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

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора биологических наук

Новосибирск – 2017

Работа выполнена в лаборатории эволюционной биоинформатики и теоретической генетики ФГБНУ “Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук” (ИЦиГ СО РАН), г. Новосибирск, Россия

Научный консультант: академик РАН,
доктор биологических наук, профессор
Колчанов Николай Александрович

Официальные оппоненты: **Макеев Всеволод Юрьевич**,
член-корреспондент РАН, доктор физико-математических наук, заведующий отделом вычислительной системной биологии,
ФГБУН Институт общей генетики РАН им. Н.И. Вавилова, г. Москва

Дубина Михаил Владимирович,
академик РАН, доктор медицинских наук, профессор, руководитель отдела молекулярно-генетических и нанобиологических технологий,
ФГБОУ ВО ПСПбГМУ им. И. П. Павлова Минздрава России, г. Санкт-Петербург

Бажан Сергей Иванович,
доктор биологических наук,
заведующий теоретическим отделом,
ФБУН ГНЦ вирусологии и биотехнологии «Вектор» Роспотребнадзора, п. Кольцово Новосибирской обл.

Ведущее учреждение: ФГУ ФИЦ Биотехнологии РАН, г. Москва

Защита состоится “___” _____ 2017 г. на утреннем заседании диссертационного совета Д 003.011.01 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук в ИЦиГ СО РАН по адресу: 630090 г.Новосибирск, пр.ак. Лаврентьева, 10. Тел/факс: (383)3634906; Факс: (383)3331278, e-mail**: dissov@bionet.nsc.ru.**

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института: **www.bionet.nsc.ru**

Автореферат разослан “____”_____ 2017 г.

Ученый секретарь диссертационного совета,
доктор биологических наук Т.М. Хлебодарова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

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 новой 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 TATA-связывающего белка к TATA-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, значимо 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 О.В. Аркова, Т.В. Аршинова, В.П. Валуев, Г.В. Васильев, Д.В. Воробьев, Д.А. Григорович, И.А. Драчкова, В.М. Ефимов, С.В.

Зубова, Л.В. Катохина, А.Э. Кель, В.Ф. Кобзев, Ф.А. Колпаков, А.Н. Колчанова, Н.А. Колчанов, С.В. Лаврюшев, М.В. Лысова, Т.И. Меркулова, Г.В. Орлова, С.Е. Пельтек, Е.Л. Перегоедова, О.А. Подколодная, Н.Л. Подколодный, П.М. Пономаренко, Ю.В. Пономаренко, Д.А. Рассказов, Л.К. Савинкова, В.В. Суслов, И.И. Титов, А.С. Фролов, Д.П. Фурман (все - ИЦиГ СО РАН, г. Новосибирск), 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), C. Overton, A.V. Mazin и S.C. Kowalczykowski (все -USA), H. Karas, H. Sclenar и E. Wingender (все - Germany), L. Milanesi (Italy), A. Sarai (Japan).

Личный вклад автора. Все представленные в диссертации результаты были получены автором самостоятельно. Роль автора в статьях, включенных в "Список публикации по теме диссертации", хотя он не был в них автором для переписки, первым или последним автором, была "компьютерный анализ данных" согласно теме диссертационной работы "Компьютерный анализ контекстно-зависимых количественных характеристик специфической биологической активности сайтов в составе геномной ДНК". Работы автора в "Списке литературы", которых нет в "Списке публикации по теме диссертации", были сделаны вне диссертационной работы в рамках исследований лаборатории эволюционной биоинформатики и теоретической генетики ФГБНУ ФИЦ ИЦиГ СО РАН.

Структура и объем работы. Диссертация включает введение, главы "Обзор литературы» и четыре главы, которые содержат материалы, методы, результаты и обсуждения как разделы этих глав, заключение, выводы, список литературы (467источников), список терминов, 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

В главе 1 введено два основных понятия - "нуклеотидная последовательность ДНК" и "количественная характеристика последовательности ДНК", - которые

используются в диссертации. В этих терминах рассмотрены базы данных по геномным последовательностям ДНК, компьютерные системы и используемые в них методы анализа данных, существующие к моменту начала диссертационной работы. На этой основе в “Заключении по обзору литературы” обоснован выбор нового направления математической биологии и биоинформатики, которое было предложено и развитого автором в рамках настоящей диссертационной работы.

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

Поскольку к началу диссертационной работы в литературе были свидетельства как в пользу прогноза количественных величин специфической биологической активности сайтов в составе геномных ДНК на основе матриц частот символов-нуклеотидов в позициях этих сайтов, так и против таких прогнозов, то введен компромисс в форме линейно-аддитивного приближения (Колчанов и др., 1998):

$$F_{\text{опыт}}(S_{\text{ДНК}}) = F_0 + \sum_{n=1}^{N} \varphi_n X_n(S_{\text{ДНК}}). \quad (1)$$

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

С помощью формулы (1) переформулировали целевую задачу предсказания количественных величин $F_{\text{опыт}}(S_{\text{ДНК}})$ специфической биологической активности сайтов в составе геномных ДНК по нуклеотидной последовательности этих сайтов в новую задачу поиска контекстно-зависимых количественных величин $\{X_n(S_{\text{ДНК}})\}_{1\le n\le N}$ ДНК, которые значимо линейно коррелируют с количественными величинами $F_{\text{опыт}}(S_{\text{ДНК}})$ заданной активности этих сайтов. Для применения этой формулы создали базу данных PROPERTY (Колчанов и др., 1998) по 38 физико-химическим и конформационным свойствам динуклеотидных шагов спирали ДНК (Рисунки 1 и 2, Таблица 1), а также ввели их усреднение в районе [a; b] сайта:

$$X_{k[a;b]}(S = \{s_1 \dots s_a \dots s_i \dots s_b \dots s_L\}) = \sum_{i=a}^{b-1} X_k(s_i s_{i+1})/(b-a-1). \quad (2)$$

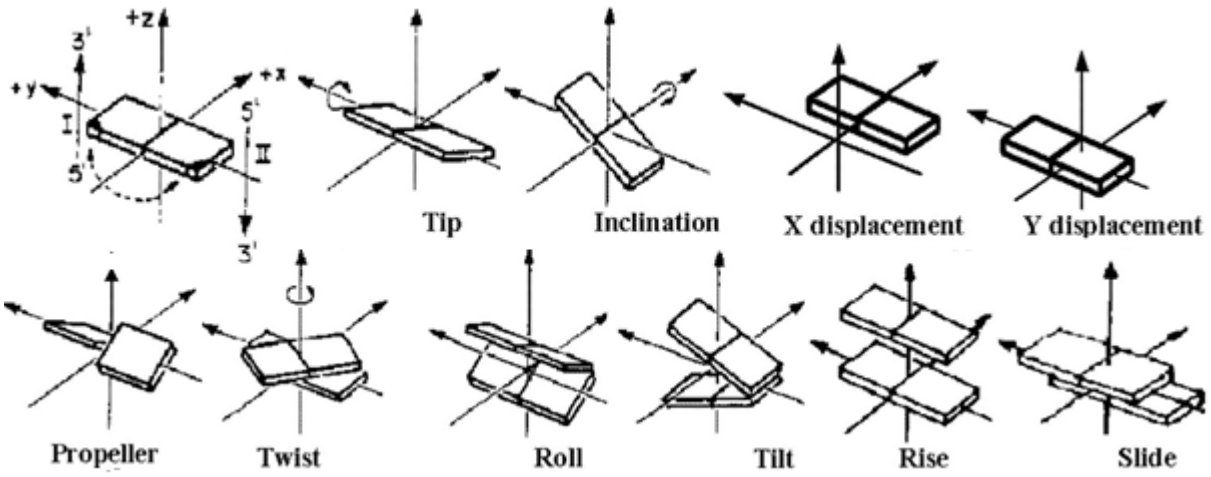

Рисунок 1 - Номенклатура конформационных количественных характеристик для B-формы спирали ДНК (Dickerson *et al.*, 1989).

Таблица 1 – Примеры свойств динуклеотидных шагов спирали ДНК, собранные в базе данных PROPERTY (Колчанов и др., 1998) в рамках диссертации.

Название свойства	Единицы	min	max	Литература
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)
Свободная энергия Гиббса, ΔG	Kcal/mol	-2.8	-0.9	(Sugimoto *et al.*, 1996)
Конформационные свойства (Рисунок 1):				
Угол кручения, *twist*	градус	27.7	40.0	(Shpigelman *et al*, 1993)
Угол перекреста, *propeller*		-17.3	-6.7	(Gorin *et al.*, 1995)
Угол изгиба оси спирали, *bend*		2.16	6.74	(Karas *et al.*, 1996)
Угол раскрытия, *roll*		-2.0	6.5	(Suzuki, Yagi, 1995)
Сдвиг по длинной оси, *slide*	ангстрем	-0.37	1.46	(Gorin *et al.*, 1995)
Шаг спирали вдоль оси, *rise*		3.16	4.08	(Karas *et al.*, 1996)
Ширина малой бороздки		4.62	6.40	
Глубина большой бороздки		8.45	9.60	

Для сайта длины 120 п.о., $\{X_{k;[a,b]}(S_{ДНК})\}_{1\le k\le 30;1\le a<b\le 120}$, с помощью формулы (2) можно вычислить 38*119*118/2=56168 вариантов количественных величин. В главе 2 описана компьютерная система bDNAvideo для поиска таких $X_{k;[a,b]}(S_{ДНК})$, которые коррелируют с простейшей бинарной биологической активностью: $F_{опыт}(S^+) \equiv 1$ для сайта S^+ связывания заданного транскрипционного фактора и $F_{опыт}(S^-) \equiv -1$ для последовательности S^- случайных равновероятных независимых нуклеотидов (здесь и далее: “случайных ДНК”). Входные данные для bDNAvideo – это две соответствующие выборки последовательностей ДНК

а)
```
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
```

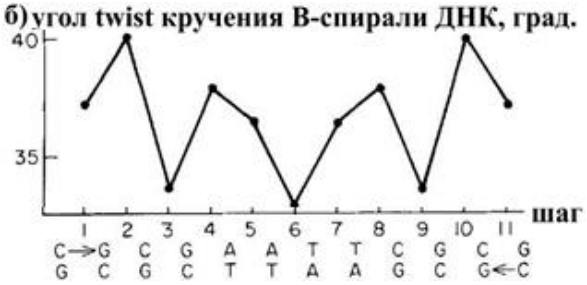

Рисунок 2 – Пример (а) документа базы данных PROPERTY (Колчанов и др., 1998) о свойстве “Угол кручения, twist” спирали ДНК и (б) график этого свойства для последовательности 5’-cgcgaattcgcg-3’.

длины 120 п.о. $\{S^+\}$ и $\{S^-\}$. В рамках теории (Fishburn, 1970) на этих выборках оценивали полезность $U(X_{k;[a,b]}; \{S^+\}; \{S^-\})$ учета в формуле (1) каждого $X_{k;[a,b]}$ независимо от всех 56167 остальных вариантов.

Анализ столь большого числа $X_{k;[a,b]}$ был вызван отсутствием данных о влиянии спирали ДНК на связывание транскрипционных факторов. Для $X_{k;[a;b]}$ строили распределения $p(X_{k;[a;b]}(S^+))$ и $p(X_{k;[a;b]}(S^-))$ для выборок $\{S^+\}$ и $\{S^-\}$ (Рисунок 3). С их помощью проверяли 4 критерия применимости дискриминантного анализа к входным данным: (i) нормальность $p(X_{k;[a;b]}(S^+))$; (ii) нормальность $p(X_{k;[a;b]}(S^-))$; (iii) достоверность различия между $p(X_{k;[a;b]}(S^+))$ и $p(X_{k;[a;b]}(S^-))$; (iv) достоверность различия между средними $M_0\{p(X_{k;[a;b]}(S^+))\}$ и $M_0\{p(X_{k;[a;b]}(S^-))\}$; а также критерия χ^2 (v) и точный критерий Фишера (vi) для бинарной дискриминации относительно порога $(M_0\{p(X_{k;[a;b]}(S^+))\}+M_0\{p(X_{k;[a;b]}(S^-))\})/2$. Каждый критерий проверяли в 100 bootstrap-испытаниях (Efron, 1979) на случайных подвыборках 50%-объема из входных выборок $\{S^+\}$ и $\{S^-\}$ (Рисунок 3). Всего 100 bootstrap-испытаний шести критериев давали 600 оценок α достоверности, которые были проектированы в шкалу [-1; 1] полезности, α→U(α), в рамках нечетких множеств (Zadeh, 1965) как:

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

где: 1.3, −28.3, 55.6 – коэффициенты сплайна, проходящего через общепринятый порог значимости $\{U_{qz} \equiv 0$ при $\alpha_{qz} \equiv 0.05\}$ и непрерывного в двух его крайних точках $\{U_{qz} \equiv -1$ при $\alpha_{qz} \equiv 0.1\}$ и $\{U_{qz}(X_k) \equiv 1$ при $\alpha_{qz}(X_k) \equiv 0.01\}$, как показано на Рисунке 4.

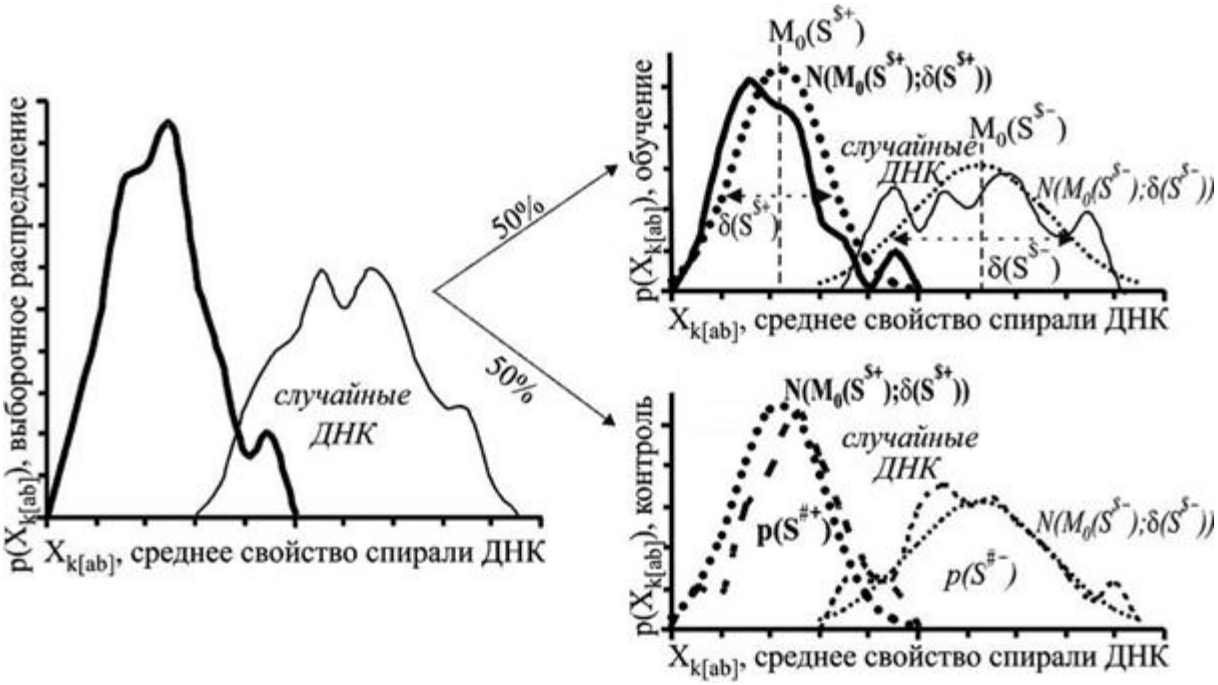

Рисунок 3 – Оценка достоверности α по критерию χ^2 для соответствия между выборочным $p(X_{k;[a;b]}(S^{@}))$ распределением и нормальным $N[M_0(X_{k;[a;b]}(S^{@}));\ \delta(X_{k;[a;b]}(S^{@}))]$ со средним M_0 и стандартным отклонением δ (здесь: @ ϵ {+, −}) в bootstrap-испытаниях (Efron, 1979) случайного разбиения выборки $\{S^{@}\}$ на обучающую $\{S^{\$@}\}$ и контрольную $\{S^{\#@}\}$ части 50%-объема.

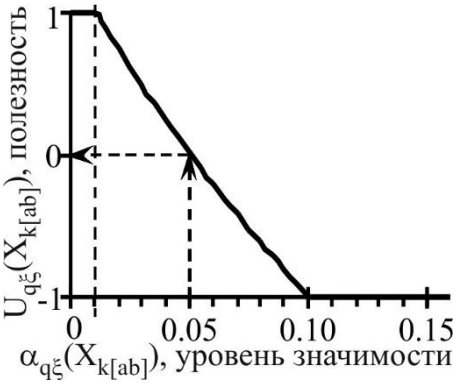

Рисунок 4 - Шкала полезности $U_{q\xi}(X_{k[ab]})$ среднеарифметической оценки $X_{k[ab]}$ для k–го свойства конформации на участке [a; b] спирали ДНК для отличия сайтов связывания транскрипционных факторов от случайных ДНК при достоверности $\alpha_{q\xi}(X_{k[ab]})$ q-го критерия в ξ-ом испытании.

Наконец, в рамках теории полезности для принятия решений (Fishburn, 1970) оценили $X_{k[a;b]}$ среднеарифметическим всех частных оценок ее полезности:

$$U\left(X_{k;[a;b]}\right) = \frac{1}{600}\sum_{q=1}^{6}\sum_{\xi=1}^{100} U_{q\xi}\left(X_{k;[a;b]}\right),. \qquad (3)$$

которое обладает следующими двумя важными асимптотическими свойствами:

<u>если</u> $U(X_{k';[a';b']})\leq 0$ **<u>то</u>** $X_{k';[a';b']}$ бесполезна
<u>если</u> $U(X_{k;[a;b]})>U(X_{k';[a';b']})>0$ **<u>то</u>** $X_{k;[a;b]}$ полезнее $X_{k';[a';b']}$ для отличия S^+ от S^-. (4)

Свойство в верхней строке формулы (4) позволяет удалять бесполезные $X_{k'[a'b']}$ с негативной полезностью $U(X_{k'[a'b']}) \leq 0$, свойство в нижней строке – выявить лучшие X^0_k с наибольшей позитивной полезностью $U(X^0_k)=\max_{ab}\{U(X_{k[ab]})\}>0$ для каждого из 38 физико-химических и конформационных свойств спирали ДНК. Список $\{X^0_k\}$ – это "выходные данные" системы bDNAvideo. Случай $\{X^0_k\} = Ø$ – отсутствие достоверных отличий между исследуемыми сайтами связывания транскрипционных факторов и случайными ДНК по k-ому свойству спирали ДНК. Биномиальное распределение оценивает вероятность исхода $\{X^0_k\}\neq Ø$, как:

$$P_{X_{k;[a;b]}}\left(U\left(X_{k;[a;b]}\right)>0\right)=\sum_{q=1+600/2}^{600} C_{600}^{q}\times 0.05^{q}(1-0.05)^{600-q}<10^{-45}, \qquad (5)$$

что с поправкой Бонферрони дает нижнюю оценку значимости исхода $\{X^0_k\}\neq Ø$:

$$P\left(U\left(X_{k;[a;b]}\right)>0\right)<10^{5}P_{X_{k;[a;b]}}\left(U\left(X_{k;[a;b]}\right)>0\right)<10^{5}\times 10^{-45}<10^{-40}. \qquad (6)$$

Результатом bDNAvideo для сравнения сайтов связывания транскрипционного фактора EN со случайными ДНК было 10 свойств спирали ДНК (Таблице 2).

Наибольшую $U(X_{14,[-10,2[})=0.989$ получил достоверно низкий средний угол roll в комплексе белок/ДНК (Suzuki et al., 1996) на участке [-10; 2] относительно центра сайта связывания EN, установленного методом футпринтинга. Он был 2.26 ± 0.20^{O} для сайтов связывания EN достоверно (Рисунок 5: $\chi^2=172.11$, $\alpha<0.0005$) меньше, чем в случайных ДНК, 2.72 ± 0.04^{O}, в согласии с известной пространственной структурой комплекса ДНК/EN (Kissinger et al, 1990).

Аналогичные результаты были получены для 1819 сайтов связывания 42 транскрипционных факторов (Таблица 3). Всего было выявлено 848 значимых

Таблица 2 – Результат сравнения 12 сайтов связывания EN и случайных ДНК.

Среднее значение конформационного свойства спирали ДНК, $X_{k[a;b]}$ (формула 2)				U, ф-ла	среднее ± ст.ош.средн.	
k	Название свойства, *обозначение*	единицы	[a; b]	(31)	EN сайт	случайные
14	Раскрытие по длинной оси, *roll*	градус	[−10; 2]	0.989	2.26±0.20	2.72±0.04
22	Температура плавления, T_M	^{O}C	[−10; 3]	0.886	64.10±3.33	73.43±0.59
24	Гибкость по большой бороздке	log-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	Угол изгиба 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 вдоль длинной оси, *slide*	ангстр.	[−1; 20]	0.657	−0.03±0.02	-0.06±0.004
35	Дисбаланс размеров бороздок	log-ед.	[−9; 5]	0.484	1.03±0.09	1.10±0.01
38	Свободная энергия Гиббса, *ΔG*	kcal/mol	[−10; 0]	0.440	−1.42±0.15	−1.61±0.02

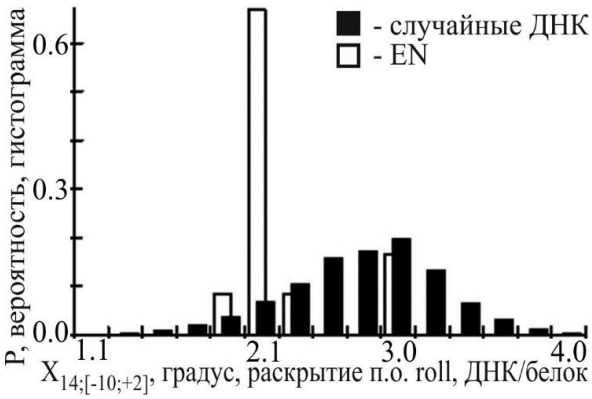

Рисунок 5 – Результат bDNAvideo для сайтов связывания транскрипционного фактора EN (□) и случайных ДНК (■): наибольшая $U(X_{14,[-10,2]})=0.989$ была у среднего *roll* в комплексе белок/ДНК (Suzuki *et al.*, 1996) на участке [−10; 2] от центра сайта по методу футпринтинга.

контекстно-зависимых количественных характеристик спирали ДНК для этих сайтов. В 516 из 848 случаев средние значения свойств для сайтов связывания транскрипционных факторов были выше, чем таковые для случайных ДНК как это показано символом “+”, в 332 случаях – ниже (“-”).

На Рисунке 6 показан результат пакета Statistica (Statsoft™, USA) для данных из Таблицы 3 (параметры: мера Евклида и метод UPGA), устойчивый ко всем 42 сочетаниям 6 методов кластеризации с 7 мерами сходства этого пакета. Правый подграф содержит преимущественно Zn-координированные и основные белки с локальным избытком электростатического заряда (80%), тогда как левый подграф содержит преимущественно β-слои и гомеодомены без такого избытка (64%). Это различие является достоверным ($\alpha<0.01$) по точному критерию Фишера.

Таблица 3. Результат bDNAvideo для сайтов связывания 42 транскрипционных факторов

№	Фактор транскрипции / Свойство B-ДНК (Рисунок 1)	основный домен												Zn-координируемый														гомеодомен									β-слой								
k		AP-1	cFos	cJun	NFE2	CAREBP1	ACREBP	CCAAT	NFIL6	Myo D	E2F	USF	UNCF1	RFX1	PPAR	EGR	GRE	PRAX	RRR	RTG3	TCOUP	GSAP1	SpA	GAGA	GYAY1	GATA4	CHTF1	CNTF1	TEF-1	TNF	HNF3	HNF	c-Myb	Ets	EIRF1	NFκB	MEF2	SER	E2F	TFIID	TFC	TFIIF1
1	Кручение спирали	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+
2	Шаг спирали					+	−	−			+	+	−		−			+	+		−	+					−	−	−			−				+	−	+			−	+
3	Угол изгиба оси	−		−		−	+			−	−	−	+		−			−	−	−	−	+	−	−		−		+	+	−			−				+	−	+			+	−
4	Наклон tip	−	−	−	−	−	−	−			−		−	−	−	−	−	−	−	−	−	−	−	−	−	−	−		+	−	+	+			−	−	+	+					
5	Наклон inclination						−	+		−	+	−	−		−											+		+	+		−	−					+			−				
…	…	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	Т плавления				+	+	+	−	−	+	+	+	+	−	+			+	+			−	+	+	+			−	−		−	−					−	+	−	−	+	−		
23	Контакт с гистоном	+	+	+		+	+	+	+	+	+	+	+	−	+			+	+	+	+	+	+	+	−			+	+	+	+	+					+	+			+	+		
…	…	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
35	Различие бороздок			+		+	−	−	+	−	+				+	−		+	+		−	+	+	+				−	−	−	+	+			−		+	−		+	−	+		
36	Энтальпия	+				−	+	+		−	−			+	+			+	+			+	−					+	+	+		+			+		+	−	+	+				
37	Энтропия	+				−	+	+		−	−				+	+		+	+			+	−	+				+	+	+		+			+		+	−	+	+				
38	Свободная энергия					−	+	+		−	−	−	+		+	+		+	+			+	−					+	+	+	+	+			+		+		+	+				

“+”/“−” – среднее значение для сайтов достоверно выше/ниже, чем для случайных ДНК.

Рисунок 6 – Граф сходства транскрипционных факторов по свойствам спирали ДНК сайтов их связывания (пакет Statistica (USA): мера Евклида, метод UPGA).

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

Представленная в предыдущей главе система bDNAvideo успешно решила задачу поиска количественных характеристик геномной ДНК, которые дискриминируют сайты S^+ связывания заданного транскрипционного фактора ($F_{опыт}(S^+)\equiv 1$) от для случайных ДНК, S^- (где: $F_{опыт}(S^-)\equiv -1$). В главе 3 описано как на основе системы bDNAvideo создали систему Activity, заменив критерии дискриминантного анализа на критерии корреляционного анализа (Рисунок 7).

Входные данные Activity являются выборкой из N пар $\{S_n; F(S_n)\}_{1\le n\le N}$, каждая из которых включает последовательность ДНК и экспериментально измеренную количественную величину биологической активности исследуемых сайтов. Для каждой контекстно-зависимой характеристики X_m проверяется пять критериев корреляции между $F(S_n)$ и $X_m(S_n)$: (i) линейная, ранговые (ii) Спирмена и (iii) Кендалла, бинарные (iv) χ^2 и (v) Фишера. Проверяется также шесть критериев применимости корреляционного анализа к входным данным: (vi) нормальность $p(X_m(S_n))$; (vii) нормальность $p(F(S_n))$; (viii) нормальность отклонений от прямой регрессии $p(\Delta_F=F(S_n)-\lambda_0-\lambda_1X_m(S_n))$; (ix) нормальность отклонений от сопряженной регрессии $p(\Delta_X=X_m(S_n)-\mu_0-\mu_1F(S_n))$; (x) независимость знаков $\{sign[\Delta_F(S_n)]\}$; (xi) независимость знаков $\{sign[\Delta_F(S_n)]\}$. Каждый из этих 11 критериев проверяется в семи bootstrap-испытаниях: (i) на всех данных; на 50%-подвыборках (ii) меньших $F(S_n)$; (iii) больших $F(S_n)$; (iv) ближайших к $M_0\{F(S_n)\}$, (v) меньших $X_m(S_n)$; (vi) больших $X_m(S_n)$; (vii) ближайших к $M_0\{X_m(S_n)\}$. Во всем остальном bDNAvideo и Activity совпадают. Выходные данные Activity являются списком контекстно-зависимых характеристик $\{X^0_k\}$ с наибольшей позитивной полезностью $U(X^0_k; F)$.

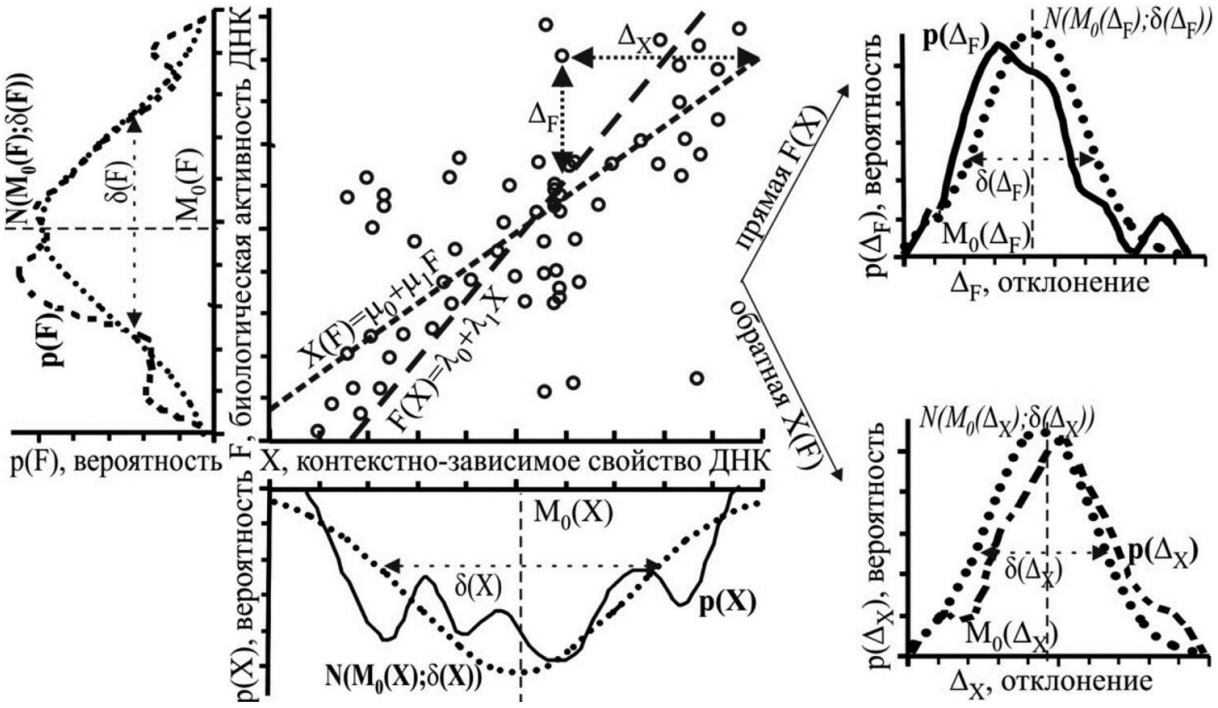

Рисунок 7 - Оценка соответствия между выборочными распределениями p(X(S)), p(F), p(Δ_X), p(Δ_F) и нормальными распределениями N[M_0(X(S)); δ(X(S))], N[M_0(F); δ(F)], N[M_0(Δ_X); δ(Δ_X)] и N[M_0(Δ_F); δ(Δ_F)] в системе Activity. M_0 и δ - среднее и стандартное отклонение; Δ – отклонение данных от их регрессий.

Работоспособность Activity была показана на примере самого изученного сайта в геномах эукариот: ТАТА-бокса (сайт связывания ТАТА-связывающего белка (TBP), ДНК-связывающей субъединицы транскрипционного фактора TFIID).

Согласно эксперименту (Powell et al., 2002), изгиб 90^O оси спирали частично денатурированной ДНК стабилизирует комплекс “TBP/TATA-бокс”. Для учета этой стадии связывания TBP с ТАТА-боксом были исследованы величины от 11.78 до 24.23 натуральных логарифмических единиц (ln-ед.) сродства -ln[K_D] TBP дрожжей к 19 нитевых олигоДНК длиной 15 нт, онДНК, которые были измерены *in vitro* в эксперименте (Соколенко и др., 1996). Восемь олигоДНК, равномерно представлявших нуклеотидный состав, были “входными данными” для Activity (обучение), 11 оставшихся – контроль. Activity анализировала в них содержание олигонуклеотидов $[\xi_1 \ldots \xi_m]_f$ длины m в сайте длины L(здесь: $1 \le m \le 4 \ll L$), взвешенных с помощью правила “чем выше вес f(i) позиции i, тем выше вклад олигонуклеотида $s_i \ldots s_{(i+m-1)} \in \xi_1 \ldots \xi_m$ в сродства TBP/ДНК”:

$$[\xi_1 \ldots \xi_m]_f(S = \{s_1 \ldots s_i \ldots s_L\}) = \sum_{\substack{i=1 \\ s_i \ldots s_{i+m-1} \in \xi_1 \ldots \xi_m}}^{L-m+1} f(i). \qquad (7)$$

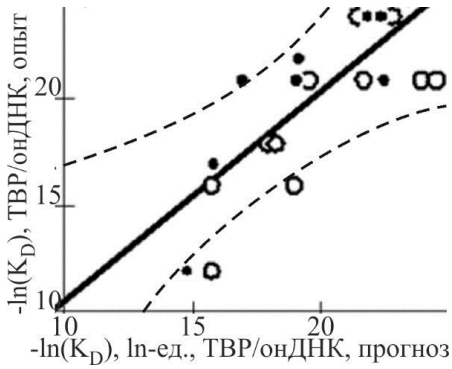

Рисунок 8 - Корреляция между предсказанными (формула 8) и экспериментальными величинами сродства ТВР к однонитевым ДНК (Соколенко и др., 1996). Обозначения: ○ – обучение; ● и линия – контроль (r=0.785, α < 0.05); пунктир – границы 95% доверительных интервалов - здесь и далее, - построенных пакетом Statistica (StatSoft, USA).

Всего было 360 вариантов веса f(i), различающихся положением наименьших и наибольших значений, а также формой монотонных переходов между ними. Они задавали (формула 7) $\approx 10^7$ вариантов $[\xi_1 \ldots \xi_m]_f$. Всего было найдено два динуклеотида TV и WR в 15-символьной номенклатуре IUPAC-IUB (1971) с самыми обоснованными корреляциями сродства ТВР/онДНК с их содержанием в ТАТА-боксе. На основе этих корреляций мы вывели формулу:

$$-ln[K_D(s_1 \ldots s_{15})] = 14.5 + 2.5[TV]_{\text{центр}}(s_1 \ldots s_{15}) + 0.9[WR]_{3'\text{часть}}(s_1 \ldots s_{15}). \quad (8)$$

Прогнозы формулы (8) достоверно (Рисунок 8: r=0.785, α<0.05) коррелируют со сродством ТВР/онДНК (Соколенко и др., 1996), на независимом контроле (●).

Кроме того, с помощью Activity исследовали величины от 11.63 до 23.54 ln-ед. сродства, $-\ln(K_D)$, ТВР дрожжей к 19 олигоДНК диной 15 п.о., днДНК (Savinkova et al., 1998). В результате выявили ширину малой бороздки спирали ДНК района [6; 9] ТАТА бокса, $X_{8;[6;9]}$ (r=0.95, $\alpha < 10^{-4}$), и содержание ТА (r=0.80, $\alpha < 10^{-3}$). На основе выявленных этих двух корреляций вывели регрессионное уравнение:

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

Прогнозы формулы (9) достоверно коррелируют со сродством ТВР/днДНК как на исследованных данных (Savinkova et al., 1998) (Рисунок 9а: r=0.96, $\alpha<10^{-4}$), так и на независимых данных (Wiley et al., 1992) (Рисунок 9б: r=0.76, α<0.05).

На независимых данных о сродстве ТВР человека к двунитевым олигоДНК длины 26 п.о., идентичным ТАТА-боксам генов человека (Савинкова и др., 2007) объединили формулы (8 и 9) с критерием ТАТА-бокса (Bucher, 1990), PWM_{TATA}:

$$-ln\left[K_{D;TATA}\right] = 10.90 - 0.23ln\left[K_{D;\text{днДНК}}\right] + 0.15PWM_{TATA} - 0.20ln\left[K_{D;\text{онДНК}}\right]. \quad (10)$$

Прогнозы формулы (10) были подтверждены (Рисунок. 10) в спланированных на их основе оригинальных экспериментах в равновесных (Savinkova et al., 2013) и в неравновесных (Drachkova et al., 2014) условиях *in vitro*.

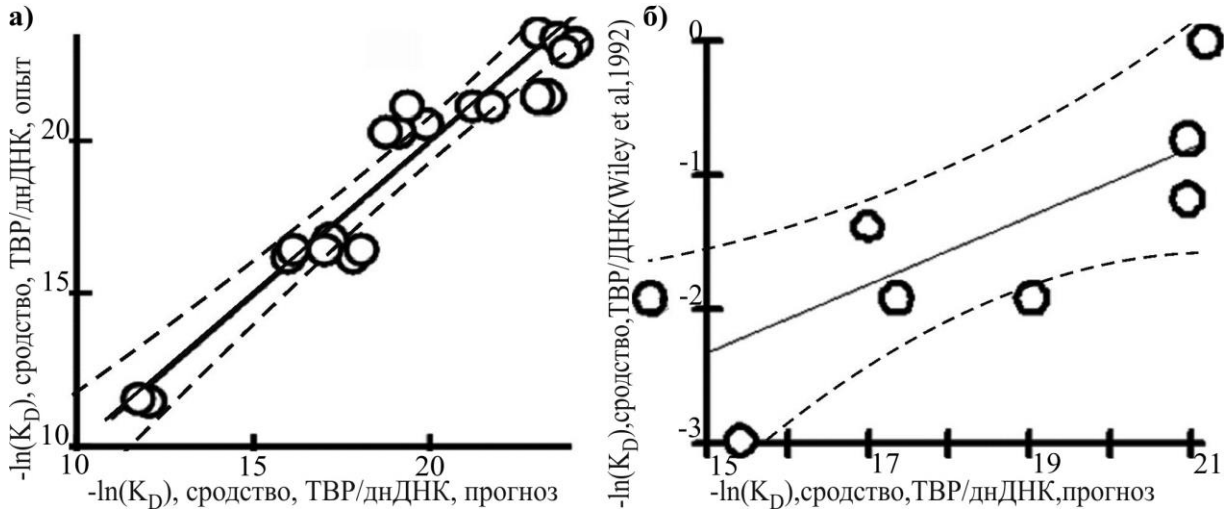

Рисунок 9 – Корреляции между предсказанным (формула 9) и измеренным сродством ТВР к двунитевым олигоДНК для (а) исследованного опыта (Savinkova et al., 1998), r=0.96, $\alpha<10^{-4}$; и для (г) независимого опыта (Wiley et al., 1992), r=0.76, $\alpha < 0.05$. Пунктир – границы 95% доверительных интервалов.

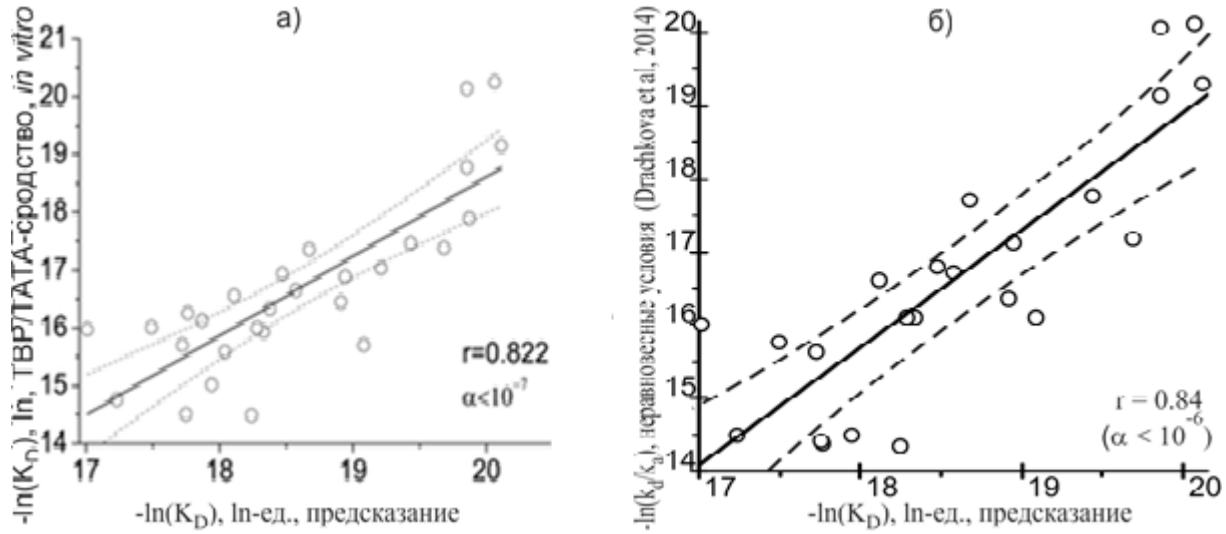

Рисунок 10 – Корреляции между предсказанным (формула 10) и измеренным сродством ТВР к промоторам генов человека при: (а) равновесных (Savinkova et al., 2013) и (б) неравновесных (Drachkova et al., 2014) условиях эксперимента *in vitro*. Пунктир – границы 95% доверительных интервалов.

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

В главе 4 на примере сайтов предмутационных повреждений геномной ДНК оценили границы применимости Activity, созданной на примере сайта связывания ТАТА-связывающего белка в промоторах генов эукариот как описано в главе 3.

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

С помощью Activity исследовали величины от 0.00 до 1.59 ln-ед. относительных частот предмутационных повреждений гуанинов, F(G), при облучении ДНК *in vitro* лазером с длиной волны 193 нм, измеренные в опыте с плазмидой pGEM7(f+) *E. coli* (Втюрина и др., 2011). В результате вывели формулу для предсказания величин таких частот по локальному окружению ±10 нт исследуемого гуанина:

$$\mathrm{F}(S_{\pm 10}(G)) = 0.69 - 0.07\mathrm{N}_{\mathrm{ANTICONS}}(S_{\pm 10}(G)) + 0.19PWM(S_{\pm 10}(G)) + \\ +0.22[YNVW]_{5'}(S_{\pm 10}(G)) + 0.07P_{23;[-7;2]}(S_{\pm 10}(G)), \quad (11)$$

где: $\mathrm{N}_{\mathrm{ANTICONS}}$ и PWM - число совпадений с антиконсенсусом ttaaagc{G}tcg-actgc, составленным из достоверно редких нуклеотидов и суммарный вес $S_{\pm 10}$(G) по позиционно-весовой матрице (приближение Больцмана) вокруг UV-повреждений G; $[\mathrm{YNVW}]_{5'}$ – содержание тетрануклеотидов YNVW перед G; и $P_{23;[-7;2]}$ – средняя частота контактов динуклеотидов с гистонами в районе [-7;2] вокруг G.

Прогнозы (формула 11) частот $\mathrm{F}(S_{\pm 10}(\mathrm{G}))$ достоверно коррелируют с данными независимого опыта (Melvin at el., 1998) с геном MIP-1α мыши (Рисунок 11: r=0.82, $\alpha<0.005$). При этом частота повреждений гуанина (формула 11) была тем выше, чем выше средняя частота контакта динуклеотидов с гистонами (Satchwell et al., 1986). Нуклеосома-подобная шарообразная упаковка генома *E. coli* была открыта в опыте (Griffith, 1976). Методом иммунноэлектронной микроскопии в ней идентифицировали белки HU (Киселева и др., 1986). Линии *E. coli*, дефектные по генам этих белков, были гиперчувствительны к UV-излучению (Li, Waters, 1998). Эволюционное родство генов HU бактерий и генов гистонов эукариот было

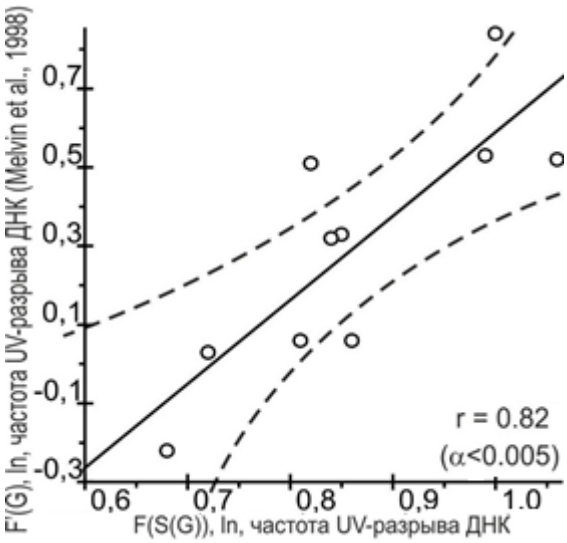

Рисунок 11 - Корреляции прогнозов (формула 11) для независимых опытов (а) с геном *MIP-1α* мыши (Melvin et al., 1998), r=0.82 ($\alpha<0.005$). Пунктир – границы 95% доверительных интервалов.

доказано *in silico* (Wong et al., 2003). Все это вместе взятое обобщили в гипотезе, что в эволюции бактерий мог отобраться способ защиты участков ДНК, уязвимых для UV-излучения, в случае их связывания с гистон-подобными белками HU.

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

С помощью Activity исследовали величины от 8 до 400 нМ константы K_M Михаэлиса и величины от 0.4 до 2.9 $м^{-1}$ каталитической константы k_{CAT} для 8-оксогуанин-ДНК гликозилазы OGG1 человека, измеренные в эксперименте (Kirpota et al., 2011) *in vitro* с синтетическими олигоДНК с числом $N_{\neq}$ нарушений комплементарности пар вокруг 8-оксогуанина (oxoG, X) в центре их (+)-нитей, $S^0(X)/S^{\#}(C)=\{s^0_{-10}...s^0_{-1}Xs^0_1...s^0_{10}/s^{\#}_{-10}...s^{\#}_{-1}Cs^{\#}_1...s^{\#}_{10}\}$ (здесь: s^0_i и $s^{\#}_{-i}$ – нуклеотиды комплементарной пары). Входными данными Activity взяли 40% олигоДНК, равномерно представлявших контекст вокруг oxoG, оставшиеся 60% олигоДНК были контролем. На основе результатов Activity были выведены эмпирические формулы для прогноза величин K_M и k_{CAT} для заданного олигоДНК:

$$K_M(s^0_{-10}...X...s^0_{+10}/s^{\#}_{-10}...C...s^{\#}_{+10}) = 0.86+72.7\times N_{\neq} -$$

$$-42.68 \times (\mathbf{max}[P_{38;[-10;+10]}\{s^0_{-10}\ldots G\ldots s^0_{+10}\}; P_{38;[-10;+10]}\{s^{\#}_{-10}\ldots C\ldots s^{\#}_{+10}\}]). \quad (12)$$

$$k_{CAT}(s^0_{-10}...X...s^0_{+10}/s^{\#}_{-10}...C...s^{\#}_{+10}) = 0.88(\mathbf{min}[P_{11;[-6;+6]}\{s^0_{-10}\ldots G\ldots s^0_{+10}\};$$

$$P_{11;[-6;+6]}\{s^{\#}_{-10}\ldots C\ldots s^{\#}_{+10}\}]) - 33.76; \quad (13)$$

где: $P_{38;[-10;+10]}$ и $P_{11;[-6;+6]}$ – среднеарифметические оценки свободной энергии Гиббса района [-10; 10] и угла кручения района [-6; 6] спирали ДНК вокруг oxoG; **max** и **min** – верхние и нижние оценки в приближении “лимитирующей стадии”.

На Рисунке 12 показаны корреляции предсказанных и измеренных констант (а) K_M (r = 0.81, $\alpha < 10^{-5}$) и (б) k_{CAT} (r = 0.67, $\alpha < 0.05$) на независимом контроле.

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

С помощью Activity исследовали величины Ф(S) от -3.40 до 0.54 ln-ед. сродства филамента RecA к 16 модельным нитям ДНК длиной 18 нт, S, измеренные в опыте (Mazin, Kowalczykowski, 1996). Десять проб, которые представляли разнообразие контекста ДНК, были “входными данными” для Activity, оставшиеся 6 проб были независимым контролем. В результате была выведена эмпирическая формула:

$$\Phi(S) = 0.54 - 1.03[DRV]_{5'}, \quad (14)$$

где: D={A,T,G}, R={A,G} и V={A, G, C} согласно 15-символьному коду (IUPAC-IUB, 1971); $[DRV]_{5'}$ - содержание тринуклеотидов DRV в 5’-половине нити.

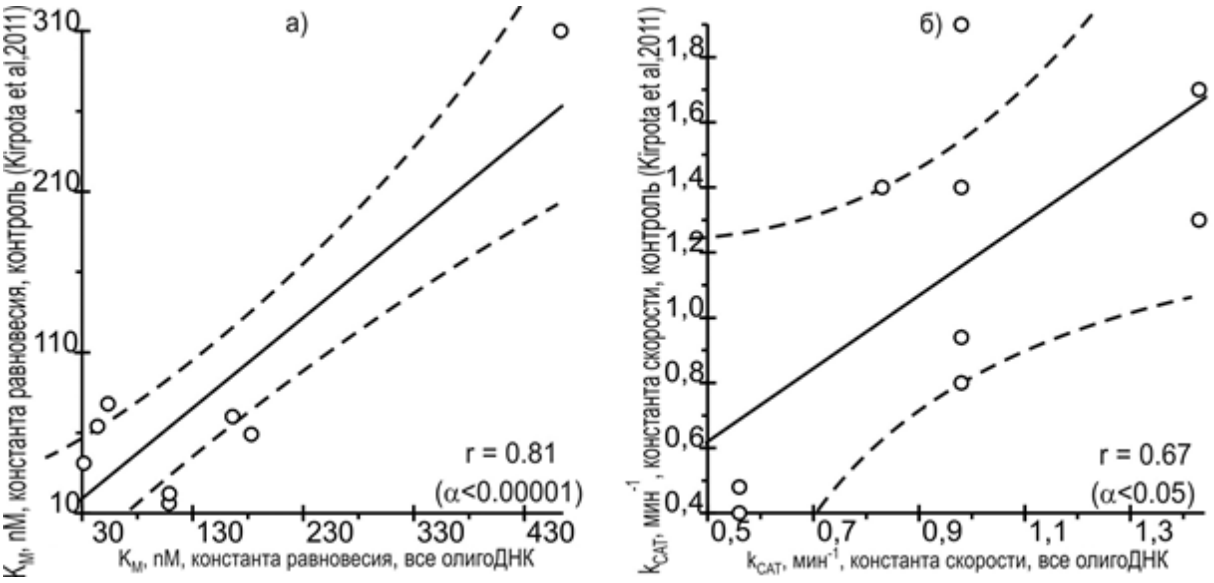

Рисунок 12 - Корреляции предсказанных и измеренных констант (а) K_M (формула (12) и (б) k_{CAT} (формула 13). Пунктир, 95% доверительные интервалы.

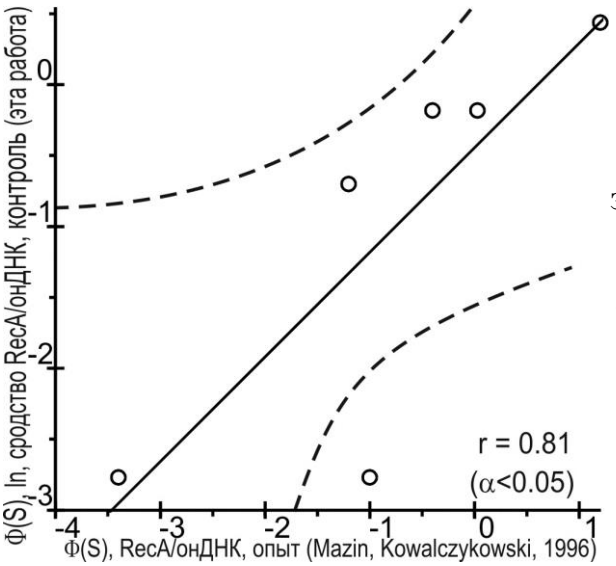

Рисунок 13 - Корреляция между предсказанными по формуле (13) и экспериментально измеренными (Mazin, Kowalczykowski, 1996) величинами сродства филамента RecA к нитям ДНК на независимом контроле. Пунктир - границы 95% доверительных интервалов для линии регрессии.

На Рисунке 13 показана корреляция предсказанных (формула 14) и измеренных величин сродства филамента RecA к нитям ДНК на независимом контроле. Согласно формуле (14), сродство RecA/онДНК уменьшается с увеличением числа тринуклеотидов DRV, соответствующих кодонам аргинина, глицина, триптофана, цистеина, лизина, серина, тирозина, аспарагина, аспарагиновой и глютаминовой кислот. В Таблице 4 показано, что эти аминокислотные остатки достоверно часто являются заряженными ($\alpha<0.05$), частыми во вторичной структуре "случайный клубок" белка ($\alpha<0.025$) и редкими в глобулярных ядрах белков ($\alpha<0.0025$).

В заключение к главе 4 представлен широкий спектр результатов применения Activity к анализу сайтов в природных и синтетических ДНК и РНК (Рисунок 14).

Таблица. 4. Проекция тринуклеотида DRV на генетический код *E. coli*

а.к.о. \ кодон:			DRV		не DRV		критерий Фишера	
№	свойство	список а.к.о.	есть	нет	есть	нет	DRV	значимость, 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

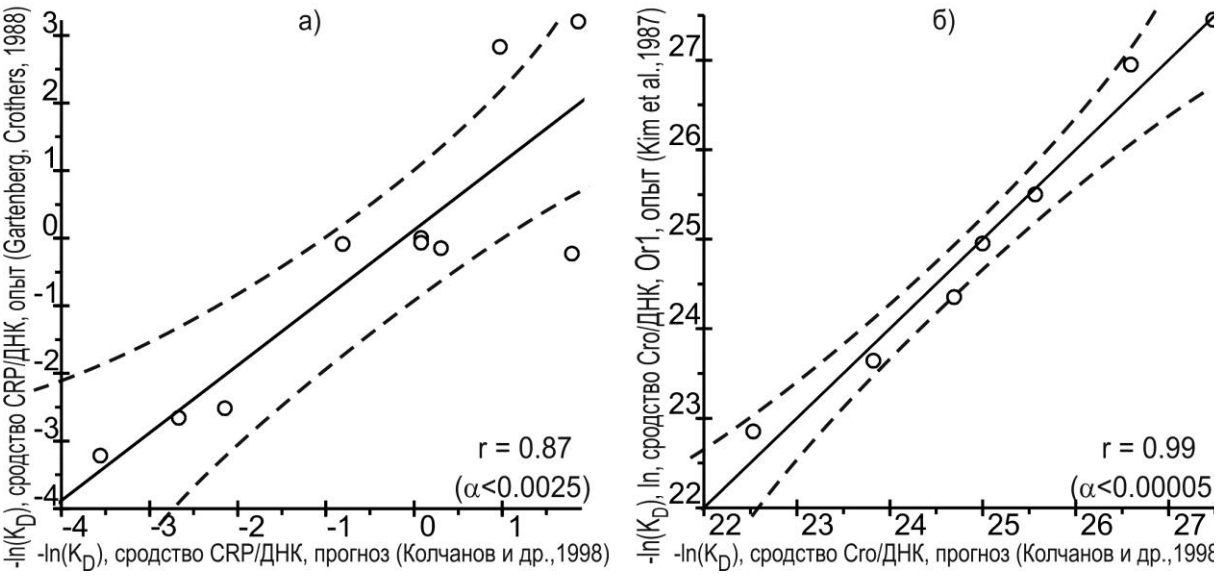

Рисунок 14 - Примеры (а) корреляция между величинами сродства белка CRP к сайтам его связывания в геноме *E. coli* (Gartenberg, Crothers, 1988) и прогнозами этих величин на основе ширины малой бороздки и шага rise спирали ДНК; (б) корреляция между величинами сродства Cro-репрессора к оператору OR1 фага λ (Kim et al., 1987) и их прогнозами на основе ширины малой бороздки, шага rise и угла roll спирали ДНК (α<0.00005). Пунктир, 95% доверительные интервалы.

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

Глава 5 начинается с представления выявленных с использованием Activity корреляций сродства MEF2/ДНК с персистентной длиной и с шириной малой бороздки спирали ДНК и корреляций сродства USF/ДНК с углом кручения и с глубиной большой бороздки спирали ДНК. Соответственно, Рисунок 15 содержит основанные на этих корреляциях прогнозы сродства транскрипционных факторов (а) MEF-2 и (б) USF к сайтам их связывания, которые были значимы на контроле.

Кроме того, с помощью Activity исследовали величины φ от -4.60 до -0.03 ln-ед. экспрессии репортерного гена *LUC* люциферазы в культуре клеток HeLa,

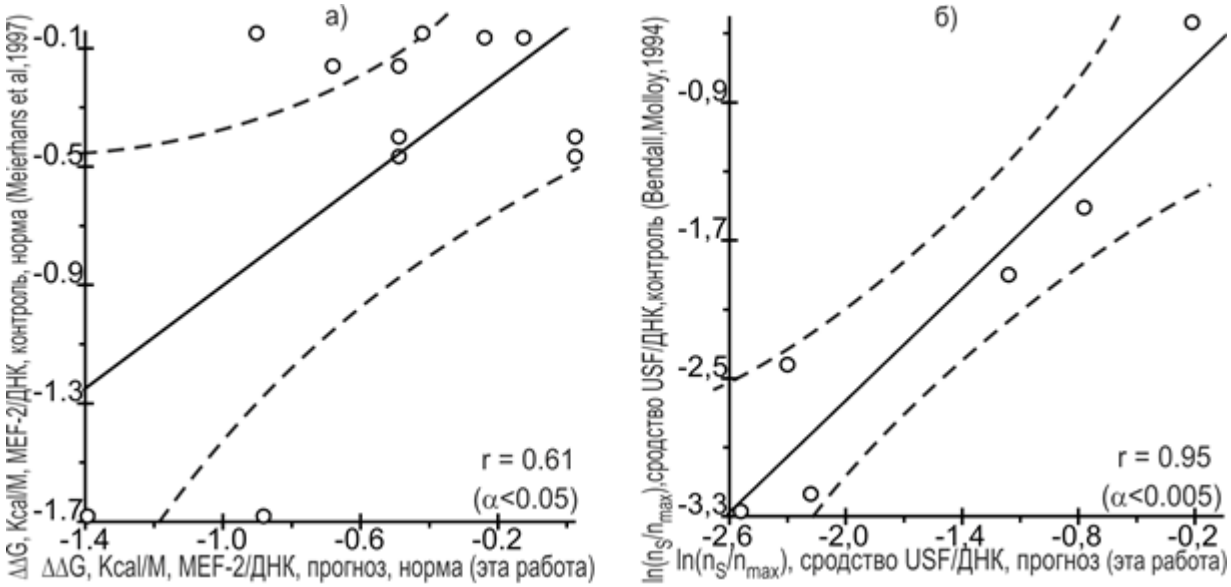

Рисунок 15. Корреляция сродства MEF2/ДНК и его прогноза на основе ширины малой бороздки и персистентной длины спирали ДНК (а), а также (б) корреляция сродства USF/ДНК с его прогнозом на основе угла кручения и глубины большой бороздки спирали ДНК. Пунктир - границы 95% доверительных интервалов.

репрессированного транскрипционным фактором YY1 при разных вариантах сайта его связывания в промоторе этого гена (Hyde-DeRuyscher et al., 1995).

В результате вывели формулу для предсказания величин YY1-зависимой репрессии генов человека на основе кручения twist спирали ДНК сайтов YY1:

$$\varphi(S) = -119.46 + 3.42P_{14;[10;21]}(S). \quad (15)$$

где: $P_{14;[10;21]}$ – среднее кручение twist спирали ДНК участка [10; 21] сайта YY1.

На Рисунке 16 представлены корреляции между предсказанными (формула 15) и измеренными величинами репортерной *LUC* активности *ex vivo* (Hyde-DeRuyscher et al., 1995) при трансфекции клеток *HeLa* (а) плазмидой pTiLUC (обучающие данные), а также (б) плазмидой pGL2 (независимый контроль).

Наконец на Рисунке 17 представлен пример применения формулы (15) в рамках экспериментально-компьютерного анализа (Vasiliev et al., 1999) нормального варианта WT интрона 6 гена *TDO2* человека в сравнении с вариантами G663A (M1) и G666T (M2) этого интрона, которые были ранее клинически связаны с рядом поведенческих расстройств (Comings et al., 1996). При этом, сначала методом “задержки в геле” синтетических олигоДНК, идентичных вариантам WT, M1 и M2, и экстракта ядер клеток печени крысы (Vasiliev et al., 1999) заметили неизвестный белок из этого экстракта, образующий комплекс с олигоДНК WT, который полностью или частично поврежден в случаях M1 и M2, соответственно.

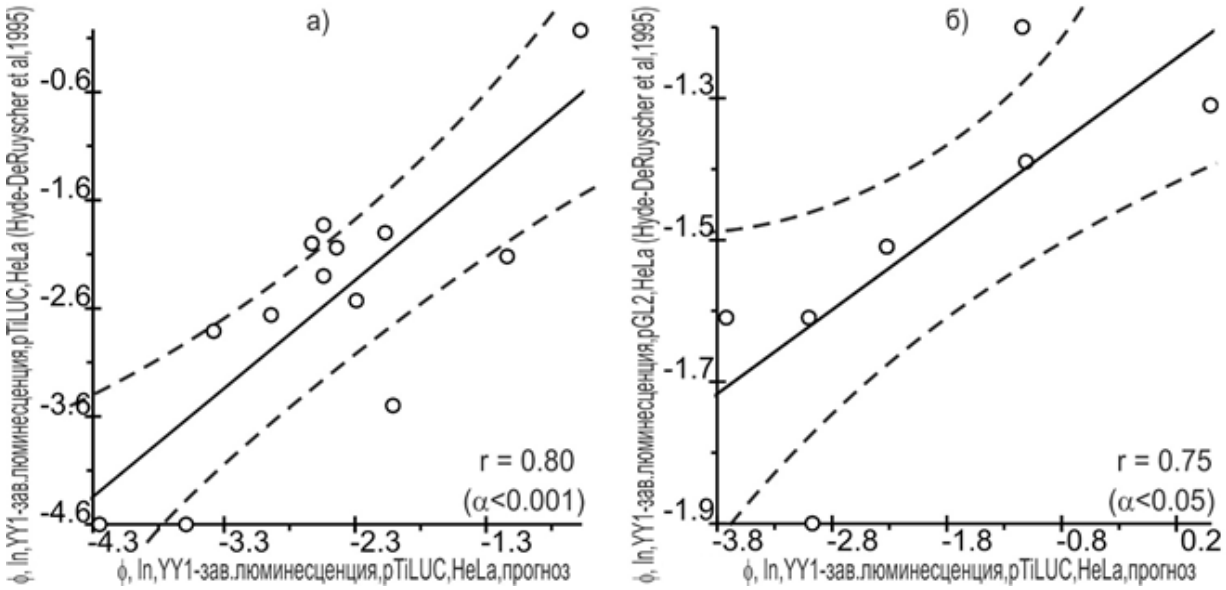

Рисунок 16 – Корреляции между предсказанными (формула 15) и измеренными величинами репортерной *LUC* активности *ex vivo* (Hyde-DeRuyscher et al., 1995) при трансфекции клеток HeLa (а) плазмидой pTiLUC (обучающие данные), и (б) плазмидой pGL2 (независимые данные). Пунктир, 95% доверительный интервал.

Затем, выявили корреляцию между оценками сродства этого неизвестного белка к олигоДНК и прогноза на основе формулы (15) для уровней YY1-зависимой репрессии репортерного гена *LUC* в случае вариантов WT, M1 и M2 в качестве сайтов связывания транскрипционного фактора YY1 в его промоторе плазмиды pTiLUC в условиях эксперимента *ex vivo* (Hyde-DeRuyscher et al., 1995) с учетом особенностей условий *in vitro* (Vasiliev et al., 1999). Наконец, на основе этой корреляции был спланирован и осуществлен эксперимент с антителами к транскрипционному фактору YY1, продемонстрировавший в норме WT наличие сайта связывания YY1, который был поврежден полностью (частично) в M1 (M2).

ЗАКЛЮЧЕНИЕ

В диссертации был предложен оригинальный подход к компьютерному анализу экспериментальных данных о влиянии контекста на специфическую биологическую активность сайтов в составе геномной ДНК с использованием теории аддитивной полезности для принятия решений и нечетких множеств.

В рамках предложенного подхода были созданы две компьютерные системы bDNAvideo и Activity для выявления контекстных, а также контекстно-зависимых конформационных и физико-химических характеристик спирали ДНК, величины которых достоверно коррелируют с экспериментально измеренными величинами специфической биологической активности сайтов в составе геномной ДНК.

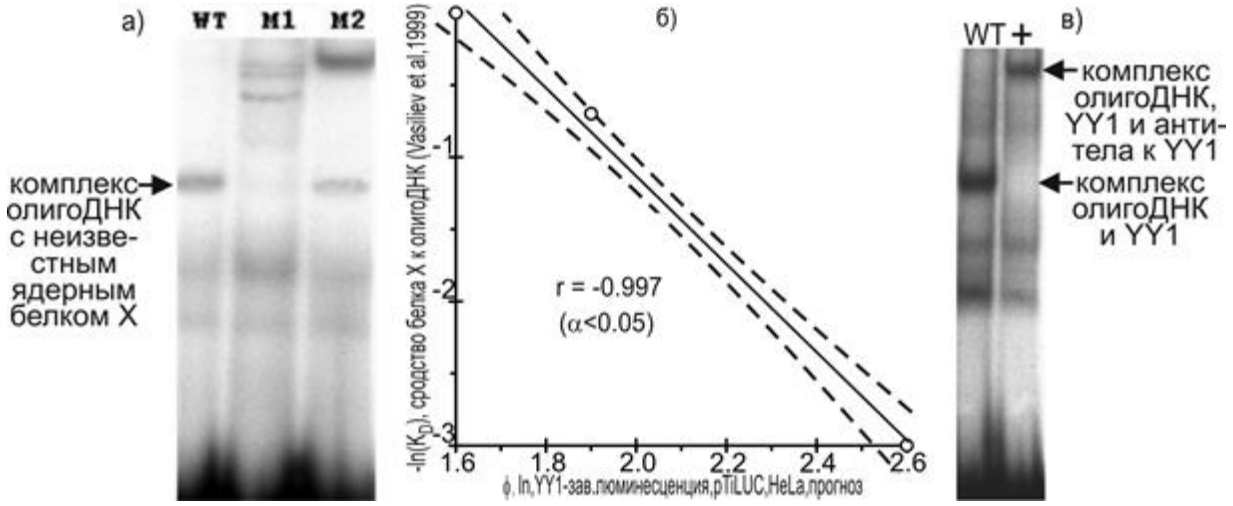

Рисунок 17- Экспериментально-компьютерный анализ нормы (WT) гена *TDO2* человека и его вариантов G663A (M1) и G666T (M2), связанных с расстройством поведения (Comings et al., 1996): (а) данные “задержки в геле” синтетических олигоДНК, идентичных WT, M1 и M2, с экстрактом ядерных белков из клеток печени крысы (Vasiliev et al., 1999); (б) корреляция между экспериментальными оценками сродства неизвестного белка из экстракта ядер клеток печени крысы к олигоДНК и прогноза на основе формулы (14) для YY1-зависимой репрессии репортерного гена *LUC* в случае WT, M1 и M2 в качестве сайтов связывания YY1 в промоторе плазмиды pTiLUC в условиях *ex vivo* (Hyde-DeRuyscher et al., 1995) с поправкой на условия *in vitro* (Vasiliev et al., 1999); (в) результат эксперимента с антителами к транскрипционному фактору YY1 доказал сайт связывания YY1 в норме (вариант WT). Пунктир, 95% доверительные интервалы. Рисунок на основе данных из статей автора (Vasiliev et al., 1999; Ponomarenko J et al., 2001)

С использованием этих компьютерных систем был проанализирован широкий круг экспериментальных данных о специфической биологической активности сайтов в составе геномной ДНК про- и эукариот. В результате был обнаружен ряд закономерностей структурно-функциональной организации и молекулярной эволюции этих сайтов, была оценена предрасположенность районов геномной ДНК к возникновению предмутационных повреждений в них и эффективности репарации этих геномных повреждений, а также была установлена молекулярная основа патогенеза аллельных вариантов интрона 6 гена *TDO2*, которые были клинически связаны с некоторыми поведенческими расстройствами человека.

Все это вместе взятое означает, что настоящая диссертация представляет новое научное направление в области математической биологии и биоинформатики.

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 ОСНОВНЫХ 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.

Главы в монографиях

32. **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.

Тезисы конференций

33. 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, 1999, 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.
34. **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.