quantitatively the auxin response upon the gene promoter sequence. In: Proceedings of The 9th Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2014 / P.M. Ponomarenko, **M.P. Ponomarenko**, Ed. N.A. Kolchanov - Novosibirsk: IC&G Press. - **2014**. - P. 129.

По материалам диссертации опубликовано 55 научных работ, из них – 30 статей в журналах из Перечня ВАК (все индексированы в РИНЦ, Scopus и Web of Science), в том числе 19 – в зарубежных журналах; а также 6 статей в сборниках научных трудов, 8 глав в монографиях, 11 тезисов конференций.

СПИСОК ЛИТЕРАТУРЫ

1. Брызгалов, Л.О. Выявление генов-мишеней транскрипционных факторов FOXA, связанных с регуляцией пролиферации. / Л.О. Брызгалов, Н.И. Ершов, Д.Ю. Ощепков, В.И. Каледин, Т.И. Меркулова // Биохимия. - 2008. - Т. 73, № 1. - С. 86 - 92.

2. Бухер, Ф. Описание промоторов эукариот в базе данных EPD. / Ф. Бухер // Мол. биол. - 1997. - Т. 31, № 4. - С. 616 - 625.

3. Васильев, Г.В. Точковые мутации в районе 663-666 п.н. интрона 6 гена триптофаноксигеназы, связанные с рядом психических расстройств, разрушают сайт связывания фактора транскрипции YY1. / Г.В. Васильев, В.М. Меркулов, В.Ф. Кобзев, Т.И. Меркулова, М.П. Пономаренко, Ю.В. Пономаренко, О.А. Подколодная, Н.А. Колчанов // Мол. биол. – 2000. - Т. 34, № 2. - С. 214 - 222.

4. Вингендер, Э. Классификация транскрипционных факторов эукариот. / Э. Вингендер // Мол. биол. - 1997. - Т. 31, № 4. - С. 584 - 600.

5. Втюрина, Н.Н. Расщепление фрагментов ДНК наносекундным лазером с длиной волны 193 нм. / Н.Н. Втюрина, С.Л. Гроховский, И.В. Филимонов, О.И. Медведков, Д.Ю. Нечипуренко, С.А. Васильев, Ю.Д. Нечипуренко // Биофизика. - 2011. - Т. 56, № 3. - С. 410 - 414.

6. Втюрина, Н.Н. Контекстные характеристики ДНК, значимые для ее повреждения ультрафиолетовым лазерным излучением с длиной волны 193 нм. / Н.Н. Втюрина, С.Л. Гроховский, А.Б. Васильев, И.И. Титов, П.М. Пономаренко, М.П. Пономаренко, С.Е. Пельтек, Ю.Д. Нечипуренко, Н.А. Колчанов // ДАН. – 2012. - Т. 447, № 2. - С. 217 - 222.

7. Гусев, В. Анализ сложности геномов. Мера сложности и классификация выявленных структурных особенностей. / В. Гусев, В. Куличков, О. Чупахина // Мол. биол. – 1991. - Т. 25, № 3. - С. 825 - 834.

8. Ершов, Н.И. Изменения транскриптома печени крысы под действием гепатоканцерогенного для этих животных 3'-МеДАБ и неканцерогенного

ОАТ. / Н.И. Ершов, В.Г. Левицкий, Д.Ю. Ощепков, О.В. Вишневский, Л.О. Брызгалов, Е.В. Антонцева, Т.И. Меркулова // Вавилов. ж. генет. селек. - 2009. - Т. 13, № 4. - С. 703 - 722.

9. Игнатьева, Е.В. Регуляция транскрипции генов липидного метаболизма: описание в базе данных TRRD. / Е.В. Игнатьева, Т.И. Меркулова, О.В. Вишневский, А.Э. Кель // Мол. биол. - 1997. - Т. 31, № 4. - С. 684 - 700.

10. Игнатьева, Е.В. Поиск новых сайтов связывания транскрипционного фактора SF-1 методом SITECON: экспериментальная проверка и анализ регуляторных районов генов-ортологов. / Е.В. Игнатьева, Н.В. Климова, Д.Ю. Ощепков, Г.В. Васильев, Т.И. Меркулова, Н.А. Колчанов // ДАН. - 2007. - Т. 415, № 1. - С. 120 - 124.

11. Игнатьева, Е.В. Выявление новых сайтов связывания транскрипционных факторов SREBP в промоторных районах генов позвоночных на основе комбинации биоинформатического и экспериментального подходов. / Е.В. Игнатьева, Т.И. Меркулова, Д.Ю. Ощепков, Н.В. Климова, Г.В. Васильев, И.И. Турнаев, В.Ф. Кобзев, Н.А. Колчанов // Вавилов. ж. генет. селек. - 2009. - Т. 13, № 1. - С. 37 - 45.

12. Игнатьева, Е.В. Регуляторная геномика – экспериментально-компьютерные подходы. / Е.В. Игнатьева, О.А. Подколодная, Ю.Л. Орлов, Г.В. Васильев, Н.А. Колчанов // Генетика. - 2015. - Т. 51, № 4. - С. 409-429.

13. Ильина, В.Л. Зависимость частоты спонтанного возникновения реверсов разных типов у ауксотрофных по аденину дрожжей от содержания аденина в среде. / В.Л. Ильина, В.И. Корогодин, Ч. Файси // Генетика. - 1987. - Т. 23, № 4. - С. 637—642.

14. Караванов, А.А. Сравнительный анализ структуры геномов Vicia faba и Vicia sativa. / А.А. Караванов, А.Б. Иорданский // Мол. биол. - 1971. - Т. 5, № 6. - С. 706 - 709.

15. Кель, А.Э. Конвергентное возникновение повторов в генах, кодирующих глобулярные белки. Анализ факторов, определяющих наличие

прямых повторов. / А.Э. Кель, Н.А. Колчанов, В.В. Соловьев // ЖОБ. - 1988. - Т. 49, № 3. - С. 343 - 354.

16. Кель, А.Э. Теоретический анализ механизмов возникновения делеций ДНК в геномах прокариот на основе прямых повторов. / А.Э. Кель, Н.А. Колчанов, В.В. Соловьев // Мол. биол. – 1989. - Т. 23, № 3. - С. 184 - 192.

17. Кель, А.Э. TRRD: база данных транскрипционных регуляторных районов генов эукариот. / А.Э. Кель, Н.А. Колчанов, О.В. Кель, А.Г. Ромащенко, Е.А. Ананько, Е.В. Игнатьева, Т.И. Меркулова, О.А. Подколодная, И.Л. Степаненко, А.В. Кочетов, Ф.А. Колпаков, Н.Л. Подколодный, А.А. Наумочкин // Мол. биол. - 1997. - Т. 31, № 4. - С. 626 - 636.

18. Кель, О.В. Композиционные регуляторные элементы: классификация и описание в базе данных COMPEL / О.В. Кель, А.Э. Кель, А.Г. Ромащенко, Э. Вингендер Н.А., Колчанов // Мол. биол. - 1997. - Т. 31, № 4. - С. 601 - 615.

19. Киселева, Е.В. Электронно-микроскопический анализ уровней структурной организации хромосомы *Escherichia coli*. / Е.В. Киселева, Е.В. Лихошвай, Н.А. Сердюкова, Н.Б. Христолюбова // ДАН СССР. – 1986. – Т. 289, № 5. – С.1235-1237.

20. Киселева, Е.В. Электронно-микроскопическое изучение структурной организации ДНК в нуклеоидах спор и мицелия стрептомицетов. / Е.В. Киселева, Н.П. Кулыба, С.И. Байбородин, З.И. Панфилова, Н.Б. Христолюбова, А.В. Козлов, Р.И.Салганик // ДАН СССР. – 1988а. – Т. 299, № 6. – С.1486-1488.

21. Киселева, Е.В. Структурная организация генома хлоропластов *Beta vulgaris L.* / Е.В. Киселева, Н.А. Дударева, А.Э. Дикалова, Н.Б. Христолюбова, Р.И.Салганик // ДАН СССР. – 1988б. – Т. 302, № 5. – С.1229-1488.

22. Колпаков, Ф.А. Высокая гетерогенность промоторов генов высших эукариот, транскрибируемых РНК-полимеразой II. / Ф.А. Колпаков, А.Э. Кель, М.П. Пономаренко, Н.А. Колчанов // ДАН. – 1997. - Т. 357, № 5. - С. 693 - 695.

23. Колпаков, Ф.А. Методы интеграции неоднородных молекулярно-генетических информационных ресурсов в электронной библиотеке

GENEEXPRESS. / Ф.А. Колпаков, Н.Л. Подколодный, С.В. Лаврюшев, Д.А. Григорович, М.П. Пономаренко, Н.А. Колчанов // Программирование. – 2000. - Т. 4, № 3. - С. 72 - 80.

24. Колчанов, Н.А. Высокая насыщенность прямые повторами генов РНК-полимеразы на основе данных контекстного анализа. / Н.А. Колчанов, В.В. Соловьев, А.А. Жарких // ДАН СССР. - 1983. - Т. 273, № 3. - С. 741 - 744.

25. Колчанов, Н.А. Контекстные методы теоретического анализа генетических макромолекул (ДНК, РНК и белков). В: Итоги науки и техники. / Н.А. Колчанов, В.В. Соловьев, А.А Жарких. - М: ВИНИТИ. - Т. 21, Мол. биол. - 1985. С. 6 - 34.

26. Колчанов, Н.А. Теоретическое исследование закономерностей структурно-функциональной организации и эволюции генетических макромолекул: автореф. дис. … докт. биол. наук: 03.00.15 / Колчанов, Николай Александрович – Новосибирск. – 1988 - 32 с.

27. Колчанов, Н.А. Регуляция транскрипции генов эукариот: базы данных и компьютерный анализ. / Н.А. Колчанов // Мол. биол. - 1997. - Т. 31, № 4. - С. 581 - 583.

28. Колчанов, Н.А. Функциональные сайты геномов про- и эукариот: компьютерное моделирование и предсказание активности. / Н.А. Колчанов, М.П. Пономаренко, Ю.В. Пономаренко, А.С. Фролов, Н.Л. Подколодный // Мол. биол. – 1998. - Т. 32, № 2. - С. 255 - 267.

29. Колчанов, Н.А. GeneExpress: интегратор баз данных и компьютерных систем, доступных по сети Интернет и предназначенных для изучения экспрессии генов эукариот / Н.А. Колчанов, М.П. Пономаренко, А.Э. Кель, Ю.В. Кондрахин, А.С. Фролов, Ф.А. Колпаков, Т.Н. Горячковская, О.В. Кель, Е.А. Ананько, Е.В. Игнатьева, О.А. Подколодная, И.Л. Степаненко, Т.И. Меркулова, В.В. Бабенко, Д.В. Воробьев, С.В. Лаврюшев, Ю.В. Пономаренко, А.В. Кочетов, Г.Н. Колесов, Н.Л. Подколодный, Л. Миланези, Э. Вингендер, Т. Хейнемайер, В.В. Соловьев, Г.К. Овертон // Биофизика. – 1999. - Т. 44, № 5. - С. 837 - 841.

30. Кузнецова, Т.Н. Анализ структуры инсулин-зависимых регуляторных контуров зрелых адипоцитов. / Т.Н. Кузнецова, Е.В. Игнатьева, В.А. Мордвинов, А.В. Катохин, М.Ю. Шаманина, Д.Ю. Ощепков, Н.А. Колчанов // Усп. физиол. наук. - 2008. - Т. 39, № 1. - С. 3 - 22.

31. Левицкий, В.Г. Компьютерный анализ нуклеосомной организации ДНК и промоторов эукариот: автореф. дис. ... канд. биол. наук: 03.00.15 / Левицкий, Виктор Георгиевич. – Новосибирск. - 2001. - 17 с.

32. Левицкий, В.Г. Анализ результатов эксперимента по массовой иммунопреципитации хроматина с помощью методов распознавания сайтов связывания транскрипционных факторов. / В.Г. Левицкий, Г.В., Васильев Д.Ю. Ощепков, Н.И. Ершов, Т.И. Меркулова // Вавилов. ж. генет. селек. - 2010. - Т. 14, № 4. - С. 685 - 697.

33. Левицкий, В.Г. Разработка методов распознавания сайтов связывания транскрипционных факторов FOXA, их экспериментальная верификация и использование для анализа данных массовой иммунопреципитации хроматина. / В.Г. Левицкий, Д.Ю. Ощепков, Н.И. Ершов, Л.О. Брызгалов, Е.В. Антонцева, Г.В. Васильев, Т.И. Меркулова, Н.А. Колчанов // ДАН. - 2011. - Т. 436, № 3. - С. 417 - 421.

34. Леман, Э. Проверка статистических гипотез. / Э. Леман. - М: Наука. - 1979. - 408 с.

35. Меркулова, Т.И. Регуляторные коды транскрипции геномов эукариот. / Т.И. Меркулова, Е.А. Ананько, Е.В. Игнатьева, Н.А. Колчанов. // Генетика. - 2013. - Т. 49, № 1. - С. 37-54.

36. Минский, М. Перцептроны. / М. Минский, С. Пайперт. - М.: Мир. - 1971. - 261 с.

37. Миронов, А.А. Предсказание ансамблей вторичных структур РНК. Кинетический анализ самоорганизации. / А.А. Миронов, Л.П. Дьяконов, А.Э. Кистер // Мол. биол. – 1984. - Т. 18, № 6. - С. 1686 - 1694.

38. Миронова, В.В. Эффективность связывания TBP с промотором ARF-генов растений коррелирует с характером влияния ARF белков на

транскрипцию (активатор/репрессор). / В.В. Миронова, Н.А. Омельянчук, П.М. Пономаренко, М.П. Пономаренко, Н.А. Колчанов // ДАН. – 2010. – Т. 433, № 4. - С. 549 – 554.

39. Нильсон, Н. Принципы искусственного интеллекта. / Н. Нильсон. – М: Радио и связь. – 1985. - 376 с.

40. Омельянчук, Н.А. Особенности нуклеотидных последовательностей зрелых микроРНК могут влиять на сродство к белкам Ago2 и Ago3 человека. / Н.А. Омельянчук, П.М. Пономаренко, М.П. Пономаренко // Мол. биол. – 2011. - Т. 45, № 2. - С. 366 – 375.

41. Ощепков, Д.Ю. Компьютерный анализ конформационных и физико-химических особенностей нуклеотидных последовательностей, расщепляемых ДНК-топоизомеразой I. / Ощепков Д.Ю., Бугреев Д.В., Колчанов Н.А., Невинский Г.А. // Мол. биол. - 2005. - Т. 39, № 3. - С. 488 - 496.

42. Ощепков, Д.Ю. Выявление новых DRE в регуляторной области генов человека, кодирующих компоненты цитозольного комплекса арил-гидрокарбонового рецептора / Д.Ю. Ощепков, Д.П. Фурман, Е.А. Ощепкова, А.В. Катохин, М.Ю. Шаманина, В.А. Мордвинов // Вавилов. ж. генет. селек. - 2009. - Т. 13, № 1. - С. 46 - 52.

43. Ощепков, Д.Ю. Компьютерный анализ конформационных и физико-химических особенностей функциональных сайтов геномной ДНК эукариот: дис. ...канд. биол. наук: 03.01.09 / Ощепков, Дмитрий Юрьевич - Новосибирск. - 2010. - 177 с.

44. Подколодная О.А. Механизмы транскрипционной регуляции эритроид-специфичных генов. / О.А. Подколодная, И.Л. Степаненко // Мол. биол. - 1997. - Т. 31, № 4. -С. 671 - 683.

45. Подколодный, Н.Л. Программный комплекс SNP-MED для анализа влияния однонуклеотидных полиморфизмов на функцию генов, связанных с развитием социально значимых заболеваний. / Н.Л. Подколодный, Д.А. Афонников, Ю.Ю. Васькин, Л.О. Брызгалов, В.А. Иванисенко, П.С. Деменков, М.П. Пономаренко, Д.А. Рассказов, К.В. Гунбин, И.В. Процук, И.Ю. Шутов,

П.Н. Леонтьев, М.Ю. Фурсов, Н.П. Бондарь, Е.В. Антонцева, Т.И. Меркулова, Н.А. Колчанов // Вавилов. ж. генет. селек. - 2013. - Т. 17, № 4/1. - С. 577 - 588.

46. Пономаренко, М.П. Контекстные сигналы и антисигналы сайта встраивания td-интрона. / М.П. Пономаренко, А.Э. Кель, А.Н. Колчанова, Н.А. Колчанов // ДАН. – 1996. – Т. 348, № 3. - С. 424 - 427.

47. Пономаренко, М.П. Компьютерное моделирование последовательностей ТАТА-боксов промоторов эукариот. / М.П. Пономаренко, Л.К. Савинкова, А.Э. Кель, Н.А. Колчанов // ДАН. - 1997а. - Т. 355, № 4. - С. 557 - 561.

48. Пономаренко, М.П. Моделирование последовательностей ТАТА-боксов генов эукариот. / М.П. Пономаренко, Л.К. Савинкова, Ю.В. Пономаренко, А.Э. Кель, И.И. Титов, Н.А. Колчанов // Мол. биол. - 1997б. - Т. 31, № 4. - С. 726 - 732.

49. Пономаренко, М.П. Компьютерный анализ конформационных особенностей ДНК ТАТА-боксов промоторов эукариот. / М.П. Пономаренко, Ю.В. Пономаренко, А.Э. Кель, Н.А. Колчанов, Х. Карас, Э. Вингендер, Х. Скленар // Мол. биол. – 1997в. - Т. 31, № 4. - С. 733 - 740.

50. Пономаренко, М.П. Предпочтительность RecA-филамента к последовательностям ДНК коррелирует с генетическим кодом. / М.П. Пономаренко, Ю.В. Пономаренко, И.И. Титов, Н.А. Колчанов, А.В. Мазин, С. Ковальчиковски // ДАН. – 1998. - Т. 363, № 1. - С. 122 - 125.

51. Пономаренко, М.П. Усреднение результатов распознавания сайтов может увеличить точность аннотации генома человека. / М.П. Пономаренко, Ю.В. Пономаренко, О.А. Подколодная, А.С. Фролов, Д.В. Воробьев, Н.А. Колчанов, Г. Овертон // Биофизика. - 1999а. - Т. 44, № 4. - С. 649 - 654.

52. Пономаренко, М.П. Вклад сигналов и антисигналов в мутационный спектр сайта встраивания td-интрона. / М.П. Пономаренко, Ю.В. Пономаренко, Н.А. Колчанов // Биофизика. - 1999б. - Т. 44, № 4. - С. 655 - 663.

53. Пономаренко, М.П. LIKENESS: система поиска в режиме “реального времени” и выравнивания пространственных структур белков. / М.П.

Пономаренко, И.Н. Шиндялов, Ф. Борн, Н.А. Колчанов // Биофизика. 1999в. - Т. 44, № 4. - С. 821 - 831.

54. Пономаренко, М.П. Содержание микроРНК в Arabidopsis thaliana коррелирует с встречаемостью тетрануклеотидов WRHW и DRYD. / М.П. Пономаренко, Н.А. Омельянчук, А.В. Катохин, Н.А. Колчанов // Инф. Вест. ВОГиС. – 2006. - Т. 10, № 2. - С. 304 - 311.

55. Пономаренко, М.П. Содержание микроРНК в Arabidopsis thaliana коррелирует с наличием в последовательностях тетрануклеотидов WRHW и DRYD. / М.П. Пономаренко, Н.А. Омельянчук, А.В. Катохин, С.А. Савинская, Н.А. Колчанов // ДАН. – 2008. - Т. 420, № 5. - С. 703 - 707.

56. Пономаренко, М.П. Выявление связи вариабельности экспрессии генов путей передачи сигналов в мозге человека со сродством TATA-связывающего белка к промоторам этих генов. / М.П. Пономаренко, В.В. Суслов, К.В. Гунбин, П.М. Пономаренко, О.В. Вишневский, Н.А. Колчанов // Вавилов. ж. генет. селек. - 2014. – V. 18, № 4/3. – С. 1219-1230.

57. Пономаренко, П.М. Пошаговая модель связывания TBP/TATA-бокс позволяет предсказать наследственное заболевание человека по точечному полиморфизму. / П.М. Пономаренко, Л.К. Савинкова, И.А. Драчкова, М.В. Лысова, Т.В. Аршинова, М.П. Пономаренко, Н.А. Колчанов // ДАН. – 2008. - Т. 419, № 6. - С. 828 - 832.

58. Пономаренко, П.М. Прогноз изменения аффинности TATA-связывающего белка к TATA-боксам в результате полиморфизмов TATA-боксов промоторов генов человека. / П.М. Пономаренко, М.П. Пономаренко, И.А. Драчкова, М.В. Лысова, Т.В. Аршинова, Л.К. Савинкова, Н.А. Колчанов // Мол. биол. – 2009. - Т. 43, № 3. - С. 512 - 520.

59. Пономаренко, П.М. Точное уравнение равновесия четырех шагов связывания TBP с TATA-боксом для прогноза фенотипического проявления мутаций. / П.М. Пономаренко, В.В. Суслов, Л.К. Савинкова, М.П. Пономаренко, Н.А. Колчанов // Биофизика. – 2010. - Т. 55, № 3. - С. 400 - 414.

60. Пономаренко, Ю.В. Компьютерный анализ конформационных и физико-химических особенностей сайтов связывания транскрипционных факторов эукариот: дис. … канд. биол. наук: 03.00.15 / Пономаренко, Юлия Владимировна. – Новосибирск. - 2002. - 225 с.

61. Рассказов, Д.А. Оценка по технологии rSNP_Guide SNPs промоторов генов APC и MLH1 человека, связанных с раком толстого кишечника. / Д.А. Рассказов, Е.В. Антонцева, Л.О. Брызгалов, М.Ю. Матвеева, Е.В. Кашина, П.М. Пономаренко, Г.В. Орлова, М.П. Пономаренко, Д.А. Афонников, Т.И. Меркулова. // Вавилов. ж. генет. селек. – 2013а. - Т. 17, № 4/1. - С. 589 - 598.

62. Рассказов, Д.А. SNP_TATA_COMPARATOR: Web-сервис применения уравнения равновесия TBP/TATA-комплекса в сравнительной оценке SNPs промоторов генов, связанных с болезнями человека. / Д.А. Рассказов, К.В. Гунбин, П.М. Пономаренко, О.В. Вишневский, М.П. Пономаренко, Д.А. Афонников // Вавилов. ж. генет. селек. – 2013б. - Т. 17, № 4/1. - С. 599 - 606.

63. Ратнер, В.А. Молекулярная генетика: принципы и механизмы. / В.А. Ратнер – Новосибирск: Наука, 1983. - 256 с.

64. Ратнер, В.А. Концепция лимитирующих генетических факторов экспрессии, организации и эволюции. / В.А. Ратнер // Генетика. - 1990. – Т. 26, № 5. – С. 789 - 803.

65. Ратнер, В.А Блочно-модульный принцип организации и эволюции молекулярно-генетических систем управления (МГСУ) / В.А. Ратнер // Генетика. - 1992. - Т. 28, № 2. - С.5 - 23.

66. Савинкова, Л.К. Взаимодействие рекомбинантного TATA-связывающего белка с TATA-боксами промоторов генов млекопитающих. / Л.К. Савинкова, И.А. Драчкова, М.П. Пономаренко, М.В. Лысова, Т.В. Аршинова, Н.А. Колчанов // Экол. генетика – 2007. - Т. V, № 2. - С. 44 - 49.

67. Савинкова, Л.К. Полиморфизмы TATA-боксов промоторов генов человека и ассоциированные с ними наследственные патологии. / Л.К. Савинкова, М.П Пономаренко, П.М. Пономаренко, И.А. Драчкова, М.В.

Лысова, Т.В. Аршинова, Н.А. Колчанов // Биохимия. – 2009. - Т. 74, № 2. - С. 149 - 163.

68. Соколенко, А.А. Взаимодействие дрожжевого TATA-связывающего белка с короткими участками промоторов. / А.А. Соколенко, И.И. Сандомирский, Л.К. Савинкова // Мол. биол. - 1996. - Т. 30, № 2. - С. 279 – 285.

69. Соловьев, В.В. Молекулярный механизм соматического гипермутагенеза в генах иммуноглобулинов. Связь соматических мутаций с повторами. Метод статистического взвешивания. / В.В. Соловьев, Н.А. Колчанов, И.Б. Рогозин // Мол. биол. – 1989. - Т. 23, № 3. - С. 783 - 794.

70. Суслов, В.В. Полиморфизмы TATA-боксов генов хозяйственно важных и лабораторных животных и растений, ассоциированные с их селекционно-ценными признаками. / В.В. Суслов, П.М. Пономаренко, М.П. Пономаренко, И.А. Драчкова, Т.В. Аршинова, Л.К. Савинкова, Н.А. Колчанов // Генетика. – 2010. - Т. 46, № 4. - С. 448 - 457.

71. Тимофеева, О.А. Ассоциированный с чувствительностью к канцерогенезу в легких точковый полиморфизм в интроне 2 гена K-ras влияет на связывание с фактором GATA-6, но не на уровень экспрессии гена. / О.А. Тимофеева, Е.В. Горшкова, З.Б. Левашова, В.Ф. Кобзев, М.Л. Филипенко, В.И. Каледин, Т.И. Меркулова // Мол. биол. 2002. - Т. 36, №. 5. - С. 817 - 824.

72. Abeel, T. Generic eukaryotic core promoter prediction using structural features of DNA. / T. Abeel, Y. Saeys, E. Bonnet, P. Rouze, Y. van de Peer // Genome Res. - 2008. - V. 18, N. 2. - P. 310 - 323.

73. Abnizova, I. Statistical information characterization of conserved non-coding elements in vertebrates. / I. Abnizova, K. Walter, R. Te Boekhorst, G. Elgar, W.R. Gilks // J. Bioinform. Comput. Biol. – 2007. - V. 5, N. 2B. - P. 533 - 547.

74. Aboussekhra, A. TATA-binding protein promotes the selective formation of UV-induced (6-4)-photoproducts and modulates DNA repair in the TATA box. / A. Aboussekhra, F. Thoma // EMBO J. – 1999. – V. 18, N. 2. – P. 433 - 443.

75. Adami, C. Evolution of biological complexity. / C. Adami, C. Ofria, T. Collier // Proc. Natl. Acad. Sci. USA. – 2000. - V. 97, N. 9. - P. 4463 - 4468.

76. Akhtar, W. TBP-related factors: a paradigm of diversity in transcription initiation. / W. Akhtar, G.J. Veenstra // Cell Biosci. - 2011. - V. 1, N. 1. - P. 23.

77. Altschul, S.F. Basic local alignment search tool. / S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman // Mol. Biol. - 1990. - V. 215, N. 3. - P. 403 - 410.

78. Ananko, E.A. GeneNet in 2005. / E.A. Ananko, N.L. Podkolodny, I.L. Stepanenko, OA Podkolodnaya, D.A. Rasskazov, D.S. Miginsky, V.A. Likhoshvai, A.V. Ratushny, N.N. Podkolodnaya, N.A. Kolchanov // Nucleic Acids Res. – 2005. - V. 33, Database issue. - P. D425-D427.

79. Angelov, D. Origin of the heterogeneous distribution of the yield of guanyl radical in UV laser photolyzed DNA. / D. Angelov, B. Beylot, A. Spassky // Biophys. J. - 2005. - V. 88, N. 4. - P. 2766 - 2778.

80. Antonarakis, S.E. beta-Thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site. / S.E. Antonarakis, S.H. Irkin, T.C. Cheng, A.F. Scott, J.P. Sexton, S.P. Trusko, S. Charache, H.H. Kazazian Jr. // Proc. Natl. Acad. Sci. USA. - 1984. - V. 81, N. 4. - P. 1154 - 1158.

81. Arkova, O.V. Obesity-related known and candidate SNP markers can significantly change affinity of TATA-binding protein for human gene promoters. / O.V. Arkova, M.P. Ponomarenko, D.A. Rasskazov, I.A. Drachkova, T.V. Arshinova, P.M. Ponomarenko, L.K. Savinkova, N.A. Kolchanov // BMC Genomics. - 2015. - V. 16, Suppl. 13. - P. S5.

82. Arnaud, E. Polymorphisms in the 5' regulatory region of the tissue factor gene and the risk of myocardial infarction and venous thromboembolism: the ECTIM and PATHROS studies. Etude Cas-Temoins de l'Infarctus du Myocarde. Paris Thrombosis case-control Study. / E. Arnaud, V. Barbalat, V. Nicaud, F. Cambien, A. Evans, C. Morrison, D. Arveiler, G. Luc, J.B Ruidavets., J. Emmerich, J.N. Fiessinger, M. Aiach // Arterioscler. Thromb. Vasc. Biol. - 2000. - V. 20, N. 3. - P. 892 - 898.

83. Auble, D.T. The dynamic personality of TATA-binding protein. / D.T. Auble // Trends Biochem. Sci. - 2009. - V. 34, N. 2. - P. 49 - 52.

84. Audit, B. From genes to genomes: universal scale-invariant properties of microbial chromosome organisation. / B. Audit, C.A. Ouzounis // J. Mol. Biol. – 2003. - V. 332, N. 3. - P. 617 - 633.

85. Axtell, M.J. Antiquity of microRNAs and their targets in land plants. / M.J. Axtell, D.P. Bartel // Plant Cell. - 2005. - V. 17, N. 6. - P. 1658 - 1673.

86. Azuma-Mukai, A. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. / A. Azuma-Mukai, H. Oguri, T. Mituyama, Z.R. Qian, K. Asai, H. Siomi, M.C. Siomi // Proc. Natl. Acad. Sci. USA. - 2008. - V. 105, N. 23. - P. 7964 - 7969.

87. Babenko, V.N. Investigating extended regulatory regions of genomic DNA sequences. / V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin, A.S. Frolov // Bioinformatics. - 1999. - V. 15, N. 7/8. - P. 644 - 653.

88. Badens, C. Characterization of a new polymorphism, IVS-I-108 (T-->C), and a new beta-thalassemia mutation, -27 (A-->T), discovered in the course of a prenatal diagnosis. / C. Badens, N. Jassim, N. Martini, J.F. Mattei, J. Elion, D. Lena-Russo // Hemoglobin. - 1999. - V. 23, N. 4. - P. 339 - 344.

89. Ballas, N. Two auxin-responsive domains interact positively to induce expression of the early indoleacetic acid-inducible gene PS-IAA4/5. / N. Ballas, L.M. Wong, M. Ke, A. Theologis // Proc. Natl. Acad. Sci. USA. – 1995. - V. 92, N. 8. - P. 3483 - 3487.

90. Banerjee, A. Structure of a repair enzyme interrogating undamaged DNA elucidates recognition of damaged DNA. / A. Banerjee, W. Yang, M. Karplus, G.L. Verdine // Nature. - 2005. - V. 434, N. 7033. - P. 612 - 618.

91. Barrick, D. Quantitative analysis of ribosome binding sites in E.coli. / D. Barrick, K. Villanueba, J. Childs, R. Kalil, T.D. Schneider, C.E. Lawrence, L. Gold, G.D. Stormo // Nucleic Acids Research. - 1994. - V. 22, N. 7. - P. 1287 - 1295.

92. Basehoar, A.D. Identification and distinct regulation of yeast TATA box-containing genes. / A.D. Basehoar, S.J. Zanton, B.F. Pugh // Cell. - 2004. - V. 116, N. 5. - P. 699 - 709.

93. Bazykin, G.A. Rate of promoter class turnover in yeast evolution. / G.A. Bazykin, A.S. Kondrashov // BMC Evol. Biol. - 2006. - V. 6. - P. E14.

94. Beckman, K.B. The free radical theory of aging matures. / K.B. Beckman, B.N. Ames // Physiol. Rev. - 1998. - V. 78. – N. 2. - P. 547 - 581.

95. Beiko, R.G. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. / R.G. Beiko, R.L. Charlebois // BMC Bioinf. - 2005. - V. 6. - P. 36.

96. Bendall, A.J. Base preferences for DNA binding by the bHLH-Zip protein USF: effects of MgCl2 on specificity, comparison with binding of Myc family members. / A.J. Bendall, P.L. Molloy // Nucleic Acids Res. - 1994. - V. 22, N. 14. - P. 2801 - 2810.

97. Benson, D.A. GenBank. / D.A. Benson, M. Boguski, D.J. Lipman, J. Ostell // Nucleic Acids Res. - 1996. - V. 24, N. 1. - P. 1 - 5.

98. Berg, O. Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity. / O. Berg // Nucleic Acids Res. – 1988. - V. 16, N. 11. - P. 5089 - 5105.

99. Berg, O.G. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. / O.G. Berg, P.H von Hippel // J. Mol. Biol. – 1987. - V. 193, N. 4. - P. 723 - 750.

100. Berikov, V.B. Regression trees for analysis of mutational spectra in nucleotide sequences. / V.B. Berikov, I.B. Rogozin // Bioinformatics. - 1999. - V. 15, N. 7/8. - P. 553 - 562.

101. Bernardi, G. The human genome and its evolutionary context. / G. Bernardi, G. Bernardi // Cold. Spring Harb. Symp. Quant. Biol. - 1986. - V. 51, Pt. 1. - P. 479 - 487.

102. Bieberstein, N.I. First exon length controls active chromatin signatures and transcription. / N.I. Bieberstein, F.C. Oesterreich, K. Straube, K.M. Neugebauer // Cell Rep. – 2012. - V. 2, N. 1. - P. 62 - 68.

103. Blackwell, T.K. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. / T.K. Blackwell, H. Weintraub // Science. - 1990. - V. 250, N. 4984. - P. 1104 - 1110.

104. Blaisdell, B. Distinctive charge configurations in proteins of the Epstein-Barr virus and possible functions. / B. Blaisdell, S. Karlin // Proc. Natl. Acad. Sci. USA. – 1988. - V. 85, N. 18. - P. 6637 - 6641.

105. Boldt, A.B. Diversity of the MBL2 gene in various Brazilian populations and the case of selection at the mannose-binding lectin locus. / A.B. Boldt, L. Culpi, L.T. Tsuneto, I.R de Souza., J.F. Kun, M.L. Petzl-Erler // Hum. Immunol. - 2006. - V. 67, N. 9. - P. 722 - 734.

106. Brindefalk, B. Evolutionary history of the TBP-domain superfamily. / B. Brindefalk, B.H. Dessailly, C. Yeats, C. Orengo, F. Werner, A.M. Poole // Nucleic Acids Res. - 2013. - V. 41, N. 5. - P. 2832 - 2845.

107. Brown, D.W. Unfolding of nucleosome cores dramatically changes the distribution of ultraviolet photoproducts in DNA. / D.W. Brown, L.J. Libertini, C. Suquet, E.W. Small, M.J. Smerdon // Biochemistry. - 1993. - V. 32, N. 40. - P. 10527 - 10531.

108. Bruner, S.D. Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA. / S.D. Bruner, D.P. Norman, G.L. Verdine // Nature. - 2000. - V. 403, N. 6772. - P. 859 - 866.

109. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. / P. Bucher // J. Mol. Biol. – 1990. - V. 212, N. 4. - P. 563 - 578.

110. Burks, C. The GenBank nucleic acid sequence database. / C. Burks, J.W. Fickett, W.B. Goad, M. Kanehisa, F.I. Lewitter, W.P. Rindone, C.D. Swindell, C.S. Tung, H.S. Bilofsky // Comput. Appl. Biosci. - 1985. - V. 1, N. 4. - P. 225 - 233.

111. Cai, S.P. A new TATA box mutation detected at prenatal diagnosis for beta-thalassemia. / S.P. Cai, J.Z., Zhang, M. Doherty, Y.W. Kan // Am. J. Hum. Genet. - 1989. - V. 45, N. 1. - P. 112 - 114.

112. Calladine, C.R. Mechanics of sequence-dependent stacking of bases in B-DNA. / C.R. Calladine // J. Mol. Biol. - 1982. - V. 161, N. 2. - P. 343 - 352.

113. Carreto, L. Expression variability of co-regulated genes differentiates Saccharomyces cerevisiae strains. / L. Carreto, M.F. Eiriz, I. Domingues, D. Schuller, G.R. Moura, M.A. Santos // BMC Genomics. - 2011. - V. 12. - P. 201.

114. Carrillo Oesterreich, F. Pause locally, splice globally. / F. Carrillo Oesterreich, N. Bieberstein, K.M. Neugebauer // Trends Cell Biol. - 2011. - V. 21, N. 6. - P. 328-335.

115. Cavalieri, V. Promoter activity of the sea urchin (Paracentrotus lividus) nucleosomal H3 and H2A and linker H1-histone genes is modulated by enhancer and chromatin insulator. / V. Cavalieri, R. Melfi, G. Spinelli // Nucleic Acids Res. - 2009. - V. 37, N. 22. - P. 7407 - 7415.

116. Chen, Q. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. / Q. Chen, G. Hertz, G.D. Stormo // Comput. App.l Biosci. – 1995. - V. 11, N. 5. - P. 563 - 566.

117. Cheng, C. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. / C. Cheng, M. Gerstein // Nucleic Acids Res. - 2012. - V. 40, N. 2. - P. 553 - 568.

118. Cheng, J. Scaling behavior of nucleotide cluster in DNA sequences. / J. Cheng, Z. Tong, L. Zhang // J. Zhejiang Univ. Sci. B. – 2007. - V. 8, N. 5. - P. 359 - 364.

119. Chou, P. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. / P. Chou, G. Fasman // Biochemistry. - 1974. - V. 13, N. 2. - P. 222 - 245.

120. Chuzhanova, N. Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. / N. Chuzhanova, M. Krawczak, L. Nemytikova, V. Gusev, D. Cooper // Gene. – 2000. - V. 254, N. 1 – 2. - P. 9 - 18.

121. Chuzhanova, N. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. / N.

Chuzhanova, E. Anassis, E. Ball, M. Krawczak, D.N. Cooper // Hum. Mutat. - 2003a. - V. 21, N. 1. - P. 28 - 44.

122. Chuzhanova, N. Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. / N. Chuzhanova, S. Abeysinghe, M. Krawczak, D.N. Cooper // Hum. Mutat. - 2003b. - V. 22, N. 3. - P. 245 - 251.

123. Clark, I.A. Genes, nitric oxide and malaria in African children. / I.A. Clark, K.A. Rockett, D. Burgner // Trends Parasitol. - 2003. - V. 19, N. 8. - P. 335 - 337.

124. Cohen, B.I. Pattern-based approaches to protein structure prediction. B.I. / Cohen, S.R. Presnell, F.E. Cohen // Methods Enzimol. - 1991. - V. 202. - P. 252-278.

125. Coleman, R.A. Evidence for functional binding and stable slid-ing of the TATA binding protein on nonspecific DNA. / R.A. Coleman, B.F. Pugh // J. Biol. Chem. - 1995. - V. 270, N. 23. - P. 13850 - 13859.

126. Colonna, V. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. / V. Colonna, Q. Ayub, Y. Chen, L. Pagani, P. Luisi, M. Pybus, E. Garrison, Y. Xue, C. Tyler-Smith, 1000 Genomes Project Consortium, G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean // Genome Biol. – 2014. - V. 15, N. 6. – P. R88.

127. Comings, D. Exon and intron variants in the human tryptophan 2,3-dioxygenase gene: potential association with Tourette syndrome, substance abuse and other disorders. / D. Comings, R. Gade, D. Muhleman, C. Chiu, S. Wu, M. To, M. Spence, G. Dietz, E. Winn-Deen, R. Rosenthal, H., Lesieur L. Rugle, J. Sverd, L. Ferry, J. Johnson, J. MacMurray // Pharmacogenetics. - 1996. - V. 6, N. 4. - P. 307 - 318.

128. Conkright, M.D. Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. / M.D. Conkright, E. Guzman, L.

Flechner, A.I. Su, J.B. Hogenesch, M. Montminy // Mol. Cell. - 2003. - V. 11, N. 4. - P. 1101 - 1108.

129. Contreras-Levicoy, J. Schizosaccharomyces pombe positive cofactor 4 stimulates basal transcription from TATA-containing and TATA-less promoters through Mediator and transcription factor IIA. / J. Contreras-Levicoy, F. Urbina, E. Maldonado // FEBS J. - 2008. - V. 275, N. 11. - P. 2873 - 2883.

130. Costa, M. Multiscale entropy analysis of biological signals. / M. Costa, A. Goldberger, C. Peng // Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys. - 2005. - V. 71, N. 2. - Pt. 1. - P. 021906.

131. Coulondre, C. Molecular basis of base substitution hotspots in Escherichia coli. / C. Coulondre, J.H. Miller, F.J. Farabaugh, W.Gilbert // Nature. - 1978. - V. 274, N. 5673. - P. 775 - 780.

132. Cox, M.M. Relating biochemistry to biology: how the recombinational repair function of RecA protein is manifested in its molecular properties. / M.M. Cox // BioEssay. - 1993. - V. 15, N. 9. - P. 617 - 623.

133. Davison, B.L. Formation of stable preinitiation complexes between eukaryotic class B transcription factors and promoter sequences. / B.L. Davison, J.M. Egly, E.R. Mulvihill, P. Chambon // Nature. - 1983. - V. 301, N. 5902. - P. 680 - 686.

134. Dayhoff, M.O. Nucleic acid sequence database. / M.O. Dayhoff, R.M. Schwartz, H.R. Chen, W.C. Barker, L.T. Hunt, B.C. Orcutt // DNA. - 1981. - V. 1, N. 1. - P. 51 - 58.

135. Dayhoff, M.O. Establishing homologies in protein sequences. / M.O. Dayhoff, W. Barker, L. Hunt // Methods Enzymol. – 1983. - V. 91. - P. 524-545.

136. de Gobbi, M. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. / M. de Gobbi, V. Viprakasit, J.R. Hughes, C. Fisher, V.J. Buckle, H. Ayyub, R.J. Gibbons, D. Vernimmen, Y. Yoshinaga, P. de Jong, J.F. Cheng, E.M. Rubin, W.G. Wood, D. Bowden, D.R. Higgs // Science. - 2006. - V. 312, N. 5777. - P. 1215 - 1217.

137. Delgadillo, R.F. The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. / R.F. Delgadillo, J.E. Whittington, L.K. Parkhurst, L.J. Parkhurst // Biochemistry. - 2009. - V. 48, N. 8. - P. 1801 – 1809.

138. Deplancke, B. The genetics of transcription factor DNA binding variation. // B. Deplancke, D. Alpern, V. Gardeux // Cell. - 2016. - V. 166, N. 3. - P. 538-554.

139. Deyneko, I.V. FeatureScan: revealing property-dependent similarity of nucleotide sequences. / I.V. Deyneko, B. Bredohl, D. Wesely, Y.M. Kalybaeva, A.E. Kel, H. Bloker, G. Kauer // Nucleic Acids Res. - 2006. - V. 34, Web-server issue. - P. W591 - W595.

140. Dickerson, R.E. Definitions and nomenclature of nucleic acid structure components. / R.E. Dickerson // Nucleic Acids Res. - 1989. – V. 17, N. 5. – P. 1797 - 1803.

141. Dickerson, R.E. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure. / R.E. Dickerson, H.R. Drew // J. Mol. Biol. - 1981. - V. 149, N. 4. - P. 761 - 786.

142. Dong, D. Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. / D. Dong, Z. Yuan, Z. Zhang // Nucleic Acids Res. - 2011. - V. 39, N. 3. - P. 837 - 847.

143. Drachkova, I.A. In vitro examining the existing prognoses how TBP binds to TATA with SNP associated with human diseases. / I.A. Drachkova, P.M. Ponomarenko, T.V. Arshinova, M.P. Ponomarenko, V.V. Suslov, L.K. Savinkova, N.A. Kolchanov // Health. - 2011. - V. 3, N. 9. - P. 577 - 583.

144. Drachkova, I. The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the TATA-binding protein. / I. Drachkova, L. Savinkova, T. Arshinova, M. Ponomarenko, S. Peltek, N.A. Kolchanov // Hum. Mutat. - 2014. - V. 35, N. 5. – P. 601 – 608.

145. Dulbecco, R. A turning point in cancer research: sequencing the human genome. / R. Dulbecco // Science. - 1986. - V. 231, N. 4742. - P. 1055 - 1056.

146. Dutta, S. Data deposition and annotation at the worldwide protein data bank. / S. Dutta, K. Burkhardt, J. Young, G.J. Swaminathan, T. Matsuura, K. Henrick, H. Nakamura, H.M. Berman // Mol. Biotechnol. - 2009. - V. 42, N. 1. - P. 1 - 13.

147. Eckdahl, T. Conserved DNA structures in origins of replication. / T. Eckdahl, J. Anderson // Nucleic Acids Res. – 1990. - V. 18, N. 6. - P. 1609 - 1612.

148. Efron, B. Bootstrap confidence levels for phylogenetic trees. / B. Efron, E. Halloran, S. Holmes // Proc. Natl. Acad. Sci. USA. – 1996. - V. 93, N. 23. - P. 13429 - 13434.

149. Ellington, A.D. In vitro selection of RNA molecules that bind specific ligands. / A.D. Ellington, J.W. Szostak // Nature. - 1990. - V. 346, N. 6287. - P. 818 - 822.

150. Engelhorn, M. In vivo interaction of the Escherichia coli integration host factor with its specific binding sites. / M. Engelhorn, F. Boccard, C. Murtin, P. Prentki, J. Geiselmann // Nucleic Acids Res. - 1995. - V. 23, N. 17. - P. 2959 - 2965.

151. Farber, R. Determination of eukaryotic protein coding regions using neural networks and information theory. / R. Farber, A. Lapedes, K. Sirotkin // J. Mol. Biol. – 1992. - V. 226, N. 2. - P. 471 - 4769.

152. Fei, Y.J. Beta-thalassemia due to a T----A mutation within the ATA box. / Y.J. Fei, T.A. Stoming, G.D. Efremov, D.G. Efremov, R. Battacharia, J.M. Gonzalez-Redondo, C. Altay, A. Gurgey, T.H. Huisman // Biochem. Biophys. Res. Commun. - 1988. - V. 153, N. 2. - P. 741 - 747.

153. Felsenfeld, G. Chromatin as essential part of the transcription mechanism. / G. Felsenfeld // Nature. - 1992. - V. 355, N. 6357. - P. 219 - 224.

154. Fernandez-Suarez, X.M. The 2014 Nucleic Acids Research Database Issue and an updated NAR online molecular biology database collection. / X.M. Fernandez-Suarez, D.J. Rigden, M.Y. Galperin // Nucleic Acids Res. - 2014. - V. 42, Database issue. - P. D1 - D6.

155. Ferre-D'Amare, A.R. Structure, function of the b/HLH/Z domain of USF. / A.R. Ferre-D'Amare, P. Pognonec, R.G. Roeder, S.K. Burley // EMBO J. - 1994. - V. 13, N. 1. - P. 180 - 189.

156. Fire, A. Interactions between RNA polymerase II, factors, and template leading to accurate transcription. / A. Fire, M. Samuels, P.A. Sharp // J. Biol. Chem. - 1984. - V. 259, N. 4. - P. 2509 - 2516.

157. Fishburn, P.C. Utility theory for decision making. / P.C. Fishburn - NY: John Wiley & Sons. - 1970.

158. Fischer E. Syntheses in the purine and sugar group (Nobel Lecture, December 12, 1902) In: Nobel Lectures. Chemistry 1901-1921 / E. Fischer - Amsterdam: Elsevier Publ. Co. - 1966. - P. 21 - 35.

159. Flatters, D. Sequence-dependent dynamics of TATA-Box binding sites. / D. Flatters, R. Lavery // Biophys. J. - 1998. - V. 75, N. 1. - P. 372 - 381.

160. Flicek, P. Ensembl 2011. / P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H.S. Riat, D. Rios, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y.A. Tang, S. Trevanion, J. Vandrovcova, A.J. Vilella, S. White, S.P. Wilder, A. Zadissa, J. Zamora, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernandez-Suarez, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, J. Vogel, S.M. Searle // Nucleic Acids Res. - 2011. - V. 39, Database issue. - P. D800 - D806.

161. Florquin, K. Large-scale structural analysis of the core promoter in mammalian and plant genomes. / K. Florquin, Y. Saeys, S. Degroeve, P. Rouze, Y van de Peer. // Nucleic Acids Res. - 2005. - V. 33, N. 13. - P. 4255 - 4264.

162. Franklin, R.E. Molecular configuration in sodium thymonucleate. / R.E. Franklin, R.G. Gosling // Nature. – 1953. - V. 171, N. 4356. - P. 740 - 741.

163. Friedel, M. DiProDB: a database for dinucleotide properties. / M. Friedel, S. Nikolajewa, J. Suhnel, T. Wilhelm // Nucleic Acids Res. – 2009. - V. 37, Database issue. - P. D37 - D40.

164. Frischknecht, H. Two new delta-globin mutations: Hb A2-Ninive delta133 (H11) Val-Ala] and a delta(+)-thalassemia mutation [-31 (A --> G)] in the TATA

box of the delta-globin gene. / H. Frischknecht, F. Dutly // Hemoglobin. - 2005. - V. 29, N. 2. - P. 151 - 154.

165. Fu, W. DISCOVER: a feature-based discriminative method for motif search in complex genomes. / W. Fu, P. Ray, E.P. Xing // Bioinformatics. - 2009. - V. 25, N. 12. - P. i321 - i329.

166. Furman, D.P. Promoters of the genes encoding the transcription factors regulating the cytokine gene expression in macrophages contain putative binding sites for aryl hydrocarbon receptor. / D.P. Furman, E.A. Oshchepkova, D.Y. Oshchepkov, M.Y. Shamanina, V.A. Mordvinov // Comput. Biol. Chem. - 2009. - V. 33, N. 6. - P. 465 - 468.

167. Gabrielian, A. Sequence complexity and DNA curvature. / A. Gabrielian, A. Bolshoy // Comput. Chem. – 1999. - V. 23, N. 3–4. - P. 263 - 274.

168. Galas, D.J. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. / D.J. Galas, A. Schmitz // Nucleic Acids Res. - 1978. - V. 5, N. 9. - P. 3157 - 3170.

169. Gamow, G. The problem of information transfer from the nucleic acids to proteins. / G. Gamow, A. Rich, M. Ycas // Adv. Biol. Med. Phys. - 1956. - V. 4, P. 23 - 68.

170. Gamow, G. Statistical correlation of protein and ribonucleic acid composition. / G. Gamow, M. Ycas // Proc. Natl. Acad. Sci. U.S.A. - 1955. - V. 41, N. 12. - P. 1011 - 1019.

171. Gartenberg, M.R. DNA sequence determinants of CAP-induced bending, protein binding affinity. / M.R. Gartenberg, D.M. Crothers // Nature. - 1988. - V. 333, N. 6176. - P. 824 - 829.

172. Gelfand, M.S. Prediction of function in DNA sequence analysis. / M.S. Gelfand // J. Comp. Biol. - 1995. - V. 2, N. 1. - P.87 - 115.

173. GenBank: distribution release 207.0 notes. April 15, 2015. – Bethesda: National Center for Biotechnology Information. National Library of Medicine. USA. – 2015.

174. Gerstenblith, M.R. Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis. / M.R. Gerstenblith, J. Shi, M.T. Landi // Pigment Cell Melanoma Res. - 2010. - V. 23, N. 5. - P. 587 - 606.

175. Gilbert, W. The nucleotide sequence of the lac operator. / W. Gilbert, A. Maxam // Proc. Natl. Acad. Sci. USA. - 1973. - V. 70, N. 12. - P. 3581 - 3584.

176. Godde, J.S. The amino-terminal tails of the core histones and the translational position of the TATA box determine TBP/TFIIA association with nucleosomal DNA. / J.S. Godde, Y. Nakatani, A.P. Wolffe // Nucleic Acids Res. - 1995. - V. 23, N. 22. - P. 4557 - 4564.

177. Gold, L. From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. / L. Gold, D. Brown, Y. He, T. Shtatland, B. Singer, Y. Wu // Proc. Natl. Acad. Sci. U.S.A. - 1997. - V. 94, N. 1. - P. 59 - 64.

178. Goni, J.R. Determining promoter location based on DNA structure first-principles calculations. / J.R. Goni, A. Perez, D. Torrents, M. Orozco // Genome Biol. - 2007. - V. 8, N. 12. - P. R263.

179. Gorin, A.A. B-DNA twisting correlates with base-pair morphology. / A.A. Gorin, V.B. Zhurkin, W.K. Olson // J. Mol. Biol. – 1995. - V. 247, N. 1. - P. 34 - 48.

180. Gotoh, O. Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. / O. Gotoh, Y. Tagashira // Biopolymers. - 1981. - V. 20, N. 5. - P. 1033 - 1042.

181. Grennan, A. Genevestigator. Facilitating web-based gene-expression analysis. / A. Grennan // Plant Physiol. – 2006. - V. 141, N. 4. - P. 1164 - 1166.

182. Griffith, J.D. Visualization of prokaryotic DNA in a regularly condensed chromatin-like fiber. / J.D. Griffith // Proc. Natl. Acad. Sci. USA. - 1976. - V. 73, N. 2. - P. 563-567.

183. Griffiths-Jones, S. miRBase: microRNA sequences, targets and gene nomenclature. / S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, A.J. Enright // Nucleic Acids Res. – 2006. - V. 34, Database issue. - P. D140 - D144.

184. Grokhovsky, S.L. Sequence-specific ultrasonic cleavage of DNA. / S.L. Grokhovsky, I.A. Il’icheva, D.Yu. Nechipurenko, M.V. Golovkin, L.A. Panchenko,

R.V. Polozov, Yu.D. Nechipurenko // Biophys. J. – 2011. - V. 100, N. 1. - P. 117 - 125.

185. Gunbin, K.V. Evolution of brain active gene promoters in human lineage towards the increased plasticity of gene regulation./ K.V. Gunbin, M.P. Ponomarenko, V.V. Suslov, F. Gusev, G.G. Fedonin, E.I. Rogaev // Mol. Neurobiol. - 2017. - DOI: 10.1007/s12035-017-0427-4.

186. Gunewardena, S. Enhancing the prediction of transcription factor binding sites by incorporating structural properties and nucleotide covariations. / S. Gunewardena, P. Jeavons, Z. Zhang // J. Comput. Biol. - 2006. - V. 13, N. 4. - P. 929 - 945.

187. Gurzadyan, G.G. Photolesions and biological inactivation of plasmid DNA on 254 nm irradiation and comparison with 193 nm laser irradiation. / G.G. Gurzadyan, H. Gorner, D. Schulte-Frohlinde // Photochem. Photobiol. - 1993. - V. 58, N. 4. - P. 477 - 485.

188. Hahn, S. Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. / S. Hahn, S. Buratowski, P.A. Sharp, L. Guarente // Proc. Natl. Acad. Sci. USA. - 1989. - V. 86, N. 15. - P. 5718 – 5722.

189. Haldane, J.B.S. The cost of natural selection. / J.B.S. Haldane // J. Genet. – 1957 – V. 55. – P. 511 – 524.

190. Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. / A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick // Nucleic Acids Res. - 2005. - V. 33, Database issue. - P. D514 - D517.

191. Hampson, S. Distribution patterns of over-represented k-mers in non-coding yeast DNA. / S. Hampson, D. Kibler, P. Baldi // Bioinformatics. – 2002. - V. 18, N. 4. - P. 513 - 528.

192. Hardenbol, P. Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA. / P. Hardenbol, J.C. Wang, M.W. van Dyke // Nucleic Acids Res. - 1997. - V. 25, N. 16 - P. 3339 - 3344.

193. Harrow, J. GENCODE: the reference human genome annotation for The ENCODE Project. / J. Harrow, A. Frankish, J.M. Gonzalez, T E. apanari, M. Diekhans, F. Kokocinski, B.L. Aken, l D. Barrel, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, e G. Mukherje, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J.M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, T.J. Hubbard // Genome Res. - 2012. - V. 22, N. 9. - P. 1760 - 1774.

194. Hawley, D. Compilation and analysis of Escherichia coli promoter DNA sequences. / D. Hawley, W. McClure // Nucleic Acids Res. – 1983. - V. 11, N. 8. - P. 2237 - 2255.

195. He, Y. Near-atomic resolution visualization of human transcription promoter opening. / Y. He, C. Yan, J. Fang, C. Inouye, R. Tjian, I. Ivanov, E. Nogales // Nature. - 2016. - V. 533, N. 7603. - P. 359-365.

196. Hein, M. Tumor cell response to bevacizumab single agent therapy in vitro. / M. Hein, S. Graver // Cancer Cell Int. - 2013. – V. 13, N. 1. – P. 94.

197. Heinemeyer, T. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. / T. Heinemeyer, X. Chen, H. Karas, A.E. Kel, O.V. Kel, I. Liebich, T. Meinhardt, I. Reuter, F. Schacherer, E. Wingender // Nucleic Acids Res. - 1999. - V. 27, N. 1. - P. 318 - 322.

198. Higgins, D.G. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. / D.G. Higgins, P.M. Sharp // Gene. - 1988. - V. 73, N. 1. - P. 237 - 244.

199. Hindorff, L.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. / L.A. Hindorff, P., Sethupathy H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio // Proc. Natl. Acad. Sci. USA. - 2009. - V. 106, N. 23. - P. 9362 - 9367.

200. Hofmann, H. Comparative analysis of the sequences of the three collagen chains alpha 1(I), alpha 2 and alpha 1(III) Functional and genetic aspects. / H. Hofmann, P. Fietzek, K. Kuhn // J. Mol. Biol. – 1980. - V. 141, N. 3. - P. 293 - 314.

201. Hogan, M.E. Importance of DNA stiffness in protein-DNA binding specificity. / M.E. Hogan, R.H. Austin // Nature. – 1987. - V. 329, N. 6136. - P. 263 - 266.

202. Hornung, G. Noise-mean relationship in mutated promoters. / G. Hornung, R. Bar-Ziv, D. Rosin, N. Tokuriki, D.S. Tawfik, M. Oren, N. Barkai // Genome Res. - 2012. - V. 22, N. 12. - P. 2409 - 2417.

203. Houbaviy, H.B. Cocrystal structure of YY1 bound to the adeno-associated Virus P5 Initiator. / H.B. Houbaviy, A. Usheva, T. Shenk, S.K. Burley // Proc. Nat. Acad. Sci. USA. – 1996. - V. 93, N. 24. - P. 13577 - 13582.

204. Hsu, J.Y. TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription. / J.Y. Hsu, T. Juven-Gershon, M.T. Marr 2nd, K.J. Wright, R. Tjian, J.T. Kadonaga // Genes Dev. - 2008. - V. 22, N. 17. - P. 2353 - 2358.

205. Hu, J. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. / J. Hu, J.W. Locasale, J.H. Bielas, J. O'Sullivan, K. Sheahan, L.C. Cantley, M.G. Vander Heiden, D. Vitkup // Nature Biotechnol. - 2013. – V. 31, N. 6. – P. 522 - 529.

206. Hu, T. Isolation and characterization of a rice glutathione S-transferase gene promoter regulated by herbicides and hormones. / T. Hu, S. He, G. Yang, H. Zeng, G. Wang, Z. Chen, X. Huang // Plant Cell Rep. – 2011. - V. 30, N. 4. - P. 539 - 549.

207. Huang, T. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. / T. Huang, B. Fan, M. Rothschild, Z.L. Hu, K. Li, S.H. Zhao // BMC Bioinformatics. – 2007. - V. 8. - P. 341.

208. Hyde-DeRuyscher, R. DNA binding sites for the transcriptional activator/repressor YY1. / R. Hyde-DeRuyscher, E. Jennings, T. Shenk // Nucleic Acids Res. – 1995. - V. 23, N. 21. - P. 4457 - 4465.

209. Imbalzano, A.N. Facilitated binding of TATA-binding protein to nucleosomal DNA. / A.N. Imbalzano, H. Kwon, M.R. Green, R.E. Kingston // Nature. - 1994. - V. 370, N. 6489. - P. 481 - 485.

210. Ioshikhes, I. Nucleosomal DNA sequence database. / I. Ioshikhes, E.N. Trifonov // Nucleic Acids Res. - 1993. - V. 21, N. 21. - P. 4857 - 4859.

211. Ioshikhes, I. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. / I. Ioshikhes, E.N. Trifonov, M.Q. Zhang // Proc. Nat.l Acad. Sci. USA. - 1999. - V. 96, N. 6. - P. 2891 - 2895.

212. Isogai, Y. Transcription of histone gene cluster by differential core-promoter factors. / Y. Isogai, S. Keles, M. Prestel, A. Hochheimer, R. Tjian // Genes Dev. - 2007. - V. 21, N. 22. - P. 2936 - 2949.

213. IUPAC-IUB commission on biochemical nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. // J. Mol. Biol. - 1971. - V. 55, N. 3. - P. 299 - 310.

214. Ivanisenko, V.A. PDBSite: a database of the 3D structure of protein functional sites. / V.A. Ivanisenko, S.S. Pintus, D.A. Grigorovich, N.A. Kolchanov // Nucleic Acids Res. – 2005. - V. 33, Database issue. - P. D183 - D187.

215. Javahery, R. DNA sequence requirements for transcriptional initiator activity in mammalian cells. / R. Javahery, A. Khachi, K. Lo, B. Zie-Gregory, S.T. Smale // Mol. Cell. Biol. – 1994. - V. 14, N. 1. - P. 116 - 127.

216. Jimenez-Montano, M. Entropy and complexity of finite sequences as fluctuating quantities. / M. Jimenez-Montano, W. Ebeling, T. Pohl, P. Rapp // Biosystems. – 2002. - V. 64, N. 1 – 3. - P. 23 - 32.

217. Johnson, P. Eukaryotic transcriptional regulatory proteins. / P. Johnson, S. McKnight // Annu. Rev. Biochem. – 1989. - V. 58. - P. 799 - 839.

218. Johnson, M. NCBI BLAST: a better web interface. / M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T.L. Madden // Nucleic Acids Res. - 2008. - V. 36, Web-server issue. - P. W5 - W9.

219. Jonsson, J. Quantitative sequence-activity models (QSAM) - tools for sequence design. / J. Jonsson, T. Norberg, L. Carlsson, C. Gustafsson, S. Wold // Nucleic Acids Res. - 1993. - V. 21, N. 3. - P. 733 - 739.

220. Juo, Z.S. How proteins recognize the TATA box. / Z.S. Juo, T.K. Chiu, P.M. Leiberman, I. Baikalov, A.J. Berk, R.E. Dickerson // J. Mol. Biol. - 1996. - V. 261, N. 2. - P. 239 - 254.

221. Jurka, J. Repbase Update, a database of eukaryotic repetitive elements. / J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz // Cytogenet. Genome Res. – 2005. - V. 110, N. 1 – 4. - P. 462 - 467

222. Juven-Gershon, T. The RNA polymerase II core promoter - the gateway to transcription. / T. Juven-Gershon, J.Y. Hsu, J.W. Theisen, J.T. Kadonaga // Curr. Opin. Cell Biol. - 2008. - V. 20, N. 3. - P. 253 - 259.

223. Kabsch, W. How good are predictions of protein secondary structure? / W. Kabsch, C. Sander // FEBS Lett. - 1983. - V. 155, N. 2. - P. 179 - 182.

224. Kabsch, W. The ten helical twist angles of B-DNA. / W. Kabsch, C. Sander, E.N. Trifonov // Nucleic Acids Res. – 1982. - V. 10, N. 3. - P. 1097 - 1104.

225. Kanehisa, M.I. Los Alamos sequence analysis package for nucleic acids and proteins. / M.I. Kanehisa // Nucleic Acids Res. – 1982. - V. 10, N. 1. - P. 183 - 196.

226. Kanehisa, M. A relational database system for the maintenance and verification of the Los Alamos sequence library. / M. Kanehisa, J.W. Fickett, W.B. Goad // Nucleic Acids Res. - 1984. - V. 12, N. 1. - P. 149 - 158.

227. Kanehisa, M. KEGG for linking genomes to life and the environment. / M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, Y. Yamanishi // Nucleic Acids Res. – 2008. - V. 36, Database issue. - P. D480 - D484.

228. Kapitonov, V.V. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. / V.V. Kapitonov, J. Jurka // PLoS Biol. – 2005. - V. 3, N. 6. - P. e181.

229. Karas, H. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. / H. Karas, R. Knuppel, W. Schulz,

H. Sklenar, E. Wingender // Comput. Applic. Biosci. - 1996. - V. 12, N. 5. - P. 441 - 446.

230. Karlin, S. New approaches for computer analysis of nucleic acid sequences. / S. Karlin, G. Ghandour, F. Ost, S. Tavare, L.J. Korn // Proc. Natl. Acad. Sci. U.S.A. – 1983. - V. 80, N. 18. - P. 5660 - 5664.

231. Karlin, S. Patterns in DNA and amino-acid sequences and their statistical significance. In: “Mathematical methods for DNA sequences.” / S. Karlin, F. Ost, B.T. Blaisdell; Ed. M.S. Waterman - Boca Raton: CRC Press. – 1989. - P. 133 - 158.

232. Karlin, S. Dinucleotide relative abundance extremes: a genomic signature. / S. Karlin, C. Burge // Trends Genet. – 1995. - V. 11, N. 7. - P. 283 - 290.

233. Karlin, S. Global dinucleotide signatures and analysis of genomic heterogeneity. / S. Karlin // Curr. Opin. Microbiol. – 1998. - V. 1, N. 5. - P. 598 - 610.

234. Karolchik, D. The UCSC Genome Browser database: 2014 update. / D. Karolchik, G.P. Barber, J. Casper, H. Clawson, M.S. Cline, M. Diekhans T.R. , Dreszer, P.A. Fujita, L. Guruvadoo, M. Haeussler, R.A. Harte, S. Heitner, A.S. Hinrichs, K. Learned, B.T. Lee, C.H. Li, B.J. Raney, B. Rhead, K.R. Rosenbloom, C.A. Sloan, M.L. Speir, A.S. Zweig, D. Haussler, R.M. Kuhn, W.J. Kent // Nucleic Acids Res. – 2014. - V. 42, Database issue. - D764 - D770.

235. Kim, J.G. Kinetic studies on Cro repressor-operator DNA interaction. / J.G. Kim, Y. Takeda, B.W. Matthews, W.F. Anderson // J. Mol. Biol. - 1987. - V. 196, N. 1. - P. 149 - 158.

236. Kim, J.L. Co-crystal structure of TBP recognizing the minor groove of a TATA element. / J.L. Kim, D.B. Nikolov, S.K. Burley // Nature. - 1993. - V. 365, N. 6446. - P. 520 – 527.

237. Kim, Y. Crystal structure of a yeast TBP/TATA-box complex. / Y. Kim, J.H. Gieger, S. Hahn, P.B. Sigler // Nature. 1993. V. 365, N. 6446. - P. 512 – 520.

238. Kinzler, K.W. Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. / K.W. Kinzler, B. Vogelstein // Nucleic Acids Res. - 1989. - V. 17, N. 10. - P. 3645 - 3653.

239. Kirpota, O.O. Thermodynamic and kinetic basis for recognition and repair of 8-oxoguanine in DNA by human 8-oxoguanine-DNA glycosylase. / O.O. Kirpota, A.V. Endutkin, M.P. Ponomarenko, P.M. Ponomarenko, D.O. Zharkov, G.A. Nevinsky // Nucleic Acids Res. - 2011. - V. 39, N. 11. - P. 4836 - 4850.

240. Kissinger, C.R. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 angstroms resolution: a framework for understanding homeodomain-DNA interactions. / C.R. Kissinger, B. Liu, E. Martin-Blanco, T.B. Kornberg, C.O. Pabo // Cell. - 1990. - V. 63, N. 3. - P. 579 - 590.

241. Klug, A. A hypothesis on a specific sequence-dependent conformation of DNA, its relation to the binding of the lac-repressor protein. / A. Klug, A. Jack, M.A. Viswamitra, O. Kennard, Z. Shakked, T.A. Steitz // J. Mol. Biol. – 1979. - V. 131, N. 4. - P. 669 - 680.

242. Kochetov, A.V. Prediction of eukaryotic mRNA traslational properties. / A.V. Kochetov, M.P. Ponomarenko, A.S. Frolov, L.L. Kisselev, N.A. Kolchanov // Bioinformatics. – 1999. - V. 15, N. 7/8. - P. 704 - 712.

243. Kolchanov, N.A. GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences in eukaryotic genome. In: “Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology; ISMB’1998, Montreal, Quebec, Canada, June 28th - July 1st, 1998” /N.A. Kolchanov, M.P. Ponomarenko, A.E. Kel, Yu.V. Kondrakhin, A.S. Frolov, F.A. Kolpakov, T.N. Goryachkovsky, O.V. Kel, E.A. Ananko, E.V. Ignatieva, O.A. Podkolodnaya, V.N. Babenko, I.L. Stepanenko, A.G. Romashchenko, T.I. Merkulova, D.G. Vorobiev, S.V. Lavryushev, Yu.V. Ponomarenko, A.V. Kochetov, G.B. Kolesov, V.V. Solovyev, L. Milanesi, N.L. Podkolodny, E. Wingender, T. Heinemeyer; Eds J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, C. Sensen - Palo Alto: AAAI Press. - 1998. – V. 6. – P. 95 - 104.

244. Kolchanov, N.A. Integrated databases and computer systems for studying eukaryotic gene expression. / N.A. Kolchanov, M.P. Ponomarenko, A.S. Frolov, E.A. Ananko, F.A. Kolpakov, E.V. Ignatieva, O.A. Podkolodnaya, T.N. Goryachkovskaya, I.L. Stepanenko, T.I. Merkulova, V.N. Babenko, J.V. Ponomarenko, A.V. Kochetov, N.L. Podkolodny, D.G. Vorobyev, S.V. Lavrushev, D.A. Grigorovich, Yu.V. Kondrakhin, L. Milanesi, E. Wingender, V.V. Solovyev, G.C. Overton // Bioinformatics. – 1999. - V. 15, N. 7/8. - P. 669 - 686.

245. Kolchanov, N.A. Transcription Regulatory Regions Database (TRRD): its status in 2002. / N.A. Kolchanov, E.V. Ignatieva, E.A. Ananko., O.A. Podkolodnaya, I.L. Stepanenko, T.I. Merkulova, M.A. Pozdnyakov, N.L. Podkolodny, A.N. Naumochkin, A.G. Romashchenko // Nucleic Acids Res. – 2002. - V. 30, N. 1. - P. 312 - 317.

246. Kolchanov, N.A. Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. / Kolchanov N.A., Merkulova T.I., Ignatieva E.V., Ananko E.A., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Merkulov V.M., Hodgman C.T. // Brief Bioinform. - 007. - V. 8, N. 4. - P. 266 - 274.

247. Kolmer, J. A contribution to the bacteriology of acute anterior poliomyelitis. / J. Kolmer, C. Brown, A. Freese // J. Exp. Med. - 1917. - V. 25, N. 6. - P. 789 - 806.

248. Konopka, A. Distance analysis helps to establish characteristic motifs in intron sequences. / A. Konopka, G. Smythers, J. Owens, J.V. Jr. Maizel // Gene Anal. Tech. – 1987. - V. 4, N. 4. - P. 63-74.

249. Korogodin, V.I. On the dependence of spontaneous mutation rates on the functional state of genes. / V.I. Korogodin, V.L. Korogodina, C. Fajszi, A.I. Chepurnoy, N. Mikhova-Tsenova, N.V. Simonyan // Yeast. - 1991. - V. 7, N. 2. - P. 105 - 117.

250. Kozak, M. Context effects and inefficient initiation at non-AUG codons in eukaryotic cell-free translation systems. / M. Kozak // Mol. Cell. Biol. – 1989. - V. 9, N. 11. - P. 5073 - 5080.

251. Kutach, A.K. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. / A.K. Kutach, J.T. Kadonaga // Mol. Cell. Biol. - 2000. - V. 20, N. 13. - P. 4754 – 4764

252. Kuznetsov, N.A. Kinetic conformational analysis of human 8-oxoguanine-DNA glycosylase. / N.A. Kuznetsov, V.V. Koval, G.A. Nevinsky, K.T. Douglas, D.O. Zharkov, O.S. Fedorova // J. Biol. Chem. - 2007. - V. 282, N. 2. - P. 1029 - 1038.

253. Kuznetsov, V.A. Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in chip-pet experiments. / V.A. Kuznetsov, Y.L. Orlov, C.L. Wei, Y. Ruan // Genome Inform. - 2007. - V. 19. - P. 83 - 94.

254. Lagrange, T. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. / T. Lagrange, A.N. Kapanidis, H. Tang, D. Reinberg, R.H. Ebright // Genes Dev. - 1998. - V. 12, N. 1. - P. 34-44.

255. Larsabal, E. Genomes are covered with ubiquitous 11 bp periodic patterns, the "class A flexible patterns". / E. Larsabal, A. Danchin // BMC Bioinformatics. – 2005. - V. 6. - P. 206.

256. Latchman, D.S. Eukaryotic transcription factors. / D.S. Latchman; 2nd Edn. - San Diego: Academic Press. - 1995.

257. Lavery, R. The molecular electrostatic potential and steric accessibility of poly (dA-dT). poly (dA-dT) in various conformations: B-DNA, D-DNA and 'alternating-B' DNA. / R. Lavery, B. Pullman, S. Corbin // Nucleic Acids Res. - 1981. - V. 9, N. 23. - P. 6539 - 6552.

258. Lawrence, C.E. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. / C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton // Science. - 1993. - V. 262, N. 5131. - P. 208 - 214.

259. Lee, T.I. Transcription of eukaryotic protein-coding genes. / T.I. Lee, R.A. Young // Annu. Rev. Genet. 2000. - V. 34. - P. 77 – 137.

260. Lempel, A. On the complexity of finite sequences. / A. Lempel, J. Ziv // IEEE Trans. Inf. Theory. – 1976. - V. 22, N. 1. - P. 75–81.

261. Levitsky, V.G. Nucleosomal DNA property database. / V.G. Levitsky, M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, N.A. Kolchanov // Bioinformatics. – 1999. - V. 15, N. 7/8. - P. 582 - 592.

262. Levitsky, V.G. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. / V.G. Levitsky, E.V. Ignatieva, E.A. Ananko, I.I. Turnaev, T.I. Merkulova, N.A. Kolchanov, T.C. Hodgman // BMC Bioinf. - 2007. - V. 8. - P. 481.

263. Levitsky, V.G. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. / V.G. Levitsky, I.V. Kulakovskiy, N.I. Ershov, D.Y. Oschepkov, V.J. Makeev, T.C. Hodgman, T.I. Merkulova // BMC Genomics. - 2014. - V. 15, N. 1. - P. 80.

264. Li, S. Escherichia coli strains lacking protein HU are UV sensitive due to a role for HU in homologous recombination. / S. Li, R. Waters // J. Bacteriol. - 1998. - V. 180, N. 15. - P. 3750 - 3756.

265. Lifton, R. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. / R. Lifton, M. Goldberg, R. Karp, D. Hogness // Cold Spring Harb. Symp. Quant. Biol. - 1978. - V. 42, Pt. 2. - P. 1047 – 1051.

266. Likhoshvai, V.A. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy. / V.A. Likhoshvai, Y.G. Matushkin // FEBS Lett. – 2002. - V. 516, N. 1 – 3. - P. 87 - 92.

267. Lin, Y. Optimization of a versatile in vitro transcription assay for the expression of multiple start site TATA-less promoters. / Y. Lin, T.A. Ince, K.W. Scotto // Biochemistry. - 2001. - V. 40, N. 43. - P. 12959 - 12966.

268. Lin, Z. The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. / Z. Lin, W.S. Wu, H. Liang, Y. Woo, W.H. Li // BMC Genomics. - 2010. - V. 11. - P. 581.

269. Lipman, D.J. Comparative analysis of nucleic acid sequences by their general constraints. / D.J. Lipman, J. Maizel // Nucleic Acids Res. – 1982. - V. 10, N. 8. - P. 2723 - 2739.

270. Liu, Z.B. Soybean GH3 promoter contains multiple auxin-inducible elements. / Z.B. Liu, T. Ulmasov, X. Shi, G. Hagen, T.J. Guilfoyle // Plant Cell. - 1994. - V. 6, N. 5. - P. 645 - 657.

271. Long, Y.S. Identification of the transcriptional promoters in the proximal regions of human microRNA genes. / Y.S. Long, G.F. Deng, X.S. Sun, Y.H. Yi, T. Su, Q.H. Zhao, W.P. Liao // Mol Biol Rep. - 2011. - V. 38, N. 6. - P. 4153 - 4157.

272. Luk, E. Stepwise histone replacement by SWR1 requires dual activation with histone H2A.Z and canonical nucleosome. / E. Luk, A. Ranjan, P.C. Fitzgerald, G. Mizuguchi, Y. Huang, D. Wei, C. Wu // Cell. - 2010. - V. 143, N. 5. - P. 725 - 736.

273. Lunter, G. A nucleotide substitution model with nearest-neighbour interactions. / Lunter G., Hein J. // Bioinformatics. – 2004. - V. 20, Suppl. 1. - P. i216 - i223.

274. MacLeod, M.C. Identification of a DNA structural motif that includes the binding sites for Spi, p53 and GA-binding protein. / M.C. MacLeod // Nucleic Acids Res. - 1993. - V. 21, N. 6. - P. 1439 - 1447.

275. Makeev, V. The third nucleotide of the Gly coding triplet remembers the periodicity of the collagen chain. / V. Makeev, V.G. Tumanyan, N.G. Esipova // FEBS Lett. – 1995. - V. 366, N. 1. - P. 33 - 36.

276. Makeev, V. Search of periodicities in primary structure of biopolymers: a general Fourier approach. / V. Makeev, V.G. Tumanyan // Comput. Applic. Biosci. – 1996. - V. 12, N. 1. - P. 49 - 54.

277. Maloney, M.D. Safety and efficacy of ultraviolet-a light-activated gene transduction for gene therapy of articular cartilage defects. / M.D. Maloney, J.J. Goater, R. Parsons, H. Ito, R.J. O'Keefe, P.T. Rubery, M.H. Drissi, E.M. Schwarz // J. Bone Joint Surg. Am. - 2006. - V. 88, N. 4. - P. 753 - 761.

278. Marklund, E.G. Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. / E.G. Marklund, A. Mahmutovic, O.G. Berg,

P. Hammar, D. van der Spoel, D. Fange, J. Elf // Proc. Natl. Acad. Sci. USA. - 2013. - V. 110, N. 49. - P. 19796 - 19801

279. Martianov, I. RNA polymerase II transcription in murine cells lacking the TATA binding protein. / I. Martianov, S. Viville, I. Davidson // Science. - 2002. - V. 298, N. 5595. - P. 1036-1039.

280. Marx, J.L. Putting the human genome on the map. / J.L Marx. // Science. - 1985. - V. 229, N. 4709. - P. 150 - 151.

281. Maxam, A.M. A new method for sequencing DNA. / A.M. Maxam, W. Gilbert // Proc. Natl. Acad. Sci. U.S.A. - 977. - V. 74, N. 2. - P. 560 - 564.

282. Mazin, A.V. The specificity of the secondary DNA binding site of RecA protein defines its role in DNA strand exchange. / A.V. Mazin, S.C. Kowalczykowski // Proc. Natl. Acad. Sci. U.S.A. - 1996. - V. 93, N. 20. - P. 10673 - 10678.

283. McKusick, V.A. The human genome through the eyes of a clinical geneticist. / V.A. McKusick // Cytogenet. Cell. Genet. - 1982. - V. 32, N. 1 – 4. - P. 7 - 23.

284. McDevitt, M.A. Sequences capable of restoring poly(A) site function define two distinct downstream elements. / M.A. McDevitt, R.P. Hart, W.W. Wong, J.R. Nevins // EMBO J. - 1986. - V. 5, N. 11. - P. 2907 - 2913.

285. Meierhans, D. The N-terminal methionine is a major determinant of the DNA binding specificity of MEF-2C. / D. Meierhans, R.K. Allemann // J. Biol. Chem. - 1998. - V. 273, N. 40. - P. 26052 - 26060.

286. Meierhans, D. High affinity binding of MEF-2C correlates with DNA bending. / D. Meierhans, M. Sieber, R.K. Allemann // Nucleic Acids Res. - 1997. - V. 25, N. 22. - P. 4537 - 4544.

287. Melvin, T. Guanine is the target for direct ionisation damage in DNA, as detected using excision enzymes. / T. Melvin, S.M. Cunniffe, P. O'Neill, A.W. Parker, T. Roldan-Arjona // Nucleic Acids Res. - 1998. - V. 26, N. 21. - P. 4935 - 4942.

288. Metropolis, N. The Monte Carlo method. / N. Metropolis, S. Ulam // J. Am. Stat. Assoc. – 1949. - V. 44, N. 247. - P. 335 - 341.

289. Meysman, P. Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. / P. Meysman, T.H. Dang, K. Laukens, R. De Smet, Y. Wu, K. Marchal, K. Engelen // Nucleic Acids Res. - 2011. - V. 39, N. 2. - P. e6.

290. Mhaskar, D.N. On the molecular basis of transition mutations. Frequency of forming 2-aminopurine-cytosine base mispairs in the G X C----A X T mutational pathway by T4 DNA polymerase in vitro. / D.N. Mhaskar, M.F. Goodman // J. Biol. Chem. - 1984. - V. 259, N. 19. - P. 11713 - 11717.

291. Mironova, V.V. How multiple auxin responsive elements interact in plant promoters: evidences from a reverse problem solution. / V.V. Mironova, N.A. Omelyanchuk, M.S. Savina, P.M. Ponomarenko, M.P. Ponomarenko, V.A. Likhoshvai, N.A. Kolchanov // J. Bioinform. Comput. Biol. - 2013. - V. 11, N. 1. - P. 1340011.

292. Mirzabekov, A.D. Localization of chromatin proteins within DNA grooves by methylation of chromatin with dimethyl sulphate. / A.D. Mirzabekov, A.F. Melnikova // Mol. Biol. Rep. - 1974. - V. 1, N. 7. - P. 379 - 384.

293. Mogno, I. TATA is a modular component of synthetic promoters. / I. Mogno, F. Vallania, R.D. Mitra, B.A Cohen. // Genome Res. - 2010. - V. 20, N. 10. - P. 1391 - 1397.

294. Molina, C. Genome wide analysis of Arabidopsis core promoters. / C. Molina, E. Grotewold // BMC Genomics. - 2005. - V. 6. - P. 25.

295. Mulligan, M. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. / M. Mulligan, D. Hawley, R. Entriken, W.R. McClure // Nucleic Acids Res. – 1984. - V. 12, N. 1, Pt. 2. - P. 789 - 800.

296. Nakamura, M. Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. / M. Nakamura, T. Tsunoda, J. Obokata // Plant J. - 2002. - V. 29, N. 1. - P. 1 - 10.

297. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. / NCBI Resource Coordinators // Nucleic Acids Res. - 2013. - V. 41, Database issue. - P. D8 - D20.

298. Needleman, S.B. A general method applicable to the search for similarities in the amino acid sequence of two proteins. / S.B. Needleman, C.D. Wunsch // J. Mol. Biol. – 1970. - V. 48, N 3. - P. 443 - 453.

299. Nikolajewa, S. BioBayesNet: a web server for feature extraction and Bayesian network modeling of biological sequence data. / S. Nikolajewa, R. Pudimat, M. Hiller, M. Platzer, R. Backofen // Nucleic Acids Res. - 2007. - V. 35, Web-server issue. - P. W688-W693.

300. Nicolay, S. Low frequency rhythms in human DNA sequences: a key to the organization of gene location and orientation? / S. Nicolay, F. Argoul, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, A. Arneodo // Phys. Rev. Lett. – 2004. - V. 93, N. 10. - P. 108101.

301. Neidle, S. DNA structure, recognition. / S. Neidle - NY: IRL Press. - 1994. - 108 p.

302. Ni, Y. Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. / Y. Ni, A.W. Hall, A. Battenhouse, V.R. Iyer // BMC Genet. - 2012. – V. 13. – P. 46.

303. Niemann, S. Analysis of a genetic defect in the TATA box of the SOD1 gene in a patient with familial amyotrophic lateral sclerosis. / S. Niemann, W.J. Broom, R.H. Brown Jr. // Muscle Nerve. - 2007. - V. 36, N. 5. - P. 704 - 707.

304. Nirenberg, M.W. Nobel Prizes for Medicine, 1968. / M.W. Nirenberg, H.G. Khorana, R.H. Holley // Nature. - 968. - V. 220, N. 5165. - P. 324 - 325.

305. Nussinov, R. Sequence context of oligomer tracts in eukaryotic DNA: biological and conformational implications. / R. Nussinov, A. Sarai, G. Smythers, R.L. Jernigan // J. Biomol. Struct. Dyn. – 1988. - V. 6, N. 3. - P. 543 - 562.

306. Ohlschlegel, H. The patentability of the human genome. / H. Ohlschlegel // Naturwissenschaften. - 1981. - V. 68, N. 8. - P. 423.

307. Ohkuma, Y. Engrailed, a homeodomain protein, can repress in vitro transcription by competition with the TATA box-binding protein transcription factor IID. / Y. Ohkuma, M. Horikoshi, R.G. Roeder, C. Desplan // Proc. Natl. Acad. Sci. USA. – 1990. - V. 87, N. 6. - P. 2289 - 2293.

308. Oliver, J. SEGMENT: identifying compositional domains in DNA sequences. / J. Oliver, R. Roman-Roldan, J. Perez, P. Bernaola-Galvan // Bioinformatics. – 1999. - V. 15, N. 12. - P. 974 - 979.

309. Omelina, E.S. Analysis and recognition of the GAGA transcription factor binding sites in Drosophila genes. / E.S. Omelina, E.M. Baricheva, D.Y. Oshchepkov, T.I. Merkulova // Comput. Biol. Chem. – 2011. - V. 35, N. 6. - P. 363 - 370.

310. Orkin, S.H. ATA box transcription mutation in beta-thalassemia. / S.H. Orkin, J.P. Sexton, T.C. Cheng, S.C. Goff, P.J. Giardina, J.I. Lee, H.H. Kazazian Jr. // Nucleic Acids Res. - 1983. - V. 11, N. 14. - P. 4727 - 4734.

311. Orlov, Y. Complexity: an internet resource for analysis of DNA sequence complexity. / Y. Orlov, V. Potapov // Nucleic Acids Res. – 2004. - V. 32, Web-server issue. - P. W628 - W633.

312. Orlov, Y. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. / Y. Orlov, R. Te Boekhorst, I.I. Abnizova // J. Bioinform. Comput. Biol. – 2006. - V. 4, N. 2. - P. 523 - 536.

313. Oshchepkov, D.Y. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. / D.Y. Oshchepkov, E.E. Vityaev, D.A. Grigorovich, E.V. Ignatieva, T.M. Khlebodarova // Nucleic Acids Res. - 2004. - V. 32, Web-server issue. - P. W208 - W212.

314. Oshchepkov, D.Y. In silico prediction of transcriptional factor-binding sites. / D.Y. Oshchepkov, V.G. Levitsky // Methods Mol. Biol. - 2011. - V. 760. - P. 251-267.

315. Ou, S.H. Role of flanking E box motifs in human immunodeficiency virus type 1 TATA element function. / S.H. Ou, L.F. Garcia-Martinez, E.J. Paulssen, R.B. Gaynor // J. Virol. - 1994. - V. 68, N. 11. - P.7188 - 7199.

316. Ouyang, Z. Hierarchical structure analysis describing abnormal base composition of genomes. / Z. Ouyang, J. Liu, Z. She // Phys. Rev. E. - 2005. - V. 72, N. 4, Pt. 1. - P. 041915.

317. Paraskevopoulou, M.D. BiDaS: a web-based Monte Carlo BioData Simulator based on sequence/feature characteristics. / M.D. Paraskevopoulou, I.S. Vlachos, E. Athanasiadis, G. Spyrou // Nucleic Acids Res. - 2013. - V. 41, Web-server issue. - P. W582 - W586.

318. Pareek, C.S. Sequencing technologies and genome sequencing. / C.S. Pareek, R. Smoczynski, A. Tretyn // J. Appl. Genet. - 2011. - V. 52, N. 4. - P. 413 - 435.

319. Parker, C.S. A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene. / C.S. Parker, J. Topol // Cell. - 1984. - V. 37, N. 1. - P. 273 - 283.

320. Pater, M. Comparative analysis of GS and BK virus genomes. / M. Pater, A. Pater, G. di Mayorca // J. Virol. – 1979. - V. 32, N. 1. - P. 220 - 225.

321. Pellegrini, L. Structure of serum response fac-tor core bound to DNA. / L. Pellegrini, S. Tan, T.J. Richmond // Nature. - 1995. - V. 376, N. 6540. - P. 490 - 498.

322. Penner, C.G. Transcription factor GATA-1-multiprotein complexes and chicken erythroid development. / C.G. Penner, J.R. Davie // FEBS Lett. - 1994. - V. 342, N. 3. – P. 273 - 277.

323. Perier, R.C. The eukaryotic promoter database (EPD). / R.C. Perier, V. Praz, T. Junier, C. Bonnard, P. Bucher // Nucleic Acids Res. - 2000. - V. 28, N. 1. - P. 302 - 303.

324. Perez, A. Towards a molecular dynamics consensus view of B-DNA flexibility. / A. Perez, F. Lankas, F.J. Luque, M. Orozco // Nucleic Acids Res. - 2008. - V. 36, N. 7. - P. 2379 - 2394.

325. Pfeffer, S. Identification of virus-encoded microRNAs. / S. Pfeffer, M. Zavolan, F. Grasser, M. Chien, J.J. Russo, J. Ju, B. John, A.J. Enright, D. Marks, C. Sander, T. Tuschl // Science. – 2004. - V. 304, N. 5671. - P. 734 - 736.

326. Pitarque, M. Identification of a single nucleotide polymorphism in the TATA box of the CYP2A6 gene: impairment of its promoter activity. / M. Pitarque, O. von Richter, B. Oke, H. Berkkan, M. Oscarson, M. Ingelman-Sundberg // Biochem. Biophys. Res. Commun. - 2001. - V. 284, N. 2. - P. 455 - 460.

327. Pollock, R. A sensitive method for the determination of protein-DNA binding specificities. / R. Pollock, R. Treisman // Nucleic Acids Res. - 1990. - V. 18, N. 21. - P. 6197 - 6204.

328. Poncz, M. beta-Thalassemia in a Kurdish Jew. Single base changes in the T-A-T-A box. / M. Poncz, M. Ballantine, D. Solowiejczyk, I. Barak, E. Schwartz, S. Surrey // J. Biol. Chem. - 1982. - V. 257, N. 11. - P. 5994 - 5996.

329. Ponjavic, J. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. / J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, A. Sandelin // Genome Biol. - 2006. - V. 7, N. 8. - P. R78.

330. Ponomarenko, J.V. Conformational and physicochemical DNA features specific for transcription factor binding sites. / J.V. Ponomarenko, M.P. Ponomarenko, A.S. Frolov, D.G. Vorobyev, G.C. Overton, N.A. Kolchanov // Bioinformatics. – 1999a. - V. 15, N. 7/8. - P. 654 - 668.

331. Ponomarenko, J.V. Sequence-dependent B-helix DNA features common for transcription factor superclasses. In: The First Cold Spring Harbor Workshop "Bridging the Gap between Sequences and Functions", September 7 - 9, 1999b, New York, USA" / J.V. Ponomarenko, M.P. Ponomarenko, O.A. Podkolodnaya, A.S. Frolov; Ed. M. Zhang - Cold Spring Harbor: CSHL Press. – 1999b. - P. 12.

332. Ponomarenko, J.V. SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. / J.V. Ponomarenko, G.V. Orlova, M.P. Ponomarenko, S.V. Lavryushev, A.S. Frolov, S.V. Zybova, N.A. Kolchanov // Nucleic Acids Res. - 2000a. - V. 28, N. 1. - P. 205 - 208.

333. Ponomarenko, J.V. Diffusion additive to TATA-box recognition score detects the non-contextual TBP-binding site at -30 position of TATA-less promoters. In: "Proceedings of WSES/MIUE/HNA International Conference: Mathematics and Computers in Biology and Chemistry (MCBC'2000), December 20-22, 2000, Montego Bay, Jamaica" / J.V. Ponomarenko, M.P. Ponomarenko, A.S. Frolov, I. Zvolsky; Ed. N. Mastorakis - NY: WSES Press Publ. - 2000b. - P. 901 - 906.

334. Ponomarenko, J.V. ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another. / J.V. Ponomarenko, D.P. Furman, A.S. Frolov, N.L. Podkolodny, G.V. Orlova, M.P. Ponomarenko, N.A. Kolchanov, A. Sarai // Nucleic Acids Res. - 2001a. - V. 29, N. 1. - P. 284 - 287.

335. Ponomarenko, J.V. rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. / J.V. Ponomarenko, T.I. Merkulova, G.V. Vasiliev, Z.B. Levashova, G.V. Orlova, S.V. Lavryushev, O.N. Fokin, M.P. Ponomarenko, A.S. Frolov, A. Sarai // Nucleic Acids Res. - 2001b. - V. 29, N. 1. - P. 312 - 316.

336. Ponomarenko, J.V. rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. / J.V. Ponomarenko, G.V. Orlova, T.I. Merkulova, E.V. Gorshkova, O.N. Fokin, G.V. Vasiliev, A.S. Frolov, M.P. Ponomarenko // Hum. Mutat. - 2002a. - V. 20, N. 4. - P. 239 - 248.

337. Ponomarenko, J.V. Mining DNA sequences to predict sites which mutations cause genetic diseases. / J.V. Ponomarenko, T.I. Merkulova, G.V. Orlova, O.N. Fokin, E.V. Gorshkova, M.P. Ponomarenko // Knowledge-Based Systems J. - 2002b. - V. 15, N. 4. - P. 225 - 233.

338. Ponomarenko, J.V. SELEX_DB: a database on in vitro selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data. / J.V. Ponomarenko, G.V. Orlova, A.S. Frolov, M.S. Gelfand, M.P. Ponomarenko // Nucleic Acids Res. - 2002c. - V. 30, N. 1. - P. 195 - 199.

339. Ponomarenko, J.V. Annotation of potential transcription factor binding sites using rSNP_Guide with the models of experimentally characterized altered TF sites. In: “EuroQSAR 2002: Designing drugs and crop protectants” / J.V. Ponomarenko, G.V. Orlova, T.I. Merkulova, E.V. Gorshkova, V.P. Valuev, M.P. Ponomarenko Eds. M. Ford, D. Livingstone, J. Dearden, H. Van de Waterbeemd - Oxford: Blackwell Publ. Ltd (UK). - 2003. - P. 347 - 351.

340. Ponomarenko, J.V. rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. / J.V. Ponomarenko, T.I. Merkulova, G.V. Orlova, O.N. Fokin, E.V. Gorshkova, A.S. Frolov, V.P. Valuev, M.P. Ponomarenko // Nucleic Acids Res. - 2003. - V. 31, N. 1. - P. 118 - 121.

341. Ponomarenko, J.V. Mining genome variation to associate genetic disease with mutation alterations and ortho/paralogous polimorphysms in transcription factor binding site. / J.V. Ponomarenko, G.V. Orlova, T.M. Merkulova, G.V. Vasiliev, M.P. Ponomarenko // Int. J. Artif. Intell. Tools. - 2005. - V. 14, N.4. - P. 599 - 620.

342. Ponomarenko, M.P. Generating programs for predicting the activity of functional sites. / M.P. Ponomarenko, A.N. Kolchanova, N.A. Kolchanov // J. Comput. Biol. - 1997a. - V. 4, N. 1. - P. 83 - 90.

343. Ponomarenko, M.P. Search for DNA conformational features for functional sites. Investigation of the TATA box. In: Pac. Symp. Biocomput. / M.P. Ponomarenko, J.V. Ponomarenko, A.E. Kel, N.A. Kolchanov; Eds. R. Altman, A.K. Dunker, L. Hunter, T.E. Klein - Singapore: World Sci. - 1997b. - V. 2, P. 340 - 351.

344. Ponomarenko, M.P. Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. / M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, N.L. Podkolodny, L.K. Savinkova, N.A. Kolchanov, G.C. Overton // Bioinformatics. - 1999a. - V. 15, N. 7/8. - P. 687 - 703.

345. Ponomarenko, M.P. Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. / M.P. Ponomarenko, J.V. Ponomarenko, A.S. Frolov, O.A. Podkolodnaya, D.G. Vorobyev, N.A. Kolchanov, G.C. Overton // Bioinformatics. - 1999b. - V. 15, N. 7/8. - P. 631 - 643.

346. Ponomarenko, M.P. A database on DNA sequence/activity relationships: application to phylogenetic footprinting. In: Proceedings of the Fourth Conference on Bioinformatics of Genome Regulation and Structure: BGRS'2004 / M.P. Ponomarenko, J.V. Ponomarenko; Eds. N.A. Kolchanov, E. Borovskikh, G. Chirikova, D. Afonnikov, S. Lavryushev - Novosibirsk: IC&G Press. - 2004. - V. 1. - P. 166-169.

347. Ponomarenko, M. Degrees of freedom. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, V. Babenko, A. Kochetov, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013a. - V. 2. – P. 290 - 292.

348. Ponomarenko, M. Cladogenesis. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, K Gunbin, A. Doroshkov, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013b. - V. 2. – P. 21 - 24.

349. Ponomarenko, M. Hogness box. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, V. Mironova, K. Gunbin, L. Savinkova; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013c. - V. 3. – P. 491 - 494.

350. Ponomarenko, M. Unique DNA. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, G. Orlova, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013d. - V. 7. – P. 259 - 262.

351. Ponomarenko, M. Initiation factors. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, L. Savinkova, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013e. - V. 4. – P. 83 - 85.

352. Ponomarenko, M. Heat shock proteins. In: Brenner's Encyclopedia of Genetics. / M. Ponomarenko, I. Stepanenko, N. Kolchanov; Eds. S. Maloy, K. Hughes - 2nd edn. - San Diego: Academic Press, Elsevier Inc. – 2013f. - V. 3. – P. 402 - 405.

353. Ponomarenko, M.P. Abundances of microRNAs in human cells can be estimated as a function of the abundances of YRHB and RHHK tetranucleotides in these microRNAs as an ill-posed inverse problem solution. / M.P. Ponomarenko, V.V. Suslov, P.M. Ponomarenko, K.V. Gunbin, I.L. Stepanenko, O.V. Vishnevsky, N.A. Kolchanov // Front Genet. – 2013g. - V. 4, N. 2. - P. 122.

354. Ponomarenko, M.P. Candidate SNP markers of gender-biased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. / M.P. Ponomarenko,

O. Arkova, D. Rasskazov, P. Ponomarenko, L. Savinkova, N. Kolchanov // Front. Immunol. - 2016. – V. 7. – P. 130.

355. Ponomarenko, P.M. Sequence-based prediction of transcription upregulation by auxin in plants. / P.M. Ponomarenko, M.P. Ponomarenko // J. Bioinform. Comput. Biol. - 2015. - V. 13, N. 1. - P. 1540009.

356. Powell, R.M. Comparison of TATA-binding protein recognition of a variant and consensus DNA promoters. / R.M. Powell, K.M. Parkhurst, L.J. Parkhurst // J. Biol. Chem. - 2002. - V. 277, N. 10. - P. 7776 – 7784.

357. Pudimat, R. A multiple-feature framework for modelling and predicting transcription factor binding sites. / R. Pudimat, E.G. Schukat-Talamazzini, R. Backofen // Bioinformatics. - 2005. - V. 21, N. 14. - P. 3082-3088.

358. Qin, T.T. SOS response and its regulation on the fluoroquinolone resistance. / T.T. Qin, H.Q. Kang, P. Ma, P.P. Li, L.Y. Huang, B. Gu // Ann. Transl. Med. – 2015. – V. 3, N. 22. – P. 358.

359. Quandt, K. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. / K. Quandt, K. Frech, H. Karas, E. Wingender, T. Werner // Nucleic Acids Res. – 1995. - V. 23, N. 23. - P. 4878 - 4884.

360. Raney, B.J. ENCODE whole-genome data in the UCSC genome browser (2011 update). / B.J. Raney, M.S. Cline, K.R. Rosenbloom, T.R. Dreszer, K. Learned, G.P. Barber, L.R. Meyer, C.A. Sloan, V.S. Malladi, K.M. Roskin, B.B. Suh, A.S. Hinrichs, H. Clawson, A.S. Zweig, V. Kirkup, P.A. Fujita, B. Rhead, K.E. Smith, A. Pohl, R.M. Kuhn, D. Karolchik, D. Haussler, W.J. Kent // Nucleic Acids Res. - 2011. - V. 39, Database issue. - P. D871 - D875.

361. Raser, J.M. Control of stochasticity in eukaryotic gene expression. / J.M. Raser, E.K. O'Shea // Science. - 2004. - V. 304, N. 5678. - P. 1811 - 1814.

362. Reijnen, M.J. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. / M.J. Reijnen, F.M. Sladek, R.M. Bertina, P.H. Reitsma // Proc. Natl. Acad. Sci. USA. - 1992. - V. 89, N. 14. - P. 6300 - 6303.

363. Rice, C.M. The EMBL data library. / C.M. Rice, R. Fuchs, D.G. Higgins, P.J. Stoehr, G.N. Cameron // Nucleic Acids Res. - 1993. - V. 21, N. 13. – P. 2967 - 2971.

364. Richmond, T.J. The structure of DNA in the nucleosome core. / T.J. Richmond, C.A. Davey // Nature. - 2003. - V. 423, N. 6936. - P. 145 - 150.

365. Roberts, R.W. In vitro selection of nucleic acids and proteins: what are we learning? / R.W. Roberts, W.W. Ja // Curr Opin Struct Biol. - 1999. - V. 9, N. 4. - P. 521 - 529.

366. Robison, K. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. / K. Robison, A.M. McGuire, G.M. Church // J. Mol. Biol. - 1998. - V. 284, N. 2. - P. 241 - 254.

367. Rogozin, I.B. Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection. / I.B. Rogozin, V.V. Solovyov, N.A. Kolchanov // Biochim. Biophys. Acta. - 1991. - V. 1089, N. 2. - P. 175 - 182.

368. Rosenquist, T.A. Cloning and characterization of a mammalian 8-oxoguanine DNA glycosylase. / T.A. Rosenquist, D.O. Zharkov, A.P. Grollman // Proc. Natl Acad. Sci. U.S.A. - 1997. - V. 94, N. 14. - P. 7429 - 7434.

369. Roulet, E. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. / E. Roulet, I. Fisch, T. Junier, P. Bucher, N. Mermod // In Silico Biol. – 1998. - V. 1, N. 1. - P. 21 - 28.

370. Saito, I. Photoinduced DNA cleavage via electron transfer: demonstration that guanine residues located 5’ to guanine are the most electron-donating sites. / I. Saito, M. Takayama, H. Sugiyama, K. Nakatani // J. Am. Chem. Soc. - 1995. - V. 117, N. 23. - P. 6406 - 6407.

371. Salganik, R.I. Structure of the plant mitochondrial genome and light-regulated transcription of the mitochondrial cenes. In: Nuclear Structure and Function / Salganik R.I., Dudareva N.A., Popovsky A.V., Kiseleva E.V., Rozov S.M.; Eds. J.R. Harris, I.B. Zbarsky – NY: Plenum Press. – 1990 - P. 19-22.

372. Salganik, R.I. Structural organization and transcription of plant mitochondrial and chloroplast genomes. / Salganik R.I., Dudareva N.A., Kiseleva E.V. // Electron Microsc Rev. - 1991. - V. 4, N 2. - P. 221-247.

373. Sarai, A. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. / A. Sarai, Y. Takeda // Proc. Natl. Acad. Sci. USA. – 1989. - V. 86, N. 17. - P. 6513 - 6517.

374. Satchwell, S.C. Sequence periodicities in chicken nucleosome core DNA. / S.C. Satchwell, H.R. Drew, A.A. Travers // J. Mol. Biol. - 1986. - V. 191, N. 4. - P. 659 - 675.

375. Savinkova, L.K. Quantitative computer-assisted analysis of the TATA-binding protein affinity for complementary duplexes of synthetic oligodeoxyribonucleotides. In: Proceedings of the First Conference on Bioinformatics of Genome Regulation and Structure: BGRS'98 / L.K. Savinkova, A.A. Sokolenko, V.A. Rau, V.F. Kobzev, M.P. Ponomarenko, J.V. Ponomarenko, N.A. Kolchanov Eds. N.A. Kolchanov, E. Borovskikh, G. Chirikova, D. Afonnikov, S. Lavryushev - Novosibirsk: IC&G Press. - 1998. - V. 1. - P. 165-169.

376. Savinkova, L.K. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. / L.K. Savinkova, I.A. Drachkova, T.V. Arshinova, P.M. Ponomarenko, M.P. Ponomarenko, N.A. Kolchanov // PLoS ONE. - 2013. - V. 8, N. 2. - P. e54626.

377. Sax, C.M. Lens-specific activity of the mouse alpha A-crystallin promoter in the absence of a TATA box: functional and protein binding analysis of the mouse alpha A-crystallin PE1 region. / C.M. Sax, A. Cvekl, M. Kantorow, R. Gopal-Srivastava, J.G. Ilagan, N.P. Ambulos Jr, J. Piatigorsky// Nucleic Acids Res. - 1995. - V. 23, N. 3. - P. 442 - 451.

378. Schastak, S. Flexible UV light guiding system for intraocular laser microsurgery. / S. Schastak, Y. Yafai, T. Yasukawa, Y.S. Wang, G. Hillrichs, P. Wiedemann // Lasers Surg. Med. - 2007. - V. 39, N. 4. - P. 353 - 357.

379. Schmidt, M.C. Yeast TATA-box transcription factor gene. / M.C. Schmidt, C.C. Kao, R. Pei, A.J. Berk // Proc. Natl. Acad. Sci. U.S.A. - 1989. - V. 86, N. 20. - P. 7785 - 7789.

380. Schneider, T. Information content of binding sites on nucleotide sequences. / T. Schneider, G. Stormo, L. Gold, A. Ehrenfeucht // J. Mol. Biol. – 1986. - V. 188, N. 3. - P. 415 - 431.

381. Schroer, B. Molekularbiologisch-genetische Charakterisierung des RhoA-Promotors bezuglich kardiovaskularer Erkrankungen: PhD … Natural Sciences / Bianca Schroer. - Munster: Westfalischen Wilhelms-Universitat Munster (Germany). - 2010. - 119 p.

382. Schug, J. TESS: Transcription Element Search Software on the WWW. Technical Report CBIL-TR-1997-1001-v0.0. / J. Schug, G.C. Overton - Philadelphia: UPen Press. – 1997.

383. Shannon, C.E. A mathematical theory of communication. / C.E. Shannon // Bell Syst. Tech. J. – 1948. - V. 27, Pt. I. - P. 379–423; - Pt. II. - P. 623–656.

384. Shepherd, J.C. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. / J.C. Shepherd // J. Mol. Evol. – 1981. - V. 17, N. 2. - P. 94 - 102.

385. Shine, J. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. / J. Shine, L. Dalgarno // Eur. J. Biochem. – 1975. - V. 57, N. 1. - P. 221 - 230.

386. Shows, T.B. Human genome organization of enzyme loci and metabolic diseases. / T.B. Shows // Isozymes Curr. Top. Biol. Med. Res. - 1983. - V. 10. - P. 323 - 339.

387. Shpigelman, E.S. CURVATURE: software for the analysis of curved DNA. / E.S. Shpigelman, E.N. Trifonov, A. Bolshoy // Comput. Appl. Biosci. - 1993. - V. 9, N. 4. - P. 435 - 440.

388. Shulzaberger, R. Using sequence logos and informational analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and

SELEX. / R. Shulzaberger, T.D. Schneider // Nucleic Acids Res. - 1999. - V. 27, N. 3. - P. 882 - 887.

389. Sievers, F. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. / F. Sievers, A. Wilm, D.G. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J.D. Thompson, D.G. Higgins // Mol. Syst. Biol. – 2011. - V. 7. - P. 539.

390. Skoblov, M. Human RFP2 gene promoter: unique structure and unusual strength. / M. Skoblov, K. Shakhbazov, D. Oshchepkov, D. Ivanov, A. Guskova, D. Ivanov, P. Rubtsov, V. Prasolov, N. Yankovsky, A. Baranova // Biochem. Biophys. Res. Commun. - 2006. - V. 342, N. 3. - P. 859 - 866.

391. Starr, D.B. DNA bending is an important component of site-specific recognition by the TATA binding protein. / D.B. Starr, B.C. Hoopes, D.K. Hawley // J. Mol. Biol. - 1995. - V. 250, N. 4. - P. 434 - 446.

392. Staden, R. Sequence data handling by computer. / R. Staden // Nucleic Acids Res. – 1977. - V. 4, N. 11. - P. 4037 - 4051.

393. Staden, R. Further procedures for sequence analysis by computer. / R. Staden // Nucleic Acids Res. – 1978. - V. 5, N. 3. - P. 1013 - 1016.

394. Staden, R. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. / R. Staden // Nucleic Acids Res. - 1982a. - V. 10, N. 9. - P. 2951 - 2961.

395. Staden, R. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. / R. Staden // Nucleic Acids Res. - 1982b. - V. 10, N. 15. - P. 4731 - 4751.

396. Staden, R. Computer methods to locate signals in nucleic acid sequences. / R. Staden // Nucleic Acids Res. - 1984a. - V. 12, N. 1, Pt. 2. - P. 505 - 519.

397. Staden, R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. / R. Staden // Nucleic Acids Res. - 1984b. - V. 12, N. 1, Pt. 2. - P. 551 - 567.

398. Staden, R. The current status and portability of our sequence handling software. / R. Staden // Nucleic Acids Res. – 1986. - V. 14, N. 1. - P. 217-231.

399. Stormo, G.D. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. / G.D. Stormo, T.D. Schneider, L. Gold, t A. Ehrenfeuch // Nucleic Acids Res. – 1982. - V. 10, N. 9. - P. 2997 - 3011.

400. Stormo, G.D. Quantitative analysis of the relationship between nucleotide sequence and functional activity. / G.D. Stormo, T.D. Schneider, L. Gold // Nucleic Acids Res. - 1986. - V. 14, N. 16. - P. 6661 - 6679.

401. Stormo, G. Identifying protein-binding sites from unaligned DNA fragments. / G. Stormo, G.W. 3rd. Hartzell // Proc. Natl. Acad. Sci. USA. – 1989. - V. 86, N. 4. - P. 1183 - 1187.

402. Sugimoto, N. Improved thermodynamic parameters, helix initiation factor to predict stability of DNA duplexes. / N. Sugimoto, S. Nakano, M. Yoneyama, K. Honda // Nucleic Acids Res. – 1996. - V. 24, N. 22. - P. 4501 - 4505.

403. Suslov, V.V. SNPs in the HIV-1 TATA box and the AIDS pandemic. / V.V. Suslov, P.M. Ponomarenko, V.M. Efimov, L.K. Savinkova, M.P. Ponomarenko, N.A. Kolchanov // J. Bioinform. Comput. Biol. - 2010a. - V. 8, N. 3. - P. 607 - 625.

404. Suslov, V.V. Possibility spaces and evolution. / V.V. Suslov, M.P. Ponomarenko, N.A. Kolchanov // Paleontological J. - 2010b. - Vol. 44, N. 12. - P. 1491 – 1499.

405. Suzuki, M. Stereochemical basis of DNA bending by transcription factors. / M. Suzuki, N. Yagi // Nucleic Acids Res. - 1995. - V. 23, N. 12. - P. 2083 - 2091.

406. Suzuki, M. Role of base-backbone, base-base interactions in alternating DNA conformations. / M. Suzuki., N. Yagi, J.T. Finch // FEBS L. - 1996. - V. 379, N. 2. - P. 148 - 152.

407. Tachibana, H. Location of the cooperative melting regions in bacteriophage fd DNA. / H. Tachibana, A. Wada, O. Gotoh, M. Takanami // Biochim. Biophys. Acta. – 1978. - V. 517, N. 2. - P. 319 - 328.

408. Takeda, Y. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. / Y. Takeda, A. Sarai, V. Rivera // Proc. Natl. Acad. Sci. USA. – 1989. - V. 86, N. 2. - P. 439 - 443.

409. Takeuchi, N. Evolution of complexity in RNA-like replicator systems. / N. Takeuchi, P. Hogeweg // Biol. Direct. – 2008. - V. 3. - P. 11.

410. Tatarinova, T. Skew in CG content near the transcription start site in Arabidopsis thaliana. / T. Tatarinova, V. Brover, M. Troukhan, N. Alexandrov // Bioinformatics. – 2003. - V. 19, Suppl. 1. - P. i313 - i314.

411. Tatarinova, T.V. GC3 biology in corn, rice, sorghum and other grasses. / T.V. Tatarinova, N.N. Alexandrov, J.B. Bouck, K.A. Feldmann // BMC Genomics. - 2010. - V. 11. - P. 308.

412. Takihara, Y. A novel mutation in the TATA box in a Japanese patient with beta +-thalassemia. / Y. Takihara, T. Nakamura, H. Yamada, Y. Takagi, Y. Fukumaki // Blood. - 1986. - V. 67, N. 2. - P. 547 - 550.

413. Tamames, J. Estimating the extent of horizontal gene transfer in metagenomic sequences. / J. Tamames, A. Moya // BMC Genomics. – 2008. - V. 9. - P. 136.

414. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. / The International Human Genome Sequencing Consortium // Nature. - 2004. - V. 431, N. 7011. - P. 931 - 945.

415. Thogersen, J. Reproductive death of cancer cells induced by femtosecond laser pulses. / J. Thogersen, C.S. Knudsen, A. Maetzke, S.J. Jensen, S.R. Keiding, J. Alsner, J. Overgaard // Int. J. Radiat. Biol. - 2007. - V. 83, N. 5. - P. 289 - 299.

416. Tillo, D. G+C content dominates intrinsic nucleosome occupancy. / D. Tillo, T.R. Hughes // BMC Bioinf. - 2009. - V. 10. - P. 442.

417. Tirosh, I. A genetic signature of interspecies variations in gene expression. / I. Tirosh, A. Weinberger, M. Carmi, N. Barkai // Nat. Genet. - 2006. V. 38, N. 7. - P. 830 - 834.

418. Tirosh, I. The pattern and evolution of yeast promoter bendability. / I. Tirosh, J. Berman, N. Barkai // Trends Genet. - 2007. - V. 23, N. 7. - P. 318 - 321.

419. Tomita, M. ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. / M. Tomita, M. Wada, Y. Kawashima // J. Mol. Evol. – 1999. - V. 49, N. 2. - P. 182 - 192.

420. Tora, L. The TATA box regulates TATA-binding protein (TBP) dynamics in vivo. / L. Tora, H.T. Timmers // Trends Biochem. Sci. - 2010. - V. 35, N. 6. - P. 309 - 314.

421. Trifonov, E.N. Sequence-dependent deformational anisotropy of chromatin DNA. / E.N. Trifonov // Nucleic Acids Res. – 1980. - V. 8, N. 17. - P. 4041 - 4053.

422. Trifonov, E.N. The pitch of chromatin DNA is reflected in its nucleotide sequence. / E.N. Trifonov, J.L. Sussman // Proc. Natl. Acad. Sci. USA. – 1980. - V. 77, N. 7. - P. 3816 - 3820.

423. Trifonov, E.N. Gnomic: A dictionary of genetic codes. / E.N. Trifonov, V. Brendel. - NY: VCH Publishers. – 1987 - 279 p.

424. Trifonov, E.N. The multiple codes of nucleotide sequences. / E.N. Trifonov // Bull. Math. Biol. - 1989. - V. 51, N. 4. - P. 417-432.

425. Trifonov, E.N. Making sense of the human genome. In: Structure and Methods / E.N. Trifonov; Eds. R. Sarma, M. Sarma - NY: Adenine Press. – 1990. - V. 1. - P. 69 - 77.

426. Troyanskaya, O. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. / O. Troyanskaya, O. Arbell, Y. Koren, G. Landau, A. Bolshoy // Bioinformatics. – 2002. - V. 18, N. 5. - P. 679 - 588.

427. Tsai, F.T. Structural basis of preinitiation complex assembly on human pol II promoters. / F.T. Tsai, P.B. Sigler // EMBO J. - 2000. - V. 19, N. 1. - P. 25 - 36.

428. Tuerk, C. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. / C. Tuerk, L. Gold // Science. - 1990. - V. 249, N. 4968. - P. 505 - 510.

429. Ulmasov, T. ARF1, a transcription factor that binds to auxin response elements. / T. Ulmasov, G. Hagen, T.J. Guilfoyle // Science. - 1997. - V. 276, N. 5320. - P. 1865 - 1868.

430. van Werven, F.J. Distinct promoter dynamics of the basal transcription factor TBP across the yeast genome. / F.J. van Werven, H.A. van Teeffelen, F.C. Holstege, H.T. Timmers // Nat. Struct. Mol. Biol. - 2009. - V. 16, N. 10. - P. 1043 - 1048.

431. Valuev, V.P. ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved in vitro. / V.P. Valuev, D.A. Afonnikov, M.P. Ponomarenko, L. Milanesi, N.A. Kolchanov // Nucleic Acids Res. - 2002. - V. 30, N. 1. - P. 200-202.

432. Vasiliev, G.V. Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY1 transcription factor binding site. / G.V. Vasiliev, V.M. Merkulov, V.F. Kobzev, T.I. Merkulova, M.P. Ponomarenko, N.A. Kolchanov // FEBS Lett. – 1999. - V. 462, N. 1/2. - P. 85 - 88.

433. Vernoux, T. The auxin signalling network translates dynamic input into robust patterning at the shoot apex. / T. Vernoux, G. Brunoud, E. Farcot, V. Morin, H. Van den Daele, J. Legrand, M. Oliva, P. Das, A. Larrieu, D. Wells, Y. Guedon, L. Armitage, F. Picard, S. Guyomarch, C. Cellier, G. Parry, R. Koumproglou, J.H. Doonan, M. Estelle, C. Godin, S. Kepinski, M. Bennett, L. De Veylder, J. Traas // Mol. Syst. Biol. - 2011. - V. 7. - P. 508.

434. Villescas-Diaz, G. Sequence context dependence of tandem guanine:adenine mismatch conformations in RNA: a continuum solvent analysis. / G. Villescas-Diaz, M. Zacharias // Biophys. J. – 2003. - V. 85, N. 1. - P. 416 - 425.

435. Vlasic, I. recA730-dependent suppression of recombination deficiency in RecA loading mutants of Escherichia coli. / I. Vlasic, A. Simatovic, K. Brcic-Kostic // Res. Microbiol. - 2011. – V. 162, N. 3. – P. 262-269.

436. Vologodskii, A.V. Theoretical melting profiles and denaturation maps of DNA with known sequence: fdDNA. / A.V. Vologodskii, M.D. Frank-Kamenetskii // Nucleic Acids Res. – 1978. - V. 5, N. 7. - P. 2547 - 2556.

437. Wain-Hobson, S. Preferential codon usage in genes. / S. Wain-Hobson, R. Nussinov, R. Brown, J. Sussman // Gene. - 1981. - V. 13, N. 4. - P. 355 - 364.

438. Walcher, C.L. Bipartite promoter element required for auxin response. / C.L. Walcher, J.L. Nemhauser // Plant Physiol. - 2012. - V. 158, N. 1. - P. 273 - 282.

439. Wang, A. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. / A. Wang, G. Quigley, F. Kolpak, J.L. Crawford, J.H. van

Boom, G. van der Marel, A. Rich // Nature. – 1979. - V. 282, N. 5740. - P. 680 - 686.

440. Wang, A. Molecular structure of the octamer d(GGCCGGCC): modified A-DNA. / A. Wang, S. Fujii, J.H. van Boom, A. Rich // Proc. Natl. Acad. Sci. USA. – 1982. - V. 79, N. 13. - P. 3968 - 3972.

441. Wang, Y. Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity. / Y. Wang, R. Jensen, W. Stumph // Nucleic Acids Res. - 1996. - V. 24, N. 15. - P. 3100 - 3106.

442. Watanabe, M. Molecular analysis of a series of alleles in humans with reduced activity at the triosephosphate isomerase locus. / M. Watanabe, B.C. Zingg, H.W. Mohrenweiser // Am. J. Hum. Genet. - 1996. - V. 58, N. 2. - P. 308 - 316.

443. Watson, J.D. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. / J.D. Watson, F.H. Crick // Nature. – 1953. - V. 171, N. 4356. - P. 737 - 738.

444. Werstuck, G. Controlling gene expression in living cells through small molecule-RNA interactions. / G. Werstuck, M.R. Green // Science. - 1998. - V. 282, N. 5387. - P. 296 - 298.

445. West, S.C. The processing of recombination intermediates: mechanistic insights from studies of bacterial proteins. / S.C. West // Cell. - 1994. - V. 76, N. 1. - P. 9 - 15.

446. Wieczorek, E. Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. / E. Wieczorek, M. Brand, X. Jacq, L. Tora // Nature. - 1998. - V. 393, N. 6681. - P. 187 - 191.

447. Wilkins, M.H.F. Molecular structure of deoxypenthatose nucleic acids. / M.H.F. Wilkins, A.R. Stokes, H.R. Wilson // Nature. – 1953. - V. 171, N. 4356. - P. 738 - 740.

448. Wing, R.M. Crystal structure analysis of a complete turn of B-DNA. / R.M. Wing, H.R. Drew, T. Takano, C. Broka, S. Tanaka, K. Itakura, R.E. Dickerson // Nature. - 1980. - V. 287, N. 5784. - P. 755 - 758.

449. Wolner, B.S. TATA-flanking sequences influence the rate and stability of TATA-binding protein and TFIIB binding. / B.S. Wolner, J.D. Gralla // J. Biol. Chem. - 2001. - V. 276, N. 9. - P. 6260 - 6266.

450. Wong, J.T. Histone-like proteins of the dinoflagellate Crypthecodinium cohnii have homologies to bacterial DNA-binding proteins. / J.T. Wong, D.C. New, J.C. Wong, V.K. Hung // Eukaryot. Cell. - 2003. - V. 2, N. 3. - P. 646-650.

451. Wootton, J. Analysis of compositionally biased regions in sequence databases. / J. Wootton, S. Federhen // Methods Enzymol. – 1996. - V. 266. - P. 554 - 571.

452. Wright, W.E. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. / W.E. Wright, M. Binder, W. Funk // Mol. Cell Biol. - 1991. - V. 11, N. 8. - P. 4104 - 4110.

453. Wright, K.J. TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. / K.J. Wright, M.T. Marr 2nd, R. Tjian // Proc. Natl. Acad. Sci. USA. - 2006. - V. 103, N. 33. - P. 12347 - 12352.

454. Wu, G. Timing of mutation in hemagglutinins from influenza A virus by means of unpredictable portion of amino-acid pair and fast Fourier transform. / G. Wu, S. Yan // Biochem. Biophys. Res. Commun. – 2005. - V. 333, N. 1. - P. 70 - 78.

455. Wu, J. dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. / J. Wu, M. Wu, L. Li, Z. Liu, W. Zeng, R. Jiang // Database (Oxford). – 2016 - V. 2016. – P. baw024.

456. Wu, K.S. Influence of interleukin-1 beta genetic polymorphism, smoking and alcohol drinking on the risk of non-small cell lung cancer. / K.S. Wu, X. Zhou, F. Zheng, X.Q. Xu, Y.H. Lin, J. Yang // Clin. Chim. Acta. - 2010. - V. 411, N. 19 – 20. - P. 1441 - 1446.

457. Xing, H. Mechanism of hsp70i gene bookmarking. / H. Xing, D.C. Wilkerson, C.N. Mayhew, E.J. Lubert, H.S. Skaggs, M.L. Goodson, Y. Hong, O.K. Park-Sarge, K.D. Sarge // Science. - 2005. - V. 307, N. 5708. - P. 421 - 423.

458. Xuan, J. Next-generation sequencing in the clinic: promises and challenges. / J. Xuan, Y. Yu, T. Qing, L. Guo, L. Shi // Cancer Lett. - 2013. - V. 340, N 2. - P. 284 - 295.

459. Yakovchuk, P. RNA polymerase II and TAFs undergo a slow isomerization after the polymerase is recruited to promoter-bound TFIID. / P. Yakovchuk, B. Gilman, J.A. Goodrich, J.F. Kugel // J. Mol. Biol. - 2010. - V. 397, N. 1. - P. 57 - 68.

460. Yamamoto, Y.Y. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. / Y.Y. Yamamoto, H. Ichida, T. Abe, Y. Suzuki, S. Sugano, J. Obokata // Nucleic Acids Res. - 2007. - V. 35, N. 18. - P. 6219 - 6226.

461. Yang, C. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. / C. Yang, E. Bolotin, T. Jiang, F.M. Sladek, E. Martinez // Gene. - 2007. - V. 389, N. 1. - P. 52 - 65.

462. Yang, M.Q. Genome-wide detection of a TFIID localization element from an initial human disease mutation. / M.Q. Yang, K. Laflamme, V. Gotea, C.H. Joiner, N.E., Seidel C. Wong, H.M. Petrykowska, J. Lichtenberg, S. Lee, L. Welch, P.G. Gallagher, D.M. Bodine, L. Elnitski // Nucleic Acids Res. - 2011. - V. 39, N. 6. - P. 2175 - 2187.

463. Yarden, G. Characterization of sINR, a strict version of the Initiator core promoter element. / G. Yarden, R. Elfakess, K. Gazit, R. Dikstein// Nucleic Acids Res. - 2009. - V. 37, N. 13. - P. 4234 - 4246.

464. Zadeh, L.A. Fuzzi sets. / L.A. Zadeh // Information and Control. - 1965. - V. 8. - P. 338 - 353.

465. Zanegina, O. Conserved features of complexes of TATA-box binding proteins with DNA. / O. Zanegina, E. Aksianov, A.V. Alexeevski, A. Karyagina, S. Spirin // J. Bioinform. Comput. Biol. - 2016. - V. 14, N. 2. - P. 1641007.

466. Zhang, Z. Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression. / Z. Zhang, J. Parsch // Mol. Biol. Evol. – 2005. - V. 22, N. 10. - P. 1945 - 1947.

467. Zhao, Y.Y. Role of C/A polymorphism at -20 on the expression of human angiotensinogen gene. / Y.Y. Zhao, J. Zhou, C.S. Narayanan, Y. Cui, A. Kumar // Hypertension. - 1999. - V. 33, N. 1. - P. 108 - 115.

СПИСОК ТЕРМИНОВ, ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

а.к.о.	−**а**мино**к**ислотный **о**статок (единица длины нити белка)
В-ДНК	−**В**-форма двойной спирали **ДНК**
днДНК	−**дн**уклеиновая ДНК
ДНК	−**д**езоксирибо**н**уклеиновая **к**ислота
кДНК	−**к**одирующая **ДНК**
М	−**м**оль на литр (единица измерения концентрации вещества)
миРНК	−**ми**кро**РНК**
мРНК	−**м**атричная **РНК**
моль (mol)	−моль (единица измерения количества вещества)
нт	−**н**уклео**т**ид (единица длины РНК и/или однонитевой ДНК)
олигоДНК	−короткий синтетический **олиго**нуклеотид **ДНК**
олигоРНК	−короткий синтетический **олиго**нуклеотид **РНК**
онДНК	−**о**дно**н**итевая ДНК
п.о.	−**п**ара **о**снований (единица длины двунитевой ДНК)
пре-мРНК	−**пре**дшественник **м**атричной **РНК** (первичный транскрипт гена, кодирующего белок)
РНК	−**р**ибо**н**уклеиновые **к**ислота
случайные ДНК	последовательности **случайных** равновероятных независимых нуклеотидов **ДНК**
тРНК	−**т**ранспортная **РНК**
ТФ	−**т**ранскрипционный **ф**актор
шт	−**шт**ук, единица измерения количества однотипных объектов
Ago2, Ago3	−белки семейства ***Argo****naute* в составе комплекса рибонуклеопротеиновый комплекс-репрессор (RISC, сокращение от английского “**R**NA-**I**nduced **S**ilencing **C**omplex”)
ANSI	−сокращение от “American National Standards Institute” англ. яз.
AP-1	−суперсемейство транскрипционных факторов с общим названием “белок-активатор 1”, являющихся гетеромерами, субъединицами которых могут быть белки семейств c-Fos, c-Jun, ATF, JDP (сокращение английского “**A**ctivator **P**rotein **1**”)
ARF	−семейство транскрипционных факторов ответа на ауксин у растений(сокращение от английского “**A**uxin **R**esponse **F**actors”)

ATF −семейство транскрипционных факторов с общим названием “активирующий фактор транскрипции”, которое состоит из семи генов ATF1-7, относящееся к суперсемейству AP-1 (сокращение от английского “**A**ctivating **T**ranscription **F**actor”)

BREu и BREd − B-регуляторные элементы промотора (сокращение от англ. “**B** **R**egulatory **E**lements located **u**pstream and/or **d**ownstream TATA-box, respectively” англ. яз.), которые являются двумя сайтами связывания инициирующего транскрипционного фактора TFIIB, локализованными непосредственно перед и после TATA-бокса промоторов генов у эукариот, соответственно, и определяющими направление транскрипции 5’→BREu→(TBP/TATA)→BREd→3’ вследствие контекстного различия между ними: BREu имеет консенсус (c/g)(c/g)(a/g)CGCC, тогда как у BREd нет консенсуса и есть лишь GC-обогащение в сравнении с TATA-боксом.

CEBP −общее обозначение продуктов семейства генов, кодирующих, по меньшей мере, семь транскрипционных факторов, первый из который был открыт как белок, который связывает очень частый сигнал усиления транскрипции “CCAAT-бокс” в промоторах генов (сокращение от англ. “**C**CAAT **E**nhancer-**B**inding **P**rotein”)

c-Fos −транскрипционный фактор, продукт гена *FOS*, впервые найденный у цыпленка (сокращение от англ. “**c**hicken **– FOS** ”)

c-Jun −транскрипционный фактор, продукт гена *Jun*, который был впервые выделен как клеточный аналог белка трансформации ретровируса цыпленка ASV17 (сокращение от английского “**c**hicken retrovirus ASV17 **– JUN** ”)

c-Myb −транскрипционный фактор, впервые выделенный в 1979 г. как регулятор дифференцировки миелобластов у цыпленка (сокращение от английского “**c**hicken **– My**elo**b**last ”)

COMPEL −база данных по композиционным элементам регуляции транскрипции генов эукариот, созданная в ИЦиГ СО РАН (сокращение от английского “**COMP**osite **EL**ement ”)

COUP −транскрипционный фактор, открытый в 1986 г. как ядерный белок, который связывает промотор гена овальбумина цыпленка между -90 и -70 его позициями перед стартом транскрипции (сокращение от англ. “**C**hicken **O**valbumin **U**pstream **P**romoter”)

CP1 −транскрипционный фактор, открытый в 1987 г. в качестве белка, связывающего центромеру *Saccharomyces cerevisiae* и имеющего высококонсервативного гомолога у человека (сокращение от английского “**C**entromere-binding **P**rotein **1**”)

CRE −ДНК-сигнал ответа на циклоаденазинмонофосфат, цАМФ, (сокращение от английского “**c**AMP **R**esponse **E**lement”)

CREB −общее обозначение продуктов семейства генов, кодирующего семь транскрипционный факторов, первый из которых был открыт в 1987 г. как ядерный белок регуляции экспрессии гена стоматостатина в качестве ответа на циклоаденазинмонофосфат, цАМФ, (сокращение от англ. “**c**AMP **R**esponse **E**lement-**B**inding”)

CRE-BP1 −синоним CREB (сокращение от английского “**c**AMP **R**esponse **E**lement-**B**inding **P**rotein **1**”)

CRF − циркулирующая рекомбинантная форма ВИЧ-1, найденная более чем у одного пациента (сокращение от английского “**C**irculating **R**ecombinant **F**orm”)

D(ξ) −размах переменной ξ (разница между максимумом и минимумом)

DPE −нижерасположенный элемент промотора (сокращение от англ. “**D**ownstream **P**romoter **E**lement”)

E2 −общее название продуктов и изоформ транскрипционных факторов, кодируемых семейством генов, первый из которых был найден в 1990 г. как продукт открытой рамки считывания E2 в геноме вируса папиломы быка и был авторепрессором, а также репрессором раннего промотора P1 этого вируса (сокращение от английского “**E2** open reading frame of the bovine papillomavirus”)

E2F −общее обозначение продуктов семейства, состоящего из 9 генов, кодирующих ключевые для клеточного цикла транскрипционные факторы, первый из которых был открыт как белок, связывающий ранний промотор E2 аденовируса (сокращение от английского “adenovirus **E2** promoter binding **F**actor”)

EMBL −Европейская Молекулярно-Биологическая Лаборатория (сокращение от английского “**E**uropean **M**olecular **Bi**ology **L**aboratory”)

EMSA −метод экспериментальных измерений “задержка в геле” (сокращение от англ. “**E**lectrophoretic **M**obility **S**hift **A**ssay”)

EN −транскрипционный фактор из семейства гомеодомен-содержащих генов регуляции пути сегментации в эмбриогенезе и морфогенеза дрозофилы, которое было открыто в 1988 г. (сокращение от английского “**EN**grailed homeodomen ”)

ER −общее обозначение изоформ α и β транскрипционного фактора “рецептор эстрогена” из семейства ядерных рецепторов гормонов “G белки” (сокращение от английского “**E**strogen **R**eceptor”)

Ets −общее обозначение продуктов одного из самых многочисленных семейств E26 генов ткане-специфических транскрипционных факторов, названного по обозначению линии вируса лейкемии, с использованием которой был открыт первый ген этого семейства (сокращение от английского “**E**-**t**wenty **s**ix”)

Fpg −8-оксогуанин-ДНК гликозилаза бактерий для распознавания в ДНК 7,8-дигидро-8-оксогуанина (охоG) и инициации ряда последовательных шагов репарации охоG→G. OGG1. Fpg не является гомологом функционально сходного ему фермента OGG1 у человека; Fpg и OGG1 не имеют также 3D-сходства

GAGA −транскрипционный фактор, который был открыт в 1988 г. как регулятор экспрессии гена Ultrebuthorax в эмбриогенезе дрозофилы, связывающий тракты $(\mathbf{GA})_n$, что авторы открытия этого транскрипционного фактора зафиксировали в его названии

GAL4 −транскрипционный фактор, который был открыт в 1979 г. на *Saccharomyces cerevisiae* в качестве белка-регулятора (активатора) путей метаболизма галактозы (сокращение от английского “**GAL**actose pathway regulator **4**”)

GATA −общее обозначение продуктов семейства, включающего, по меньшей мере, 6 генов, которые кодируют транскрипцилонные факторы, названные по инвариантному тетрануклеотиду GATA в составе сайтов их связывания белков в составе геномной ДНК

GR −транскрипционный фактор “рецептор глюкокортикодов” (например, кортизола), кодируемый геном из семейства ядерных рецепторов внутриклеточных стероидных гормонов (сокращение от английского “**G**lucocorticoid **R**eceptor”)

HeLa −лабораторная линия клеток рака матки человека.

HNF1 (HNF3)	−транскрипционный фактор из семейства генов ядерных факторов гепатоцитов печени (сокращение от английского “**H**epatocyte **N**uclear **F**actor **1** (**3**)”)
HSF	−транскрипционный фактор ответа на тепловой шок (сокращение от английского “**H**eat **S**hock **F**actor”)
hsp 70	−белок теплового шока с молекулярной массой 70 кДа (сокращение от английского “**H**eat **S**hock **P**rotein”)
InDel	−общее обозначение двух типов точечных мутаций: вставок (**In**sertion) и делеций (**Del**etion) фрагментов геномных ДНК и РНК
Inr	−участок ДНК вокруг старта транскрипции эукариот (инициатор)
in silico	−компьютерное моделирование биологического эксперимента, по аналогии с латынью *in vivo* (опыт в живом организме) и *in vitro* (опыт в пробирке). Введено в 1989 г. Педро Мирамонтесом (Pedro Miramontes) в его устном докладе о применении теории клеточных автоматов фон Неймана к компьютерному моделированию молекулярной эволюции.
IRF1	−транскрипционный фактор, который был открыт в 1988 г. как ядерный белок, регулирующий транскрипцию гена интерферона β мыши посредством связывания с промотором этого гена (сокращение от английского “**I**nterferon **R**egulatory **F**actor-**1**”)
IT	− информационно-технологический (сокращение от “**I**nformation **T**echnology”, англ. яз.)
K-ras	−ген, принадлежащий семейству генов **ras** и впервые выделеный из клеток саркомы Кирстена (от английского “**K**irsten’s **ras**”)
Lrp	−белок регуляции ответа на лейцин бактерии *Escherichia coli* (сокращение от англ. “**L**eucine-responsive **r**egulatory **p**rotein”)
$M_0(\xi)$	−среднеарифметическое переменной величины ξ.
mAb	−моноклональное антитело (акроним от английского “**m**onoclonal **A**nti**b**ody”)
MEF-2	−общее название продуктов семейства генов, которые кодируют транскрипционные факторы регуляции развития клеток мышц и ответа на гипертонический стресс (сокращение от английского “**M**yocyte **E**nhancer **F**actor−**2**”)
Mg	−магний

MyoD −транскрипционный фактор регуляции дифференцировки клеток мышц (сокращение от английского “**Myo**genic **D**ifferentiation”)

$N[M_0(\xi);\sigma(\xi)]$ −выборочная оценка нормального распределения переменной ξ со среднеарифметическим $M_0(\xi)$ и вариансой $\sigma(\xi)$.

NF-1 −транскрипционный фактор, открытый в 1985 г. как ядерный белок, который связывается с промотором гена *IGHM*, кодирующего мю-цепь иммуноглобулинов, и с терминатором репликации аденовируса (сокращение от англ. “**N**uclear **F**actor **1**”)

NF-E2 −эритроид-специфический ядерный фактор 2 (сокращение от английского “**N**uclear **F**actor **E**rythroid-derived **2**”)

NF-IL6 −транскрипционный фактор, кодируемый геном из семейства CEBP, который был открыт как ядерный фактор регуляции экспрессии гена интерлейкин-6 (сокращение от английского “**N**uclear **F**actor for **I**nter**L**eukin-**6**”)

NF-κB −транскрипционный фактор, являющийся главным регулятором (называемый “мастер-ген”) иммунного ответа на инфекцию (сокращено от английского “**N**uclear **F**actor **kappa**-light-chain-enhancer of activated **B** cells”)

NGS −высокопроизводительнное секвенирование (сокращение от англ. “**N**ext-**G**eneration **S**equencing)

OCT −общее название продуктов семейства, включающего, по меньшей мере, 8 генов, которые кодируют транскрипционные факторы в форме октамера, связывающего инвариантный октонуклеотид "ATTTGCAT" в составе тканеспецифических промоторов, открытый в 1984 г. (сокращение от английского “**OCT**amer”)

OGG1 −8-оксогуанин-ДНК гликозилаза человека (суперсемейство эндонуклеазы III), фермент для распознавания 7,8-дигидро-8-оксогуанина (oxoG) в ДНК и первого в ряду последовательных шагов репарации oxoG→G, который не является гомологом функционально сходного фермента Fpg у бактерий и не имеет пространственного сходства с этим бактериальным ферментом.

oxoG −7,8-дигидро-8-оксогуанин, один из самых часто встречамых результатов окисления гуанина (G) в ДНК, который имеет сильный мутагенный потенциал в силу комплементарности как oxoG:C, так и oxoG:A с последующей заменой в итоге следующей репликации, в целом: G:C→oxoG:A→T:A. Репарацию oxoG в G инициирует фермент 8-оксогуанин-ДНК гликозилаза OGG1 у человека или Fpg у бактерий, не являющиеся гомологами и имеющие разные пространственные структуры.

PR −транскрипционный фактор “рецептор прогестерона” из семейства ядерных рецепторов внутриклеточных стероидных гормонов (сокращение от английского “**P**rogesterone **R**eceptor”)

r −коэффициент линейной корреляции Пирсона, называемой также “простая корреляция” в случае сравнения двух переменных.

RAR −общее название продуктов и их изоформ семейства из трех генов (α, β и γ) транскрипционных факторов “рецептор ретиноевой кислоты” (сокращение от английского “**R**etinoic **A**cid **R**eceptor”)

RecA −многофункциональный белок у *Escherichia coli*, филаментация которого, в частности, на флангах бреши в нити ДНК, возникшей на месте вырезанного из нее предмутационного повреждения этой нити, запускает экспрессию генов SOS-системы репарации ДНК.

RFX −транскрипционный фактор, открытый в 1990 г. как регуляторный белок, который связывается с высококонсервативным X-боксом генов класса II главного комплекса гистосовместимости (сокращение от английского “**R**egulatory **F**actor binding to the **X**-box of MHC class II genes”)

RNAPII −РНК-полимераза II

RXR −общее название продуктов семейства из трех генов, кодирующих транскрипционные факторы “ретиноевый-X рецептор” α, β и γ, каждый из которых является ко-репрессором в составе его гетеродимера с другим ядерным рецептором X для совместной с ним репрессии соответствующих генов в случае одновременного отсутствия как ретиноевой кислоты, так и субстрата активации этого ядерного рецептора-партнера X (сокращение от английского “**R**etinoic **X** **R**eceptor”)

shotgun −стратегия секвенирования протяженного района геномной ДНК путем сегментации его на короткие перекрывающиеся фрагменты оптимальной длины от 30 п.о. до 350 п.о., называемые “контиги”, каждый из которых, затем, независимо клонируют, секвенируют и, наконец, из всех расшифрованных таким образом фрагментов собирают нуклеотидную последовательность всего исследуемого района геномной ДНК, результат чего часто неустойчив в случае насыщенности этого района прямыми совершенными повторами.

SNP −однонуклеотидный полиморфизм (сокращение от английского “**S**ingle **N**ucleotide **P**olymorphism”)

SP1 −транскрипционный фактор, специфичный для промоторов генов ранних стадий дифференцировки, который был открыт в 1983 г. как компонент №1 экстракта ядер клеток HeLa, активирующий транскрипцию с раннего промотора SV40 (сокращение от английского “**S**pecific for **P**romoter component № **1**”)

SRF −транскрипционный фактор, который регулирует экспрессию генов клеточного цикла, апоптоза, роста и дифференцировки клеток (сокращение от английского “**S**erum **R**esponse **F**actor”)

SV40 −обезьяний вирус 40 (сокращение от английского “**s**imian **v**irus **40**”)

$t_{\alpha;\nu}$ −Табличное значение t-критерия Стьюдента для заданных уровня значимости α и количества степеней свободы ν

T3R −транскрипционный фактор “рецептор тироидного гормона Т3” (сокращение от англ. “activated thyroid hormone **T3 R**eceptor”)

TBP −TATA-связывающий белок (сокращение от английского “**T**ATA-**b**inding **p**rotein”, первоначально “**T**ATA-**b**inding **p**olypeptide”), ДНК-связывающая субъединица РНК-полимеразы II (RNAPII)

TCF-1 −транскрипционный фактор, который был открыт в 1991 г. как специфический для Т-лимфоцитов (сокращение от английского “**T** lymphocyte **C**ell-specific **T**ranscription **F**actor **1**”)

TDO2 −ген, кодирующий триптофан 2,3-диоксигеназу (EC 1.13.11.11)

TESS −одно из наиболее часто используемых инструментальных средств распознавания неизвестной локализации сайтов связывания транскрипционных факторов по известной последовательности геномной ДНК (сокращение от англ. “**T**ranscription **E**lement **S**earch **S**oftware”), который был создан (Schug, Overton, 1997) в Пенсильванском университете (Филадельфия, США)

TRRD −база данных по районам регуляции транскрипции генов эукариот (сокращение от англ. “**T**ranscription **R**egulatory **R**egion **D**atabase”), созданная в ИЦиГ СО РАН (Новосибирск, Россия)

TTF-1 −транскрипционный фактор, специфический для мозга и легких, а также для тироидной системы (сокращение от английского “**T**hyroid **T**ranscription **F**actor **1**”)

UPGA −метод кластер-анализа “невзвешенное среднее сходство для пары (вершин-потомков при их объединении в одну вершину-предок)” (сокращение от английского “**U**nweighted **P**air-**G**roup **A**veraging”), который является предустановленным разработчиками пакета Statistica (StatSoft™, Tulsa, USA) как очень часто используемый.

USF −общее обозначение продуктов семейства генов, кодирующих транскрипционные факторы, первый из которых был открыт в 1985 г. как белок-активатор главного позднего промотора аденовируса, (сокращение от английского “**U**pstream adenovirus major late promoter TATA box **S**timulatory **F**actor”)

vs −против, в сравнении (сокращение от латыни “**v**ersu**s**”)

WPGA −метод кластер-анализа “взвешенное среднее сходство для пары (вершин-потомков при их объединении в одну вершину-предок)” (сокращение от английского “**W**eighted **P**air-**G**roup **A**veraging”)

WPGC −метод кластеризации “взвешенное центрированное сходство для пары (вершин-потомков при их объединении в одну вершину-предок)” (сокращение от англ. “**W**eighted **P**air-**G**roup **C**entroid”)

WT −дикий тип (сокращение от английского “**w**ild **t**ype”)

yEGFP −адаптированный для дрожжей зеленый флуоресцентный белок (сокращение с англ. “**y**east **E**nhanced **G**reen **F**luorescent **P**rotein”)

YY1 −активирующий/репрессирующий транскрипционный фактор 1 (сокращение от английского “**Y**in-**Y**ang **1**”, “Инь-Янь 1”)

Zn	−цинк
α	−уровень значимости (для статистического критерия)
$\delta(\xi)$	−стандартизированная ошибка среднего переменной величины ξ
$\Delta\Delta G$	−изменение свободной энергии Гиббса
$\Delta(\xi)$	−функция-индикатор: {Δ(истина)=1; Δ(ложь)=0}
ν	−число степеней свободы (для статистического критерия)
τ	−коэффициент ранговой корреляции Кендалла
$\sigma(\xi)$	−стандартизированное отклонение переменной величины ξ
3D	−пространственный (сокращение с англ. “**three-d**imensional”)
Ø	−пустое множество