

*Да будут славословия Богу в устах их
и меч обоюдоострый в руке их*

Биоинформатика

обсуждение задач, подходов и методов

Предисловие

Основой при подготовке этой книги послужил опыт преподавания в Иркутском Технологическом Университете и методическое пособие, разработанное к этим курсам. В дальнейшем текст книги был значительно расширен, и многие формальные стороны изложения были опущены. Как компенсация, в книге удалось коснуться многих тем молекулярной биологии, и включить в материал некоторые из моих неопубликованные результатов.

Нет смысла здесь пересказывать детали моей биографии и говорить про мотивы, побудившие подготовить эту книгу. Но мне необходимо выразить благодарность членам своей семьи, и некоторым моим друзьям, поддерживавшим меня все это время.

Сергей Феранчук,

Иркутск, 2018-2019 г.

Оглавление

1 Введение	5
2 Структурная биоинформатика	10
2.1 Математика и физика в структурной биоинформатике	10
2.2 Уровни представления молекулярных систем	13
Уровень классической механики	13
Уровень статистической физики	14
Уровень квантовой механики	16
2.3 Квантовые расчеты: модели и методы	19
2.4 Полноатомное представление молекулярных систем: модели и методы	23
Силовое поле	23
Парциальные заряды	25
Силы Ван-дер-Ваальса	28
Учет влияния растворителя	30
Гидрофобные взаимодействия	33
Молекулярная динамика	35
Анализ нормальных мод	37
Моделирование Монте-Карло	39
2.5 Структура и сворачивание белка	40
Баланс энергии при сворачивании белка	40
Методы предсказания структуры белков	42
Восстановление путей сворачивания белка	45
2.6 Модели взаимодействия биомолекул	50
3 Системная биоинформатика	57
3.1 Исторический очерк математических методов в биологии	57
3.2 Иерархия объектов в системной биоинформатике	59
3.3 Основные понятия молекулярной биологии клетки	61
3.4 Термины, используемые при постановке экспериментов и обработке данных	67
3.5 Молекулярные методы исследования клетки	71
Цели и направления при исследовании клетки	71
Анализ протеома клетки	72
Обработка экспериментов по секвенированию при изучении процессов в клетке .	75
Дифференциальная экспрессия генов	81
Ошибки и погрешности при изучении экспрессии генов	85

Исследование систем регуляции в клетке	88
3.6 Аннотация и анализ публикаций	93
Подходы к автоматическому анализу текстов	93
Ошибки в аннотации, причины и механизмы их накопления	95
3.7 Обработка данных в медицине	97
Медицинские измерения и их интерпретация	97
Подходы к получению доказательств в медицине	98
Некоторые из терминов, относящихся к экономическим отношениям в фармацевтике	100
Некоторые из гипотез, рассматриваемые в современной медицине	100
Некоторые из терминов, относящихся к характеристикам лекарственных средств и продуктов питания	101
Краткие комментарии к методам и терминологии в прикладных медицинских исследованиях	102
Традиции медицины и их эволюция	103
3.8 Математические модели в биологии	105
Понятия из теории дифференциальных уравнений	105
Обзор и частные случаи прикладных моделей	108
Нейробиология и модели сетей нейронов	110
Модель сети нейронов с двумя типами возбуждения	114
Элементы теории фракталов	118
Биоразнообразие и модели распределения численности в экологии	121
3.9 Молекулярная филогенетика и метагеномика	124
Анализ микробных сообществ	124
Филогенетические деревья	127
Эволюция патогенов	129
Болезнь байкальской губки	134
Эволюция человека	136
3.10 Вместо заключения	139
4 Библиография	145
4.1 Структурная биоинформатика	145
4.2 Системная биоинформатика	150

1 Введение

Учебный курс по биоинформатике отличается от учебных курсов по классическим дисциплинам по естественным наукам. Первое из отличий - эта область знаний развивается чрезвычайно быстро, многие из материалов успевают устареть за время обучения. Второе отличие - этот предмет находится на стыке многих дисциплин, таких как молекулярная биология, биохимия, биофизика, информационные технологии. И эти дисциплины, в свою очередь, имеют корни в науках "классической традиции", от медицины до высшей математики и теоретической физики. И тем не менее, необходимость в преподавании и разработке учебных курсов по биоинформатике, в широком смысле, безусловно существует. Научные идеи, подобно людям, имеют времена рождения, зрелости и угасания. И если научная школа не ищет и не принимает общения со другими сообществами, смежными и далекими, то со временем эти воззрения остаются лишь частью истории.

Настоящий курс ориентирован, в первую очередь, на студентов и читателей, имеющих базовые знания по молекулярной биологии. Поэтому при изложении материала вводятся по мере необходимости понятия из курсов математики, физики и химии, без углубленного обсуждения этих понятий. Отбор материала для курса был продиктован, отчасти, опытом работы авторов в этих областях. Однако в основном целью при отборе и компоновке материала было выбрать из всего многообразия методов и результатов, опубликованных в последние десятилетия в периодических научных изданиях, стержневые понятия и утверждения, и попытаться угадать направления, которые не потеряли бы значимость и при дальнейшем развитии молекулярной биологии. Этот выбор не является простым; достаточно внимательно взглянуть на историю развития естественных наук в XIX и XX веках.

Слово "биоинформатика", как название научной дисциплины, по происхождению сходно со словом "информатика". В этой дисциплине изучаются методы и подходы, являющиеся применением методов информатики - науки об обработке информации - и используемые для решения задач биологии. Появление этой дисциплины произошло вслед за достаточным развитием компьютерных технологий, и происходившем в это же историческое время развитием понятий и методов молекулярной биологии.

Обнаруженное свойство молекул ДНК кодировать информацию, которая используется при развитии клеток и многоклеточных организмов, обуславливает возможность применения технологий информатики к изучению объектов биологии - науки о жизни. И при этом, понятия молекулярной биологии подразумевают согласование биологии с уровнем описания вещества на уровне молекул и атомов, который используется в физике и химии. И в связи с этим, в биоинформатике несложно отделить область исследований, в которой используется представление объектов биологии на уровне атомов и молекул. Эту область принято называть *структурной*

биоинформатикой. Необходимым атрибутом задач структурной биоинформатики является использование информации о положении изучаемых объектов в трехмерном пространстве, таким образом привязывая постановку этих задач к методам физики и химии.

Однако в прикладных задачах, которые в настоящее время возможно и необходимо ставить и решать в молекулярной биологии, детализация объектов до уровня атомов и молекул не всегда возможна. В этих, более крупных, масштабах, информация, закодированная в ДНК, рассматривается без непосредственной связи с молекулярными процессами, в которых происходит преобразование этой информации. Набор приемов, используемых при обработке информации такого рода, и история развития алгоритмов и программных инструментов в этой области, имеют много общего с информационными технологиями в целом и историей их развития. Потому, по аналогии с понятием *информационная система*, которое используется для обобщенного определения задач, решаемых программистами, уместно ввести термин *системная биоинформатика* для обозначения всего спектра этих задач биоинформатики, как это показано на рис. 1.1.

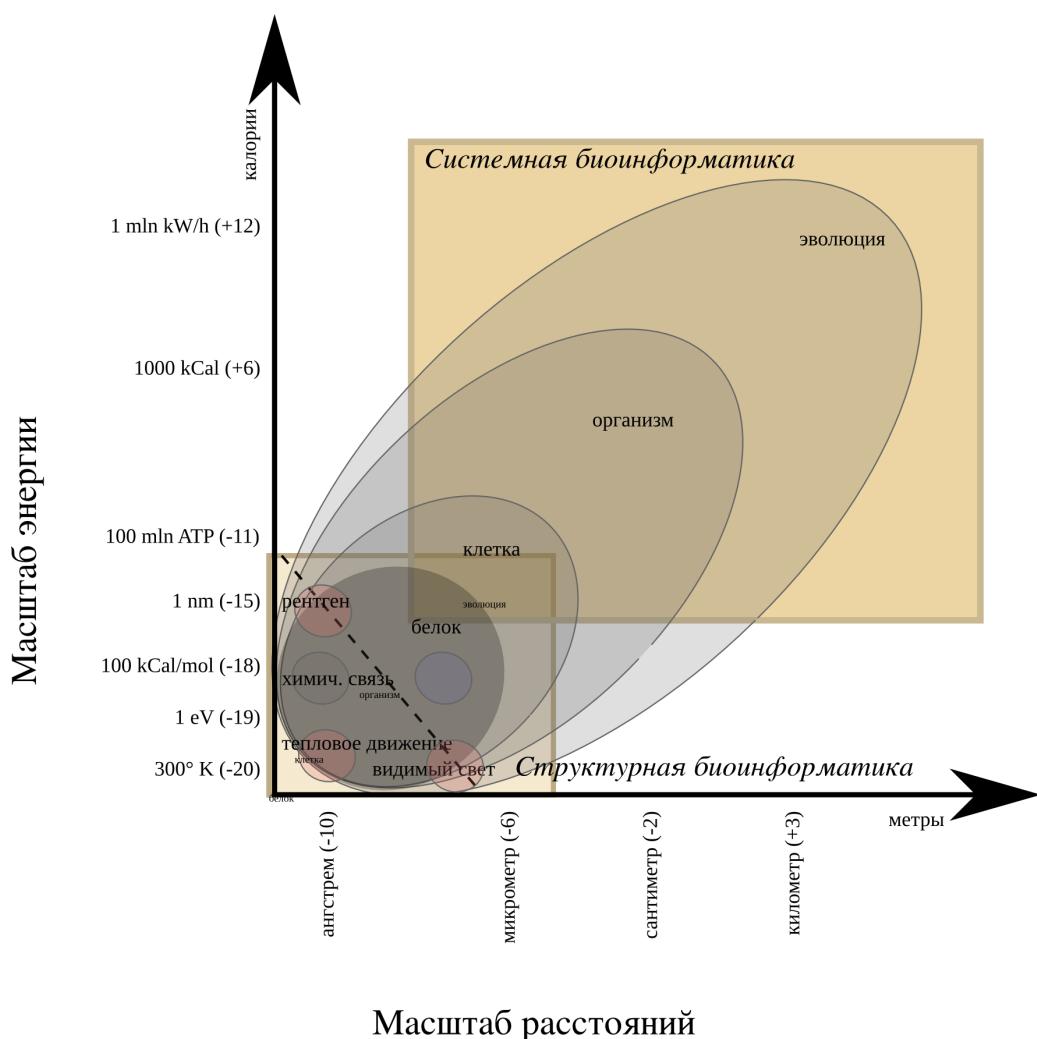


Рис. 1.1: Иллюстрация разделения масштабов при решении задач вычислительной биологии.

Существующий в настоящее время набор инструментов биоинформатики для решения задач разного масштаба весьма велик, как и объем знаний накопленных в молекулярной биологии. Более того, часто оказывается возможным совместное применение методов, использующих разные уровни детализации объекта, как это проиллюстрировано на рис. 1.2. И, когда по мере накопления знаний в науке о жизни, проясняются сомнения и вопросы, стоявшие перед человечеством за все время его истории, иногда в найденных ответах становится заметен почерк одного и того же Мастера. Но, как и умение узнавать стиль художника не тождественно знанию всех его произведений, при развитии этой науки появляются все новые свидетельства о глубине детализации и неожиданности принципов, лежащих в основе каждой из изучаемых систем.

Но к биоинформатике, в современном понимании, относятся многочисленные прикладные методы и темы исследований, часто не имеющие никаких совпадающих понятий и принципов. При таком различии в тематиках, можно иногда обнаружить трудности при общении ученых из разных научных школ и научных групп. Также, в такой ситуации возможно предположить, что при развитии молекулярной биологии в отдельных темах происходит смещение акцентов значимости направлений развития, и даже накопление заблуждений. На рис. 1.2 это проиллюстрировано с помощью сравнения интенсивности исследований, проводимых в отдельных тематиках в рамках некоторого масштаба биологических объектов, в 2003 и в 2013 годах.

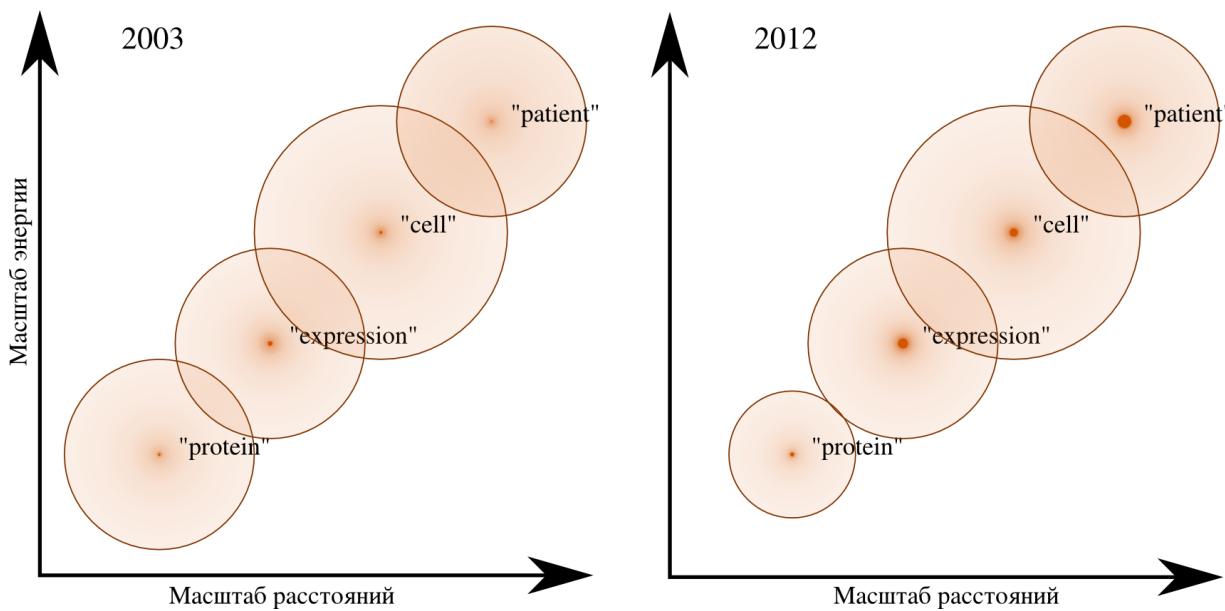


Рис. 1.2: Иллюстрация степени развития и смещения акцентов в темах, изучаемых в вычислительной биологии

Радиус круга показывает охват тематик в рамках исследования, а интенсивность цвета - степень разработки тематик, упорядоченных в порядке важности. В 2013 году, по сравнению с 2003, в каждой из областей и масштабов исследований интересы в большие степени сосредоточены в нескольких узких темах.

Изображение построено на основе текстовой обработки публикаций по биомедицинской тематике в базе данных Medline

Какие из частей и фрагментов современной науки следует считать "основанными на камне", на этот вопрос непросто ответить. Тем не менее, при подборе материала, такой вопрос, неявно, служил ограничением и ориентиром для выбора и изложения предмета. И расстановка акцентов при обсуждении задач и направлений биоинформатики не всегда соответствует интенсивности использования методов биоинформатики в современных исследованиях. Так, на рис. 1.2, как сгущение интенсивности цвета на изображениях, проиллюстрировано все большее сосредоточение акцентов вокруг задач изучения тканей опухоли методами дифференциальной экспрессии и статистических методов при обработке клинических исследований; эти задачи обсуждаются в книге недостаточно подробно. И напротив, публикаций, относящихся к задачам структурной биоинформатики, становится все меньше, но задачам структурной биоинформатики посвящена отдельная глава в курсе.

Смещение акцентов и направлений исследований в молекулярной биологии в течении первого десятилетия XXI века, показанное на рис. 1.2, легко связать с появлением новых подходов к постановке экспериментов, в первую очередь методов высокопроизводительного секвенирования. Пакеты программ и алгоритмы, непосредственно предназначенные для обработки измерений в узкоспециальных методиках постановки экспериментов, как, например, методы восстановления координат атомов в белковой молекуле на основе дифракционных спектров, также не обсуждаются в настоящем вводном курсе. Но и арсенал методов в некоторых быстро развивающихся прикладных разделах системной биоинформатики зачастую оказывается замкнутым и не имеющим надежных связей со смежными областями знаний.

Согласно с целями курса, в структурной биоинформатике более заметна преемственность и связь подходов с другими естественнонаучными дисциплинами. Однако падение популярности методов структурной биоинформатики отчасти обусловлена малым количеством приложений этих методов в прикладных задачах биологии. И, напротив, рост популярности упомянутого круга задач системной биоинформатики обусловлен важностью и удобством использования такого рода подходов в прикладных задачах.

С серией прорывов в молекулярных методах анализа живых систем, точное и полное описание стало возможно для многих прикладных задач из разных разделов биологии, и ученые, среди которых и специалисты по биоинформатике, подобно первоходцам, с энтузиазмом взялись за освоение новых территорий. Многие из предположений, выдвинутых в прошлые десятилетия и в прошлые века при поисках подходов к этим задачам, оказались ошибочными. Малая часть из методов, развивавшихся за прошлые времена в физико-математических дисциплинах, оказалась необходима в молекулярной биологии. Но все же, прослеживая преемственность методов биоинформатики с классическими дисциплинами, остается возможность помнить путь назад и пройти по нему. Не сказано ли: *Ты обращаешь человека в тление, и Ты говоришь - "возвратитесь, сыны Адама!"*.

Это может показаться неожиданным, но невозможно поспорить с тем, что продолжающийся расцвет наук в рамках европейской традиции, частью которого и являются прорывы в биоинформатике, имеет корни в христианстве, в интерпретации принятой в западной части Римской империи. И Сын Человеческий начинал свою проповедь, прочитав в Назарете книгу Исаии:

Дух Господень на Мне; ибо Он помазал Меня благовествовать нищим, и послал Меня исцелять сокрушенных сердцем, проповедовать пленным освобождение, слепым прозрение, отпустить измученных на свободу, проповедовать лето Господне благоприятное. И не является ли то, с чего начиналось, в том числе, развитие наук, лучшим чем то, в каком состоянии эти науки находятся в наши дни, несмотря на открывающиеся перспективы все новых территорий в молекулярной биологии?

2 Структурная биоинформатика

2.1. Математика и физика в структурной биоинформатике

При работе в области структурной биоинформатики несомненно следует иметь достаточно образования в классических дисциплинах, на которых основаны современные расчетные методы. В широком смысле, эти дисциплины - теоретическая физика, включая ее математический аппарат, и вычислительная математика. В настоящем учебнике вопросы касающиеся этих дисциплин объяснены на поверхностном уровне, и в рамках этого учебника нет возможности для более углубленного их изложения. Поэтому авторы посчитали нужным вкратце упомянуть во введении темы, которые следует более внимательно изучить в этих дисциплинах, и некоторые учебные курсы, в которых они изложены.

В традициях российской научной школы, изложение теоретической физики обычно следует канве "от общего к частному", и классическим курсом по теоретической физике здесь следует назвать "курс теоретической физики" Ландау и Лифшица (Ландау и Лифшиц 1958; Ландау и Лифшиц 1960; Ландау и Лифшиц 1963; Ландау и Лифшиц 1964). Первые части этого курса, в порядке следования, это "Механика", "Теория поля", "Квантовая механика", "Квантовая электродинамика", "Статистическая физика". Основы разделов физики, излагаемые в каждом из этих томов (кроме квантовой электродинамики), важны для понимания методов в структурной биоинформатике.

Среди классических курсов теоретической физики, используемых в англоязычных вузах и доступных в русских переводах, следует упомянуть "Фейнмановские лекции по физике" (Feynman и др. 1964; Фейнман и др. 1965), записи лекций американского физика Ричарда Фейнмана, и "Методы теоретической физики" Филлипа Морса и Германа Фешбаха (Morse и Feshbach 1958; Морс и Фешбах 1958; Морс и Фешбах 1960). Более широкий обзор учебных курсов по теоретической физике находится за рамками данного учебника и за рамками опыта авторов. Однако каждый из трех упомянутых курсов использует свою схему и стилистику изложения и объяснения материала, и потому не следует считать что условием вхождения в структурную биоинформатику является блестящее знание какого либо из этих учебников. Однако при работе в биоинформатике зачастую возникают вопросы, ответы на которые можно найти, только освоив материал, излагаемый в этих курсах, как, впрочем, и во многих других курсах физики.

Создание современной теоретической физики в XX веке стало возможно во многом из-за появления математического аппарата, с помощью которого стало возможным записать фундаментальные законы физики в форме точных уравнений. Среди разделов высшей математики, необходимой для полноценной работы в биоинформатике и для понимания основ теоретической

физики, следует перечислить, в порядке важности, следующие темы: владение понятиями вектора и матрицы и обращения с этими объектами, излагаемое обычно в разделах аналитической геометрии и линейной алгебры; основы интегрального и дифференциального исчисления, понятие о дифференциальных уравнениях; теория функций комплексного переменного. Отдельно выделяется раздел методов теории вероятностей и математической статистики; владение этими методами необходимо при решении большей части задач современной молекулярной биологии.

Российская школа классической математики, по наблюдениям историков науки, является одной из трех или четырех независимых и развивающихся математических традиций. Многих ученых из этой школы можно упомянуть как безусловно авторитетных для всех современных математиков, как, например, академика Колмогорова, академика Понtryагина, академика Гельфандса, и безусловно целый ряд других имен. И потому многие из курсов высшей математики на русском языке можно считать подходящими для изучения упомянутых выше разделов и предметов.

Вычислительная математика является относительно молодой наукой, она начала развиваться с 50х и 60х годов XX века. Помимо того, это прикладная наука, в ней намного меньше общих понятий и намного больше частных решений. Тем не менее в вычислительной математике сформировались некоторые общие принципы при подходах к решению задач линейной алгебры, численному решению дифференциальных уравнений и вычислению интегралов, задачам оптимизации и распознавания образов. Изложение этих принципов можно найти в учебных курсах по вычислительной математике, однако предпочтения и специализация авторов курсов в этой науке более заметны. И также, именно из вычислительной математики произошла бурно развивающаяся наука о разработке программ и алгоритмов, так называемая "наука о компьютерах" (computer science). Тут, оставляя вне рамок данного пособия большую часть работ и изобретений в этой области, все же следует упомянуть классические учебники по программированию: учебник Никлауса Вирта (Jensen и Wirth 1975; Вирт и Йенсен 1982), разработчика языка Паскаль, учебник по языку Си Брайна Кернигана и Денниса Ритчи (Керниган и Ритчи 1992; Kernighan и Ritchie 1978), учебник по языку Си++ Бьерна Страуструпа (Stroustrup 1985) - несмотря на появление все новых, более удобных, средств разработки программ, в этих учебниках можно найти некоторые общие принципы, не изменившиеся со времен появления науки о компьютерах.

С другой стороны, со времен появления информатики, технологии программирования непрерывно развиваются, появляются новые понятия, и процесс программирования становится все более легким, удобным и эффективным. Также и в смежных дисциплинах, в математической статистике, в численных методах и в теоретической физике, появляются универсальные термины, допускающие совместное использование разных подходов в этих областях.

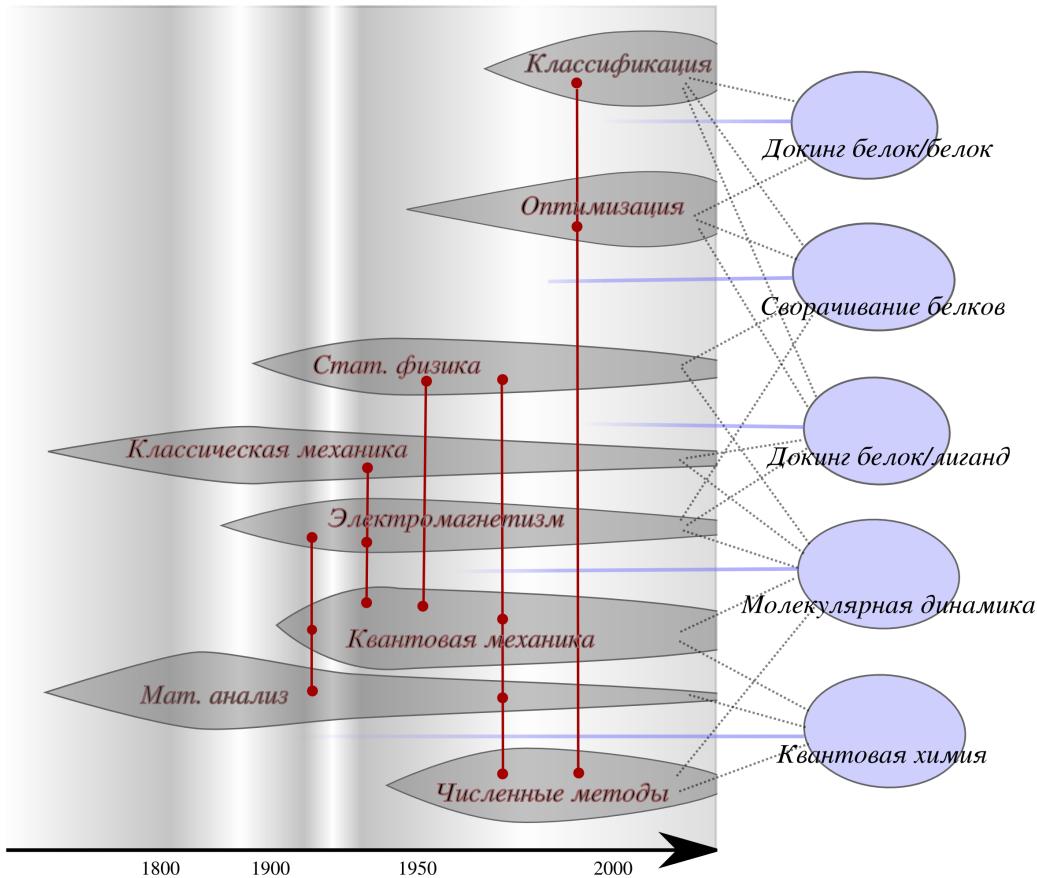


Рис. 2.1: Иллюстрация истории развития идей и дисциплин физики и математики, используемых в структурной биоинформатике.

Горизонтальная ось - шкала исторического времени. В правой части рисунка показаны темы структурной биоинформатики, обсуждаемые в пособии, и их связь с физико-математическими дисциплинами. Красные линии показывают время интенсивного обмена идеями между дисциплинами.

Временные границы показаны с долей условности, как и ширина полос, которые обозначают интенсивность разработки дисциплины. Начало развития обозначенных тем структурной биоинформатики может быть сориентировано по указанным годам: 1969 - выход работы A. Pople (Hehre и др. 1969) и начало разработки пакетов по квантовой химии; 1977 - выход работы M. Karplus (McCammon и др. 1977) и начало разработки пакетов по мол. динамике; 2000 - реорганизация журнала *Journal of Computer-Aided Molecular Design*; 1994 - начало проведения экспериментов CASP; 2001 - начало проведения экспериментов CAPRI;

В настоящем пособии не изложен подробно материал физико-математических дисциплин и понятия молекулярной биологии, на которых основаны подходы структурной биоинформатики, и также в пособии в полном объеме не изложены и сами темы структурной биоинформатики, показанные на рисунке 2.1. Ответы на многие вопросы, важные для практической работы и для глубокого понимания материала, следует искать в современных мультимедийных лекционных курсах и публикациях в международных журналах. Также, изложение этих тем можно найти в книгах и пособиях на русском языке, таких как (Холмодуров и др. 2003), (Хохлов и др. 2009),

(Андринов 2013), (Финкельштейн и Птицын 2012), (Гуреев и др. 2018). Цель, поставленная при подготовке материала в пособии, состояла, напротив, в том чтобы отойти от изложения этих тем в рамках традиций, принятых в кругу профильных специалистов, и подготовить обзор наиболее важных идей в структурной биоинформатике, для акцентирования внимания читателя на моменты согласия и моменты различия в этих идеях и методах.

2.2. Уровни представления молекулярных систем

Уровень классической механики

Механика, один из первых по времени создания разделов физики, строится на понятии «материальных точек». В молекулярных моделях, построенных на основе законов механики, атомы описываются как такие «материальные точки». В механике, для описания движения системы, используется понятие сил, действующих между точками. После того как заданы законы взаимодействия атомов, то есть силы, возможно в рамках законов механики рассчитать движение системы атомов на основе решения уравнения Ньютона.

Второй закон Ньютона, известный еще из курса физики в средней школе, выражает связь ускорения тела с действующей на него силой F , по формуле $F = ma$, где m — масса тела. По форме это так называемое «дифференциальное уравнение»; фактически, исчисление бесконечно малых, включающее понятие дифференцирования, зародилось и развивалось одновременно с рождением механики как раздела физики. В этих терминах, ускорение a , входящее в уравнение Ньютона, является второй производной по времени от координаты тела: $a = \frac{d^2x}{dt^2} = F(x)$. Результатом решения такого дифференциального уравнения будет функция $x(t)$, траектория движения тела.

Для решения дифференциальных уравнений, помимо законов движения, необходимо задать начальное положение системы. По правилам, выводимым в теории дифференциальных уравнений, уравнения, выраженные через вторые производные координат, требуют задания двух начальных параметров для каждой из координат. В случае уравнений молекулярной механики, для расчета траекторий необходимо задать начальные координаты и скорости всех атомов в системе.

Аналитические решения уравнений движения возможно получить только для случая взаимодействия двух тел (материальных точек). Эти решения известны из астрономии - это циклические движения планет по орбитам, и гиперболические траектории небесных тел, ненадолго попадающих в область притяжения Солнца. В астрономии, взаимодействием планет между собой можно пренебречь для описания их орбит, и потому движения небесных тел можно свести к задаче двух тел. Для задачи трех и более тел, решения уравнений Ньютона нельзя, в общем случае, рассчитать аналитически. Однако из структуры уравнений движения можно доказать, что эти решения тоже можно свести к циклическим движениям, подобным движению планет по орбитам. Но период времени, за который происходит цикл такого движения, может быть несопоставимо большим.

При развитии классической механики и теории дифференциальных уравнений, были разработаны изящные и эффективные способы аналитического анализа уравнений движения механических систем. Так, в астрономии, учет притяжения между планетами позволил, по малым отклонениям наблюдаемых траекторий планет, предсказать существование еще одной планеты в солнечной системе - планеты Нептун, которая затем и была обнаружена астрономами. Кометы, небесные тела, обращающиеся по эллиптическим орбитам и на лишь на малое время попадающие в область близкую к Солнцу, с древности привлекали внимание; появление кометы вносит малое, но непредсказуемое возмущение в движение небесных тел, орбиты которых иначе оставались бы безусловно неизменными. Однако молекулярные системы все же слишком сложны, чтобы использовать аналитические методы исследования уравнений движения, и эти уравнения следует решать численно.



Рис. 2.2: **Парад планет**

Парад планет - явление, когда несколько планет оказывается по одну сторону от Солнца в небольшом секторе, близко друг к другу на небесной сфере.

Парад планет 1982 года на марке КНР (Wikipedia)

Уровень статистической физики

При представлении моделей молекулярной системы следует учитывать, что рассматриваемая молекулярная система является частью некоторой большей системы. Как пример отношения большей и меньшей систем, можно привести движения белковых молекул которые происходят в цитоплазме клетки. Взаимодействие молекул системы с внешней средой может быть учтено на основе статистических свойств внешней среды, следуя законам *статистической физики*.

Статистическая физика позволяет определить понятия относящиеся к усредненным характеристикам системы, когда количество частиц в системе велико (рис. 2.3). Как пример, температура является такой статистической характеристикой, опосредованно выражющей среднюю энергию движения частиц в системе. Из уравнений движения замкнутой механической системы следует закон о сохранении полной энергии системы. Однако из-за взаимодействия с внешней средой полная энергия системы может изменяться.

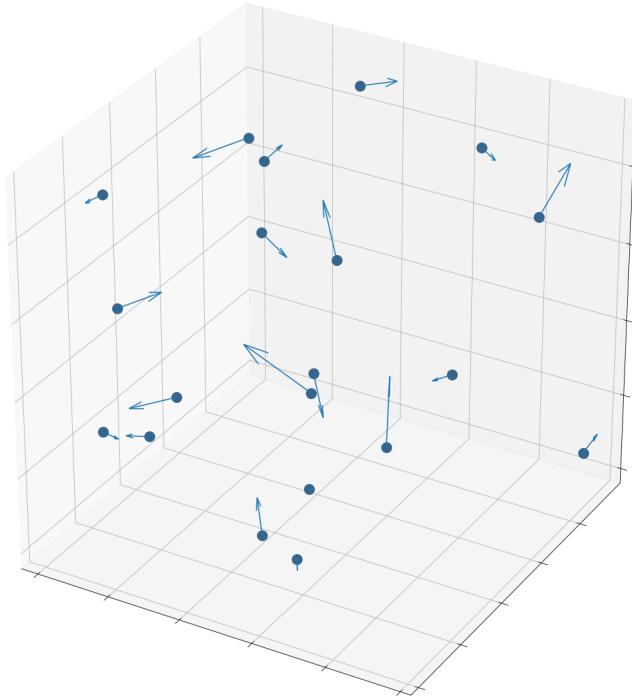


Рис. 2.3: Идеальный газ - простейшая модель статистической физики
Изображение построено с помощью пакета *matplotlib* на языке программирования *python*

Аппарат статистической физики применим в полной мере к системам, находящимся в состоянии, близком к так называемому *статистическому равновесию*. В этом случае, взаимодействие системы с внешней средой можно описать на основе *распределения Больцмана*, называемого также распределением Больцмана-Гиббса. Это распределение связывает вероятность состояния системы, одного из возможных, с энергией E этого состояния: $p = A \exp(-E/kT)$. Параметр kT в этой формуле имеет размерность энергии и выражает связь системы с внешней средой. С учетом множителя k , называемого *постоянной Больцмана*, значение T - это температура, одинаковая в системе и во внешней среде.

Понятие статистического равновесия связано с понятием *энтропии* - меры неупорядоченности системы. Понятие упорядоченности можно пояснить, используя модель идеального газа (рис. 2.3), большого количества молекул в замкнутом объеме. В каком бы из положений не находились молекулы газа в начале моделирования, по мере эволюции системы они будут все более равномерно распределены внутри этого замкнутого объема. При установлении статистического равновесия, распределение скоростей молекул может быть рассчитано на основе распределения Больцмана (рис. 2.4), какие бы скорости не имели молекулы в начале моделирования.

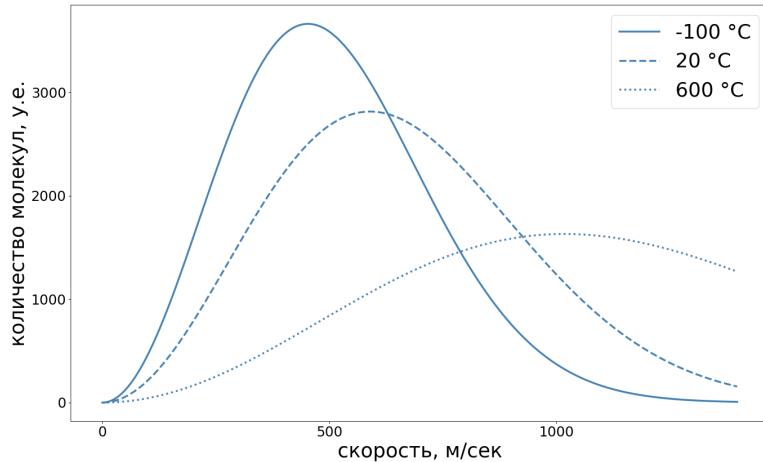


Рис. 2.4: Распределение скоростей молекул в идеальном газе для разных температур, выведенное на основе распределения Больцмана.

Горизонтальная ось отмечает скорость молекул (в метрах/сек), вертикальная ось - количество молекул с такой скоростью.

Распределение вероятностей для скорости молекул в идеальном газе известно как распределение Максвелла.

Изображение построено с помощью пакета `matplotlib` на языке программирования `python`

Как это подразумевалось при определении формулы Больцмана, в статистической физике рассматривается совместно совокупность состояний системы, так называемый *ансамбль* состояний. Значение энтропии системы, при таком определении, будет выражаться формулой $S = -\sum_i p_i \log(p_i)$, где p_i - вероятность каждого из состояний в ансамбле. Согласно *второму началу термодинамики*, по мере эволюции системы величина энтропии всегда будет возрастать; введение понятия энтропии позволяет охарактеризовать процессы установления статистического равновесия, как, например, в упомянутой модели идеального газа.

Следует отметить, что принцип возрастания энтропии системы можно ввести только совместно с понятием о влиянии внешнего окружения на процессы в системе; движение замкнутой системы будет описываться законами механики, и, в рамках уравнений механики, закон возрастания энтропии вывести нельзя. Это замечание обозначает парадоксальность и неполную согласованность всего подхода статистической физики; и, тем более, наибольший интерес для исследования представляют неуравновешенные и не полностью уравновешенные системы, где подход статистической физики неприменим в полной мере из-за несоответствия модели и объекта.

Уровень квантовой механики

На квантовом уровне описания, замкнутая система, как, например, система электронов в атоме, может находиться в одном из устойчивых *стационарных состояний*, подобно орбитам небесных тел. Спектральные линии (рис. 2.5), при таком описании, означают переходы между квантовыми

состояниями, при взаимодействии с возмущением извне - электромагнитным полем.

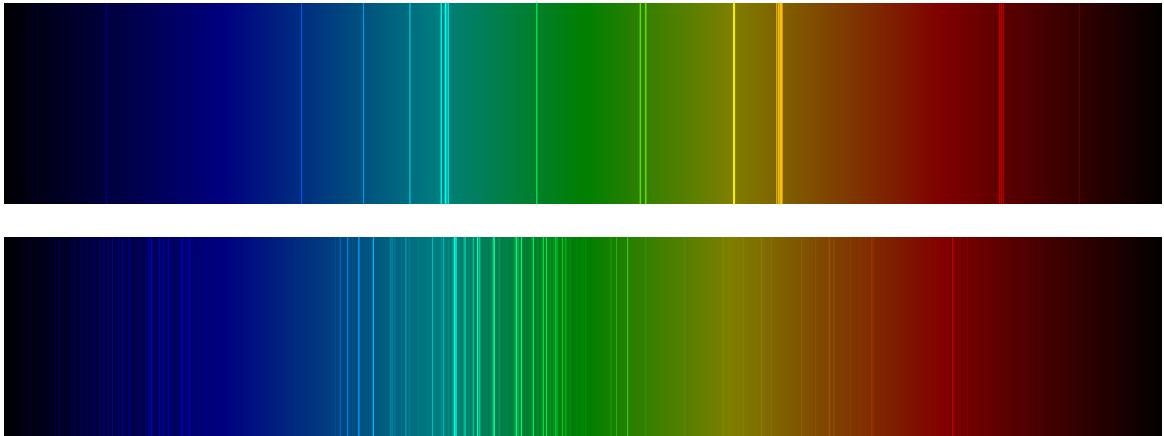


Рис. 2.5: Сравнение спектров углерода (вверху) и железа (внизу) в видимой области.
Изображения спектральных линий скопированы из разделов проекта Wikipedia.

Основным уравнением квантовой механики, описывающей распределение волновых функций электронов в молекулярной системе, является уравнение Шредингера. Методы математического анализа и теория дифференциальных уравнений существенно развились к началу XX века, что и сделало возможным вывод точных уравнений для описания квантовых систем. Уравнение Шредингера является, в широком смысле, дифференциальным уравнением, обобщающим уравнения Ньютона. Его точное аналитическое решение возможно только для простейших моделей.

Атом водорода, сходный с системой из двух тел в классической механике, это одна из моделей, для которой уравнение Шредингера удается решить точно. В отличии от орбит небесных тел, взаимодействием между электронами в атоме нельзя пренебречь. Стационарные состояния определяются совместно для всей системы электронов в атоме. Но даже для атома с двумя электронами такую задачу нельзя решить аналитически.

При определении квантового состояния используются понятие *волновой функции*; в одной из возможных интерпретаций, волновая функция описывает совместное распределение электронов в пространстве. Каждому из состояний соответствует некоторое значение энергии системы; в устойчивой замкнутой системе состояния распределены по дискретным уровням энергии.

При обобщении квантового описания системы, в терминах волновых функций, до масштаба макроскопических явлений, возникает необходимость ввести понятие *квантовой редукции* - когда при наблюдении за квантовым объектом изменяется его состояние. Так, например, понятие электронного облака, условно, подразумевает что электроны "размазаны" по всему пространству вокруг атома, как это описывается их волновыми функциями. Однако при наблюдении за электроном, с помощью фотонов с достаточно высокой энергией, электрон оказывается локализован в одной из точек пространства в "облаке", и факт такого наблюдения приводит к изменениям во всей молекуле. В более общей интерпретации, состояния квантовой системы, без присутствия наблюдателя, могут описываться как *смешанные*, когда система "размазана" или "распределена" по совокупности ее возможных состояний. Эффект наблюдения выражается в редукции системы

к одному из состояний; вероятность редукции к определенному состоянию определяется через абсолютную величину волновой функцию.



Рис. 2.6: Кот Шредингера.

Иллюстрация к мысленному эксперименту, описанному Шредингером и демонстрирующему парадоксальность принципов квантовой механики.

В этом мысленном эксперименте рассматривают черный непроницаемый для наблюдателя ящик, в который посажен живой кот. В течении эксперимента кот может погибнуть, но в рамках обобщения квантовой механики до макроскопических объектов, возможно представить смешанное квантовое состояние, сочетающее в себе живого кота и мертвого кота.

Рисунок Михаила Малоземова.

Вопросы неполноты и парадоксальности квантовой механики были замечены учеными при разработке основ этой теории (рис. 2.6). При развитии методологии и экспериментальных техник, подтвердились и углубились постановки задач, вызывавшие сомнение при создании теории. Эти эффекты привели к появлению новых направлений в физике; как пример, можно привести исследования в области *квантовые компьютеры и квантового шифрования*, связанные с понятием *квантовой запутанности* (quantum entanglement). Однако даже в этой фундаментальной и математически безупречной области знаний вопросов по прежнему остается больше чем ответов, и приложения молекулярной биологии, основанные на квантовой физике, также неявно включают парадоксы, заложенные в этом основании.

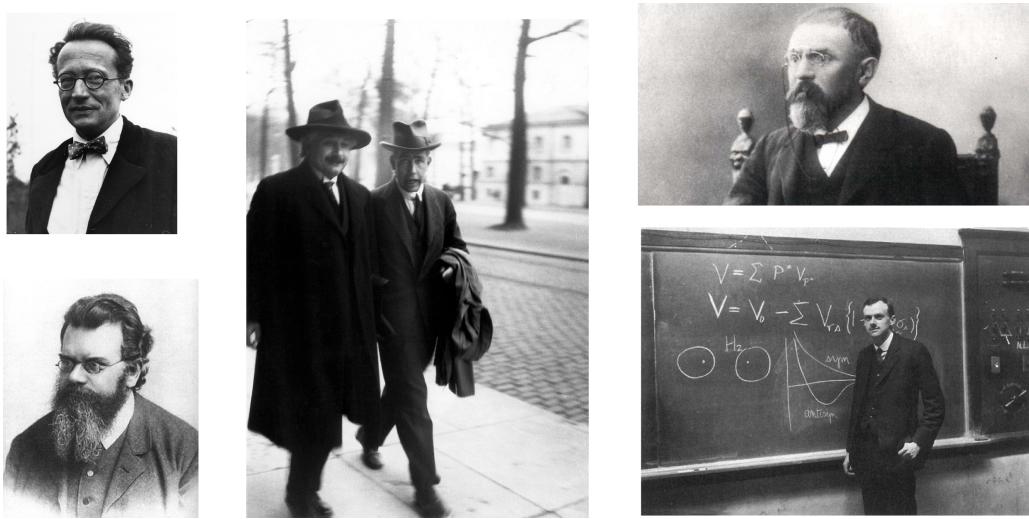


Рис. 2.7: Некоторые из физиков и математиков.

Эрвин Шредингер (1887-1961); Людвиг Больцман (1844-1906); Альберт Эйнштейн (1877-1955);

Нильс Бор (1885-1962); Анри Пуанкаре (1854-1912); Пол Дирак (1902-1984).

Фото по материалам сайтов physicstoday.scitation.org, www.physik.uni-frankfurt.de, www.dfi.dk, oasisomeoasis.blogspot.com, www-history.mcs.st-andrews.ac.uk

2.3. Квантовые расчеты: модели и методы

Квантовый переход

Стационарная квантовая система может находиться в состояниях с разной энергией, из дискретного набора возможных состояний. Так называемое *нестационарное уравнение Шредингера* является обобщением описания квантовых систем, для учета изменений состояния системы во времени, однако детальный анализ изменений в квантовых системах возможен только в некоторых частных случаях. Понятие *квантового перехода* является простой моделью для учета изменений состояния системы (рис. 2.8); квантовый переход, между двумя стационарными состояниями, возможен при внесении извне возмущений в систему.

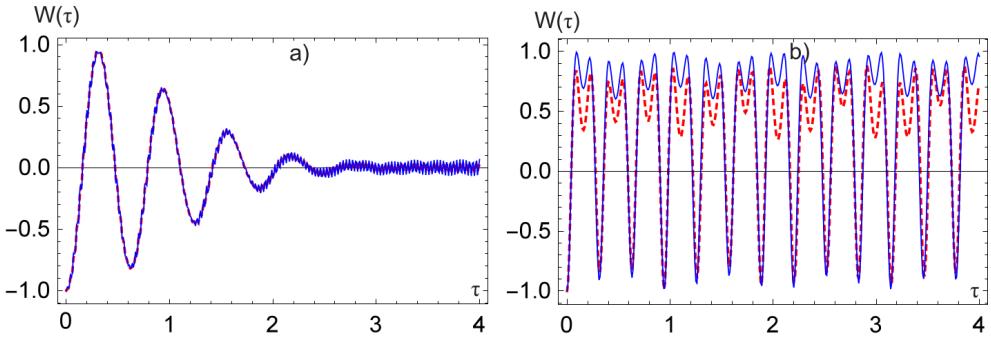


Рис. 2.8: Моделирование квантового перехода, как решения нестационарного уравнения Шредингера.

Графики соответствуют приближенным решениям так называемой "проблемы Раби", модели квантового перехода в двухуровневом атоме при взаимодействии с фотоном. Решениями модели являются колебания (осцилляции) между основным и возбужденным состояниями.

Две части рисунка соответствуют двум наборам параметров модели. Синие линии - численное решение модели, красные пунктирные линии - аналитическое решение, полученное на основе подхода "операторного метода" (I. Feranchuk, Komarov и др. 1995). по материалам работы (I. Feranchuk, Leonov и др. 2016).

Приближенные методы Для описания квантовых систем разработан ряд подходов, позволяющих найти в ряде случаев приближенное решение уравнения Шредингера; среди них следует упомянуть подход *теории возмущений* и *квазиклассическое приближение*. Подходы из класса методов Хартри-Фока ("самосогласованного поля") и подходы, использующие понятие "функционала плотности", основаны на поиске состояния системы с минимальной энергией. Эти подходы, с достаточной универсальностью, позволяют рассчитывать системы из многих частиц; однако, как компенсация, при этом исключена возможность рассматривать возбужденные состояния системы и квантовые переходы.

Модель атома Решением точного уравнения Шредингера для атома водорода являются серии волновых функций, соответствующих разным уровням энергии. Атом водорода состоит из ядра и одного электрона. Более высокие уровни энергии системы соответствуют волновым функциям, в которых плотность электронного облака локализована дальше от ядра, переходя в пределе к состоянию при котором электрон настолько удален от ядра, что электрон и ядро можно рассматривать как две независимые частицы. Для самых низких уровней энергии существует система обозначений для описания волновых функций: s - самый низкий (нулевой) уровень энергии, p - первый уровень энергии, d - второй уровень энергии. По форме, эти волновые функции можно записать как произведение полиномиальных функций от расстояния до ядра r и от угловых координат, и экспоненциально убывающей зависимости от r ; так, для первого уровня энергии, зависимость от r , для одной из волновых функций, записывается как $R(r) = \frac{1}{2\sqrt{2}}(2-r)e^{-r/2}$.

При описании многоэлектронного атома можно приближенно представить полную волновую функцию через комбинацию волновых функций отдельных электронов. Для каждого из электронов, его волновая функция предполагается подобной одной из волновых функций атома

водорода. На качественном уровне, использование этой модели позволило объяснить многие свойства химических соединений, включая периодичность элементов в таблице Менделеева. Однако для количественных расчетов при этом необходим учет взаимодействия между электронами, как, например, через введение дополнительного параметра в представление волновых функций электронов (рис. 2.9).

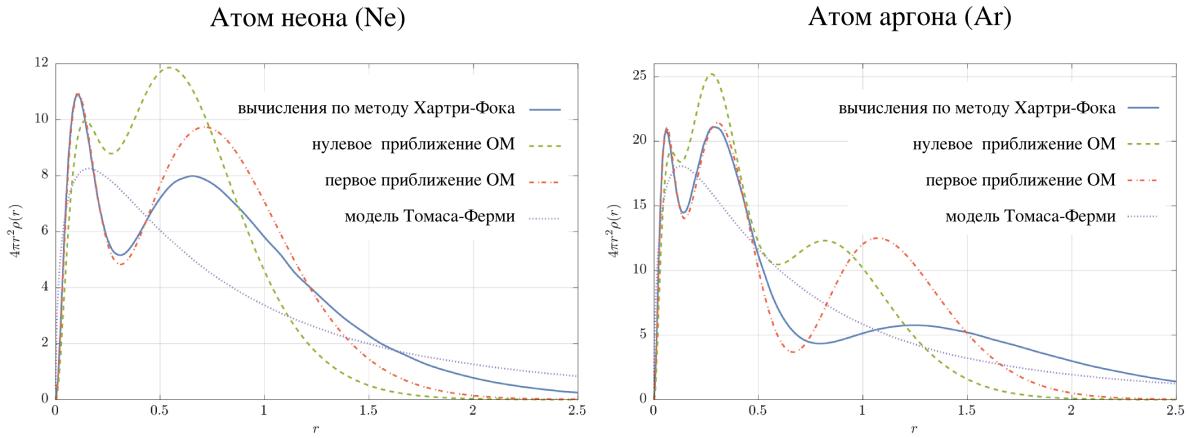


Рис. 2.9: Зависимость электронной плотности от расстояния до ядра, для атомов аргона и неона, рассчитанная несколькими методами

Приближения, показанные линиями 2 и 3 рассчитаны на основе операторного метода (ОМ) решения уравнения Шредингера, описанного в серии работ, включающей (*I. Feranchuk, Komarov и др. 1995*).

по материалам статьи (*Skoromnik и др. 2017*).

При квантовых расчетах для учета взаимодействия электронов в молекулах, необходимы расчеты шестимерных интегралов, с использованием волновых функций электронов в паре взаимодействующих атомов. Для того, чтобы выражение для шестимерного интеграла можно было записать аналитически и избежать численного интегрирования, используются модифицированные выражения для волновых функций электронов, по сравнению с точными волновыми функциями электрона в атоме водорода. Выбор подходящей формы волновых функций для каждого из атомов (*базисных функций*) необходим при квантово-химических расчетах; к наиболее популярным наборам базисных функций относятся функции "STO-nG" и так называемые "базисы Попла" (Hehre и др. 1969). Возможность аналитических упрощений при расчете интегралов в этих базисных наборах достигается за счет замены зависимости вида e^{-r} на зависимость вида e^{-r^2} в формулах для радиального распределения волновой функции.

Модели молекул

Для расчета молекул на квантовом уровне, в пакетах программ квантовой химии, где реализованы расчеты по методам Хартри-Фока и функционала плотности, используются достаточно сложные алгоритмы для повышения точности расчетов, устойчивости при проведении итераций, экономии памяти и времени вычислений. Среди этих пакетов следует назвать Gaussian, коммерческий проект с богатой функциональностью, и Gamess, распространяемый бесплатно и ориентированный на большие объемы вычислений в суперкомпьютерных центрах, а также более

простые, специализированные и/или универсальные пакеты, такие как NWChem, ORCA, MPQC и др.

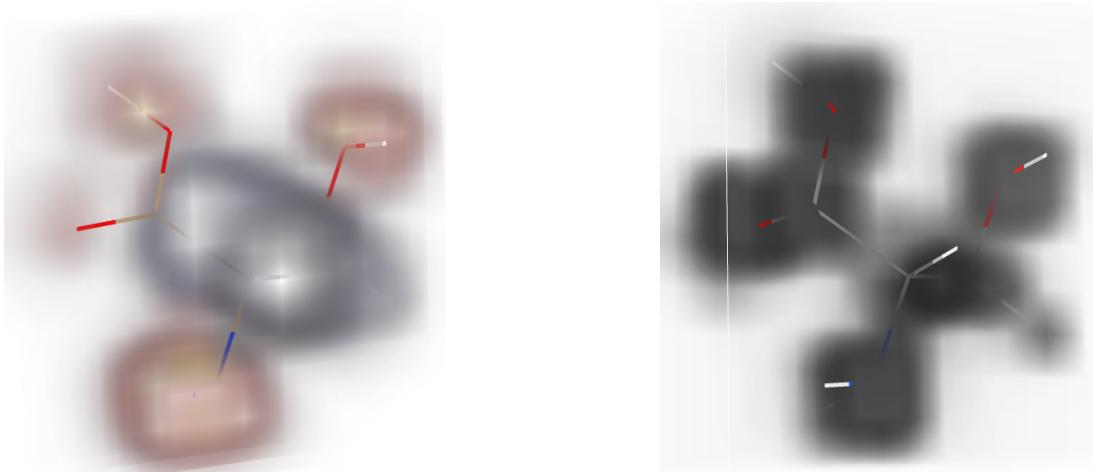


Рис. 2.10: Распределение электронной плотности в молекуле серина, рассчитанное с помощью метода Хартри-Фока

Слева: Распределение электронной плотности для двух из орбиталей: седьмой(синим) и двенадцатой(красным).

Справа: Суммарное распределение электронной плотности.

Расчеты были произведены с помощью пакета NWChem, для визуализации был использован пакет UCSF Chimera.

Результат расчета электронной плотности молекулы серина с помощью метода Хартри-Фока показан на рис. 2.10. Также, при квантовых расчетах молекулярных систем используют приемы, позволяющие моделировать протекание несложных химических реакций, и решать некоторые другие подобные задачи. Сложность вычислений при квантовых расчетах молекулярных систем возрастает пропорционально четвертой степени количества атомов в системе, и даже с использованием высокопроизводительных компьютеров, так удается рассчитывать только системы из нескольких сотен атомов.

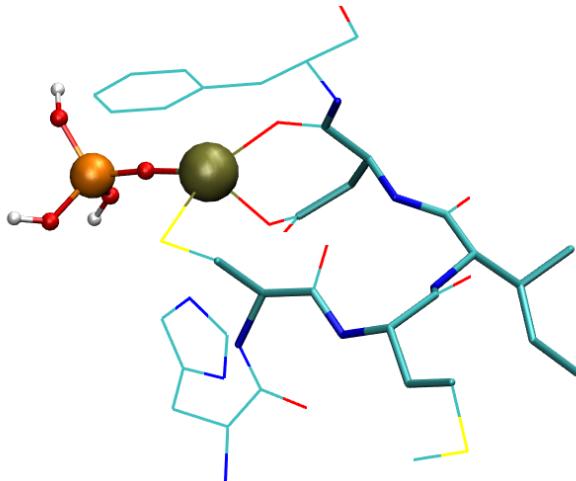


Рис. 2.11: Расчет структуры пептида в комплексе с атомом металла и кремниевой кислотой.

Структура комплекса (106 атомов) была оптимизирована с помощью метода Хартри-Фока. Показана модель пептида HCMIDF, из серии пептидов, обсуждаемых в (Marchenkov и др. 2018). Расчет произведен с использованием пакета Gaussian, рисунок подготовлен с использованием пакета VMD

Для моделирования молекулярных систем, включающих белковые молекулы, возможен прием, где моделирование основной части системы проводится с помощью моделей классической механики, кроме определенного фрагмента, где расчет проводится с помощью моделей квантовой физики. Однако многие и многие задачи химии и биологии, включая моделирование квантовых переходов в тяжелых металлах, моделирование реакций, где следует учитывать взаимодействие с молекулами воды в растворе, остаются вне возможностей численного исследования. И даже, по мере усложнения постановки задач в рамках возможностей методов, все больше проявляются проблемы численной неустойчивости при решении и существенной неточности расчетов.

2.4. Полноатомное представление молекулярных систем: модели и методы

Силовое поле

При представлении молекулярных систем на уровне описания классической механики, *силовым полем* называют согласованный набор параметров, которые используются для расчета сил, действующих в системе. Критерием при согласовании значений параметров является соответствие рассчитанного поведения моделей с поведением, оцененным из экспериментальных наблюдений. В силовых полях, рассчитанных из физических принципов, учитываются свойства молекул, полученные на основе квантовых эффектов. Так, например, при применении закона Кулона, все атомы рассматривают как частично поляризованные, и величины частичных зарядов атомов определяют исходя из распределения электронной плотности.

Типы взаимодействий, обычно учитываемые при моделировании биологических молекул,

проиллюстрированы на рисунке 2.12. Атомы в белках и в большинстве других биомолекул связаны через ковалентные химические связи, и при классификации типов взаимодействий в такой системе в первую очередь следует разделить взаимодействия между ковалентно связанными атомами (*bonded interactions*) и между несвязанными атомами (*non-bonded interactions*).

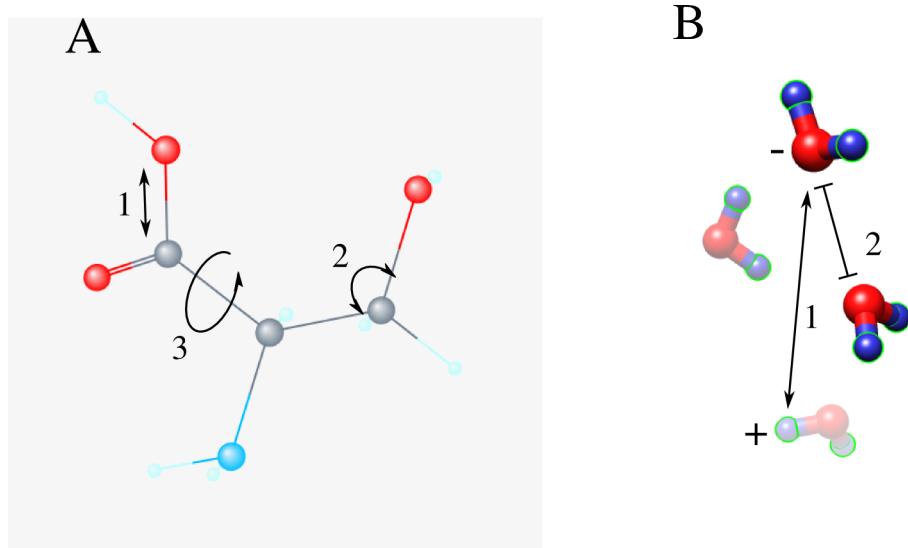


Рис. 2.12: Схема сил межатомных взаимодействий при моделировании системы атомов

A. Степени свободы и силы упругости ковалентных связей, на примере молекулы серина:

(1) растяжение (2) вращение (3) вращение двугранных углов

B. Нековалентные взаимодействия, на примере молекул воды:

(1) взаимодействия зарядов (2) близкодействующие силы Ван-дер-Ваальса

При моделировании молекулярной системы, оставаясь в рамках законов механики, невозможно рассчитать появление или разрыв ковалентных связей между атомами; топологию ковалентных связи в системе определяют при подготовке системы к моделированию. Ограничения на подвижность атомов, накладываемые наличием ковалентных связей, обычно учитываются через введение понятия упругости ковалентных связей при отклонении от положения равновесия. Линейная связь между силой упругости и отклонением от положения равновесия, известная как закон Гука ($F = -kx$) является наиболее простым способом определить силу упругости, как первое приближение в разложении функции силы в ряд Тейлора. Для молекулярной системы, возможные деформации ковалентных связей в первом приближении можно свести к трем составляющим: деформации длин связей, деформация плоских углов между атомами и поворот двугранных углов, как это показано на рис. 2.12 (A). Для каждой ковалентной связи, коэффициент k в формуле для силы упругости в более точных моделях квантовой механики определяется распределением электронной плотности в окрестности связанных атомов. Также, среди трех перечисленных групп параметров, при описании поворотов двугранных углов обычно следует учитывать взаимодействия второго порядка, поскольку часто в первом приближении молекула может свободно изменять конформацию за счет поворота двугранных углов. В этом случае начинают играть роль взаимодействия со смежными атомами, за счет перекрывания электронных

облаков, и другие эффекты.

В приложении к биологии, наиболее интересными для моделирования молекулярными системами являются белки, а также комплексы белков с другими активными соединениями. При моделировании белков, подходы к расчету силовых полей можно упростить, поскольку белки состоят из однотипных аминокислотных остатков, таких типов остатков всего 20. Поэтому возможно рассчитать параметры силового поля для ковалентных связей каждого из 20 остатков, и этот набор параметров позволит проводить моделирование большей части требуемых белков. В наиболее точных подходах, универсальные силовые поля рассчитывают на основе уравнений квантовой химии и адаптируют для достижения оптимального соответствия результатов моделирования и экспериментальных измерений. Как примеры таких силовых полей, следует привести силовое поле Amber, разработанное для моделирования белков и других биомолекул в одноименном пакете программ; силовое поле Charmm, также разработанное для моделирования белков в одноименном пакете программ. Эти силовые поля существуют в нескольких версиях, подобно традиции, принятой при разработке программного обеспечения.

Парциальные заряды

Простой моделью для описания квантового эффекта перераспределения электронной плотности является модель поляризации атома во внешнем электрическом поле, схематически изображенная на рис. 2.13. В атомах, составляющих белковые молекулы, этот эффект сложнее описать наглядно и явно, однако при расчетах следует учитывать, что атомы в таких молекулах являются частично поляризованными.

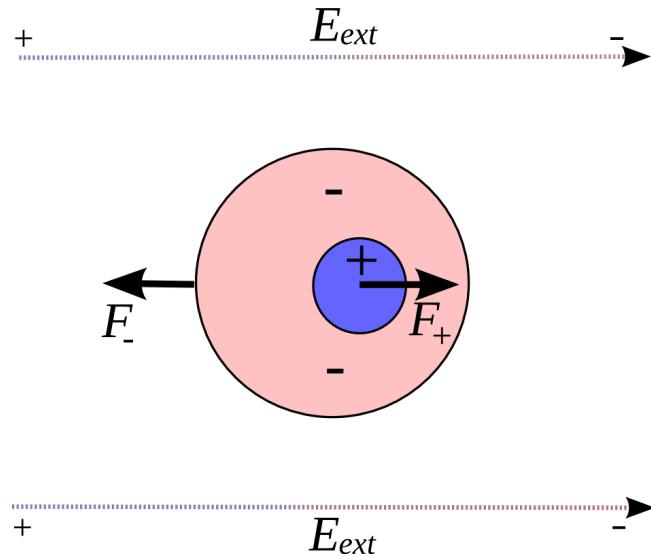


Рис. 2.13: Поляризация атома во внешнем электрическом поле

Положительно заряженное ядро атома схематично показано синим; отрицательно заряженные электроны, окружающие ядро ("электронное облако"), схематично показаны как розовая окружность. E_{ext} — направление напряженности внешнего поля, F — направления смещения положительного и отрицательного зарядов атомов.

Остатки в белках являются в большей части электрически нейтральными, как и другие химические соединения, изучаемые в молекулярном моделировании. Однако эффект перераспределение электронной плотности в молекулах можно учесть, приписав каждому из атомов в системе так называемый *парциальный заряд*. В таком приближении, заряд атома предполагается не в точности равным заряду электрона e , или величине Ze (Z - целое число), а некоторой долей положительного или отрицательного заряда: $q = xe$, величина x обычно находится в пределах от -1 до 1. Тогда силы взаимодействия между атомами можно рассчитать на основе закона Кулона, то есть сила притяжения или отталкивания между двумя зарядами q_1 и q_2 пропорциональна величине каждого из зарядов и обратно пропорциональна квадрату расстояния r между ними ($F = q_1 q_2 / r^2$). Векторное направление силы направлено вдоль линии соединяющей заряженные точки, а знак силы определяется знаками зарядов, так что заряды разных знаков притягиваются, а заряды одного знака отталкиваются.

Для расчета параметров силовых полей, в первую очередь для расчета величин парциальных зарядов ядер, возможно использовать квантовые модели, когда по координатам атомов в молекуле рассчитывается распределение электронной плотности (рис. 2.10), а затем это распределение приближенно описывается с помощью задания парциальных зарядов ядер. Также, для решения подобных задач разработан целый ряд эмпирических и полу-эмпирических методов, в которых в уравнений квантовой механики введено большое количество допущений, для быстрой, по объему вычислений, оценки решений. Рисунок 2.14 иллюстрирует различия в величинах парциальных зарядов, при использовании некоторых из упомянутых методов расчета.

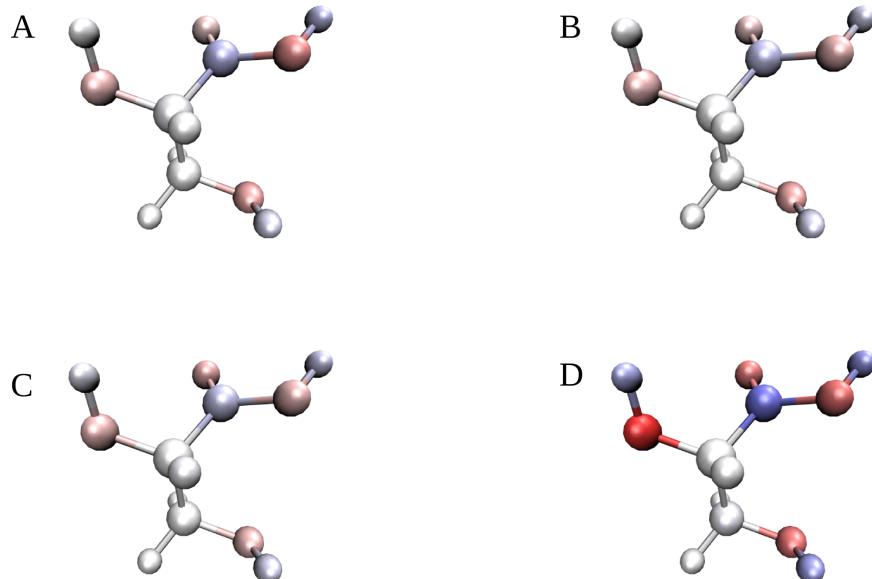


Рис. 2.14: Сравнение методов расчета парциальных зарядов, для молекулы серина
 (A) Расчет электронной плотности методом Хартри-Фока, редуцированный до уровня парциальных зарядов; (B) метод Gasteiger; (C) метод Mulliken; (D) метод Am1_bcc;
 Для всех методов, цветовая схема соответствует диапазону парциальных зарядов от -1 до 1, в единицах элементарного заряда.
 Описания методов опубликованы в (Mulliken 1955; J. Gasteiger и Marsili 1980; Jakalian и др. 2000; Bultinck и др. 2002). Расчеты были произведены с помощью пакетов NWChem, OpenBabel и AmberTools, для визуализации был использован пакет VMD.

При качественном анализе моделей молекулярных систем следует учитывать ряд особенностей, вытекающих из квантовой природы взаимодействий. Так, например, вторичную структуру белка стабилизируют водородные связи, образующиеся между положительно заряженным атомом водорода и отрицательно заряженным атомом кислорода в основной цепи белка. Однако оценки показывают, что при реальных значениях поляризации атомов кислорода и водорода энергия водородной связи оказывается в разы меньше, чем наблюдается в эксперименте, и увеличение энергии водородной связи происходит за счет квантовых эффектов (рис. 2.15).

Коррекция этих эффектов возможна при подготовке согласованного силового поля, как, например, при завышении абсолютных величин зарядов водорода и кислорода. Класс силовых полей, где парциальные заряды некоторых атомов искусственно увеличиваются чтобы добиться правильного значения энергии взаимодействия, называют *сильно поляризованными* силовыми полями.

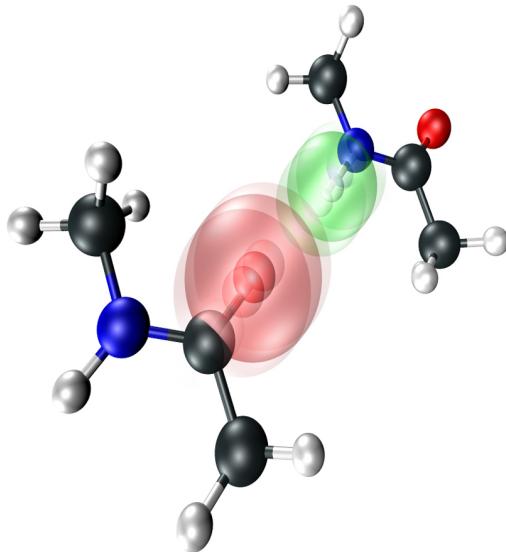


Рис. 2.15: Квантовые эффекты при образовании водородной связи в белке.

Рисунок построен по результатам экспериментального измерения колебаний молекул по методике "broadband two-dimensional infrared spectroscopy (2DIR)".

Рисунок и описание эксперимента опубликованы в (De Marco и др. 2014).

Подобные эффекты могут наблюдаться и в других взаимодействиях; образование ковалентной связи между белком и лигандом, приводящее к необратимому связыванию этих молекул, является примером, выражющим необходимость коррекции энергии, выражающей степень аффинности белка и лиганда. Необратимые эффекты и существенные поправки к энергии встречаются часто при исследовании взаимодействия белков-ферментов с ингибиторами; это наблюдение, в частности, показывает ограниченность подходов молекулярного моделирования при подборе ингибиторов для разработки новых лекарственных средств.

Силы Ван-дер-Ваальса

При использовании парциальных зарядов атомов, эти заряды рассчитываются и используются во всех фазах моделирования. Однако степень поляризации атомов и распределение электронной плотности динамически изменяются при смещениях атомов в белке в процессе моделирования. Существуют подходы к построению силовых полей, когда распределение электронной плотности учитывается с помощью более сложных приближений, чем приближение парциальных зарядов. Так, например, в силовом поле Amoeba в пакете Amber, атомы представляются как диполи - системы из двух зарядов разной полярности. Однако один из эффектов динамического изменения поляризации атомов учитывается в большинстве моделей молекулярных систем. Речь идет о так называемых *силах Ван-дер-Ваальса*. Этот физический эффект основан на свойствах атомов индуцировать взаимную поляризацию на близких расстояниях, как показано на рисунке 2.16. Характерное межатомное расстояние, на котором важно действие сил Ван-дер-Ваальса, составляет около 5 нанометров. При взаимном сближении для нейтральных атомов оказывается энергетически выгодным, чтобы их ядра сместились относительно электронных облаков, так что из-за

превращения атомов в диполи возникает сила притяжения между ними.

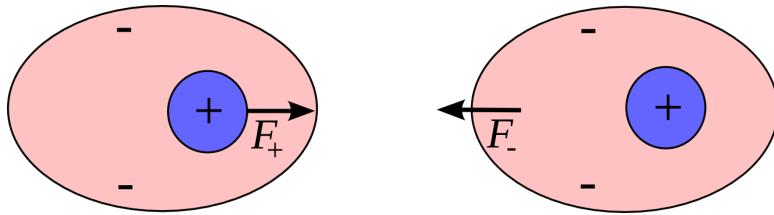


Рис. 2.16: Иллюстрация механизма возникновения сил Ван-дер-Ваальса

Состояние двух близко расположенных атомов, при котором происходит одновременное смещение электронов в обоих атомах, имеет более низкую энергию. Потому такая поляризация происходит при любом сближении атомов; от направления, в котором смещаются электроны, энергия взаимодействия не зависит и выбор направления происходит спонтанно.

Точный учет эффекта Ван-дер-Ваальса позволяет рассчитать, что потенциальная энергия при таком взаимодействии атомов будет обратно пропорциональна шестой степени расстояния, что можно записать как $F_6 = A_6/r^6$. Коэффициент пропорциональности A_6 в этой формуле, как правило, определяется в рамках подготовки самосогласованного силового поля, и может быть выражен через два параметра, характеризующие каждый из двух типов атомов, участвующих во взаимодействии.

К формуле для сил Ван-дер-Ваальса обычно добавляют второй член, характеризующий отталкивание ковалентно не связанных атомах на близких расстояниях, при некотором перекрытии их электронных облаков. Теоретически выведенной формулы, описывающей зависимость силы отталкивания между атомами от расстояния между ядрами, не существует, однако в вычислительных экспериментах обычно принимается, что энергия такого взаимодействия обратно пропорциональна двенадцатой степени расстояния. Таким образом, общую формулу для энергии взаимодействия несвязанных атомов обычно записывают как $F = A_6/r^6 - A_{12}/r^{12}$ - эта форма потенциала обычно называется *потенциалом Леннарда-Джонса*.

В молекулярном моделировании, вклад сил Ван-дер-Ваальса в общий баланс энергии обычно существенен на этапе стабилизации структуры; в состояниях, близких к равновесию, этот вклад мал. Модель попарных взаимодействий электронных оболочек, используемая в модели сил Ван-дер-Ваальса, является лишь простейшим приближением к описанию согласованной поляризации всех атомов в системе. Но модель сил Ван-дер-Ваальса используется в силовых полях в большей степени из-за простоты этого подхода, по сравнению с более сложными по форме, но не слишком эффективными силовыми полями, такими как силовое поле Amoeba.

Учет влияния растворителя

В классической теории электростатики, формула Кулона для взаимодействия между зарядами используется в модифицированной форме, чтобы учесть эффект поляризации вещества, находящегося в пространстве между зарядами: $F = q_1 q_2 / \epsilon r^2$. Здесь ϵ - так называемый коэффициент поляризации вещества, предполагается что $\epsilon > 1$. За счет такой поляризации электрическое поле ослабевает и сила взаимодействия становится меньше.

Молекулы воды, которыми окружен исследуемый белок или другая молекулярная система, являются полярными молекулами. С одной стороны молекулы воды находится отрицательно заряженный атом кислорода, а с другой — положительно заряженные атомы водорода. Поэтому поляризация некоторого объема воды состоит в том, что равновесное положение молекул воды в этом объеме несколько смешается, так что предпочтительным оказывается ориентация молекул по направлению электрического поля. Этот эффект оказывается более сильным, чем просто поляризация нейтральных атомов, и измерения показывают, что вода ослабляет электрическое поле приблизительно в 70 раз (рис. 2.17). Поэтому расчет сил взаимодействия в молекулярной системы, должен обязательно включать в себя вклад поляризации растворителя.

Внутренняя часть белка также ослабляет электрическое поле за счет поляризации атомов, и диэлектрическая проницаемость внутренней части белковой глобулы составляет около 4-5, по сравнению с коэффициентом 70 для воды.

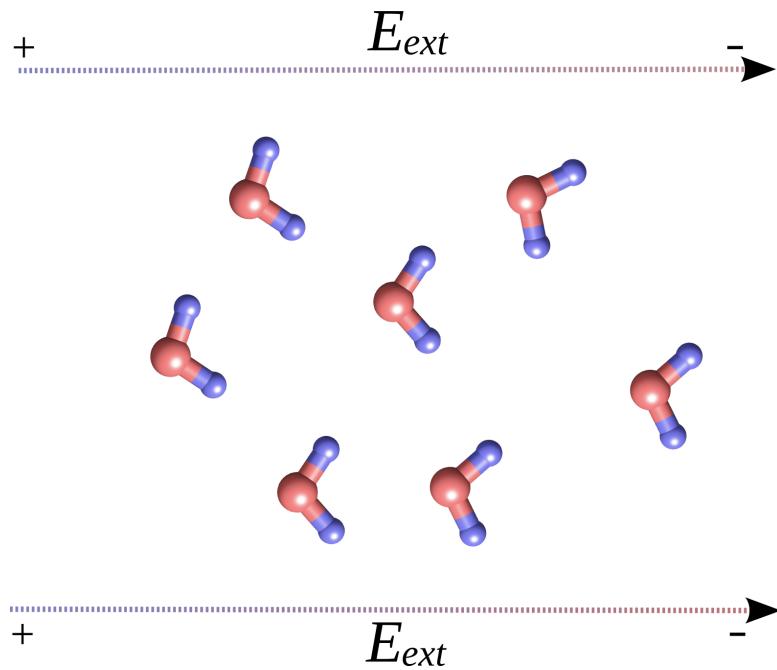


Рис. 2.17: Иллюстрация поляризации водного окружения за счет изменения ориентации молекул воды

В отсутствии электрического поля молекулы воды ориентированы в случайных направлениях, однако под действием поля предпочтительной ориентацией молекул становится направление, определяемое действием поля.

Уравнение поляризации диэлектрической среды, в котором учтена возможность перераспределения ионов в электрическом поле, называют *уравнением Пуассона-Больцмана*. Это уравнение выводится на основе уравнения Пуассона $\Delta\psi = \rho/\epsilon$ для потенциала ψ и плотности зарядов ρ , ϵ - диэлектрическая проницаемость среды. Далее при выводе уравнения Пуассона-Больцмана используют формулу Больцмана для распределения вероятностей частиц, $p = A \exp(-E/kT)$; здесь E - энергия частицы, T - температура системы, A - нормировочный параметр. Таким образом производится согласование распределения ионов в поле потенциала ψ и плотности зарядов ρ . По форме полученное уравнение является дифференциальным уравнением в частных производных, и алгоритм его решения достаточно ресурсоемок. Однако энергия молекулярной системы, где для учета вклада растворителя использовано решение уравнения Пуассона-Больцмана, по предположениям о степени полноты описания, наиболее близка к энергии, которую можно измерить в эксперименте, поэтому это уравнение решают для уточнения значений сил и зарядов в молекулярной системе. Такая коррекция потенциала электрического поля может быть не всегда заметна в малых масштабах расстояний (рис. 2.18). В физической модели баланса энергий при сворачивании белка, многие из частей невозможны рассчитать с приемлемой точностью. Однако, в такой модели, учет влияния растворителя вносит существенный вклад в энергию электростатических взаимодействий в белках, и такого рода поправки являются решающими при сворачивании белков.

Одним из приложений, в которых используется решение уравнения Пуассона-Больцмана, является расчет кинетических параметров, через которые можно определить степень окисления и восстановления боковых цепей некоторых аминокислотных остатков в зависимости от кислотности среды. В частности, показатель кислотности среды (pH), при котором в боковой цепи гистидина отщепляется один или оба атома водорода, близок к показателю кислотности нейтральной среды, и при моделировании молекулярной системы важен точный расчет электростатического потенциала для определения состояния ионизации остатков гистидина.

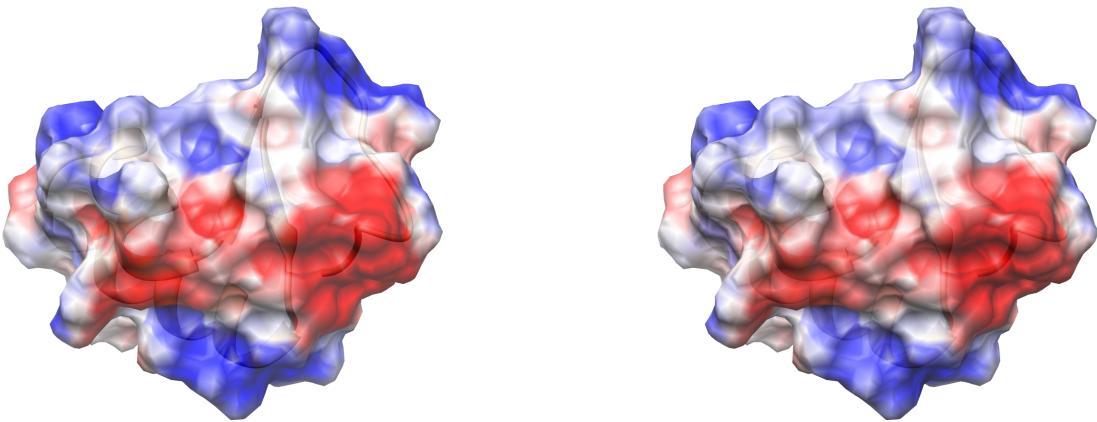


Рис. 2.18: Иллюстрация результатов расчета уравнения Пуассона-Больцмана

Распределение электростатического потенциала на поверхности белка крамбин (crambin), рассчитанное на основе парциальных зарядов атомов (слева) и с учетом влияния растворителя, на основе решения уравнения Пуассона-Больцмана (справа). Учет влияния растворителя выражается в незначительном ослаблении напряженности поля, и местами заменен в деталях оттенка поверхности белка.

Расчет электростатического потенциала на модели справа произведен с помощью пакета Delphi. Для визуализации был использован пакет UCSF Chimera. Использованная цветовая схема соответствует напряженности поля в диапазоне от -3 до 3 $\text{kcal mol}^{-1} \text{e}^{-1}$.

При моделировании движения молекулярной системы, когда силы взаимодействия необходимо рассчитывать на каждом шаге интегрирования, решение уравнения Пуассона-Больцмана требует чрезвычайно много вычислительных ресурсов. Обойти эту проблему можно двумя способами. Первый способ состоит в добавлении в систему явным образом необходимого количества молекул воды. Тогда явно можно рассчитать положение каждой молекулы воды и учитывать явным образом силы, действующие между атомами молекулы воды и атомами белка. Поскольку при моделировании положения и ориентации молекул воды будет автоматически учтен эффект поляризации воды, такая модель в целом будет учитывать влияние растворителя при расчете сил, действующих в молекулярной системе.

Второй способ основан на использовании так называемого *обобщенного приближения Борна* (Generalized Born approximation). Этот метод основан на том, что взаимодействие между парой зарядов в диэлектрической среде, которая представляет собой шар с некоторой постоянной

диэлектрической проницаемостью и находится внутри бесконечного объема с бесконечной диэлектрической проницаемостью, можно решить аналитически. Диэлектрическая проницаемость воды таким образом условно считают бесконечностью, белок приближенно представляют как шар с некоторой постоянной диэлектрической проницаемостью и при расчете взаимодействия между атомами белка ослабляющий эффект окружающего белок растворителя учитывается по аналитическим формулам, следующим из рассматриваемой модели.

Существование нескольких подходов для оценки вклада растворителя в баланс энергии позволяет, на конкретных моделях белковых систем и их комплексов, сравнить результаты, полученные с помощью этих подходов. И такое сравнение показывает (например, (Hou и др. 2011)), что эти подходы являются зачастую рассогласованными, и расчет энергии с использованием модели Пуассона-Больцмана, теоретически более точной, приводит к результатам, менее сходным с экспериментально наблюдаемыми, чем использование менее точных, в теории, моделей на основе обобщенного приближения Борна.

Гидрофобные взаимодействия

Эффекты статистической физики в задаче сворачивания белка позволяют учесть влияние растворителя и объяснить так называемые *гидрофобные взаимодействия* в белках. Понятие «гидрофобность» («водобоязнь») в этом случае относится к аминокислотным остаткам в белках. При оценках степени поляризации атомов в аминокислотных остатках (раздел 2.4) легко показать, что атомы азота и кислорода поляризуются значительно сильнее, чем атомы углерода. Аминокислотные остатки (20 типов) можно разделить на остатки, у которых боковые цепи состоят из слабо поляризованных атомов, и на те, чьи боковые цепи содержат полярные атомы кислорода и азота. И более того, некоторые остатки при физиологических показателях кислотности воды являются ионизированными, за счет потери протона у атома азота, или присоединения дополнительного протона к атому кислорода.

Известно, что молекулы, состоящие только из неполярных атомов, имеют низкие показатели растворимости, и такие вещества имеют склонность самопроизвольно отделяться от растворителя, как, например, капельки масла. Это явление можно охарактеризовать параметром, называемым гидрофобностью. Так же, по степени гидрофобности, можно условно разделить аминокислоты на гидрофобные (как, например, лейцин или валин) и гидрофильные (как, например, глутамин или лизин).

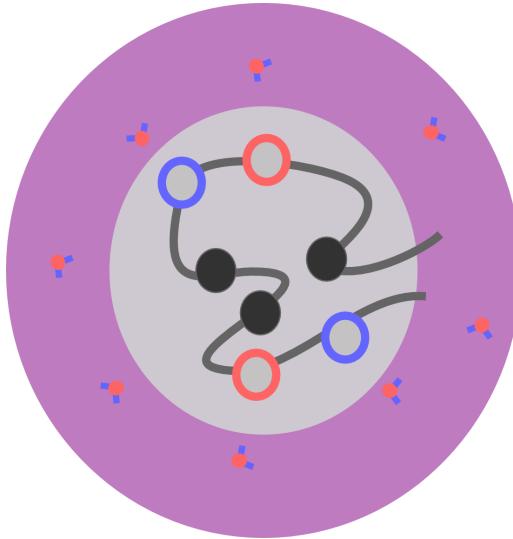


Рис. 2.19: Иллюстрация понятия гидрофобных взаимодействий в белке.
Гидрофильные остатки (показаны красным и синим) имеют тенденцию находиться на поверхности белковой глобулы, а гидрофобные остатки (показаны черным) - внутри глобулы.

При несложном анализе экспериментально определенных структур белков можно обнаружить, что в уложенной структуре белка гидрофобные остатки имеют свойство группироваться внутри белковой глобулы, а гидрофильные остатки имеют свойство находиться в большей степени на поверхности белка. Это можно охарактеризовать как так называемые гидрофобные взаимодействия, когда незаряженные части белка имеют свойство находиться рядом, аналогично тому, как маленькие капельки масла в воде взаимодействуют между собой, сливаясь в большую каплю (рис. 2.19). Этот эффект имеет коллективную природу, и объясняется тем, что молекулы воды, находящиеся вблизи части белковой глобулы, где на поверхности находятся неполярные атомы, обязательно будут иметь проигрыш в энергии по сравнению с молекулами, находящимися вдалеке от белковой глобулы.

Молекула воды является диполем и в любом объеме эти диполи в среднем ориентированы так, чтобы отрицательно заряженный кислород одной молекулы находился вблизи одного из положительно заряженных водородов другой молекулы. Если на поверхности белка находится заряженный любой полярностью атом, всегда есть возможность переориентировать окружающие молекулы воды так, чтобы вблизи этого атома находились заряды противоположного знака, без заметного проигрыша в энтропии. Если же на поверхности белка находится незаряженный атом, то при любой ориентации находящихся вблизи молекул воды нельзя избежать проигрыша в энергии. Таким образом, энергетически более выгодной оказывается конформация макромолекулы, когда на поверхности белка находятся заряженные атомы.

При моделировании сворачивания белка и предсказаниях структуры белков, исходя из объема вычислительной сложности расчетов, возможно использовать представление белка на уровне аминокислотных остатков, а не отдельных атомов. По аналогии с системами атомов, в таких расчетах также может быть введен принцип попарного взаимодействия частиц и понятие силового

поля. Учет гидрофобных взаимодействий, необходимый при исследовании сворачивания белков, следует выражать в таких моделях как взаимодействие пары неполярных остатков друг с другом. В соответствующем силовом поле, параметры взаимодействия можно оценить, рассчитав количество пар неполярных остатков белка, находящихся в близком контакте между собой и закрытых от растворителя. Однако недостаточная адекватность модели, как обратная сторона описанного подхода, иллюстрирует проблемы, возникающие при попытках определить точный смысл параметров в эмпирических силовых полях.

Молекулярная динамика

Численные методы, применяемые при моделировании молекулярных систем, развились из универсальных методов численного решения дифференциальных уравнений, относящихся к классу методов численного интегрирования. Простейший из этих методов известен как метод Ньютона; он состоит в последовательном расчете состояний системы в фиксированные моменты времени, отстоящие друг от друга на небольшой интервал, называемый шагом интегрирования. Более сложные методы используют несколько предыдущих состояний системы для расчета следующего состояния, для увеличения устойчивости и точности решения.

При расчетах молекулярных систем, необходимым является согласование уравнений Ньютона с подходами статистической физики. На этом шаге вводятся понятия температуры и давления, и уравнения движения корректируются с помощью введения случайных возмущений на каждом шаге численного интегрирования. Понятие шага интегрирования, однако, остается и в методах, применяемых при расчете молекулярных систем. При моделировании белковых молекул этот параметр для достижения максимальной точности обычно выбирается в диапазоне 0.5-2 фемтосекунды.

Учет эффектов статистической физики при расчетах траектории системы возможно произвести через введение случайных возмущений на каждом шаге моделирования. Температура молекулярной системы быть рассчитана из распределения молекул по скоростям, с использованием формулы Больцмана. При проведении моделирования движения системы следует корректировать координаты и скорости молекул системы, чтобы давление и температура рассматриваемой системы соответствовали заданным параметрам давления и температуры для внешней среды. Для такой коррекции, уравнения Ньютона, описывающие движение механической системы, используются в модифицированной форме, введением в уравнения так называемого *шумового члена*. В такой форме уравнение движения системы известно как *уравнение Ланжевена* и относится к *стохастическим дифференциальным уравнениям*.

Времена, характеризующее наиболее важные процессы в системах биомолекул, такие как сворачивание белка или образование комплекса белков, относятся в масштабу микросекунд и выше. Но методы численного решения уравнений движения молекулярных систем позволяют рассчитать траекторию системы лишь в пределах до 20-50 наносекунд. Ограничения применимости этих методов обусловлены как объемом требуемых вычислительных ресурсов, так и накоплением ошибок при численном решении уравнений движения: так, для моделирования системы в течении 20 наносекунд необходимо произвести 10 млн, или более, шагов интегрирования.

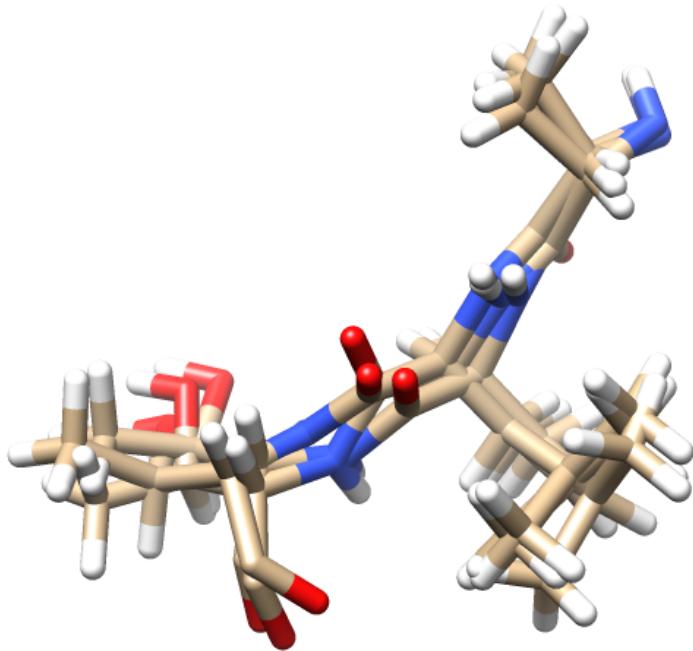


Рис. 2.20: Пример расчета траектории полипептида

Показаны три совмещенные модели аминокислотной цепи *ALA-LEU-THR*, рассчитанные как фазы траектории, полученной при моделировании движения белка по правилам молекулярной динамики. Изображенные состояния молекулы разделены интервалами времени, равными 0.3 пикосекунды.

Ковалентные связи в пептиде изображены как "палочки" (sticks), атомы находятся в местах соединения связей. Красным цветом показаны атомы азота, синим - атомы кислорода, золотым - атомы углерода, белым - атомы водорода.

Молекулярно-динамическое моделирование проведено с помощью пакета *Amber*, рисунок построен с помощью пакета *UCSF Chimera*. Для моделирования была использована экспериментально определенная структура белка Актиноксантина (код 1acx в банке данных структур белков); изображен фрагмент цепи белка из трех аминокислот.

Пример расчета траектории пептида показан на рис. 2.20. Расчеты траекторий белков в микроскопических масштабах времени, полученные с помощью молекулярного моделирования, позволяют во многих случаях получить значимые выводы о свойствах белковых систем. Разработка пакета программ, реализующего такие расчеты для белковых систем, требует значительных усилий программистов и физиков. Среди существующих пакетов программ, помимо пакета *Gromacs*, относящегося к свободному ПО, следует упомянуть пакеты *Amber* и *Charmm*, для использования которых следует получить лицензию.

Среди многих приемов, используемых при моделировании белков, следует упомянуть о необходимости минимизации энергии системы перед проведением моделирования, для коррекции неточно заданных координат атомов. При проведении такой коррекции, изменение координат в системе производят в направлении действия сил, без введения понятия скорости.

Сила электростатических взаимодействий, выражаемая законом Кулона, достаточно медленно убывает с расстоянием, и нет универсального способа оценить порог расстояния, после которого следует пренебречь взаимодействиями зарядов. Однако в молекулярном моделировании обычно явным образом задают значение *порога обрезания* (cutoff distance), после которого электростатические взаимодействия между атомами не учитывают. Значение этой величины обычно имеет порядок 11-14 ангстрем. При таком приближении молекулярно-динамические расчеты становятся менее точными, но значительно более быстрыми, поскольку это дает возможность разбить систему на относительно независимые части и проводить расчеты параллельно на нескольких ядрах или узлах компьютера.

Результатом расчета по молекулярно-динамическому моделированию является траектория молекулярной системы, и следующим шагом является интерпретация полученной траектории. Одним из приемов при обработке траекторий является возможность интерпретировать рассчитанные фазы траекторий, в терминах статистической физики, как *ансамбль состояний*. Такая возможность является следствием так называемой *эргодической гипотезы*, согласно которой, уравновешенная система при моделировании находится в каждом из возможных ее состояний такую часть времени, какова вероятность этого состояния в полном ансамбле состояний системы.

Оценка ансамбля состояний молекулярной системы позволяет рассчитать многие из ее статистических свойств. В некоторых случаях при анализе используют упорядочение и группировку наблюдаемых кадров траектории, как, например, для детекции нескольких возможных фаз состояния системы, и условий переключения между ними. Возможны и другие подходы для сведения результатов расчетов до уровня содержательных выводов, касающихся динамических свойств изучаемой системы.

Анализ нормальных мод

Методы аналитической механики, в рамках уравнений Ньютона, во многих случаях позволяют провести приближенный анализ поведения системы без необходимости численного решения точных уравнений движения. Из этих методов, при анализе молекулярных систем широко используется так называемый *анализ нормальных мод*.

Силы упругости, которые линейно зависят от координат атомов, приводят к наиболее простым для аналитического решения уравнениям движения; в общем случае точные уравнения движения системы являются нелинейными и более сложными для решения. Однако если предположить, что система при движении мало отклоняется от положения равновесия, то возможно упростить точные уравнения движения, оставив первый линейный член разложения функций силы в ряд Тейлора. Метод нормальных мод основан на анализе уравнений движения системы в линейном приближении и используется для расчета малых колебаний системы вокруг положения равновесия.

Шарик, прикрепленный к пружине, и качающийся маятник - это примеры систем, в которых происходят колебания вокруг положения равновесия. Сила упругости, с которой пружина действует на шарик, прямо пропорциональна отклонению шарика от равновесного положения.

В этом случае, как и в случае маятника, уравнение движения можно решить аналитически и решением будет гармоническая зависимость координаты x от времени t , то есть колебания выражаемые формулой $x = A \sin(\omega t + \phi)$. В этой формуле амплитуда A и фаза ϕ зависят от начального положения системы, а частота колебаний ω определяется массой частицы и степенью жесткости силы, которая возвращает частицу в положение равновесия (например, коэффициентом упругости пружины). Энергия при таком движении сохраняется, переходя из кинетической энергии движения частицы в потенциальную энергию и обратно.

Движение системы атомов или других частиц, связанных между собой через силы взаимодействия, находящуюся в одном из возможных устойчивых положений равновесия, также можно свести к совокупности колебательных движений, при условии что все эти колебания мало отклоняют систему от ее положения равновесия. Этот расчет можно провести с помощью методов матричной алгебры, и результатом будет набор так называемых «нормальных мод», направлений согласованных колебаний частиц в системе. Каждая нормальная мода будет характеризоваться своей частотой колебательных движений.

При исследовании движения белка методом нормальных мод с помощью линейной зависимости выражают не только упругие ковалентные связи, но и электростатические взаимодействия и силы Ван-дер-Ваальса, разложенные в ряд Тейлора вблизи положения равновесия. Как результат расчета нормальных мод, наибольший вклад в динамические характеристики белков будут вносить направления нормальных мод с наименьшими частотами колебаний. Это следует из подходов статистической механики, где выводится принцип о равномерном распределении энергии по всем возможным независимым направлениям движения системы. Из уравнений колебательного движения легко получить, что полную энергию колебаний, в случае простого одномерного движения, можно записать как $E = kA^2$. Поскольку у колебаний с более высокой частотой коэффициент жесткости упругой силы k больше, то амплитуда A будет наибольшая у самых медленных колебаний. И, таким образом, медленные колебания будут проявляться в большем изменении координат атомов при коллективных согласованных движениях частей белка.

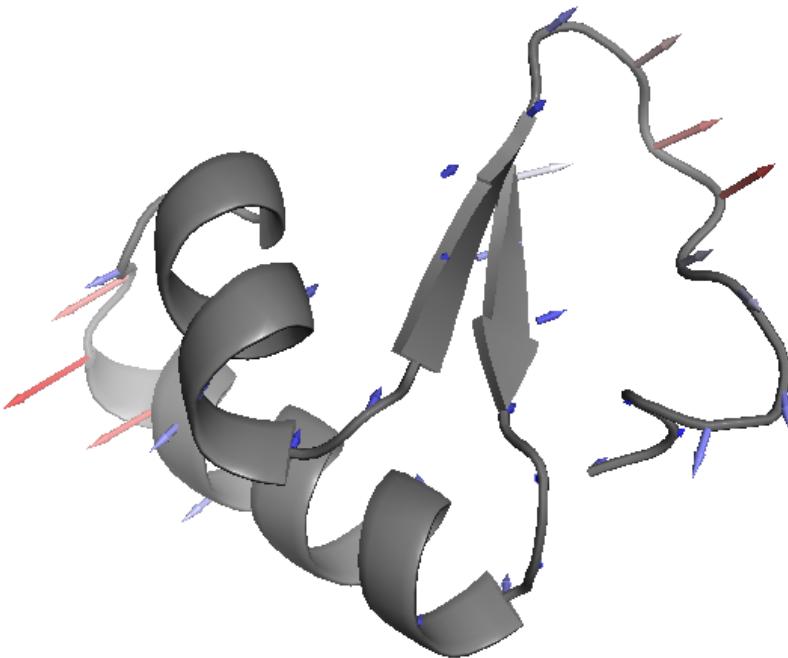


Рис. 2.21: Пример расчета нормальных мод белка

Показана структура белка *crambin* (крамбин), стрелки показывают направления и амплитуду колебаний согласно седьмой нормальной моде.

При каноническом анализе нормальных мод, первые 6 мод относятся к поступательным и вращательным движениям белка как целого, и седьмая (или первая "нетривиальная") нормальная мода описывает колебательные движения, происходящие с наибольшей амплитудой.

Расчет нормальных мод проведен с помощью пакета *bio3d*, рисунок построен с помощью пакета *rutmol*.

Для белковой молекулы, изображенной на рис. 2.21, анализ нормальных мод позволил рассчитать характерные направления колебательных движений (изображенные стрелками на рисунке). Безусловно, такой расчет является грубым приближением; в более точных моделях следует учитывать тепловые движения атомов белка, в основной цепи и в боковых цепях остатков. Однако, при усреднении этих тепловых движений, направления коллективного движения, выделенные с помощью анализа нормальных мод, будут проявляться в большей степени, чем другие направления.

Но, из постановки задачи в анализе нормальных мод следует, что применения этого метода ограничены исследованиями малых колебаний системы в окрестности положения равновесия. Более сложные структурные перестройки в белковых молекулах рассчитать с помощью этого подхода невозможно.

Моделирование Монте-Карло

Класс методов вычислительной математики, известный как "*методы Монте-Карло*", используется, в общем случае, для решения задач численного интегрирования, различных по постановке и

из многих предметных областей. Название "Монте-Карло" происходит от имени города - столицы княжества Монако, одного из центров игорного бизнеса с давней историей. И, также, схемы расчетов в методах Монте-Карло основаны на использовании псевдо-случайных чисел.

При оценке отношения потраченных вычислительных ресурсов к полученной точности результата, в задачах где возможно сравнение методов Монте-Карло и детерминированных алгоритмов, методы, основанные на псевдо-случайных числах, имеют принципиальные ограничения в эффективности. Однако эти методы используются во многих задачах для проведения примерных и прикидочных расчетов, отчасти из-за простоты реализации и возможности их интуитивно понятной интерпретации.

При моделировании молекулярных систем, возможно использовать методы из класса Монте-Карло для расчета ансамбля состояний, наряду с молекулярно-динамическим моделированием. В этих подходах, каждое следующее состояние системы рассчитывают на основании предыдущего, через внесение малого изменения системы в некотором случайно выбранном направлении. Энергия системы в каждом из состояний может быть оценена с использованием силовых полей, как это описано в разделе 2.4. Эффекты статистической физики учитывают через введение параметра температуры, когда в рамках алгоритма возможно принять новое состояние системы или отклонить его. Вероятность перехода в новое состояние, если его энергия хуже чем у предыдущего, тем больше, чем больше температура системы.

Анализ ансамбля состояний, и все методы его оценки, в наибольшей степени важен для поиска кластера состояний, соответствующего наиболее выгодному устойчивому положению системы. Но одна из проблем при моделировании такого рода состоит в том, что все состояния системы, полученные при расчетах, могут являться *метастабильными*; расчеты при других начальных параметрах приводят к сходимости в окрестности другого состояния, существенно отличающегося от других найденных вариантов.

2.5. Структура и сворачивание белка

Баланс энергии при сворачивании белка

Понятие энтропии и уравнения статистической физики позволяют, на качественном уровне, описать явление сворачивания белков. Время, в течение которого можно моделировать молекулярную систему, составляет не больше сотен наносекунд, в то время как процесс сворачивания занимает не меньше нескольких микросекунд, и невозможно средствами молекулярной динамики полностью восстановить процесс сворачивания белка, за исключением нескольких простых моделей. Тем не менее, теоретические расчеты и эксперименты позволяют составить качественную картину процесса сворачивания. Эти результаты, в частности, показывают, что выигрыш в энергии свернутой структуры по сравнению со свободно плавающей цепью аминокислотных остатков составляет порядка сотен килокалорий на моль, и не всякая произвольно заданная цепь аминокислотных остатков может принять определенную форму, характерную для белков, существующих в клетке.

Если рассчитать потенциальную энергию модели белка с учетом всех сил, действующих в модели, в первую очередь электростатических взаимодействий, то абсолютная величина выигрыша в энергии будет в десятки и сотни раз больше чем наблюдаемая в эксперименте энергия сворачивания. Разница объясняется уменьшением энтропии свернутой структуры по сравнению с развернутой. А именно, из уравнений статистической физики следует, что при уравновешивании системы стремится к минимуму величина так называемой *свободной энергии* F , которая выражается через механическую энергию системы E , энтропию S и температуру T по формуле $F = E - TS$. Белок в свернутом состоянии имеет более высокую степень упорядоченности, чем свободно плавающая цепочка, и энтропия у белка в свернутом состоянии меньше чем в развернутом. В то же время, в свернутом состоянии белок находится в потенциальной яме, так что энергия системы меньше чем в развернутом состоянии. Согласно определению свободной энергии, энтропийный и энергетический (энталпийный) вклады в эту энергию уравновешивают друг друга. При возрастании температуры T энтропийный член приобретает больший вес по сравнению с энталпийным, чем объясняется разворачивание белка при нагревании.

Из этого рассуждения следует, что каждый из белков имеет собственный эволюционно консервативный механизм сворачивания, обеспечивающий преобладание энергетического члена над энтропийным для конкретного белка и конкретной температуры. Выигрыш в энергии может достигаться за счет правильных сочетаний положительно заряженных атомов с отрицательно заряженными, в частности за счет образования вторичной структуры белка и взаимодействия с растворителем.

Для большинства белков возможен только качественный анализ механизма сворачивания, поскольку, в отличие от энергетического члена, энтропия является функцией ансамбля всех возможных состояний системы, и расчет энтропийного члена практически невозможен численно. Если постановка задачи требует численной оценки энергии сворачивания или какого-либо другого взаимодействия в системе, обычно рассчитывают энергетический член в функции свободной энергии, а энтропийный член приблизительно оценивают как пропорциональный энергетическому члену, возможно с учетом дополнительных эмпирических поправок.

Задачу предсказания третичной структуры белка по его первичной последовательности нельзя считать решенной. В физическом механизме сворачивания белка заключен парадокс, называемый парадоксом Левинталя. Суть этого парадокса в том, что множество вариантов укладки, в которые может свернуться белок, велико, и если рассматривать эти варианты как статистический ансамбль, в котором белок достигает положения минимума свободной энергии, то время уравновешивания системы до этого минимального уровня, с учетом времени обхода каждого из состояний, будет чрезвычайно долгим, в то время как сворачивание белков в природе происходит за время, находящееся в пределах секунды.

Среди сил, удерживающих белок в свернутом состоянии, главную роль играют гидрофобные (рис. 2.22) и электростатические взаимодействия между частично поляризованными и/или ионизированными атомами. Электростатические взаимодействия можно условно разделить на водородные связи, стабилизирующие элементы вторичной структуры, и взаимодействия между другими парами атомов в несмежных остатках, в основном между поляризованными атомами в

боковых цепях.

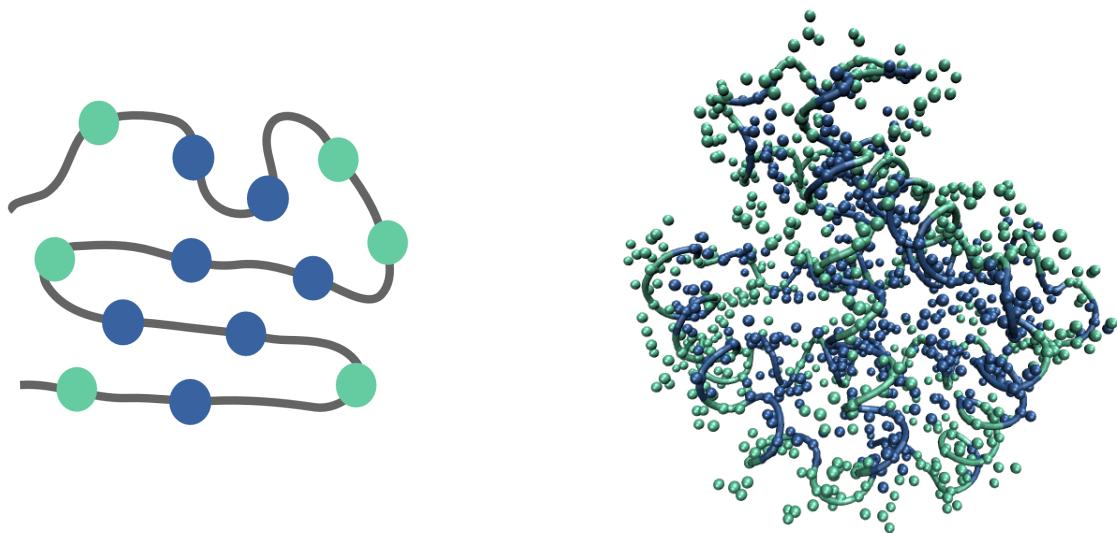


Рис. 2.22: Иллюстрация принципа выбора предпочтительной укладки белка на основании гидрофобности остатков

Слева: Схема укладки цепочки из гидрофильных (светлые кружки) и гидрофобных (темные кружки) элементов в глобулу. В предпочтительных типах укладки гидрофобные элементы должны группироваться внутри глобулы

Справа: Структура белка с цветовой схемой остатков согласно их гидрофобности. Темным цветом показаны гидрофобные остатки, светлым - гидрофильные. Цвет атомов соответствует цвету аминокислотного остатка, к которому они принадлежат.

Использована структура белка гемоглобина человека (код 1a3n). Для визуализации белка был использован пакет VMD.

Если гидрофобные взаимодействия и электростатические взаимодействия между несмежными атомами являются в большей степени характеристиками конкретного типа укладки белка целиком, то возникновение вторичной структуры белка является в большей степени свойством локальных фрагментов полипептидной цепи. Одним из следствий этого наблюдения является возможность с высокой достоверностью определять свойства вторичной структуры белка по первичной последовательности с помощью вычислительных методов.

Методы предсказания структуры белков

Задачу предсказания вторичной структуры белка можно сформулировать как задачу классификации, когда каждый из остатков в белке необходимо отнести к одному из трех классов: α -спираль, β -складка, и изгиб цепи (петля), соединяющий упорядоченные участки. И подходы к задачам классификации и распознавания образов, широко используемые во многих прикладных задачах, оказываются эффективными и в задаче предсказания вторичной структуры белка. Из этих подходов, первый и наиболее развитый метод предсказания вторичной структуры основан на моделях искусственных нейронных сетей (ИНС). Тренировки ИНС и решение задачи

классификации в этом случае основана на информации о смежных остатках белка, и на информации о наиболее частых заменах аминокислот в гомологичных белках, для этих остатков. Задачу классификации можно поставить также для более детального разделения вторичной структуры на восемь типов, для разделения остатков на погруженные в глобулу (buried) и находящиеся на поверхности (exposed), и для отделения неупорядоченных (disordered) участков белка.

На качественном уровне можно прояснить подобную классификацию остатков, используя характерные свойства отдельных аминокислотных остатков или коротких фрагментов белка. Так, например, гидрофобные остатки с большей вероятностью будут находиться внутри глобулы; остаток пролина не может входить в состав α -спирали. Период оборота α -спирали составляет около 3.5 аминокислотных остатка, и если в некотором фрагменте последовательности гидрофильные остатки встречаются через 3 или через 4, можно предположить, что фрагмент цепи уложен в спираль и одной из сторон эта спираль ориентирована вдоль внешней стороны глобулы. Однако, хорошо проверены выводы о том, что произвольная последовательность аминокислот с очень малой вероятностью примет форму белковой глобулы, и что для каждого типа белков, закодированных в геномах, сворачивание происходит по сценарию, определяемому типом укладки всего белка или домена. Эти выводы указывают на ограничения в достоверности классификации, основанной на эмпирических подходах.

Наиболее часто используемыми методами моделирования собственно третичной структуры являются методы моделирования по гомологии, когда третичная структура неизвестного белка строится по экспериментально определенной структуре некоторого белка, взятого как шаблон, на основе парного выравнивания аминокислотных последовательностей моделируемого белка и шаблона.

Процесс построения структуры в этом случае будет заключаться в изменении формы и длины коротких фрагментов структуры и моделировании положения боковых цепей. Предпочтительно при этом, чтобы перестраиваемые короткие фрагменты находились не в местах регулярной вторичной структуры белка, а в местах петель и неупорядоченных участках. В задаче моделирования петель следует выбрать петлю с минимальной энергией, при этом функционал энергии необходимо строить подобно функционалу энергии в молекулярно-динамическом моделировании, на основе заранее заданного силового поля для атомов белка. Целью при моделирования петель не является составление или расширение элементов регулярной укладки в белке, в большинстве случаев петли - это относительно неупорядоченные фрагменты структуры.

Положение боковых цепей остатков при известных координатах основной цепи белка также является задачей, решаемой в рамках моделирования по гомологии. При этом ориентация боковых цепей также может быть задана через ориентации двугранных углов, задающих вращение атомов боковой цепи. Часто для этих двугранных углов существуют предпочтительные ориентации, связанные с взаимодействием смежных электронов в атомах. Обработка экспериментально полученных структур показывает, что всевозможные ориентации боковых цепей каждого из остатков можно свести к относительно небольшому количеству вариантов, и для предсказания координат атомов боковых цепей следует перебрать эти ориентации для каждого из остатков. В

большинстве случаев этот перебор можно существенно ускорить, пользуясь фактом независимости ориентации одной группы боковых цепей относительно всех других групп.

Метод моделирования по гомологии оказывается эффективным для предсказания структуры большой части белков. Кроме того, было замечено, на основе анализа совокупности расшифрованных белковых структур, что часто белки несходные по последовательности и имеющие разные функции обладают схожим типом укладки. Таким образом, можно приблизительно классифицировать известные варианты укладки белков. По результатам анализа десятков и сотен тысяч расшифрованных структур, большую часть из них можно отнести к одному из характерных типов укладки. Существуют базы данных, например, SCOP и CATH, где вручную или автоматически белки классифицированы по типам укладки; количество типов укладки (*superfamily*) в базе CATH сейчас оценивается в пределах 7 тыс., на основании автоматической классификации белковых доменов. В связи с этим, для предсказания структуры белка иногда достаточно найти подходящий шаблон структуры, даже если для него нельзя найти близкородственных белков с известной структурой.

Тем не менее, в ряде случаев необходимо предсказать структуру белка из первых принципов (*de novo*). К решению этой задачи существует несколько подходов. Попытки моделировать процесс сворачивания с помощью молекулярной динамики оказывается успешным для простых белков, но неэффективным в большинстве случаев, из-за проблем, возникающих при превышении пределов времени моделирования. Наиболее эффективные из существующих подходов к предсказанию структур белков *de novo* основаны на использовании так называемых "библиотек шаблонов", коротких (от 4 до 10 остатков) фрагментов известных структур белков, среди которых на основании сходства с последовательностью моделируемого белка выбираются шаблоны для конструирования его структуры. Степень достоверности при подборе таких шаблонов невелика, и потому для достижения приемлемых результатов в предсказаниях структур используют много вариантов шаблона для каждого фрагмента модели. Это приводит к конструированию большого количества структур-макетов, из которых следует выбрать наиболее подходящую модель на основании эмпирических критериев, проверенных на известных структурах белков. До стадии выбора наилучших из возможных предсказаний, структуры-макеты уточняются и оптимизируются. При таком конструировании, из-за необходимости перебора слишком большого количества вариантов, обычно используются методы Монте-Карло, то есть случайное блуждание и случайный выбор направления поиска.

Так называемый "эксперимент CASP", серия мероприятий, проводимых для оценки достоверности методов предсказания структур белков, разработанных различными научными группами, состоит в выборе наиболее достоверных из предсказаний на основании сравнения с экспериментально определенными структурами некоторых белков, которые не разглашаются до окончания очередного этапа "эксперимента". Наибольший интерес при этом представляет эффективность методов предсказания *de novo*; но методами предсказания *de novo* возможно получить приемлемые результаты только для белков небольшого размера, не более 200 аминокислотных остатков. Среди белков, представляющих интерес при постановке экспериментов по определению структуры, существенную часть составляют белки большого размера с неизвестной укладкой, состоящие

из одного домена или из нескольких доменов. Но в эксперименте CASP в качестве целевых заданий обычно выбираются белки с доменами небольшого размера, где ожидается, что предсказания, полученные некоторыми из методов, могут быть сходны с экспериментальной структурой.

Восстановление путей сворачивания белка

Задача выбора оптимальной укладки цепочки, состоящей из произвольно следующей последовательности гидрофобных и гидрофильных остатков, так что в оптимальной укладке максимальное количество гидрофильных остатков должно оказаться на поверхности, является алгоритмически сложной (*NP-hard*), что согласуется с парадоксом Левинталя, упомянутым ранее.

Сворачивание белков, однако, происходит парадоксально быстро; изучение кинетики сворачивания дает основания сравнить этот процесс с так называемыми *фазовыми переходами*, такими как спонтанное возникновение магнитных свойств в кристаллах железа.

Фазовые переходы нельзя описать в рамках классической статистической физики, но методы статистической физики могут быть расширены на для описания некоторых неравновесных систем; так, например, для качественного описания фазового перехода в ферромагнетике используется так называемая *модель Изинга*. В рамках подходов неравновесной статистической физики, процесс укладки белка имеет сходство с процессом, называемым *коагуляцией*, как, например, образованием крупных капель из частиц паров в дождевом облаке (рис. 2.23).

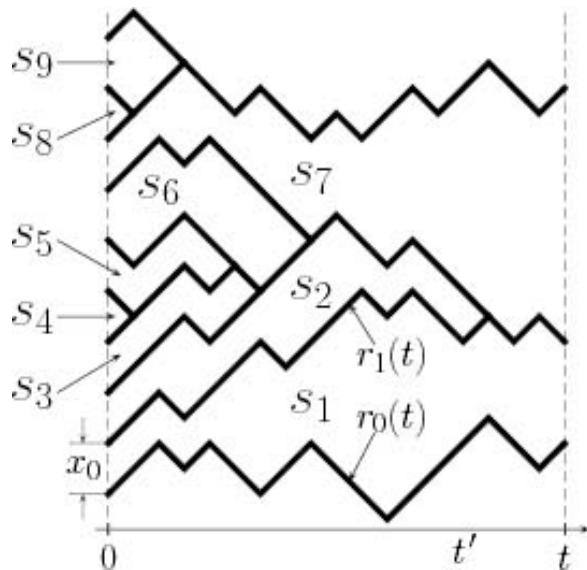


Рис. 2.23: Схематическое представление кинетики процесса коагуляции.

Рисунок взят из доклада Colm Connaughton "Nonequilibrium Statistical Mechanics of Cluster-cluster Aggregation" (2011).

В процессе полимеризации белка играют роль те же взаимодействия, которые участвуют в стабилизации окончательной структуры белка. К таким взаимодействиям следует в первую очередь отнести водородные связи, стабилизирующие элементы вторичной структуры белка. Один и тот же короткий фрагмент полипептидной последовательности может в разных белках формировать разный тип вторичной структуры. Укладка фрагмента во вторичную структуру может

происходить спонтанно, и энергия, удерживающая вторичную структуру фрагмента, является недостаточной для достаточно долгого времени жизни структуры в отсутствии дополнительных внешних факторов, стабилизирующих укладку. Поэтому неправильно образованные укладки фрагмента будут со временем распадаться, а правильно образованные фрагменты укладки будут объединяться в кластеры, подобно тому как объединяются частицы пара при их коагуляции в крупную каплю. В отличии от частиц пара, участки белка при объединении могут образовать разные варианты структуры, с большей или меньшей степенью выигрыша в энергии. Некоторые варианты укладки главной цепи белка приводят к выигрышу в энергии независимо от типов боковых цепей, и потому при сравнении структур известных белков можно выделить характерные "структурные мотивы", повторяющиеся в белках с различной укладкой (рис. 2.24).

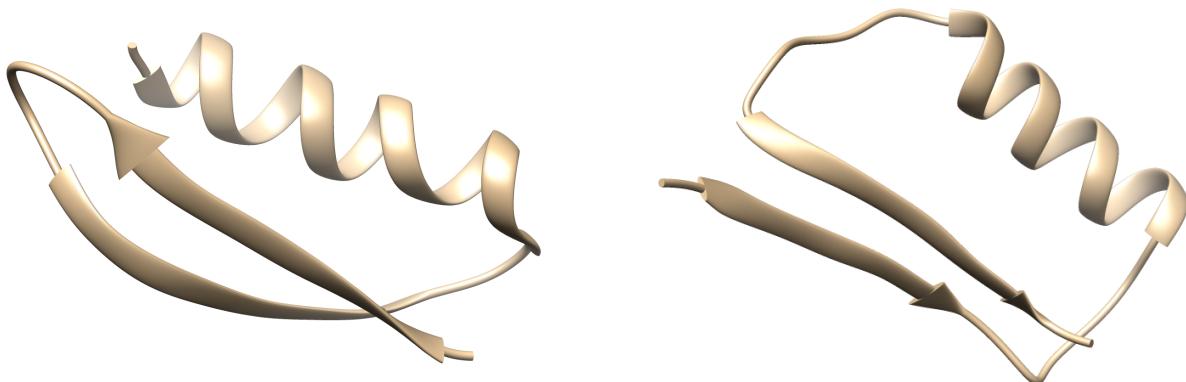


Рис. 2.24: Некоторые характерные структурные мотивы при укладке белка

Слева - бета-шилька (*beta-hairpin*), включенная в мотив $\beta\beta\alpha$. Справа - мотив $\beta\alpha\beta$ с параллельной ориентацией бета-складки.

Использованы структуры белков 5bti (*immunoglobulin G-binding protein G*) и 8tim (*triose phosphate isomerase*). Для визуализации структур был использован пакет UCSF Chimera.

В теории сворачивания белка вводят понятие “путь сворачивания”, объясняющее динамику образования структуры белка из несвернутой цепи. Молекулярно-динамические и кинетические эксперименты показывают, что даже для простейших белков при сворачивании у них существует несколько альтернативных путей. Детальный анализ структур и последовательностей белков показывает, что взаимодействие между некоторыми частями белков на этапе образования структуры является более предпочтительным, чем взаимодействие между другими смежными частями. На основе этих представлений можно ввести понятие об иерархии взаимодействия фрагментов в укладке белка, как показано на рисунке 2.25

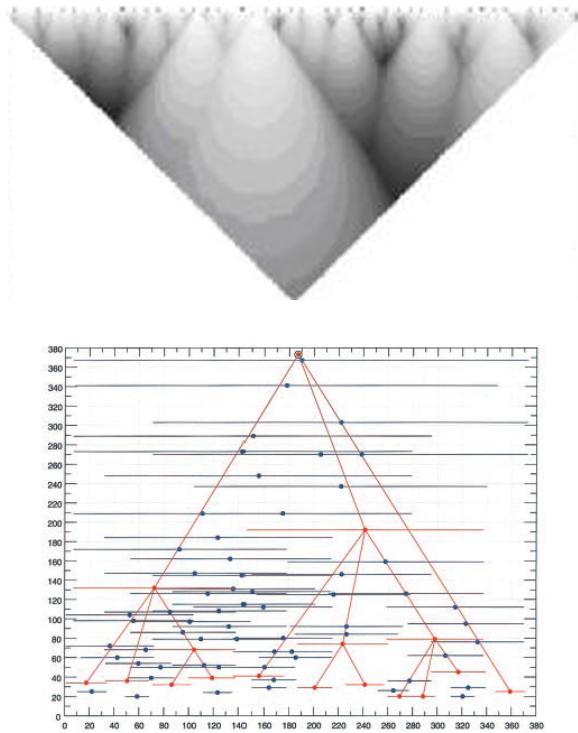


Рис. 2.25: Иерархия в укладке белков, построенная с помощью различных методов.
Рисунки взяты из статей (Козырев, Ю.П. и др. 2010; C.-J. Tsai и др. 2000)

Элементы вторичной структуры, как и устойчивые структурные мотивы, могут комбинироваться в разных сочетаниях в известных структурах белков (рис. 2.26). Но, несмотря на сложность оценки энергии белка вычислительными методами, часто оказывается возможным с помощью оценок энергии выбрать правильную комбинацию структурных мотивов из нескольких возможных альтернатив.

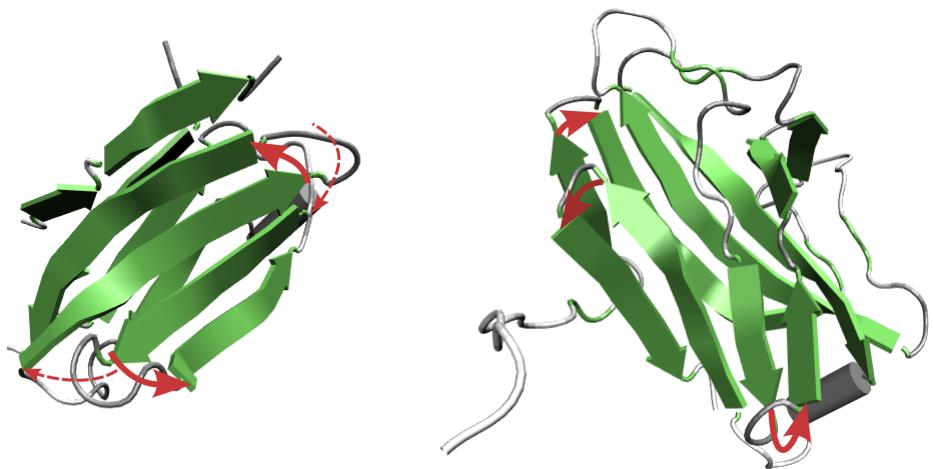


Рис. 2.26: Два белка с укладкой типа "бета-сэндвич"

Несмотря на сходство в типе укладки, порядок следования бета-штилек при образовании двух слоев "сэндвича" у белков различается.
Использованы структуры белков 3d33 и 5bxg. Для визуализации структур был использован пакет VMD.

Порядок объединения фрагментов в структуре типа "бета-сэндвич" оказалось возможно восстановить с помощью моделирования. Для достижения правильного выбора, в оценках функционала энергии оказалось необходимо учесть взаимодействие между остатками, степень гидрофобности ориентированных "вовне" остатков, а также степень подвижности петель, соединяющих фрагменты, уложенные в бета-листы. Степень подвижности петли определялась через количество остатков в петле и расстояние между ее концами. Учет подвижности петель позволил уточнить энтропийный член в оценках свободной энергии.

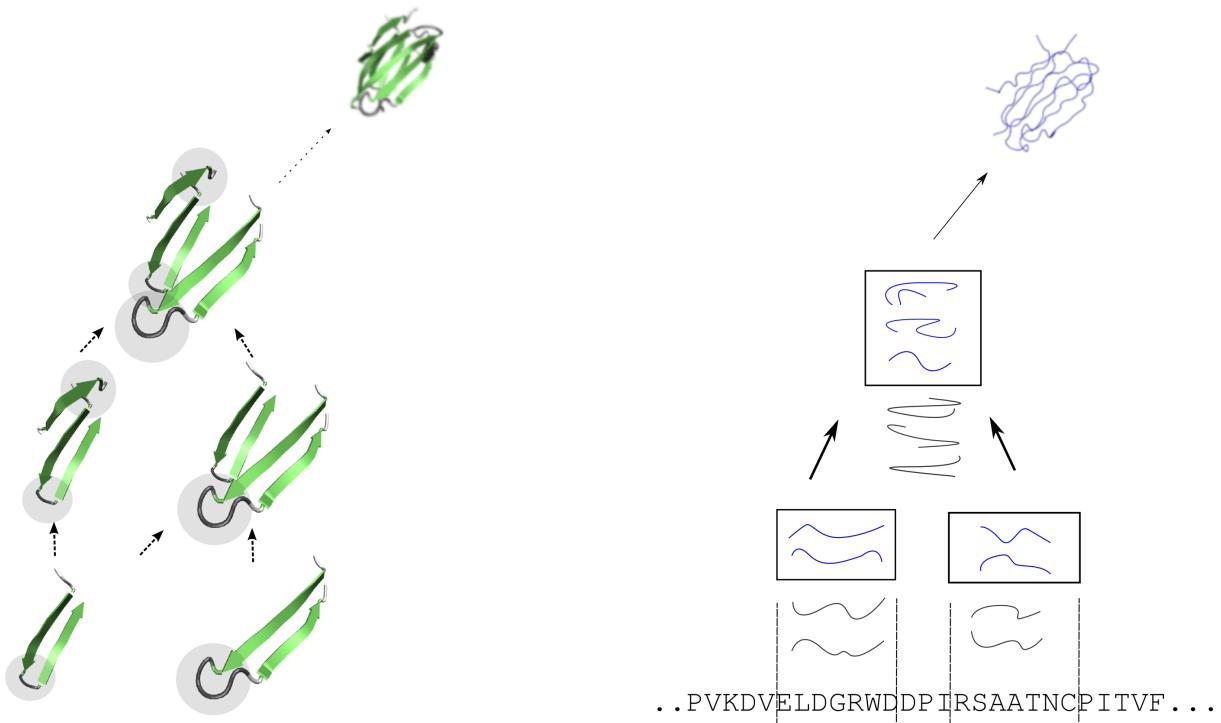


Рис. 2.27: Иллюстрации к описанию метода восстановления путей сворачивания белка
Слева: Принцип поиска правильного порядка при объединении фрагментов белковой цепи, на примере структуры 3d33 (рис. 2.26). Участки, соответствующие петлям, отмечены темным.
Справа: Принцип восстановления иерархии укладки белка. Для каждого фрагмента цепи выбираются лучшие по энергии варианты структуры.

В описанном выше подходе, следует проводить расчет энергии для каждого из вариантов объединения фрагментов белка. Потому, для возможности восстановить укладку полного белка за приемлемое время, наибольшее значение имеет вопрос о порядке перебора комбинаций фрагментов. Количество комбинаций фрагментов при проведении перебора возможно ограничить, рассматривая лишь участки белка, смежные в аминокислотной последовательности (рис 2.27). Это принципиально сокращает количество рассматриваемых комбинаций, в то время как оценка всех возможных комбинаций при полном переборе несоразмерно велика.

Описанную схему восстановления укладки оказалось возможно обобщить до уровня алгоритма предсказания структуры, применимого к существенной части белков, у которых длина

последовательности ограничена примерно 100-120 остатками. Методы классификации позволяют довольно точно предсказать положение элементов вторичной структуры, таких как бета-складчатые участки, в последовательности белка. Далее, смежные по последовательности элементы возможно объединять, на основе встречающихся в базе структур взаимных ориентаций элементов вторичной структуры, подобных изображенным на рис. 2.24. Для смежных элемента вторичной структуры следует рассматривать несколько вариантов их взаимных ориентаций, но при этом возможно сохранить преимущества описанной выше схемы расчетов, если на каждом шаге отделять лишь несколько наиболее предпочтительных по энергии укладок.

В подобном подходе можно заметить приложение принципа "разделяй и властвуй", когда сокращение количества расчетов при составлении структуры целого белка достигается за счет разделения его последовательности на части. Но при этом необходимым условием такого разделения является взаимная совместимость частей внутри одного фрагмента. Подобный подход к составлению алгоритмов известен под названием *динамическое программирование*. Алгоритм попарного выравнивания последовательностей, упоминаемый ниже в разделе 3.4, также относится к этому классу. По аналогии с такого рода разделением, было также предложено разделение нейронов на слои, как это описано в разделе 3.8 ("Модель сети нейронов с двумя типами возбуждения"). Совместно с другими введенными в этой модели допущениями, это принцип позволил бы объяснить, почему переработка информации в нервной системе происходит настолько быстро, по сравнению с объемом сохраняемых знаний.

Но впрочем, примерное объяснение физических принципов сворачивания белка, изложенное выше, и даже замечаемое в некоторых случаях сходство предсказанных структур с наблюдаемыми в эксперименте, имеют мало отношения к задачам по исследованию белков, которые встают во многих прикладных задачах молекулярной биологии. Укладка каждого из белков в геноме остается устойчивой в течении многих этапов эволюции, и последовательность каждого из белков несет свою "смысловую нагрузку", связанную с функциями белка в клетке.

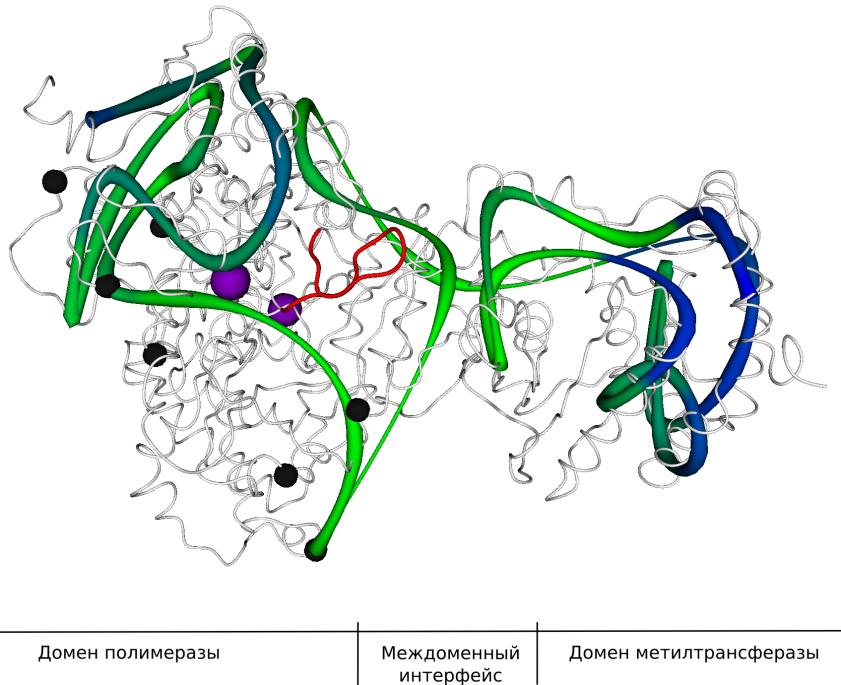


Рис. 2.28: Неструктурный белок 5 (NS5) вируса клещевого энцефалита

По структуре, белок можно разделить на два домена, и в обоих доменах, по их устройству, может производиться катализ определенных химических реакций. Основная функция большего из доменов - катализ процесса копирования (полимеризации) РНК вируса.

Цветная трубка, проведенная по результатам анализа нормальных мод, иллюстрирует механизм взаимодействия доменов белка, через участок междоменного интерфейса (*inter-domain interface*).

по материалам (Potapova, S.I. Feranchuk и др. 2018)

Белок NS5 вируса клещевого энцефалита, показанный на рис. 2.28 - один из возможных примеров, демонстрирующих различие между объяснением принципа сворачивания белков и исследованием функций каждого из белков. Укладка аминокислотной цепи, такая что образующаяся молекула белка может катализировать полимеризацию цепи вирусной РНК - устойчива и сохраняется при эволюции, даже несмотря на существенное различие в последовательностях у вирусов из различных родов.

2.6. Модели взаимодействия биомолекул

Полноценная живая клетка содержит десятки тысяч типов молекул, и живая природа клетки проявляется в том, что за счет взаимодействия этих молекул поддерживается постоянство состава клетки при различных влияниях внешней среды. Для составления моделей живой системы необходимо строить не только модели движения отдельных белков и низкомолекулярных соединений, но и модели их взаимодействия. Попытка моделировать взаимодействие биомолекул через молекулярно-динамическое моделирование системы, содержащей эти молекулы, неэффективна.

Причина кроется в том, что, подобно тому, как при сворачивании белка существуют локальные минимумы в профиле энергии белка, соответствующие метастабильным состояниям, так и при взаимодействии двух молекул возможно образование относительно неустойчивых метастабильных комплексы этих молекул, но время жизни этих комплексов обычно больше чем реалистичное время молекулярно-динамического моделирования.

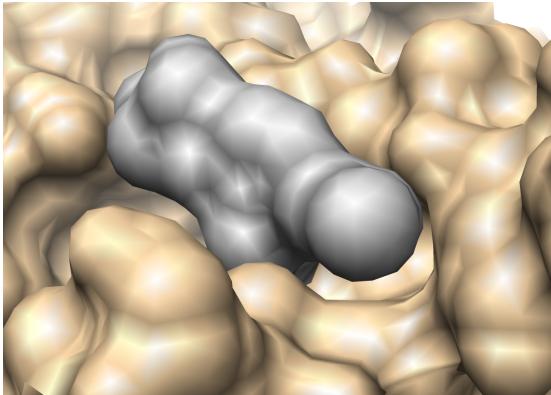


Рис. 2.29: Схема, иллюстрирующая метод докинга при построении моделей комплексов биомолекул

credits: U.S. Navy photo by Mass Communication Specialist 2nd Class Peter D. Blair

использована структура протеазы вируса клещевого энцефалита (Potapova, S.I Feranchuk и др. 2012).

Основным методом, позволяющим строить устойчивые модели комплексов биомолекул, является метод, называемый *докинг* (docking). Термин "докинг" введен по аналогии с технологией, когда корабль при ремонте помещают в "док", совпадающий по внутренней форме с формой корабля. Так и построение моделей комплексов молекул основывается на взаимном дополнении формы этих молекул, подобно схеме, изображенной на рисунке 2.29.

Решающим при выборе окончательной модели является функционал энергии. В случае докинга энергия связи рассчитывается как разность энергии комплекса и суммы энергии каждой из молекул, рассмотренной вне комплекса. Для расчета функционала энергии при докинге рассматривают широкий спектр методов, начиная от простейших эмпирических оценок, включая расчет энергии с помощью силовых полей, подобных силовым полям молекулярной динамики, и заканчивая расчетами с использованием уравнения Пуассона-Больцмана и иногда даже квантовыми расчетами. Выбор метода расчета энергии в основном определяется объемом вычислений, которые возможно потратить на расчет модели.

При подборе оптимальной взаимной ориентации двух молекул также в общем случае, следует учесть, что при образовании комплекса молекулы могут изменить в большей или меньшей степени свою конформацию по сравнению с конформацией, которую каждая из молекул принимает при независимости от другой молекулы. Для разделения методов докинга, которые учитывают либо не учитывают внутреннюю подвижность молекул, применяют термины *гибкий докинг* (flexible docking) и *жесткий докинг* (rigid docking). Также часто разделяют степени свободы при внутреннем движении молекул и проводят поиск оптимального положения докинга при изменении лишь части из внутренних степеней свободы каждой из молекул.

Конкретные методы докинга существенно отличаются при поиске модели взаимодействия белка с низкомолекулярным соединением (лигандом) и между двумя белками. Соответственно, разделяют методы докинга белка и лиганда (protein-ligand docking) и методы белок-белкового докинга (protein-protein docking).

Расчеты по докингу белка и лиганда становятся более реалистичными, если задан участок в структуре белка, куда следует встроить молекулу лиганда. Обычно в качестве такого участка выбирают некоторую полость, подобно изображенной на рисунке 2.30 (слева). Наличие полости является предпочтительным, чтобы аффинность связывания белка и лиганда была достаточной. В наилучшем из ожидаемых вариантов, подобранный, с использованием докинга, молекула лиганда при связывании с белком-мишенью будет изменять динамические свойства белка, как, например, ингибировать (подавлять) активность фермента. Поиск сайта связывания является отдельной задачей, которая часто может быть решена за счет дополнительных экспериментальных процедур, например, *сайт-направленного мутагенеза* (site-directed mutagenesis), когда изучают активность белков с внесенными в аминокислотную последовательность точечными мутациями.

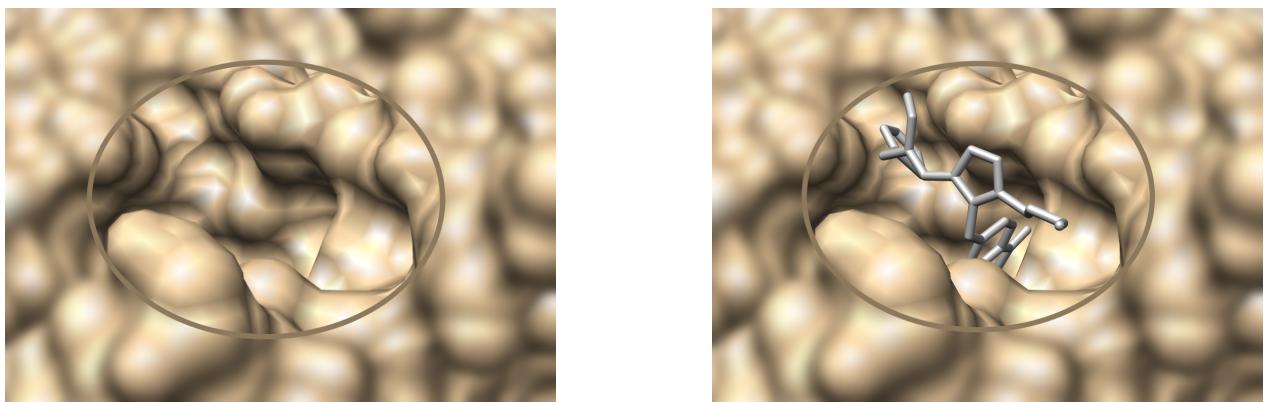


Рис. 2.30: Карман докинга

Слева: Пример полости в структуре белка, которая может являться сайтом связывания лиганда

Справа: Положение лиганда в сайте связывания белка

Поверхность белка, за исключением области полости, на рисунке изображена размытой. Использована структура протеазы вируса клещевого энцефалита (Potapova, S.I Feranchuk и др. 2012).

Процедура докинга состоит в нахождении правильной ориентации лиганда в сайте связывания белка, подобно изображеному на рисунке 2.30. При поиске ориентации лиганда двугранные углы между атомами лиганда обычно считаются подвижными.

Алгоритм нахождения оптимальной ориентации лиганда имеет целью найти и согласовать между собой наиболее энергетически выгодные взаимодействия атомов лиганда и атомов белка, подобно изображенным на рисунке 2.31 (слева). Обычно при проведении докинга белка с лигандом используют *гибкий докинг*, причем в первую очередь учитывают внутренние степени свободы молекулы лиганда. Также возможен учет подвижности некоторых боковых цепей рецептора вблизи активного сайта.

Постановка задачи по докингу белка с лигандом смыкается с постановками задач в фармацевтике, где целью является разработка лекарственных соединений. Обзор методов и подходов, используемых при подборе лекарственных соединений, выходит за рамки настоящего курса; игры, которые принято вести в экономике и по правилам конкурентной борьбы, добавляют специфику к исследованиям в этой области. Как краткое перечисление подходов при разработке лекарств, следует упомянуть технологию скрининга активности химических соединений, методы QSAR (*quantitative structure-activity relationship*), другие методы аналитической химии, и, безусловно, эксперименты по испытанию препаратов, разных уровней и направлений.

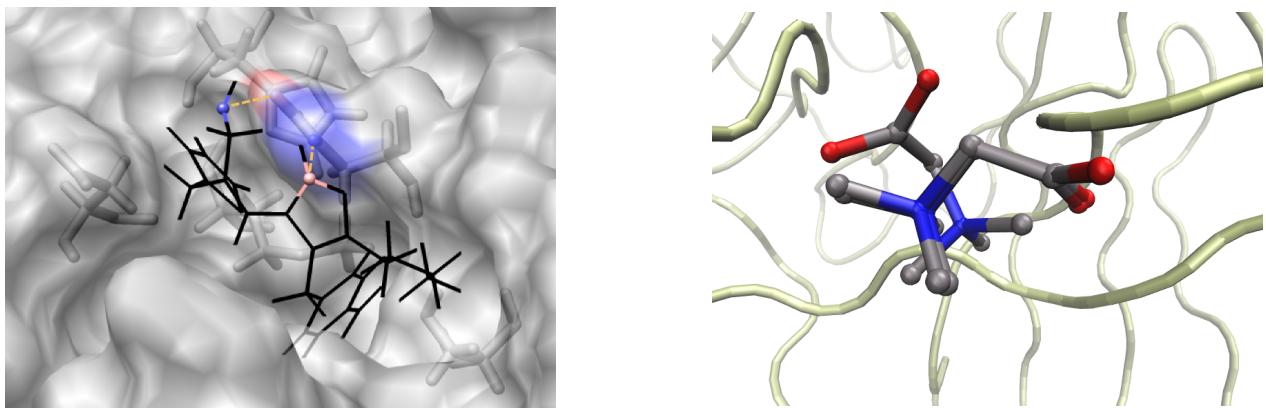


Рис. 2.31: Докинг лигандов

Слева: Пример взаимодействия атомов лиганда с атомами белка

Справа: Выбранные при докинге положения лиганда в кармане связывания

использована структура протеазы вируса клещевого энцефалита (Potapova, S.I Feranchuk и др. 2012).

В контексте задач фармацевтики, технология *виртуального скрининга* состоит в переборе лигандов и проведении молекулярного докинга, с целью выбора соединений, наиболее подходящих для регуляции активности белка-мишени. В алгоритмах докинга не сформировалось "золотого стандарта", как следует выполнять оптимизацию положения лиганда и рассчитывать энергию взаимодействия. Однако при разработке алгоритмов докинга обычно одним из требований является оптимизация времени расчетов, поскольку процедуры докинга необходимы в технологии виртуального скрининга, где количество соединений при переборе может достигать миллиона и более.

При ограничениях на время расчетов, в алгоритмах докинга возможно использовать лишь эмпирические и статистические силовые поля, настроенные и протестированные на ограниченной выборке структур молекулярных комплексов. Но общей проблемой таких силовых полей и таких подходов является недостаточная универсальность, и неадекватные оценки энергии связывания для молекул, не включенных в тестовую выборку. И потому точность таких подходов невысока, и результаты, полученные при виртуальном скрининге, требуют дополнительной верификации.

В рамках расчетов по молекулярной динамике возможно оценить энергию связывания лиганда более точно, основываясь на универсальных силовых полях и оценивая дополнительные

вклады в энергию связывания, такие как влияние растворителя. В частности, в пакете молекулярной динамики Amber разработана технология расчета энергии связывания, называемая *MM-PBSA* (molecular mechanics / Poisson-Boltzmann surface area). Принципы расчета по этой технологии показаны на рис. 2.32.

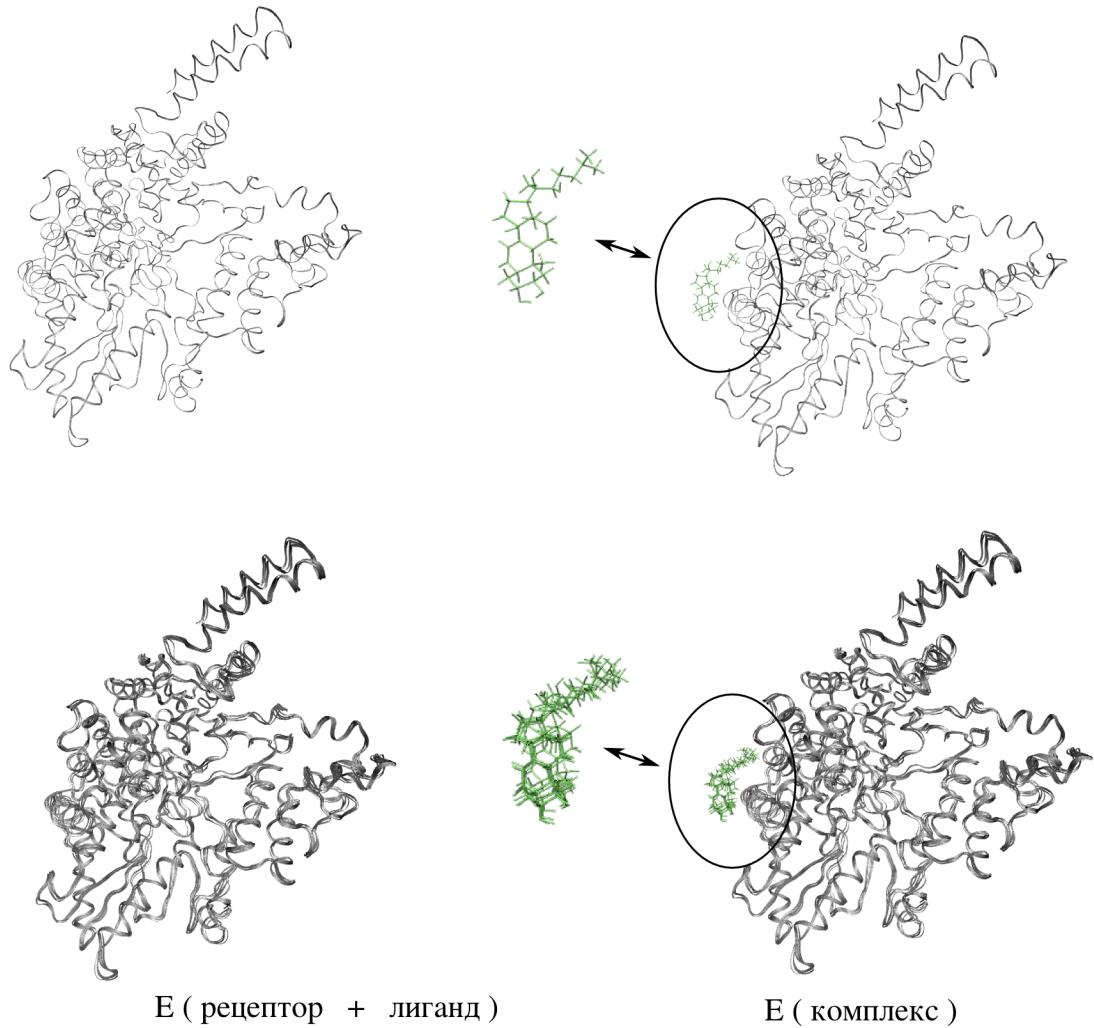


Рис. 2.32: Схема расчетов по методу ММ-ПБСА

Вверху - молекула рецептора и молекула лиганда, составляющие комплекс.

Внизу - комплекс в разных фазах моделирования, разделенный на фазы молекулы рецептора и фазы молекулы лиганда.

Энергия связывания лиганда рассчитывается как разность между энергией комплекса и энергией каждой из двух молекул; статистическое усреднение оценок проводится по кадрам траектории молекулярной динамики.

Образование комплекса можно представить как "потенциальную яму" в "энергетическом ландшафте", и потому энергия связывания в абсолютных единицах является отрицательной величиной.

Показана нейраминидаза вируса гриппа в комплексе с молекулой ponasterone, активного компонента левзеи сафлоровидной

В постановке расчетов, показанной на рис. 2.32, существенный вклад в оценку энергии вносит

учет влияния растворителя. Для учета растворителя возможно использовать решение уравнения Пуассона-Больцмана (аббревиатура PBSA), либо обобщенное приближение Борна (аббревиатура GBSA). Но впрочем, даже при наиболее точных методах оценки энергии, таких как технология MM-PBSA, часто в результатах встречаются неточности и существенные искажения.

Для иллюстрации и сравнения численных значений энергии связывания, при возможности оценить эту величину в эксперименте, обычно энергия около -20 килокалорий/моль достаточна для образования устойчивого комплекса. Энергия, рассчитанная при молекулярном докинге, для молекул не вошедших в тестовый набор, редко оказывается ниже чем -7 килокалорий/моль. В методах MM-PBSA оценки для проверенных комплексов ближе к адекватным, с ошибкой в пределах ± 5 килокалорий / моль, однако при расчетах для произвольно составленных комплексов часты выбросы и ошибки в величине энергии, в том числе существенные расхождения между оценками PBSA и GBSA.

* * *

В белок-белковом докинге (рис. 2.33) в общем случае неизвестными являются шесть координат описывающих взаимную ориентацию белков в пространстве. Кроме того, в процессе взаимодействия белки могут изменять свою внутреннюю конформацию. Возможность рассчитывать комплексы белков важно для молекулярной биологии, но задача белок-белкового докинга в общем случае трудно поддается решению.

Методы белок-белкового докинга также используют расчеты энергии взаимодействия молекул, основанные на физических или статистических силовых полях. Как и в докинге белка с лигандом, в белок-белковом докинге существенно облегчают решение задачи знание сайтов, которыми белки взаимодействуют друг с другом. Многие методы белок-белкового докинга используют методы оптимизации, такие как метод Монте-Карло и генетические алгоритмы, с целью эффективного перебора пространства взаимных ориентаций. Использование информации о сходстве с известными комплексами белков также способствует поиску решения этой задачи.

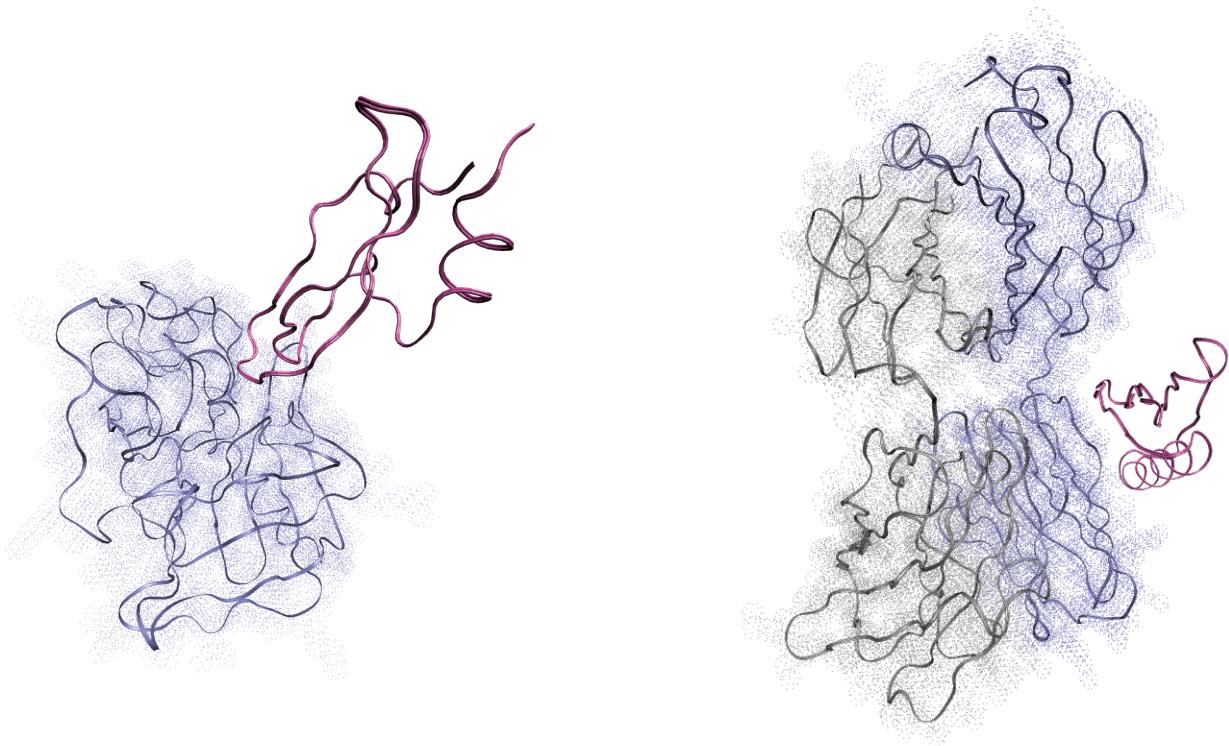


Рис. 2.33: Примеры белковых комплексов

Слева: Трипсин в комплексе с ингибитором трипсина, модельная структура для сравнения методов докинга белков. Ингибитор трипсина показан розовым.

*Справа: Антитело в комплексе с антигеном; ориентацию антигена в таких комплексах мало вероятно определить с помощью методов докинга. Антиген, фрагмент белка L бактерии *Peptostreptococcus Magnus*, показан розовым. Антитело составлено из двух белковых цепей, "тяжелой" и "легкой".*

Показаны структуры комплексов 1bzx, 1hez. Изображение построено с использованием VMD.

Эффективность методов белок-белкового докинга изучается в "эксперименте" CAPRI, когда исследовательским группам предлагается предсказать взаимную ориентацию белков. Это мероприятие, по организации сравнения методов, подобно мероприятию CASP по предсказанию структуры белков. Однако если в CASP точность предсказания структур оказывается относительно удовлетворительной, в "эксперименте" CAPRI, предсказанные ориентации молекул существенно реже оказываются близки к ориентации молекул, обнаруженной в эксперименте. И даже для ориентаций, близких к корректной, модели комплексов остаются, как правило, частично несогласованными, из-за проблем при подборе правильной ориентации боковых цепей обоих белков в участке, где происходит контакт.

3 Системная биоинформатика

3.1. Исторический очерк математических методов в биологии

Коротко описывая историю развития математических методов, используемых в алгоритмах биоинформатики, как это представлено на рисунке 3.1, следует, во-первых, отделить методы линейной алгебры и теории обыкновенных дифференциальных уравнений, которые относятся к традиционному курсу высшей математики, изучаемому в большинстве высших учебных заведений. Истоки используемых в биоинформатике методов комбинаторики, теории графов и статистики, лежат в более узких и частных, но давно разрабатываемых разделах традиционной математики. Впрочем, работы Р.Фишера (1890-1962) по классической статистике были выполнены для решения прикладных задач математической экологии, как они были поставлены в 30-х годах прошлого века, и потому применение методов статистики в вычислительной биологии имеет давнюю историю.

С именами А.А. Маркова (1856-1922), А.П. Колмогорова (1903-1987) связаны некоторые фундаментальные понятия в методах статистики и комбинаторики, используемых в биоинформатике. Но набор инструментов и подходов, обозначенный на схеме как "Функционал правдоподобия / EM (Expectation Maximization) получил свое развитие с выходом в 1977 году работы (Dempster и др. 1977), в которой был описан метод оценки параметров вероятностных моделей, содержащих скрытые переменные. В этой, достаточно универсальной, постановке задачи по подбору параметров математических моделей, алгоритм максимизации функционала правдоподобия оказался применим ко широкому кругу задачам биоинформатики из различных областей. Вероятностные процессы, содержащие скрытые переменные, известны как "цепи Маркова", и потому многие прикладные алгоритмы биоинформатики содержат в своей аббревиатуре или кратком описании упоминание имени А.А. Маркова.

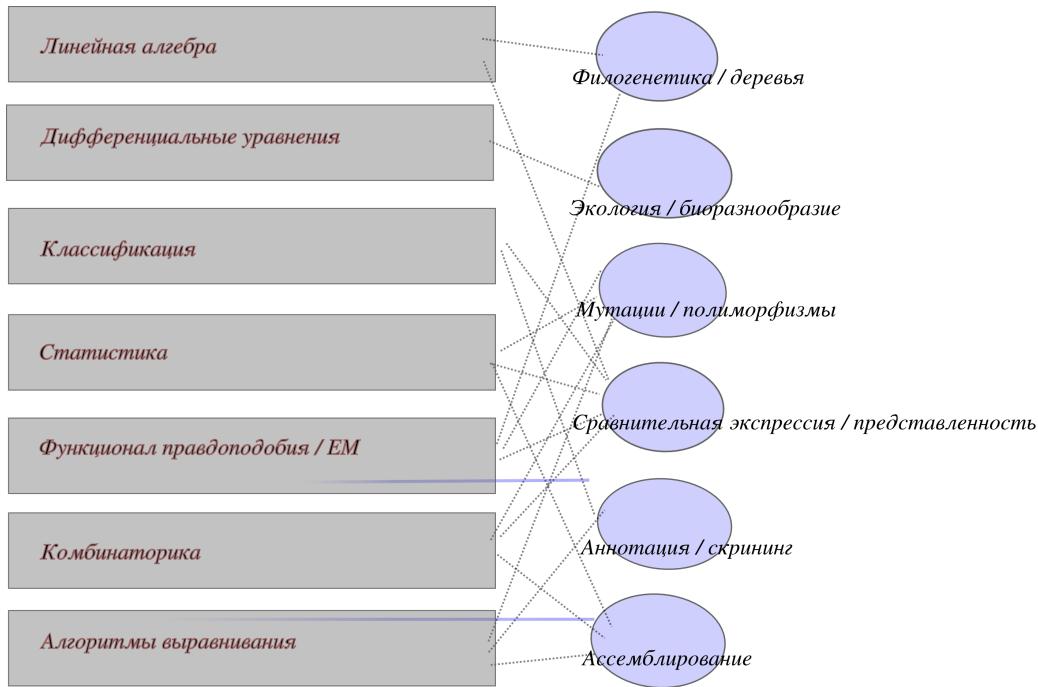


Рис. 3.1: Иллюстрация связи методов математики и разделов применения системной биоинформатике.

В правой части рисунка показаны темы системной биоинформатики, обсуждаемые в курсе, и используемые в них подходы и методы математики и информатики.

Методы решения задач классификации, интенсивно разрабатываемые в современной прикладной информатике и также широко используемые в задачах биоинформатики, также можно сгруппировать на основании подходов, лежащих в их основе. Подходов к решению задач классификации относительно немного; наиболее популярным в последние десятилетия является подход на основе метода опорных векторов, разрабатываемого с 1963 г. В. Вапником с соавторами (см. (Вапник и Червоненкис 1974)).

Алгоритмы выравнивания биологических последовательностей, отдельно помеченные на схеме 3.1, лежат в основе значительной части прикладных методов биоинформатики. В развитии этих алгоритмов следует упомянуть несколько ключевых идей, появление каждой из которых способствовало значительному прогрессу при решении прикладных задач. Это, во-первых, алгоритм поиска оптимального выравнивания двух последовательностей аминокислот (Needleman и Wunsch 1970), реализуемый в рамках алгоритмической теории называемой *динамическим программированием*. Также, использование преобразования Барроуза — Уилера (Burrows и Wheeler 1994) сделало возможным быстрый и эффективный поиск гомологичных фрагментов в базах последовательностей ДНК, на основании попарного выравнивания с последовательностью-шаблоном.

Среди алгоритмов и идей, оказавшихся важными для прогресса в других разделах биоинформатики, следует также упомянуть алгоритм ассемблирования фрагментов ДНК (Idury и Waterman

1995) с использованием понятия *графа де Брайна*, а также алгоритмы для построения филогенетических деревьев, развитие которых началось в конце 1960х (Fitch и Margoliash 1967; Felsenstein 1981). Но впрочем, задачи построения филогенетических деревьев, множественного выравнивания последовательностей, быстрого поиска гомологов в базах данных неизбежно требуют введения эмпирических приближений для ускорения времени расчетов, и потому результаты расчетов в таких задачах не всегда являются оптимальными по формальным критериям.

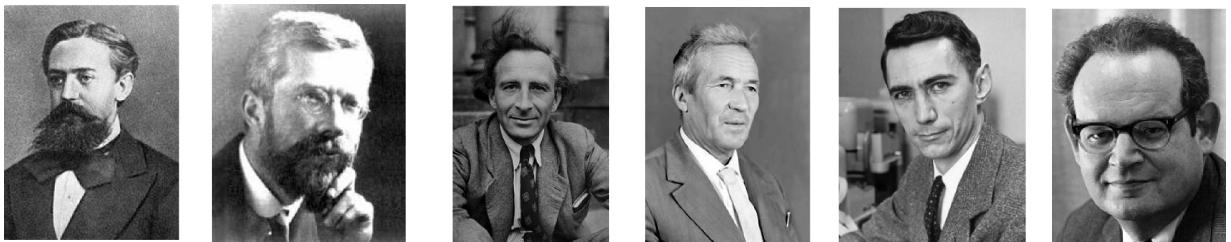


Рис. 3.2: Некоторые из ученых, работы которых лежат в основании подходов в системной биоинформатике

А.А. Марков (1856-1922), Р. Фишер (1890-1962), Н.В. Тимофеев-Ресовский (1900-1981), А.П. Колмогоров (1903-1987), К. Шеннон (1916-2001), Б. Мандельброт (1924-2010).

Использование дифференциальных уравнений для описания характерных свойств в поведении экосистем, начатое в 1920х годах (модель "хищник-жертва") и развитое в России в научной школе, основанной Н.В. Тимофеевым-Ресовским, обычно не подразумевает возможности детализации моделей экосистем до молекулярного уровня. Но при необходимости введения эмпирических упрощений в расчеты основанные на молекулярном представлении сложных биологических объектов, следует подчеркнуть значимость результатов, полученных при таком подходе. И, наконец, в обзоре подходов для анализа биологических систем следует упомянуть теорию фракталов; понятия фрактала и фрактальной размерности, введенные в работах Б.Мандельброта 1960х-1970х годов, являются еще в большей степени универсальным методом описания сложных систем, без возможности проследить связь наблюдаемых эффектов с молекулярным представлением объектов.

3.2. Иерархия объектов в системной биоинформатике

При перечислении дисциплин и предметных областей биологии, где используются подходы, которые можно объединить словом "биоинформатика", следует использовать, во-первых, масштабы и уровень детализации изучаемых объектов. Но также, развитие многих прикладных методов

и пакетов программ в биоинформатике связано с развитием экспериментальных методов и методик постановки экспериментов. В частности, бурное развитие вычислительных методов, используемых для обработки больших объемов нуклеотидных последовательностей, можно связать с появлением в 2000-2010 гг. так называемых технологий *секвенирования нового поколения (высокопроизводительного секвенирования)*, позволяющих эффективно "считывать" информацию, содержащуюся в ДНК.

При появлении этих технологий, развитие получили, в первую очередь, методы изучения живой клетки и систем регуляции в клетке на молекулярном уровне (таблица 3.1). Дополнительными экспериментальными методами при изучении живой клетки стали масс-спектрометрия и технология, использующая гибридизацию фрагментов ДНК в так называемых *микрочипах*. Методы обработки экспериментов по масс-спектрометрии развились в дисциплины, называемые *метаболомикой* и *протеомикой*.

Таблица 3.1: Уровни детализации в задачах биоинформатики

Уровень детализации	Предметная область	Дисциплина
Молекулы	Пути метаболизма	Метаболомика, Протеомика
Гены	Системы регуляции в клетке	Транскриптомика, Геномика
Клетки в организме	Иммунная система, нервная система	Математическая иммунология, Нейробиология
Организм с точки зрения медицины	Обмен веществ, циркуляция крови, и.т.п.	Статистическая поддержка методов лечения
Взаимодействие между организмами	Экологическая система	Метагеномика, Математическая экология
Эволюция организмов	Сети переноса генов	Молекулярная филогения

Но использование экспериментов основанных на методах секвенирования оказалось возможно также на более высоких уровнях детализации (таблица 3.1), в *метагеномике* и *математической филогении*. И, напротив, методы протеомики продолжают развиваться параллельно с появлением всех новых приборов и методик постановки экспериментов по масс-спектрометрии, и область применения результатов, полученных в протеомике, остается ограниченной, из-за принципиальных сложностей, возникающих при попытке получить большие объемы достоверной информации о составе молекул в биоматериале на основании измерений массы молекул с помощью масс-спектрометрии.

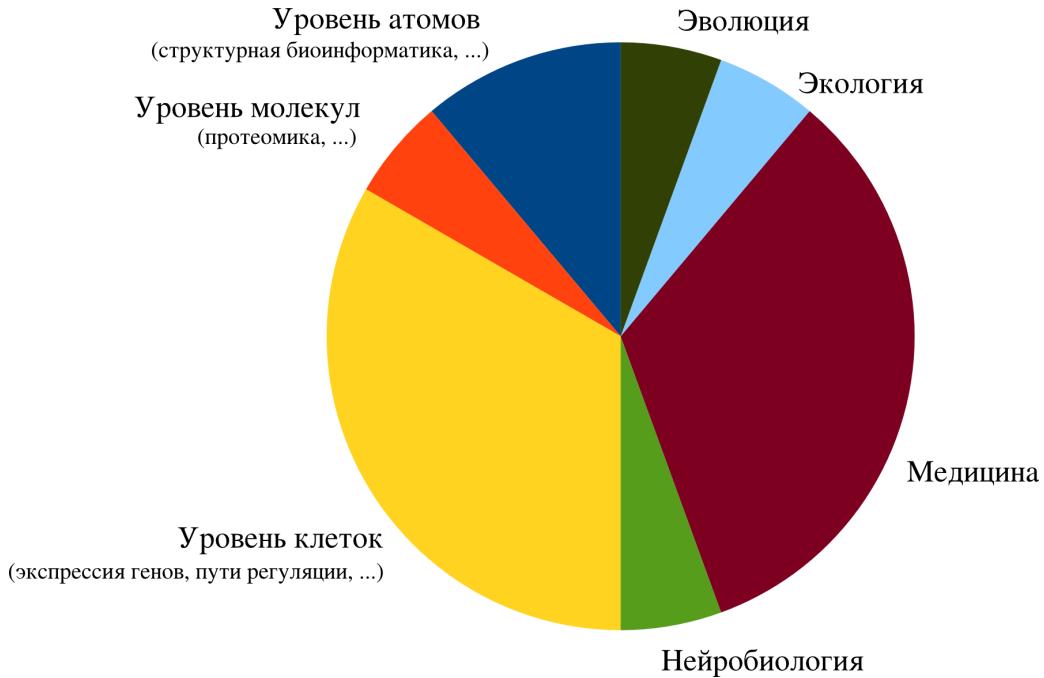


Рис. 3.3: Интенсивность исследований в биомедицине

Примерное разделение ресурсов между направлениями исследований, оцененное по количеству публикаций в базе биомедицинской литературы Pubmed за последнее десятилетие.

Изучение систем регуляции клетки, по количеству исследований, развивается наиболее интенсивно, среди перечисленных направлений. Видимо, от этих исследований ожидается наибольшая практическая польза в приложениях к медицине. Потому молекулярные методы исследования клетки представлены в пособии лишь как краткий обзор, по причине быстрого развития методов и экспериментальных наблюдений в этой области. Статистические методы в медицинских приложениях также не раскрыты подробно в пособии, по причине быстрого развития этого направления и разнородности используемых методов и решаемых задач.

3.3. Основные понятия молекулярной биологии клетки

ДНК - "Дезоксирибонуклеиновая кислота" - молекула, по структурной химической формуле составленная, как цепь, из однотипных нуклеотидных остатков. Органические молекулы из класса дезоксирибонуклеотидов содержат химическую группу (дезоксирибозу, из класса сахаров, и остаток фосфорной кислоты) и могут соединяться в цепь (полинуклеотид) за счет образования ковалентной связи между фосфорной кислотой и дезоксирибозой смежного остатка. При этом, дезоксирибонуклеотиды различаются химической группой, ковалентно связанной с дезоксирибозой, и потому полинуклеотидная цепь ДНК может быть составлена из нескольких типов дезоксирибонуклеотидов, следующих в произвольном порядке.

В живой природе чаще всего встречаются цепи ДНК, составленные из четырех типов нуклеотидных остатков, обозначаемых обычно как А (Аденин), С (Цитозин), Г (Гуанин), Т (Тимин).

Длина цепей ДНК может быть очень большой; в силу особенностей структуры дезоксирибонуклеотидов две цепи ДНК могут образовывать *двойную спираль*. Структура двойной спирали является устойчивой, когда между смежными нуклеотидами в остатках двух цепей образуются водородные связи. Условием образования водородных связей является *комплементарность* химических групп в смежных нуклеотидах: остаток аденина комплементарен остатку тимина, остаток цитозина комплементарен остатку гуанина.

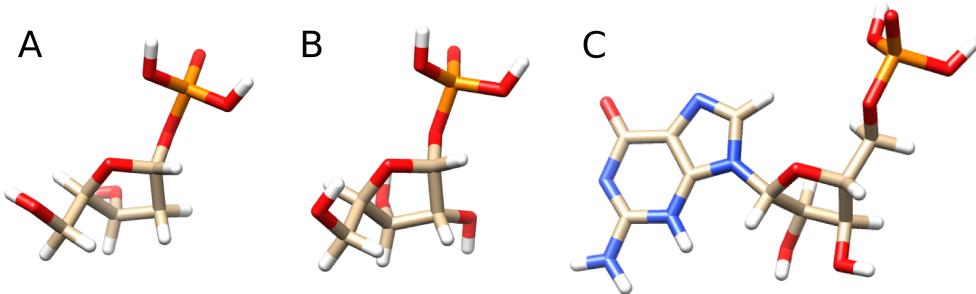


Рис. 3.4: Некоторые соединения - компоненты нукleinовых кислот

A - Рибоза, связанная с остатком фосфорной кислоты

B - Деоксирибоза, связанная с остатком фосфорной кислоты (Deoxyribose 1-phosphate)

C - Нуклеотид гуанозин (5'-Guanyllic acid; Guanosine monophosphate). Гуанин - азотистое основание класса пуринов - соответствует химической группе в левой части молекулы.

К молекулам нуклеотидов, в первую очередь к аденоzinу и гуанозину, может быть добавлен еще один или два дополнительных остатка фосфорной кислоты. Реакция присоединения остатка фосфорной кислоты (*фосфорилирование*) к нуклеотиду аденоzinу, и реакция его отщепления, происходящая с высвобождением энергии, часто используются при энергетическом обмене в клетке. Для обозначения нуклеотидов с дополнительными фосфатными группами, используются термины *аденоzinидифосфат* (АДФ, ADP) и *аденозинтрифосфат* (АТФ, ATP).

РНК - "Рибонуклеиновая кислота" - молекула, по структурной химической формуле являющаяся полинуклеотидом подобным ДНК. *Рибонуклеотиды*, из которых составлены молекулы РНК, отличаются от дезоксирибонуклеотидов дополнительным атомом кислорода в химической группе из класса сахаров. Поэтому по физико-химическим свойствам цепи РНК отличаются от цепей ДНК. Подобно цепям ДНК в живой природе, цепи РНК обычно составлены из четырех типов нуклеотидных остатков, и для цепей РНК также свойственно образование водородных связей между смежными комплементарными остатками. Но для молекулы РНК, комплементарные связи между цепями в двойной спирали менее устойчивы, чем комплементарные связи между фрагментами одной молекулы.

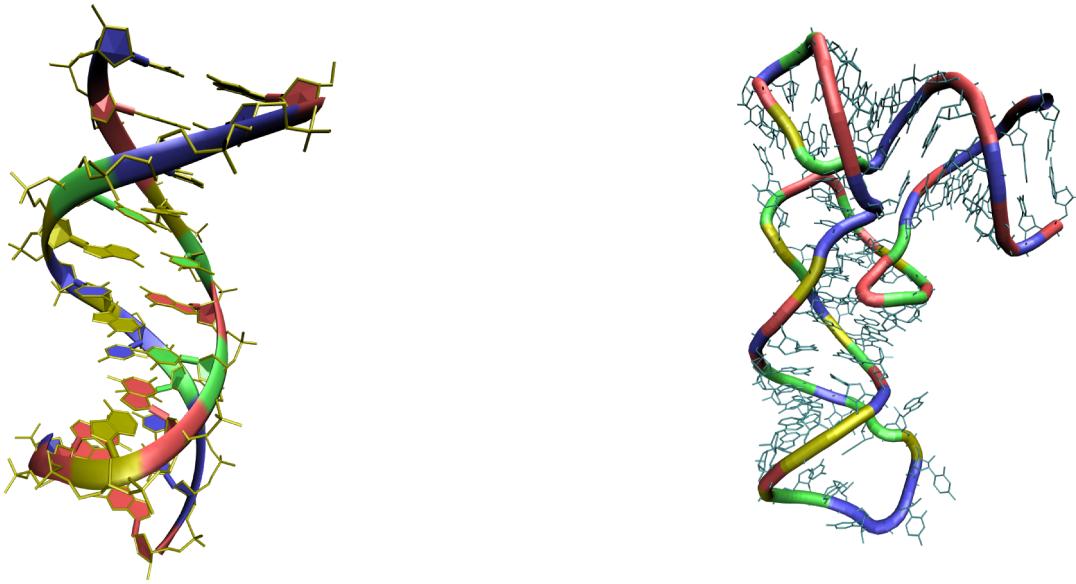


Рис. 3.5: Молекулы ДНК и РНК

Слева: две цепи молекул ДНК, образующие двойную спираль.

Справа: молекула РНК, в которой структура стабилизирована водородными связями между комплементарными нуклеотидами.

Изображены структуры ДНК и РНК (транспортная РНК tRNA-Gly), входящие в состав комплексов bclv и 4kr2.

Направление цепей в полинуклеотидах показано с помощью ленты / трубки. Ароматические кольца в боковых цепях нуклеотидов в молекуле ДНК закрашены для большей наглядности. Четырьмя цветами обозначены четыре типа нуклеотидов. Для визуализации был использован пакет VMD.

Для обозначения некоторых типов молекул РНК, используются следующие термины:

- *матричная РНК (mRNK, mRNA)* - молекула РНК, используемая как матрица, по которой при трансляции в рибосомах синтезируется белок.
- *некодирующая РНК (ncRNA)* - молекулы РНК, синтезируемые в клетке, но не используемые для трансляции
- *транспортная РНК (tRNK, tRNA)* - молекула РНК (некодирующая), которая используется при синтезе белка. В молекулах тРНК, для которых характерна типичная трехмерная структура (рис. 3.5), три нуклеотида используются для комплементарной связи с матричной РНК, а специфическая аминокислота, предварительно присоединенная к транспортной РНК, используется для продолжения цепи белка при синтезе.
- *микро РНК (miRNA)* - молекулы некодирующей РНК, которые участвуют в процессах регуляции процессов транскрипции и трансляции

Белок - молекула, составленная, как цепь, из остатков молекул, относящихся к классу аминокислот. В клетках используется 20 типов аминокислот (рис. 3.5). Для белков в клетке характерно наличие устойчивой структуры, в которую уложена цепь аминокислотных остатков. В главе

"Структурная биоинформатика" более подробно обсуждаются свойства белковых молекул и методы их анализа.

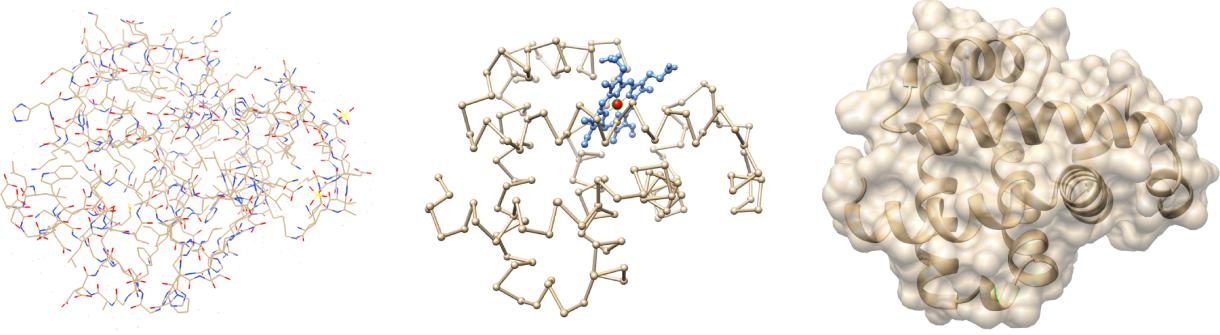


Рис. 3.6: Молекула миоглобина

Белок миоглобин, в трех представлениях:

Слева: полноатомная модель

В центре: главная цепь белка и молекула гема

Справа: представление в виде ленточной диаграммы. Полупрозрачным показана поверхность, доступная растворителю

Изображены структура окси-миоглобина 1abt. Для визуализации был использован пакет UCSF Chimera.

Некоторые из белков участвуют в катализе химических реакций в клетке; такие белки иногда называют *ферментами* или *энзимами* (*enzymes*). Для коротких цепей аминокислот, не имеющих устойчивой укладки, используют термин *пептид*.

Метаболиты - низкомолекулярные химические соединения, с массой обычно менее 1 кило-дальтон, участвующие в биохимических процессах, происходящих в клетке (рис. 3.7). В отличии от полинуклеотидов и белков, синтез метаболитов в клетке не задается непосредственно генетическим кодом. К метаболитам относятся продукты обмена веществ, гормоны и другие сигнальные молекулы, а также прочие низкомолекулярные соединения, как, например, лекарственные препараты. Иногда для обозначения низкомолекулярного соединения используют термин *лиганд*, как, например, в контексте изучения катализа химической реакции белком-ферментом. В частности, молекула нуклеотида аденоцинтрифосфата (АТФ) является лигандом многих ферментов, поскольку часто используется в энергетическом обмене клетки.

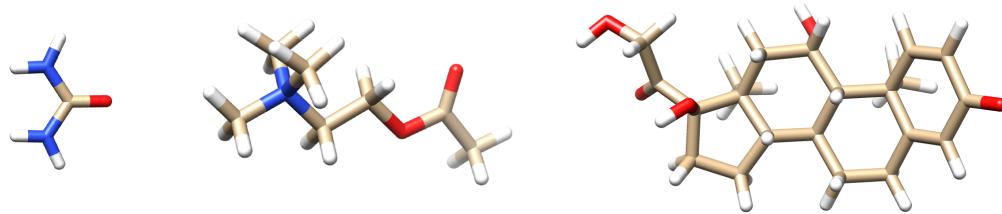


Рис. 3.7: Некоторые из химических соединений, участвующих в обмене веществ в организме

Слева: молекула мочевины

В центре: молекула ацетилхолина

Справа: молекула преднизолона

Мочевина (*urea*) - конечный продукт метаболизма белка в клетке; ацетилхолин - один из нейромедиаторов; преднизолон - синтетический лекарственный препарат, по структуре подобный гормонам, вырабатываемым корой надпочечников (кортикоидам).

Для визуализации молекул был использован пакет UCSF Chimera.

Транскрипция - полимеризация молекулы РНК, происходящая посредством катализа в комплексе белков, с использованием молекулы ДНК в качестве шаблона.

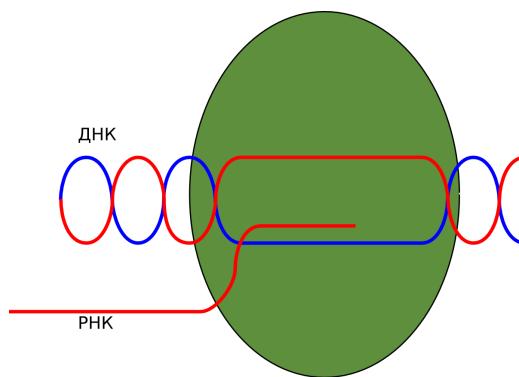


Рис. 3.8: Схема процесса транскрипции

Синтез РНК при транскрипции происходит на основании комплементарности очередного из остатков цепи шаблона и остатка, который добавляется в цепь РНК. Определение участка ДНК, на котором происходит транскрипция, происходит с участием белков, называемых *транскрипционными факторами*, через связывание этих белков со специфическими участками молекулы ДНК.

Трансляция - полимеризация молекулы белка, происходящая посредством катализа в молекулярном комплексе, называемом *рибосомой*, с использованием молекулы матричной РНК в качестве шаблона. В состав рибосомы, кроме молекул белка, входит несколько молекул некодирующей РНК (*рибосомная РНК*), с характерной устойчивой пространственной укладкой.

Метилирование ДНК - химическая модификация некоторых остатков цитозина в молекуле ДНК, состоящая замещения водорода на метильную группы у атома С5 в ароматическом кольце цитозина. В ДНК клетки метилирование, как правило, происходит у остатков цитозина за которыми в цепи следуют остатки гуанина (С-G) и регулируется специфичными ферментами.

Сплайсинг - модификация молекулы матричной РНК, происходящая посредством катализа в молекулярном комплексе, называемом *сплайсосомой*. Модификация при сплайсинге состоит в вырезании определенных фрагментов матричной РНК и сращивании смежных участков РНК. Как результат, некоторые фрагменты ДНК, скопированные в процессе транскрипции в мРНК, выпадают на стадии трансляции мРНК в белок. Фрагменты ДНК, которые остаются в мРНК после сплайсинга и преобразуются во фрагменты белковой последовательности при трансляции, называются *экзонами*, а участки ДНК, удаляемые из мРНК при сплайсинге, называются *инtronами*.

Пост-трансляционные модификации - общее название для процессов химической модификации некоторых аминокислотных остатков в белках, происходящих после синтеза белка в рибосоме. Некоторые пост-трансляционные модификации специфичны для отдельного белка или класса белков, некоторые - как, например, *убиквитинирование*, регулируются универсальными ферментами.

Ген - термин, обозначающий "элементарную" единицу информации, закодированной в ДНК и проявляющуюся в наследственных свойствах организма. Своим возникновением связан с развитием *генетики* в конце XIX - начале XX века. В молекулярной биологии термин используется, с долей условности, для обозначения фрагмента ДНК, с которого происходит транскрипция матричной РНК или некодирующую РНК, а также саму матричную / некодирующую РНК или белок, синтезируемый при трансляции матричной РНК.

При учете *экзон-инtronной* структуры последовательностей ДНК, когда в результате *альтернативного сплайсинга* одному фрагменту ДНК может соответствовать несколько мРНК, следует отличать термин *ген* от термина *изоформа*. В этом случае, для различия матричных РНК, а также белков синтезируемых при их трансляции, используют выражение *изоформы гена*.

Геном - термин, обозначающий совокупность наследственной информации, содержащейся в клетке. В большей части известных случаев, наследственная информация в клетке, к которой применимо понятие "геном", закодирована в молекулах ДНК, образующих молекулярные комплексы, которые можно заметить при достаточном увеличении как *хромосомы* в ядре клетки. В многоклеточных организмах, как правило, один и тот же геном содержится во всех клетках организма.

Среди исключений из описания приведенного выше, следует упомянуть бактерии, в которых ДНК не образует хромосом, а также некоторые вирусы, в которых информация кодируется в молекуле РНК. Кроме того, митохондрии и хлоропласты, органеллы развивающиеся внутри клеток, содержат собственный генетический материал, дополняющий ДНК в хромосомах.

3.4. Термины, используемые при постановке экспериментов и обработке данных

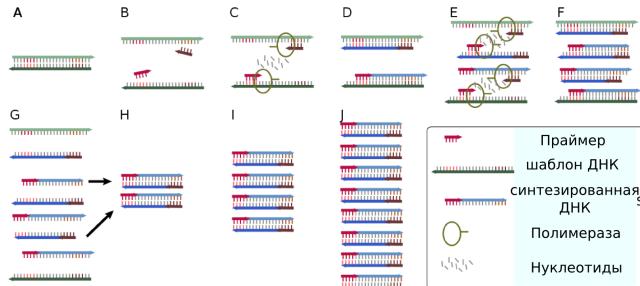
Полимеразная цепная реакция (ПЦР, PCR) - экспериментальная методика, позволяющая многократно увеличить количество определенного фрагмента ДНК, при условии наличия этого фрагмента в биоматериале. Копирование фрагмента происходит при участии фермента ДНК-полимеразы, а специфичность при выборе фрагмента в эксперименте определяется парой так называемых *праймеров* - полинуклеотидов с длиной около 10 - 20 оснований.

ДНК-микрочип (microarray) (micr) - технология, позволяющая детектировать присутствие и/или количество определенных фрагментов ДНК в биоматериале, независимо в каждой из ячеек микрочипа. Использование технологии позволяет быстро и эффективно характеризовать состав ДНК, содержащейся в биоматериале.

Секвенирование (sequencing) - экспериментальное определение последовательностей ДНК, содержащихся в биоматериале.

Постановка экспериментов по определению последовательностей ДНК предложена в 1977 г. (Sanger и др. 1977). Идея, заложенная в этой постановке, продолжает использоваться в других экспериментах, но в истории развития секвенирования разделяют методы первого, второго и третьего поколения. Ко второму поколению методов относится использование приборов, называемых *секвенаторами*, в которых автоматизированы адаптированные лабораторные методики, необходимые для определения ДНК. Первые из секвенаторов, секвенаторы второго поколения, получили название *капиллярных секвенаторов*. В третьем поколении методов, эффективность автоматизированного секвенирования оказалось возможным существенно увеличить, при условии разделения ДНК в биоматериале на относительно короткие фрагменты. Термин *олигонуклеотид* используется как синоним для понятия "фрагмент ДНК", но для олигонуклеотидов, определенных при секвенировании, используется более узкий термин *рид* ("прочтение", *read*).

Полимеразная цепная реакция

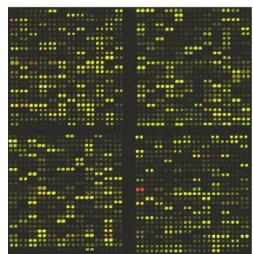


Этапы полимеразной цепной реакции



Амплификатор Eppendorf
для проведения ПЦР

ДНК-микрочипы



Матрица микрочипа
разные цвета представляют уровень
экспрессии разных генов

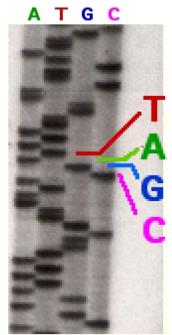


микрочип для анализа
экспрессии генов

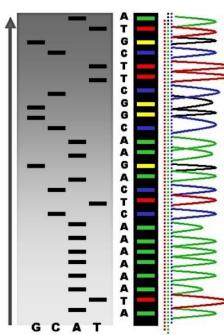


оборудование для
сканирования микрочипов
Affymetrix GeneChip

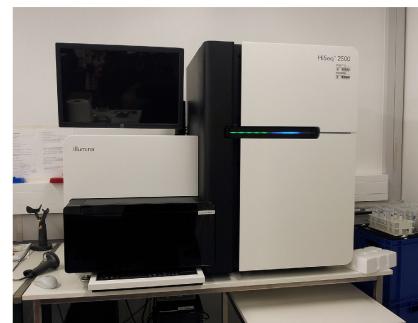
Секвенирование



гель с радиоактивными метками



радиоактивные метки =
= пики флуоресценции



Секвенатор Illumina HiSeq 2500

Рис. 3.9: Некоторые экспериментальные методы и инструменты, используемые в молекулярной биологии

Вверху: полимеразная цепная реакция осуществляется за счет циклических изменений температуры в реакционном сосуде, на каждом цикле происходит удвоение количества специфического участка ДНК между двумя праймерами.

В центре: фрагменты ДНК в биоматериале могут специфически связываться с ДНК-зондами в лунках микрочипа за счет эффекта гибридизации. Количество связанного ДНК в ячейке микрочипа определяется по интенсивности флуоресценции ячейки.

Внизу: в методе секвенирования "по Сэнгеру", обрыв цепи при синтезе ДНК на каждой из четырех дорожек может происходить в положении одного из четырех возможных нуклеотидов. В первых экспериментах, для определения последовательности нуклеотидов использовался гель-электрофорез фрагментов ДНК разной длины, помеченных радиоактивной меткой.

Авторство изображений: John Schmidt, Abizar, Schutz, WiWiki, kOchstudiO

Выравнивание (alignment) - комбинаторная задача и алгоритм расстановки вставок (*gaps, гэпов*) в две или более последовательности символов, для достижения максимальной степени сходства между символами в одном столбце, при условии что с учетом вставок последовательности имеют одинаковую длину. В описанной постановке задачи, выравниванием называется также таблица символов, где последовательности с добавленными гэпами записаны как строки.

В приложении к биологическим последовательностям, белковым или нуклеотидным, алгоритмы выравнивания позволяют восстановить мутации, состоявшие в выпадении или вставке нуклеотидных остатков в ДНК. При этом обычно подразумевается, что сравниваемые последовательности генов имеют общий ген-предок, и разделение последовательностей за счет внесения мутаций произошло в течении эволюции или развития организма.

Построение оптимального *парного выравнивания* последователей возможно, при условии задания степени сходства между символами в последовательностях. Время расчетов при построении такого выравнивания (*алгоритмическая сложность*) пропорциональна произведению длин двух последовательностей. Но, из-за степени алгоритмической сложности в задаче построения множественного выравнивания, оптимальное выравнивание трех и более последовательностей в большинстве случаев найти невозможно, и в алгоритмах решения этой задачи используются эмпирические приближения.

Также, термином "выравнивание" обозначают сам набор последовательностей с расставленными гэпами. Для наиболее точного сравнения последовательностей определенного класса часто важен опыт экспертов, или даже просто непредвзятый взгляд биоинформатика. Применение универсальных эмпирических алгоритмов множественного выравнивания может привести к неоптимальным результатам во многих частных случаях, и полученное выравнивание требуется корректировать вручную.

Ассемблирование (assembly, сборка) - класс алгоритмов и программ для обработки ридов полученных при высокопроизводительном секвенировании, с целью восстановления последовательностей ДНК, содержащихся в исходном биоматериале.

Алгоритмы ассемблирования основаны на совмещении идентичных частей в ридах, представляющих смежные участки ДНК. Задача составления длинных цепочек из ридов сводится к задаче нахождения оптимального пути в графе. При нахождении пути в графе разделяют задачу Гамильтона (поиск пути без пересечения по вершинам) и задачу Эйлера (поиск пути без пересечения по ребрам). В отличии от задачи Гамильтона, по оценкам вычислительной сложности задачи Эйлера возможен поиск оптимальных путей за приемлемое время, при объемах данных используемых для ассемблирования (100 гигабайт и более). При постановке задачи Эйлера в приложении к ассемблированию вводится понятие так называемого *k-мера (k-mer)* - фрагмента нуклеотида длиной от 21 до примерно 120, заведомо меньшей чем длина рида (от 50 до 250 и более); полученный граф, в котором k-меры соответствуют ребрам, называется *графом де Брайна*. Однако при ассемблировании с использованием графа де Брайна, для используемой аппаратной платформы предъявляются высокие требования к объему оперативной памяти, поскольку при нахождении пути Эйлера в непредсказуемом порядке используются все k-меры в исходных

данных.

Существуют некоторые методики постановки экспериментов для улучшения достоверности и длины полученных фрагментов ДНК, однако ошибки в данных, допускаемые при секвенировании, так же как и наличие мало различимых фрагментов ДНК в исходном биоматериале, ограничивают качество полученных при асSEMBЛИРОВАНИИ результатов.

Выравнивание фрагментов белка миоглобина

Homo sapiens	CIIQVLQSKH	PGDFGADA--	QG
Bos taurus	AIIHVVLHAKH	PSDFGADA--	QA
Drosophila melanogaster	VLVKVMAEKA	-----GLDAAG	QG

Принципы асSEMBЛИРОВАНИЯ / k-меры

```

accccaagctgttggggcag
aaccccaagctgttggggca
aaaccccaagctgttggggcc
aaaccccaagctgttggggccaggacacc

```

Аннотация хромосомы 1 генома человека

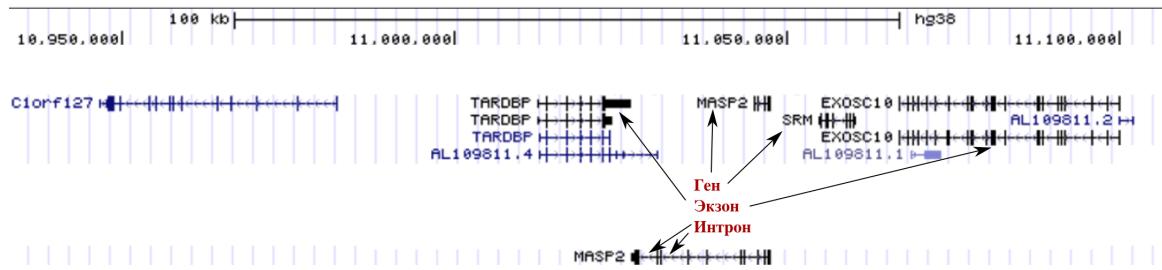


Рис. 3.10: Некоторые алгоритмы и методы биоинформатики

Вверху слева: пример множественного выравнивания. Выровненные последовательности для наглядности разбиты на блоки длиной в 10 символов, и столбцы, в которых все остатки идентичны или сходны по молекулярной массе, выделены цветом. Вставки в выравнивании принято обозначать символом "-".

Вверху справа: принцип асSEMBЛИРОВАНИЯ с использованием графа де Брайна.

Внизу: структурная аннотация фрагмента генома человека. Пометки иллюстрируют экзон-инtronную структуру генома.

при подготовке иллюстрации использован онлайн сервис "UCSC genome browser"

Аннотирование (annotation, аннотация) - в узком смысле, задача, состоящая в идентификации генов в последовательности генома. В этой задаче выделяют *структурную аннотацию* генов, где результатом является определение положения участков генома, для которых возможна транскрипция в РНК, а также сигнальных участков генома, регулирующих процессы транскрипции. Следующим шагом при анализе генома может являться *функциональная аннотация*, когда для найденных генов составляются характеристики, описывающие их функцию и роль в клеточных процессах.

В широком смысле, аннотацией биологических последовательностей или других объектов, изучаемых в молекулярной биологии, называют составление краткого описания характеризующего заданные объекты. Так, например, выделяют задачу по установлению ассоциаций между

ключевыми словами из заданного набора и некоторыми фрагментами последовательностей в базе данных белков. Другим примером можно упомянуть проект Gene Ontology (онтология генов), где в базах данных записаны связи наименования генов с так называемыми *онтологическими терминами*. Онтологические термины в рамках этого проекта группируются на три класса: *биологический процесс, молекулярная функция, компонент клетки*.

3.5. Молекулярные методы исследования клетки

Цели и направления при исследовании клетки

Геном клетки или многоклеточного организма, как последовательность нуклеотидов, возможно восстановить, следуя подробно задокументированным методикам, включающим секвенирование, ассемблирование и аннотацию генов. Но задача "прочесть" текст генома, подобно тому как читают текст книги, является намного более сложной, и эта задача далека от решения, несмотря на объем информации, накопленный в молекулярной биологии.

Интересы научных групп, занятых изучением клетки, определяются не только указанным общим направлением, но и разного рода прикладными задачами, которые оговорены при финансировании проводимых исследований. Но, несмотря на различия в направлениях и интересах ученых, для достижения взаимопонимания между ними при описании работы клетки, выработаны термины и понятия, которые позволяют согласованно описывать результаты их работы. Наиболее общие и широкие из этих понятий приведены ниже:

Путь метаболизма (metabolic pathway) - система последовательных биохимических трансформаций молекул-метаболитов, происходящая с участием белков-ферментов. Некоторые специфические процессы преобразования метаболитов лежат в основе питания клетки клетки, а также используются при ответе клетки на воздействия внешней среды, как, например, в клетках бактерий происходит компенсация к действию антибиотиков.

Путь регуляции (signaling pathway) - система последовательно взаимодействующих белков, с участием метаболитов, РНК и геномной ДНК. Обнаружено в некоторых случаях, что запуск взаимодействий в цепях такого рода происходит под действием определенных внешних факторов, приводя, как результат, к изменению состояния клетки, и составу ее РНК и белков.

Регуляция транскрипции - система взаимодействия транскрипционных факторов со специфическими участками ДНК, в ходе которой происходит синтез дополнительных транскрипционных факторов, а также других белков и РНК. Цепи взаимодействий, запущенные в этой системе, приводят к адаптации клетки к специфическим условиям среды, а также к коррекции состояния клетки, существующей в составе многоклеточного организма.

Сеть взаимодействия генов - представление системы взаимодействующих генов в форме графа, где узлы соответствуют генам, а ребра - взаимодействию между генами. Такое представление является интуитивно понятным, однако его возможно применить к описанию биологических систем лишь с большой долей условности.

Также, для согласования результатов разнородных экспериментов, в большей части научных

публикаций используют наименования конкретных путей регуляции, общих для многих организмов. Некоторые из этих терминов и наименований, а так же степень их отношения к реальности, упоминаются в материалах всех разделов этой главы.

Анализ протеома клетки

Некоторые из многих функций белков в клетке - катализ химических реакций и преобразование метаболитов, поддержание целостности формы клетки и ее коррекция, регуляция перемещения молекул через клеточную мембрану, передача информации в форме сигналов вовне и внутрь клетки. И потому информация о составе белков в клетке (*протеоме*) часто может существенно прояснить понимание происходящих там процессов. Методы современной органической химии позволяют анализировать материал клетки и состав метаболитов. Но для анализа белков клетки необходимо использовать подходы к проведению экспериментов и расчетов, дополняющие методы органической химии.

Среди существующих подходов к экспериментам по протеомике, позволяющих с достаточной специфичностью определить состав белков, следует упомянуть иммунологические методы, такие как иммуноферментный анализ и вестерн-блоттинг, и методы масс-спектрометрии. Известна также экспериментальная методика определения последовательности белка (так называемое "секвенирование по Эдману"), однако в современной протеомике этот метод используют нечасто, поскольку его оказывается непросто увязать с так называемыми "высокопроизводительными" (*high throughput*) подходами к экспериментам по изучению клетки.

Специфичность при иммунологических методах достигается за счет механизмов специфичности иммунной системы. Так называемые *антитела*, белки, вырабатываемые клетками иммунной системы, могут специфично связываться с белками чужеродного возбудителя инфекции (*антигенами*). Антитело является комплексом из нескольких белковых цепей, и специфичность при взаимодействии антитела с антигеном достигается за счет наличия вариабельных частей в последовательностях белков, составляющих антитело.

В основе методов масс-спектрометрии заложена возможность экспериментально определить отношение массы частицы к ее заряду, называемое иногда m/Z . Из уравнений электродинамики легко вывести, что заряженная частица в однородном магнитном поле будет совершать движение по круговой траектории, или спиральной траектории с постоянным шагом и радиусом спирали. В этом выводе, радиус спирали определяется отношением массы частицы к ее заряду, который и оказывается возможно измерить с большой точностью. В масс-спектрометрах, величина заряда частицы ограничена значениями в несколько единиц элементарного заряда, и, после разрешения неоднозначности в выборе величины заряда, в результате измерений на масс-спектрометрах возможно рассчитать массы исследуемых молекул.

Методы масс-спектрометрии допускают их использование для "высокопроизводительного" определения состава белков в биологическом материале, как это схематично показано на рис. 3.11. Фракции в биологическом материале возможно разделить с использованием *хроматографии*, по степени подвижности отдельных молекул. На следующем шаге, для разделения белков

на пептиды обычно используют ферменты, такие как *трипсин*, который разрыв любой белковой цепи в местах, где встроены аминокислоты R (аргинин) и L (лизин). Разделение пептида на заряженные ионы происходит в масс-спектрометре, после чего возможно определить отношение M/Z для каждого иона.



Рис. 3.11: Идеализированное представление экспериментов по протеомике

По постановке эксперимента, пики от каждого из ионов, полученных при расщеплении пептида, позволяют однозначно идентифицировать этот пептид.

Также, несколько идентифицированных пептидов из одного белка позволяют верифицировать выбор белка.

В идеале, идентифицированные в эксперименте белки должны согласованно указывать на решение исследуемой задачи.

Пики, полученные при измерениях на масс-спектрометре, показывают значение массы молекул, содержащихся в материале. Однако одно и то же значение массы может соответствовать молекулам, различным по составу и по структуре. Последовательность пептида в схеме на рис. 3.11 возможно идентифицировать лишь при согласованности положения пиков в измеренном спектре. И также, условием идентификации белков является согласованность в составе определенных на первом шаге пептидов.

Но измеренные спектры не всегда могут быть согласованными, как это показано на рис. 3.12. Располагая информацией о составе белков в биологическом материале, и о принципе разделения белков на пептиды, при интерпретации спектра следует выбрать наилучший пептид, из числа возможных. Однако не всегда этот выбор является однозначным.

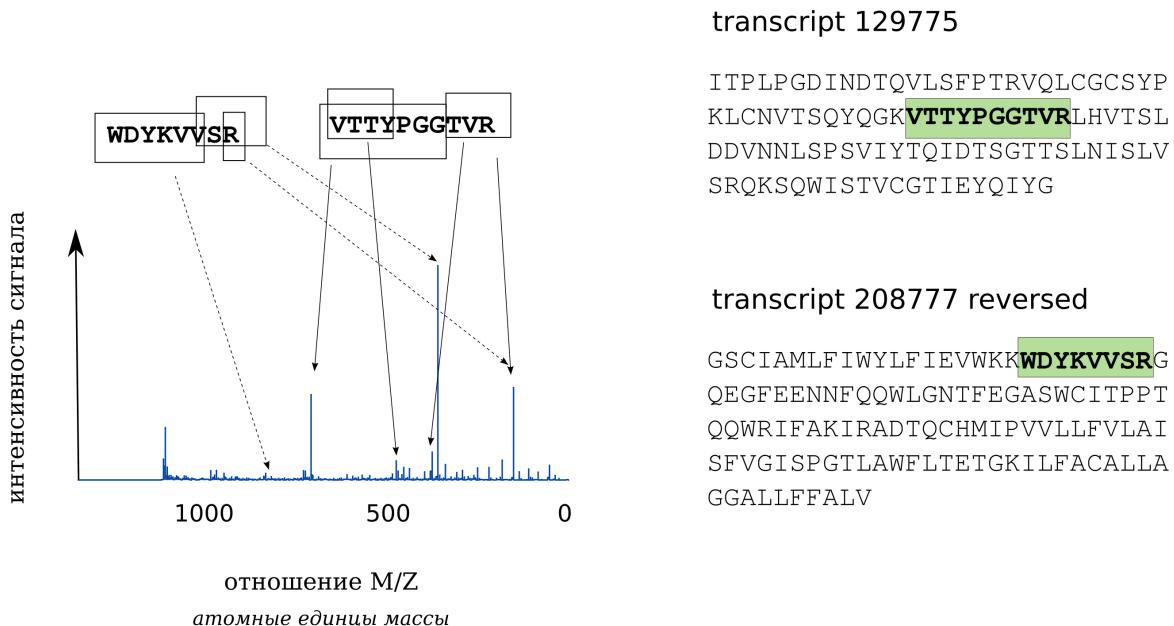


Рис. 3.12: Неоднозначности при идентификации пептидов в масс-спектрометрии

Слева: пики, полученные при измерении спектра M/Z, могут быть отнесены как разным ионам, полученным при расщеплении разных пептидов.

Справа: для уточнения идентификации пептидов используется база данных, содержащая возможные последовательности белков. Однако иногда этот набор белков слишком велик, чтобы допустить однозначный выбор.

при подготовке иллюстрации были использованы пакеты программ SearchGUI, PeptideShaker, OpenMS, на основе данных, полученных при исследованиях байкальской губки.

Для преодоления обозначенных проблем используют значительно более сложные и "ухищренные" постановки эксперимента. Среди приемов, используемых в этих экспериментах, следует упомянуть химические модификации отдельных аминокислот и использование атомов с другим молекулярным весом (изотопов). Такого рода возможности по изменению массы фрагментов и смещению положения пиков в спектре используют в этих исследованиях для дополнительной верификации и уточнения результатов. Однако нижняя граница вероятность ошибки при идентификации пептидов и белков в протеомике составляет примерно 1%. И, как правило, при использовании методов протеомики для решения прикладных задач требуется согласование этих подходов с другими экспериментальными и расчетными методами (рис. 3.13).

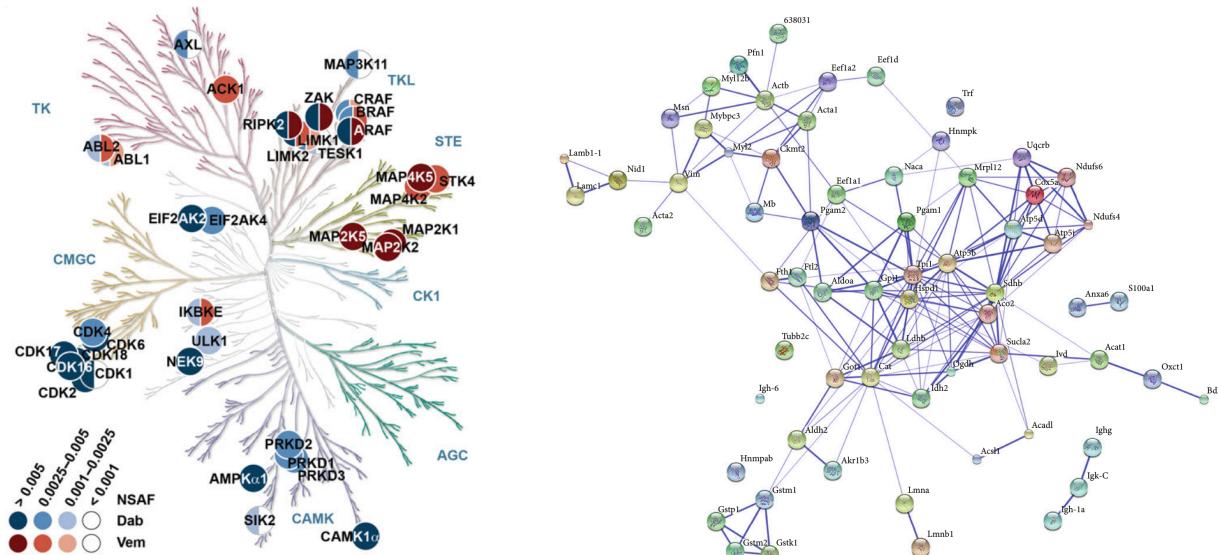


Рис. 3.13: Представление некоторых экспериментов с использованием массспектрометрии

Слева: исследование мишени препарата *dabrafenib* среди ферментов класса *kinase*. Темными оттенками показаны белки, среди которых с большей достоверностью следует выбрать искомый белок-мишень.

Справа: визуализация сети взаимодействующих белков, для которых методами протеомики было обнаружено изменение концентрации в стареющей ткани сердца. Взаимодействия белков восстановлены на основании базы данных "String".

рисунки из статей (Phadke и др. 2017; Holland и Ohlendieck 2014)

Как следствие, полученные результаты часто имеют качественный характер, и могут быть использованы лишь как подсказка, для обозначения направлений при поиске ответов и решений прикладных проблем. Впрочем, методы, основанные на секвенировании ДНК, описанные ниже, где нижнюю границу вероятности ошибки в сравнимых постановках задач можно оценить примерно в 0.1%, имеют такого же рода ограничения при их использовании, существенно не расширяя возможности по интерпретации данных в задачах прикладной молекулярной биологии.

Обработка экспериментов по секвенированию при изучении процессов в клетке

В исследуемом биологическом процессе участвуют ферменты, закодированные в ДНК как некоторые гены. Не все белки, закодированные в ДНК, присутствуют в каждой из клеток в организме, и относительное количество каждого из белков зависит от типа клетки и многих других факторов. Белки синтезируются в клетке в процессе трансляции мРНК, потому относительное содержание мРНК в клетке должно быть связано с содержанием белка, который кодирует эта мРНК. Транскрипция каждой из мРНК регулируется специфическим транскрипционными факторами, и потому регуляция биологических процессов в клетке будет выражаться в изменениях относительного содержания мРНК для определенных генов.

Среди технологий, используемых в современных методах секвенирования, есть подходы к считыванию последовательностей молекул мРНК, содержащихся в клетке, в форме коротких фрагментов ("ридов"). В итоге возможно получить большие объемы данных, в форме наборов "ридов", причем соотношение между "ридами" примерно соответствует содержанию каждой из мРНК в исходном материале. При условии, что известны последовательности всех генов в геноме, которым могли бы соответствовать мРНК, содержащиеся в клетке, относительное количество каждой из мРНК можно оценить по количеству считанных "ридов", идентичных некоторому фрагменту последовательности исходного гена.

Для оценки уровня экспрессии мРНК, при обработке таких экспериментов следует провести сравнение последовательности каждого из считанных "ридов" с последовательностью генома, так чтобы идентифицировать ген, соответствующий этому "риду". В итоге, после обработки всех исходных данных, каждому из генов будет сопоставлено некоторое количество "ридов", как это показано на рис.3.14 на примере фрагмента гена рибосомной РНК и "ридов", полученных при секвенировании образцов байкальской губки.

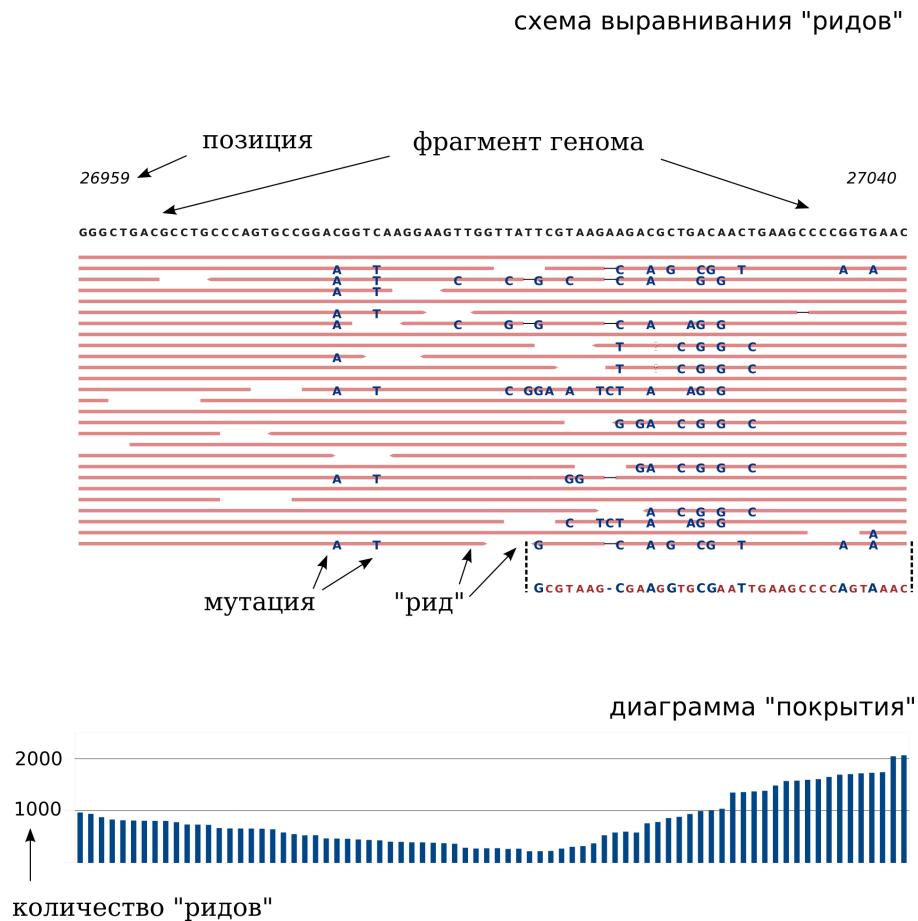


Рис. 3.14: Расчет уровня экспрессии по данным секвенирования
данные из серии экспериментов с кодом PRJNA480194

В описанной схеме расчетов предполагается, что для исследуемого организма заранее известна последовательность генома, с указанием положения кодирующих участков (генов). Последовательность генома человека получена и проанализирована в начале 2000-х годов, и с этого времени биологических видов, для которых получены последовательности геномов, становится все больше. Работа по определению последовательности генома для еще одного биологического вида может найти применение в исследовании функционирования клеток, в организме, относящемуся к этому виду.

Обсуждение алгоритмов и подходов, используемых для получения последовательности генома, не включено в материал, представленный в этой книге; эти алгоритмы лишь кратко упомянуты в таблице 3.2, среди групп алгоритмов, относящихся к обработке экспериментов по секвенированию. В таблицу также включен класс алгоритмов выравнивания, который обычно используют для проведения выравнивания ридов, где необходимым условием является ускоренное проведение расчетов, даже за счет снижения точности результата.

Таблица 3.2: Типовые преобразования данных при обработке экспериментов по секвенированию

Класс алгоритмов	Исходные данные	Результат расчетов	Некоторые из пакетов программ
Ассемблирование генома	риды ДНК	фрагменты генома (континги)	<i>Abyss, SoapDeNovo, Spades, ...</i>
Аннотация генома	континги	последовательности ДНК генов ("транскриптов")	<i>Maker, Augustus, Prokka, ...</i>
Выравнивание ридов РНК и ДНК с контингами генома	континги, риды	набор выравниваний	<i>Bowtie, BWA, ...</i>
Выравнивание ридов РНК с контингами генома	континги, риды, аннотация	набор выравниваний	<i>Tophat, ...</i>
Ассемблирование транскриптома	риды РНК	последовательности "транскриптов"	<i>Trinityrnaseq, ...</i>
Выравнивание ридов с последовательностями генов	транскрипты, риды	набор выравниваний	<i>Bowtie, Usearch, Blast, ...</i>
Выделение мутаций и полиморфизмов	набор выравниваний	позиции и свойства полиморфизмов	<i>Samtools, Vcf-tools, GATK, ...</i>
Расчет уровня экспрессии генов	набор выравниваний, аннотация	таблица уровней экспрессии генов	<i>Cufflinks, RSEM, ...</i>

Список, представленный в таблице 3.2, показывает возможности и проблемы, встающие при необходимости исследовать функции клеток, на основании результатов секвенирования. Некоторые из проблем, которые могут привести к искажению результатов и требуют разработки дополнительных классов алгоритмов, подробнее описаны ниже.

- Данные секвенирования могут содержать ошибки, а так же включения фрагментов вспомогательных последовательностей ДНК, используемых при проведении секвенирования. Проведение очистки и фильтрации является необходимым предварительным этапом при обработке "сырых" данных. При фильтрации данных используют вероятности ошибок считываения, оцениваемые, при проведении секвенирования, совместно с идентификацией каждого из нуклеотидов.
- Для оценки уровня экспрессии генов, следует оценить степень покрытия гена "ридами". Методологические затруднения возникают при сведении воедино оценок уровней экспрессии генов с различающейся длиной мРНК, и с различающейся степенью однородности покрытия (рис. 3.14).
- В многоклеточных организмах, кодирующая часть гена, представленная в мРНК как непрерывная последовательность, в геноме обычно разделена в геноме на фрагменты ("экзоны"). Это может создать затруднения при попытке сравнить "рид", полученный как участок мРНК, с последовательностью генома.
- При разделении гена на экзоны, эти экзоны могут быть составлены в клетке в мРНК несколькими способами. По "ридам" мРНК, может оказаться затруднительным выбор верного варианта мРНК как комбинации экзонов ("изоформы"), и даже оценка суммарного уровня экспрессии гена.
- Геном любого организма обычно содержит повторяющиеся участки (рис. 3.15). В частности, сразу несколько генов в геноме могут иметь общий участок. Это создает затруднения при оценке уровня экспрессии, когда выбор гена, при выравнивании с "ридом" мРНК, является неоднозначным.
- Наличие повторяющихся участков и неоднозначностей в геноме является наиболее "узким" местом при попытках асSEMBЛИРОВАТЬ последовательности мРНК по считанным "ридам", и потому оценки уровня экспрессии по последовательностям асSEMBЛИРОВАННЫХ транскриптов являются менее надежными, чем оценки, полученные при использовании последовательности генома.

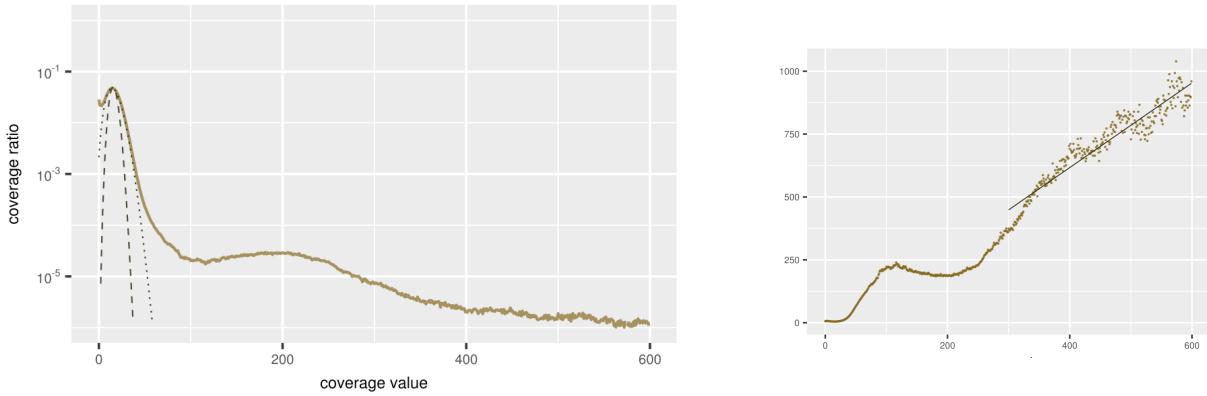


Рис. 3.15: Распределение повторяющихся участков в геноме

Слева: диаграмма степени покрытия фрагментов генома "ридами" ДНК. По горизонтальной оси - уровень покрытия, по вертикальной оси - доля участков с таким уровнем покрытия в геноме.

Правая сторона показанного распределения иллюстрирует распределение повторяющихся участков в геноме, так что при "рид" возможно выровнять на любой из этих участков.

Справа: диаграмма покрытия в преобразованных координатах, иллюстрирующая выполнение "степенного закона" для распределения повторяющихся участков. Понятие "степенного закона" обсуждается подробнее в разделе 3.8.

Использован геном модельного растения *Arabidopsis thaliana*. Рисунки из статьи (Kuzmin и др. 2019)

Геном каждого из организмов, и каждой из клеток, не вполне идентичен модельному геному биологического вида, используемому при проведении выравниваний "ридов". Различия в последовательности генома и каждого из "ридов", показанные на рис. 3.14 как "мутации", могут быть объяснены по-разному, и использованы на следующих ступенях анализа данных, обсуждение которых, в основном, выходит за рамки этой книги. Для некоторых, идеализированных и согласованных, распределений частот встречаемости нуклеотидов в позициях (рис. 3.16), появление наблюдаемых "мутаций" следует объяснять однозначно. Однако, при обработке реальных данных, для интерпретации распределений такого рода используют отдельный класс алгоритмов, также упомянутый в таблице 3.2.

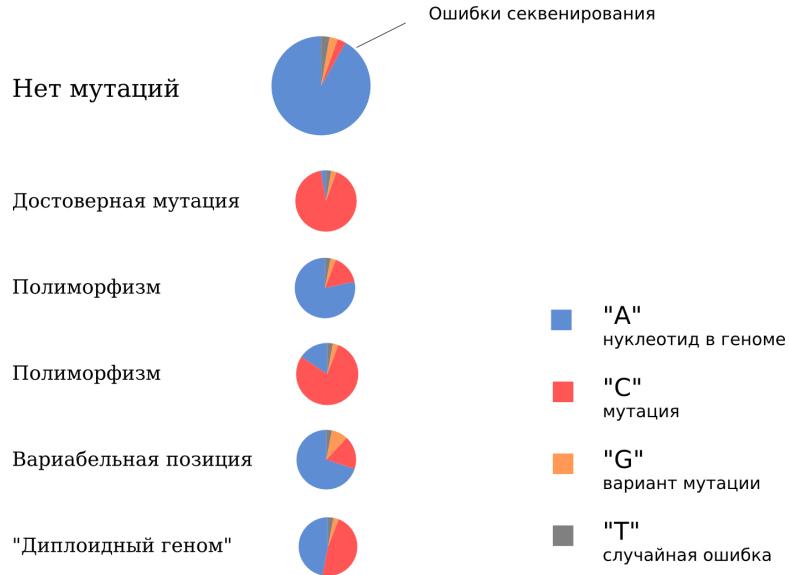


Рис. 3.16: Идеализированные распределения нуклеотидов в вариабельных позициях

Затраты на проведение секвенирования достаточно велики, но во многих исследованиях достаточно измерить относительное изменение отдельных выбранных генов. Технология полимеразной цепной реакции (ПЦР) разработана для детекции мРНК заранее определенного гена. Для отделения выбранной мРНК в этом методе используются специфичные олигонуклеотиды - *праймеры*. Модификация этой технологии, "ПЦР в реальном времени", позволяет также определить относительное количество мРНК. Этот метод может быть использован для измерения относительного уровня экспрессии определенного гена, наряду с методами измерения содержания соответствующего белка (рис. 3.17).

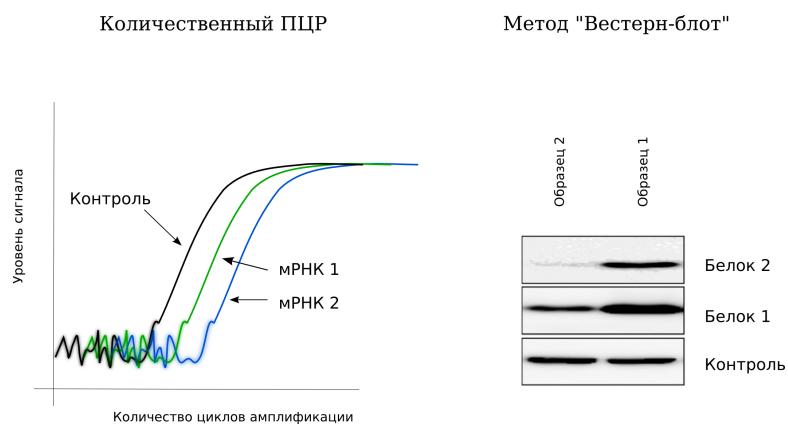


Рис. 3.17: Методы детекции изменений в экспрессии генов

Слева - схема измерения относительного содержания уровня мРНК, при удвоении количества молекул на каждом шаге температурного цикла в полимеразной цепной реакции (ПЦР).

Справа - измерение относительного содержания белков по уровню специфических антител. Исходные изображения: design-droide.com, статья ([Won и др. 2014](#)).

Идентификацию и оценку количества мРНК сразу для большого количества генов возможно провести также с использованием так называемых *микрочипов* (*microarrays*). В этой технологии, количество мРНК в биоматериале оценивается на основании степени гибридизации специфичных фрагментов ДНК в ячейках микрочипа. Ошибки и искажения в этом подходе, в целом, выше, чем при использовании секвенирования, и результаты экспериментов по двум технологиям непросто соотнести между собой. Но результат эксперимента в обоих подходах возможно свести к таблице, содержащей относительные значения содержания мРНК для каждого из выбранных генов.

Дифференциальная экспрессия генов

Традиционная и интуитивно простая постановка эксперимента по сравнению нескольких групп клеток или тканей может быть использована для определения генов, которые участвуют в регуляции биологических процессов, связанных с разделением использованных групп образцов. В экспериментах по измерению *дифференциальной экспрессии*, искомый набор генов возможно оценить по таблице, содержащей уровень экспрессии генов для каждого из образцов, с использованием моделей статистики.

Распределение генов по уровню экспрессии, полученное после обработки эксперимента, показано на рис. 3.18. Некоторые белки и соответствующие им гены представлены в клетке в большом количестве. Иллюстрации в разделе построены на основе обработки серии экспериментов по исследованию эпителия легких у больных бронхиальной астмой. И, в частности, белок ферритин, соответствующий гену FTL, оказавшимся одним из наиболее представленных генов в этом анализе, используются в клетке для накопления ионов железа.

гены с наибольшим уровнем экспрессии

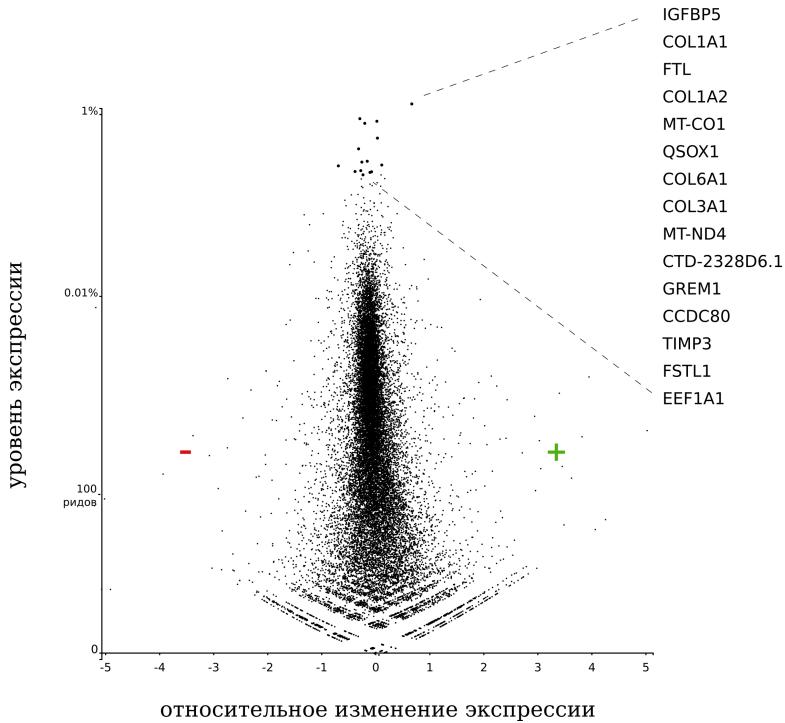


Рис. 3.18: Распределение генов по уровню экспрессии

использовано представление "md-plot" (*mean-difference plot*)

по вертикальной оси - относительное количество каждого из генов в образце, в логарифмических координатах

по горизонтальной оси - количество гена в образце, по отношению к среднему его количеству в серии образцов, в логарифмических координатах.

данные из серии экспериментов с кодом PRJNA252605

Наиболее важная из целей, стоящих при обработке используемой серии экспериментов - связать изменение экспрессии генов с фактом заболевания. Вариации в среднем уровне экспрессии генов в двух группах исследованных тканей показано на рис. 3.18 как разброс точек по горизонтальной оси.

Для наиболее представленных генов, различие в среднем уровне экспрессии между группами невелико, как это показано в верхней части распределения на рис. 3.18. Но различие между отдельными образцами, внутри каждой из групп, для выбранных генов может быть существенным, как это показано на рис. 3.19. И, в результате, следует сделать вывод о том что, хоть экспрессия генов, выбранных как наиболее представленные, может существенно изменяться в отдельных образцах, эти изменения никак не связаны с фактом заболевания. Такого рода рассуждения используется как основание для количественных оценок вероятности связи уровня экспрессии гена с разделением между группами образцов.

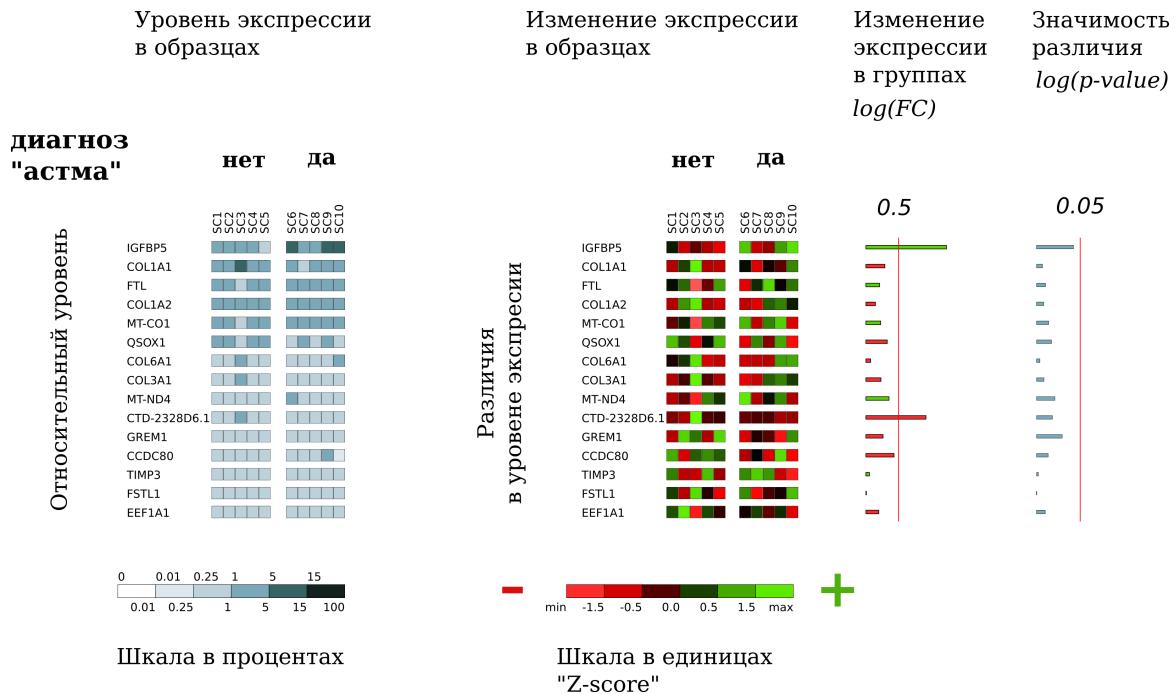


Рис. 3.19: Гены с наибольшим уровнем экспрессии
использовано представление "тепловой карты" ("heatmap")

Достоверность связи экспрессии гена с разделением образов по группам лишь косвенно соотносится с различием в среднем уровне экспрессии в группах. Различие в среднем уровне экспрессии может быть случайным, если для гена характерно неоднородное распределение в образцах, независимо от их группировки. Но косвенное соответствие между достоверностью связи и усредненным различием уровня экспрессии проявляется в наличии двух пиков в представлении таблицы экспрессии, показанном на рис. 3.20.

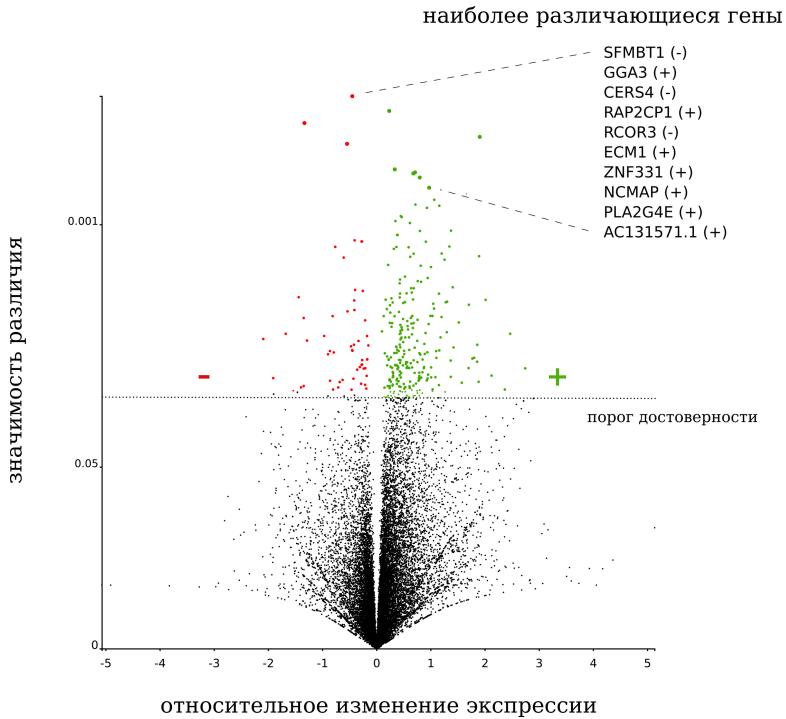


Рис. 3.20: Схема определения различающихся генов
использовано представление "volcano chart"

На рис. 3.20 проиллюстрирован принцип выбора наиболее различающихся генов. Уровень экспрессии этих генов в каждом из образцов показан на рис. 3.21. Все эти гены представлены в клетке в относительно небольшом количестве. И, как обобщение опыта работы с данными такого рода, следует отметить, что наибольший интерес в сравнительном анализе представляют гены, которые представлены в среднем хоть и в малом количестве, но полностью отсутствующие в одной из групп.

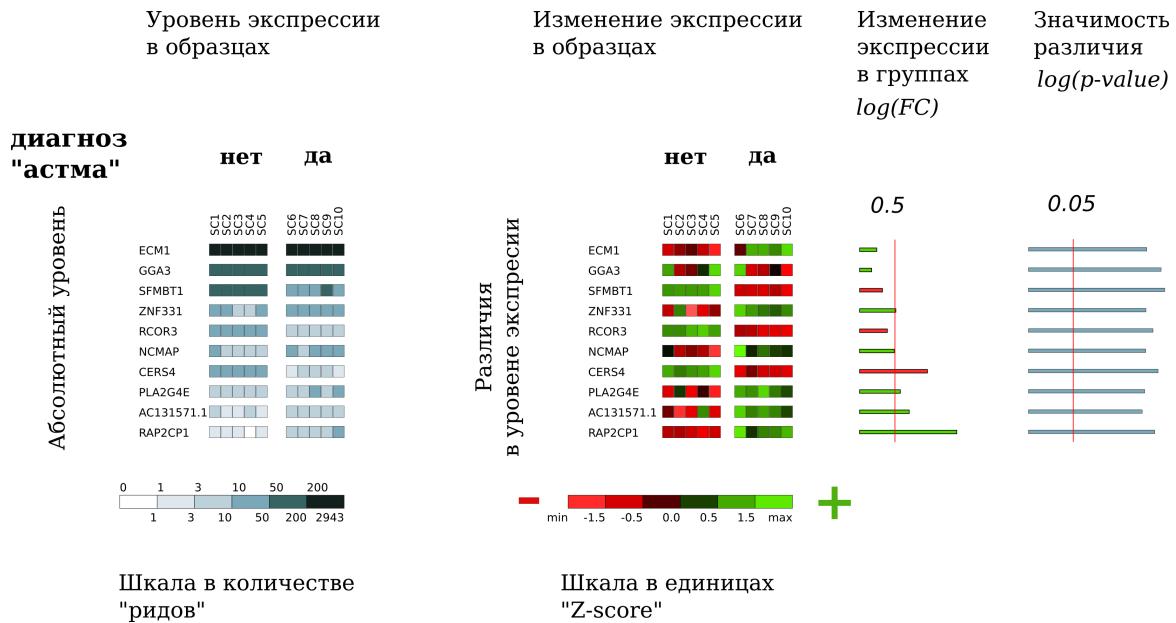


Рис. 3.21: Гены с наибольшей достоверностью различия

Достоверность разделения десяти лучших генов в приведенном примере составляет от 1×10^{-4} до 5×10^{-4} . Если учесть, что полное количество генов при этом составляет около 3×10^4 , степень разделения генов "балансирует" на грани недостоверности. Построение наиболее точных статистических моделей в пакетах и программах, используемых для исследований по дифференциальной экспрессии, начинается с моделей распределения покрытия ридов при выравнивании с геномом. Однако, даже на этом уровне, ожидаемое распределение не всегда соответствует обнаруживаемому при обработке эксперимента. И, в условиях "балансирования" на грани недостоверности, использование других статистических моделей привело бы к отделению другого списка генов, как наиболее различающихся между образцами. Эти замечания дают основание для проведения подробного обсуждения последствий возможных ошибок при изучении экспрессии генов.

Ошибки и погрешности при изучении экспрессии генов

При обработке экспериментов по секвенированию мРНК возникает достаточно много неоднозначностей, касающихся методов выравнивания и способов оценки уровня экспрессии. Некоторые из проблем, встающих при выборе метода обработки данных, описаны в предыдущих разделах. Но применение более точных расчетных методов для прояснения этих проблем не всегда оправдано. В живых клетках, используемых в серии измерений, всегда оказываются различия, не учтенные в постановке эксперимента, и эти различия ограничивает достоверность выводов о различии в уровне экспрессии любого выбранного гена.

Но если рассмотреть группу генов, и оценить достоверность совместного изменения уровня экспрессии генов в этой группе, значимость выводов о изменениях внутри группы может быть достаточно велика. И с этим можно связать распространение подхода по аннотации подобных

экспериментов на основании сравнения уровня экспрессии в заранее заданных наборах генов. В системах аннотации, каждый из наборов генов обозначен терминами, общими для многих тем, изучаемых в молекулярной биологии, что создает предпосылки к взаимопониманию между учеными из разных научных групп.

Однако возможно допустить существование разных способов группировки генов в наборы, и сам факт использования группировки приводит к увеличению достоверности утверждений в прикладных исследованиях. Поэтому, как может оказаться, некоторые из принятых способов группировки возникли по большей части по историческим причинам при развитии молекулярной биологии. И потому, возможно, некоторые понятия в языке общения молекулярных биологов имеют мало отношения к объективно существующим механизмам регуляции экспрессии генов.

Губка *Amphimedon queenslandica*, модельный организм из класса губок, привлекает внимание ученых, как объект для изучения ранних стадий развития эмбриона, для сравнению с развитием других, более высокоорганизованных, организмов. На рис. 3.22, на примере трех серий экспериментов по изучению развития эмбриона губки *Amphimedon*, проиллюстрирована неоднородность в уровне экспрессии генов в образцах из одной серии (A) и преимущества при использовании группировки генов (B). В этом примере, рассматривались группы генов, составленные на основе корреляции в уровне экспрессии, но без связи с заранее аннотированными наборами из систем по анализу дифференциальной экспрессии.

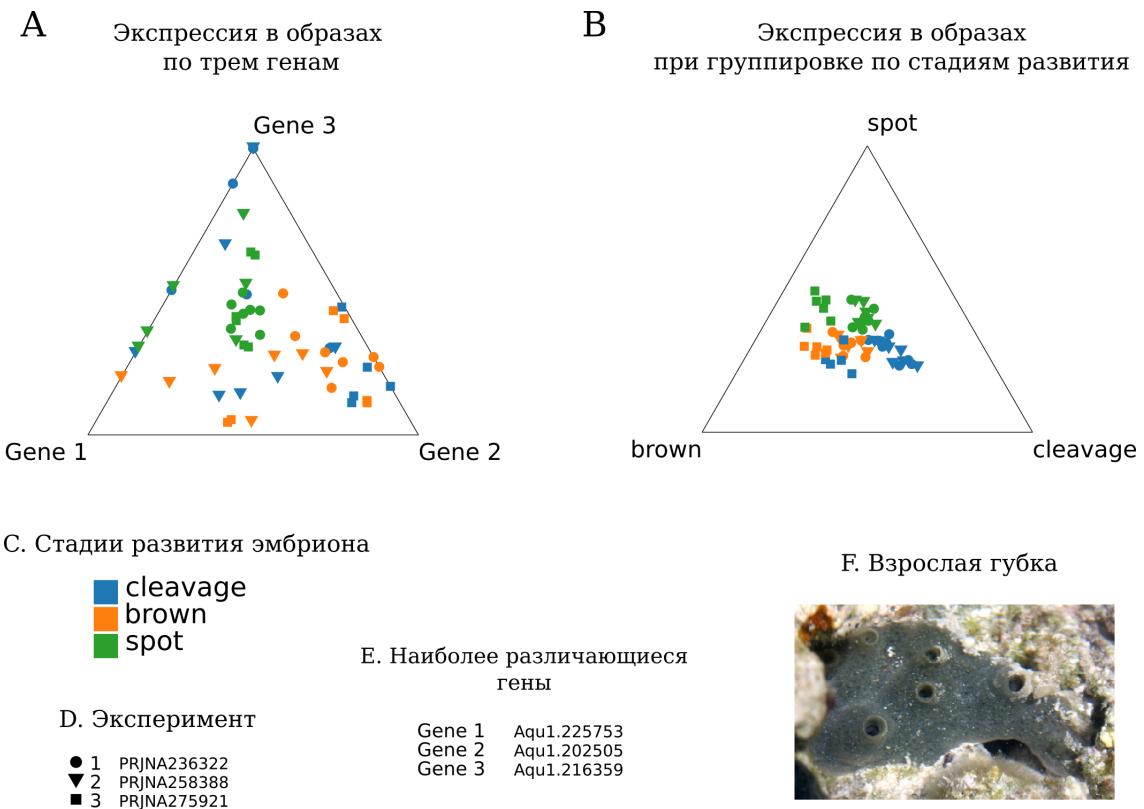


Рис. 3.22: Группировка генов для интерпретации сравнительного анализа экспрессии
Сравнение экспрессии генов в трех стадии развития эмбриона губки *Amphimedon queenslandica*.
Расчет проведен по результатам обработки трех независимых серий экспериментов Rna-seq по изучению развития эмбриона губки.

A. Тройная диаграмма, иллюстрирующая неоднородность профиля экспрессии генов в образцах, относящихся к одой стадии развития. Для диаграммы выбраны три гена, с наибольшим усредненным различием в уровне экспрессии при сравнении образцов по трем стадиям развития.

B. Тройная диаграмма, иллюстрирующая возможности по упорядочению представлений о регуляции экспрессии, допустив что в некоторой группе генов экспрессия изменяется согласовано. Для диаграммы выбрана группа из 60 генов, построенная с использованием подходов из теории графов. Выбор группы генов был основан на связи в уровне экспрессии генов в серии экспериментов, и на наиболее значимом разделении образцов по стадии развития.

Оценки статистической значимости разделения: по генам - 10^{-10} , по группе - 10^{-30} .

Для расчетов была использована библиотека *igraph* и материалы, опубликованные в (Adamska и др. 2007; Anavy и др. 2014; Levin и др. 2016).

Намерение ученых изучить механизмы развития с помощью современных методов молекулярной биологии понятно. Но результаты таких исследований, подобные опубликованным в работе (Levin и др. 2016), основанные на типовых методах аннотации экспрессии генов, с достоверностью не выше 10^{-3} , не всегда добавляют существенно новую информацию для прояснения ответов на вопросы, мотивировавшие проведение исследований. И это добавляет сомнений о реалистичности картины мира в представлении современной молекулярной биологии.

Исследование систем регуляции в клетке

В основе систем регуляции клетки лежат факты взаимодействие белков и других биомолекул. Для восстановления принципов устройства систем регуляции клетки, возможно использовать результаты экспериментов, таких как эксперименты по дифференциальной экспрессии. При этом, результаты экспериментов следует "спроектировать" на какое-либо модельное представление систем регуляции. Среди упрощений, вводимых в таких моделях, следует назвать принцип попарного взаимодействия генов. Взаимодействие возможно на уровне белков, на уровне регуляции экспрессии гена при взаимодействии белка с ДНК, не говоря о разного рода вариантах и "исключении из правил", которые возможно обнаружить при детальном анализе конкретных белков и механизмов их функционирования.

Проблемы, возникающие в задаче определения взаимодействующих генов, включают задачу отделения прямых взаимодействия от косвенных, и задачу отделения причин и следствий в наблюдаемых явлениях. Как можно пояснить эти проблемы на примере анализа дифференциальной экспрессии, по результатам обработки серии экспериментов по измерению состава транскриптома несложно заметить корреляцию в экспрессии некоторых генов. Но при этом остается открытым вопрос, связана ли наблюдаемая корреляция с фактом непосредственного взаимодействия генов, и, если факт взаимодействия был, то в каком направлении происходила регуляция.

Для уточнения фактов взаимодействия генов и направления регуляции клеточных процессов, разработано достаточно много подходов к постановке экспериментов и расчетов. Среди этих подходов, можно назвать методику постановки экспериментов по выяснению участков ДНК, являющихся мишениями для определенного транскрипционного фактора (*"ChIP-Seq"*), и методику по выяснению роли определенного гена в регуляции клетки, состоящие в подавлении его экспрессии в геноме (*"Knockout"*). Но более подробное перечисление возможных подходов выходит за рамки данного курса.

Как итоги усилий по исследованию механизмов регуляции клетки, были сформированы понятия о некоторых ключевых сигнальных путях, изменения в которых приводят к сбоям в развитии организма. На рис. 3.23 показаны результаты экспериментов по сравнительному анализу механизмов развития в многоклеточных организмах, где для нахождения сходства и различий в развитии оценивалась степень экспрессии генов в наиболее исследованных сигнальных путях.

В белках из класса транскрипционных факторов необходимой является функция связывания с двойной спиралью ДНК. Лишь небольшое число укладок белка обеспечивает эту функцию, среди всех транскрипционных факторов в геноме. Если в гене с неизвестной функцией обнаружена гомология с одним из мотивов, характерных для транскрипционных факторов, это является основанием отнести этот ген по функции к транскрипционным факторам. Гены, отнесенные к транскрипционным факторам, с общим типом укладки также образуют группу. И уровень экспрессии генов в наиболее исследованных группах транскрипционных факторов также показана в сравнительном представлении на рис. 3.23.

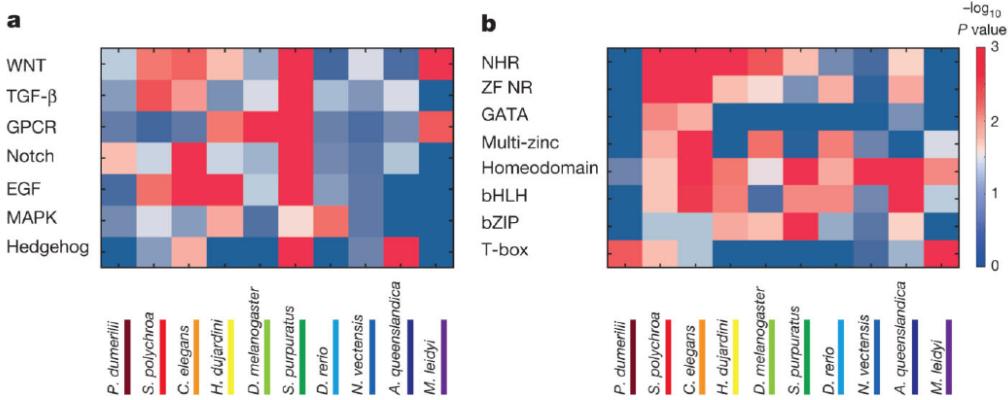


Рис. 3.23: Результаты функциональной аннотации некоторых сигнальных путей и семейств транскрипционных факторов на разных стадиях развития эмбрионов

Показанные результаты использованы для аргументации оригинальной модели развития многоклеточных организмов, названной авторами "модель обратных песочных часов".

Рисунок из работы (Levin и др. 2016).

Обозначения сигнальных путей на рис. 3.23 являются распространенными терминами в языке общения молекулярных биологов, как это показано на рис. 3.24. И действительно, даже для одноклеточных организмов в их механизмах адаптации можно заметить сценарии поведения, по сложности сравнимые с поведением высокоразвитых животных в сообществе. Клетки могут защищаться от внешних угроз, могут находиться в состоянии стресса, могут совершать "суицид", могут менять стратегию адаптации. В многоклеточных организмах, кроме этих стратегий, многие из механизмов отвечают за организацию отношений с другими клетками. Все перечисленные выше сценарии в этих клетках запускаются согласованно с состоянием организма как целого. Кроме того, возможен переход клетки в состояние "злокачественной", что в итоге приводит к развитию раковой опухоли и гибели всего организма.

В кратком пересказе, обозначения сигнальных путей можно объяснить в этих же терминах. Термин NF-кB обозначает молекулярный комплекс, который активируется в состоянии воспаления. Термин TGF (tumor growth factor) относится к сигнальному пути, характерному при перерождении клетки в раковую. Термин Jak (Janus kinase) относится к сигнальному пути, который активируется при заражении клетки внешним патогеном (вирусом). Термин Wnt относится к группе сигнальных путей, регулирующих согласованность выбора стратегий в клетках при развитии организма. Более узкий термин Bmal1 относится к сигнальному пути, обеспечивающему периодичность процессов в организме, синхронную с астрономическим временем.

Также здесь следует упомянуть термины TNF (tumor necrosis factor), относящийся к процессу апоптоза - программируемой гибели клетки, p53 - белок, тормозящий процессы переключения клетки в режим злокачественного развития, и CD4 - рецептор клеток иммунной системы, участвующий в системе различения здоровых клеток своего организма от чужеродных и зараженных клеток. Перегибы в росте числа публикаций, в которых упоминаются эти термины (рис. 3.24),

илюстрируют изменчивость интересов исследователей при развитии этой области знаний.

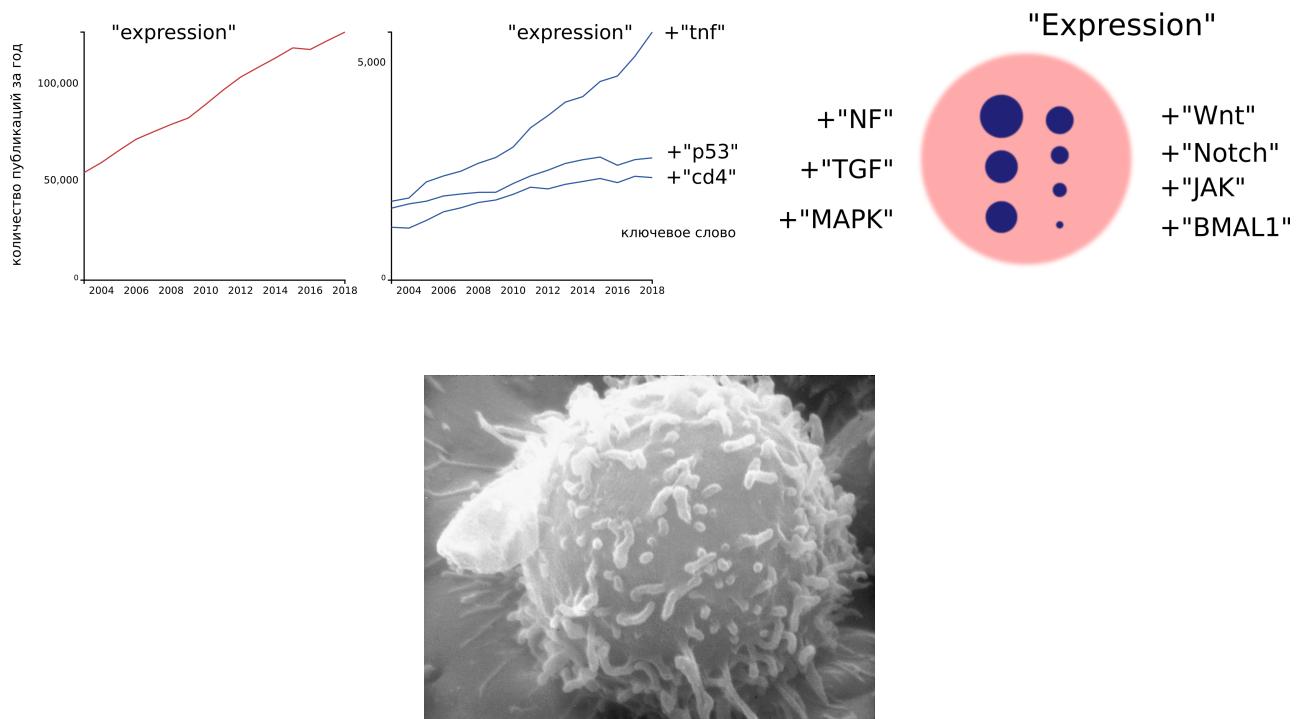


Рис. 3.24: Иллюстрация к обсуждению возможных систематических ошибок в языке общения молекулярных биологов

Вверху справа: Относительная частота упоминаний терминов, относящихся к обсуждению сигнальных путей, в библиографиях базы Pubmed, с 2013 года. Как фон, показано полное количество публикаций с упоминанием термина "экспрессия" за это время, около 700 тыс.

Вверху слева: Рост числа публикаций с упоминанием названий некоторых ключевых генов.

Внизу: Лимфоцит человека, одна из клеток иммунной системы. Изображение получено методом сканирующей электронной микроскопии, в 1976 году.

Credits: NCI

Но, в геномах многоклеточных организмов закодировано обычно более 30 тыс. генов, и при условии их изучения, соразмерного с их ролью в регуляции клетки, относительная представленность названий этих генов в научных публикациях находилась бы на уровне не выше чем для гена BMAL1 на рис. 3.24. И, возможность переоценки и смещения акцентов в интерпретации экспериментов была показана выше, при обсуждении методов обработки измерений дифференциальной экспрессии генов. Потому, характерные черты краткого списка терминов, используемых для обозначения сигнальных путей, могут лишь укрепить сомнения в адекватности языка общения молекулярных биологов.

Возможность количественных оценок значимости, безусловно, привлекает к использованию подходов по обработке экспериментов, основанных на таких оценках. Но для сложно устроенной системы регуляции в клетках, статистические модели, используемые для получения этих оценок, могут вносить существенные искажения в результаты. И также, выбор методов и технологий

обработки данных в каждом из прикладных исследований может быть не всегда наилучшим. Смещения акцентов при этом выборе могут возникнуть так же как и смещения акцентов при анализе систем регуляции. Но, несмотря на указанные недостатки, следует обратить внимание на "здоровое начало", содержащееся в обширных данных, накопленных при изучении систем клетки.

Большая часть ошибочных интерпретаций возникает при обработке экспериментов, при попытках сократить и обобщить полученные данные. Но, по принятой в современной науке традиции, "сырые данные" экспериментов также часто остаются доступными, допуская возможность их обработки с использованием других подходов. Если эксперимент был поставлен с целью проверки заведомо неадекватной гипотезы, эти данные являются мало полезными. Но все же, если даже считать уровень измеренного в эксперименте гена из семейства JAK не имеющим прямого отношения к системе защиты клетки, как это предполагалось при постановке эксперимента, эта информация также может найти применение в более широком сводном анализе.

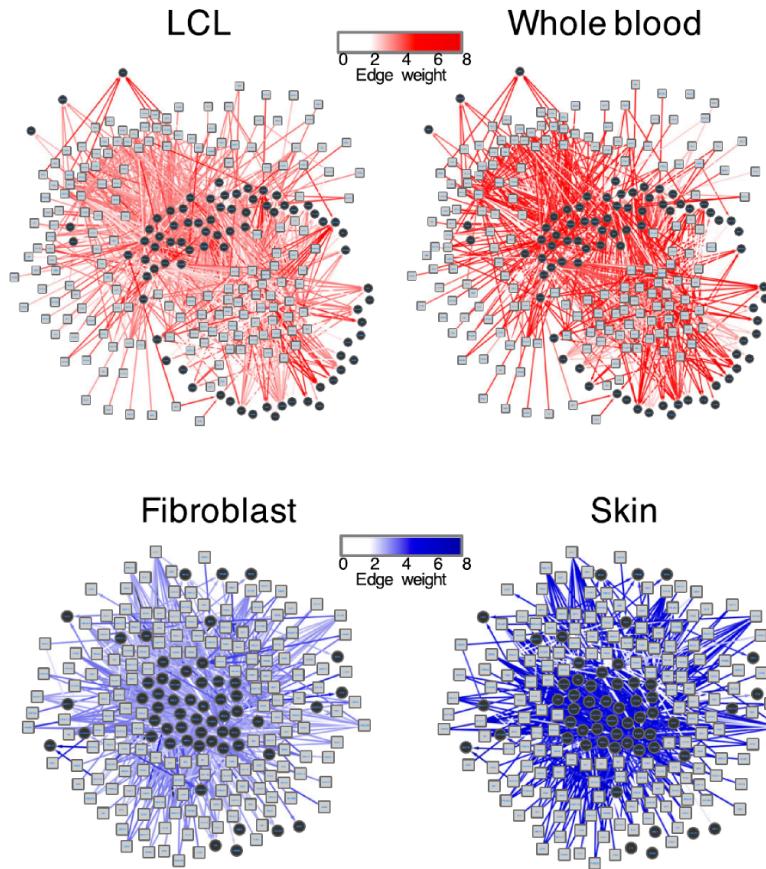


Рис. 3.25: Пример сравнения сетей регуляции генов

Слева - клеточные линии раковых опухолей; справа - здоровые ткани

Вверху - сравнение сетей регуляции для клеток крови

Внизу - сравнение сетей регуляции для клеток кожи

Светлыми прямоугольниками обозначены транскрипционные факторы, темными кружками - гены-мишени транскрипционных факторов. Цветовая интенсивность линий в графе показывает степень зависимости между элементами графа.

рисунок из статьи (Lopes-Ramos и др. 2017).

Согласно содержанию статьи, связи в графе установлены на основе обработки экспериментов по измерению экспрессии генов, и частично проверены по результатам экспериментов Chip-Seq.

Пример такого сводного анализа приведен на рисунке 3.25, где показан результат сравнения сетей регуляции генов в здоровых клетках, и в клетках, переродившихся в раковые. На рисунке заметно качественное различие в количестве взаимодействий, связывающих гены в процессах регуляции. Можно допустить также и другие способы сводного анализа, позволяющие на качественном уровне обобщить данные экспериментов. Но, поставив под сомнение количественные способы расчетов, нет оснований отдавать преимущества качественным оценкам, где заведомо нет возможности оценить достоверность выводов. Можно лишь с неизбежностью сделать вывод о том, что смысл и основание исследований в молекулярной биологии, в том числе исследований систем клетки, следует искать вне молекулярной биологии.

3.6. Аннотация и анализ публикаций

Подходы к автоматическому анализу текстов

Языки, на которых общаются люди, имеют иерархическую структуру. При осмыслиении и исследовании законов общения между людьми, возникли понятия "буква", "слово" и "предложение", чтобы обозначить наиболее устойчивые и независимые уровни, в общем объеме иерархически организованной информации. Но не любая иерархически организованная информация будет нести смысл при общении, так чтобы намерение того, кто передает сообщение, стало бы понятно.

Внутри традиций изучения каждого из языков, были выработаны правила, по которым буквы составляются в слова, и слова составляются в предложения, чтобы отличать бессмысленные сочетания слов от тех, которые, возможно, несут смысл. Но когда возникла необходимость обрабатывать тексты автоматически, при разработке программ и информационных систем, помимо использования жестких условий, разделяющих правильные или неправильные сочетания слов, получил развитие и другой подход, основанный на методах статистики.

Модели и алгоритмы, разрабатываемые для автоматического анализа текстов, возможно использовать только после проведения тестов и подбора параметров. Для подбора параметров моделей обычно используют большие объемы сходных по содержанию текстов, такую подборку текстов называют "корпус" (*corpus*). В этом контексте, набор библиографических записей с кратким описанием результатов научных исследований по биомедицинской тематике, доступный через проект Pubmed, следует считать одним из "корпусов" текстов.

При автоматической обработке научных публикаций из базы Pubmed, при обработке каждой из записей сопоставляют изложенные там результаты с некоторыми вариантами шаблонных утверждений или фраз, когда поиск информации, которую возможно выразить через одно из этих шаблонных утверждений, поставлен как цель при проведении обработки всего "корпуса". В этом контексте, поиск публикаций по ключевым словам - наиболее простой из примеров автоматической обработки, когда шаблоном является просто определенный термин.

В системе Pubmed, в "штатном" режиме работы, поиск информации происходит во взаимодействии с пользователем, читателем, который получает возможность увидеть подборку текстов, найденную по заданному ключевому слову, так чтобы самому понять смысл каждого из заголовков и выбрать среди найденных публикаций ту необходимую ему информацию, которую затруднительно записать как формальный шаблон. Но возможности человека понимать тексты по отдаленно знакомой ему тематике ограничены, и все большее применение находят результаты исследований, где результаты основаны полностью на автоматической обработке большого количества текстов, большую часть из которых никто из людей, участвующих в исследовании, не прочитывает.

В таком режиме, чем сложнее шаблон утверждения, заданный при поиске, тем больше вероятность пропустить нужную информацию, а также ошибиться, предположив наличие искомого утверждения в тексте, где его не содержится. Для фильтрации второго рода ошибок, следует использовать универсальные подходы к автоматической обработке текстов, включая проверку

условий и правил, а также статистические модели.

Hypericum perforatum is a perennial herb
that produces the anti-depression metabolite hypericin (Hyp)

John's Wort (Hypericum perforatum , HP) is hyperforin
which has antioxidant properties in dorsal root ganglion (DRG) neurons ,
due to its ability to modulate NADPH oxidase and protein kinase C .

Рис. 3.26: Иллюстрация подходов при анализе текстов

Приведенные примеры предложений подобраны в базы библиографий Pubmed; были использованы публикации (Yao и др. 2019; Naziroğlu и др. 2014).

На рисунке 3.26, в первом из примеров, фраза содержит сразу три ключевых слова. Hypericum - это латинское родовое название лекарственного растения, известного как "зверобой". Также, в этой фразе содержится название заболевания (depression), и обозначение химического соединения (hypericin). Шаблонное утверждение, которое соответствует подобранный фразе, состоит в том, что зверобой может использоваться как средство для лечения депрессии.

Содержащееся в приведенной фразе название химического соединения могло бы быть использовано для составления других шаблонных утверждений, однако это название не используется непосредственно в установленной связи между наименованием растения и заболеванием. Но, для исключения неверно установленных связей, возможно учитывать особенности используемого "корпуса" текстов. Среди текстов, где одновременно упомянуты и название растения, и термин, обозначающий заболевание, доля ложных соответствий была бы велика. Однако, в рамках неявно принятых в научном сообществе конвенций о проведении исследований, касающихся эффекта лекарственных растений, и описании их результатов, упоминание какого-либо химического соединения в тексте относит этот текст ближе к рамкам этой конвенции. И требование о наличии названия какого-либо химического соединения в тексте позволяет уменьшить долю ложных соответствий, в описанной постановке задачи анализа текстов.

Второй из примеров на рис. 3.26 иллюстрирует подход к решению задачи по анализу текстов, когда шаблоном поиска является утверждение об изменении экспрессии некоторого гена, в экспериментах по изучению эффекта лекарственных растений. В этой задаче, для исключения достаточной части ложных совпадений, и для определения направления изменения экспрессии, следует учитывать иерархическую структуру текстов. Порядок слов устанавливает структуру внутри предложения, как это помечено на рис. 3.26 тонкими стрелками. В общем случае, для восстановления смысла текста до необходимой степени, следует учитывать и другие отношения между словами. Так, в частности, использование отношения между местоимением it (он) и называнием растения (hypericum) необходимо для достоверного восстановления содержащегося в тексте утверждения об эффекте, проявляющемся на уровне экспрессии генов.

Во втором из описанных подходов, вид ошибок, когда верная информация остается пропущенной при автоматической обработке, легко пояснить, выделив в этом подходе задачу поиска в тексте слов и фраз, использованных авторами для обозначения генов. Геномы многих модельных организмов, используемых при проведении экспериментов, содержат сходные наборы генов, и для согласованности при представлении результатов исследований, для сходных генов в разных организмах принято выбирать одинаковые идентификаторы. Но в каждой из статей, право выбрать термины для обозначения генов остается за авторами. И даже в публикациях, касающихся исследования наиболее "популярных" путей регуляции, о которых упомянуто в разделе 3.5, названия генов можно записать по-разному. Например, гены, относящиеся к молекулярному комплексу NF-кВ, можно обозначить как NF-каррАВ, или NF-кВ. И, поскольку генов в публикациях упоминается достаточно много, и в публикациях, где упоминаются названия некоторых генов, авторы зачастую выбирают различные варианты обозначения, то даже для проведения расчетов на наиболее простом из уровней иерархии, объем работ по подготовке и проведению автоматического анализа текстов является существенным.

Ошибки в аннотации, причины и механизмы их накопления

Термин "онтология" относится к философии, науке о знании, и постановка задач аннотации в биоинформатике, в широком смысле, смыкается с задачей об упорядочении и "осмыслинении" знания, накопленного в молекулярной биологии. Одна из наиболее очевидных проблем, возникающая при попытках найти решение этой глобальной цели, и проявляющаяся в задачах аннотации - это неполная достоверность некоторых результатов, лежащих в основе восстанавливаемой философской системы. Проведение функциональной аннотации генов проводится на основании сравнения неизвестных генов с эталонными базами данных, где содержатся биологические последовательности с подробными и проверенными описаниями. И, когда в некоторых из таких описаний закрадываются ошибки, что неизбежно при научных исследованиях, некоторые из этих ошибок многократно копируются при составлении функциональной аннотации других геномов. И потому к результатам по аннотации следует относиться с долей условности, и выводы, полученные из этих результатов, часто требуют дополнительных обоснований.

Научное сообщество - это часть общества, и заблуждения в науке подобны заблуждениям в обществе. В обществе возможно социальное неравенство, и значительное социальное неравенство может отражать несправедливое устройство общества. Подобно этому, в современной науке некоторые из направлений исследований прорабатываются значительно подробнее чем остальные, и это может означать наличие ошибок в системе научного знания, более масштабных, чем упомянутые выше ошибки аннотации.

Ранее в истории, некоторые из результатов полученных на основе научного метода, привели к существенному расширению возможностей человека и общества, в военном деле, а так же промышленности, экономике, медицине, сельском хозяйстве. И потому, в современном обществе, если целью его движения "вперед" является стабильный рост "валового внутреннего продукта", измеряемого в денежном эквиваленте, одним из приоритетных направлений для достижения цели следует считать проведение научных исследований. Результатами работы научных групп

являются законченные и опубликованные исследования, и степень эффективности ресурсов общества, выделенных на науку, можно оценить по количеству научных публикаций. Описанный механизм объясняет рост информации, доступной научному сообществу, в формате публикаций и в других видах.

Но, возможность человека упорядочить поток информации ограничены. И, имея способность к самовоспроизведению и находясь не вполне под контролем, информация иногда начинает "жить своей жизнью". И, через такого рода механизмы, при распространении информации проявляются черты конкурентной борьбы и существенного преимущества некоторых "видов". В условиях кризиса всего общества, преимущество "победителей" возрастает, по отношению к более "слабым", в том числе и в отношении приоритетных направлений исследования, и в отношении научных работников.

При любого рода самовоспроизводстве фрагментов информации необходимыми являются затраты энергии. Источником энергии при копировании "видов" информации является та же энергия Солнца, которая питает и все живое. При описании цепочки передачи энергии в современной науке, следует упомянуть систему финансирования научных проектов, и вопрос о разделении прикладных и фундаментальных тем исследований.

В качестве примера, проиллюстрированного на рис. 3.27, можно привести сравнительную динамику количества публикаций в базе Pubmed, по запросам "protein folding" и "protein folding amyloid". Исследование механизмов сворачивания белка, на которых основаны все процессы более высокого уровня в молекулярных системах, следует безусловно отнести к фундаментальным темам биофизики. И, одной из частных задач в этой обширной теме является исследование механизма образования *амилоидных бляшек* (amyloid fibrils), накопление которых со временем привести к старческому слабоумию (*болезни Альцгеймера*). Один из генов в геноме человека кодирует пептид, для которого возможно два устойчивых варианта конформации. Один, "неверный", вариант конформации обладает свойством к самовоспроизведению, так что количество "неверно" свернутых пептидов возрастает. Такие "неверно уложенные" пептиды, накапливаясь, и составляют основу амилоидных бляшек.

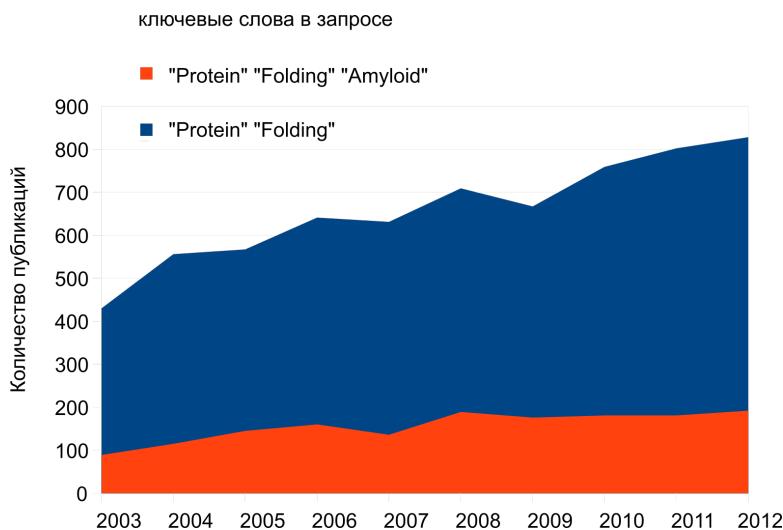


Рис. 3.27: Динамика количества публикаций, относящихся к исследованиям по сворачиванию белков

По вертикальной оси показано количество публикаций за год, по горизонтальной оси обозначен период по годам в интервале от 2003 до 2012.
График построен на основании результатов текстового поиска в базе Pubmed

Лечение болезни Альцгеймера - задача прикладная; и пример на рис. 3.27 может проиллюстрировать отношение между прикладными и фундаментальными исследованиями. Как другого рода иллюстрация, для известного специалиста в области физики белка, А.В.Финкельштейна, опубликовавшего более 100 работ, доступных в базе Pubmed, указанное отношение составляет примерно 50/2. Но все финансирование, выделенное Российской научным фондом на исследования под его руководством, согласно открытым сведениям, относится к задаче об изучении амилоидных бляшек.

В описанной цепочке передачи энергии, необходимой для самовоспроизведения фрагментов информации, необходимым звеном является участие научных работников. И иногда оказывается, что "выживание" научных понятий преобладает над свободой воли людей, относящих себя к научным работникам. Это сказывается на смещении акцентов в научных публикациях, и появляется при решении задач подобным функциональной аннотации генов, и автоматическом анализе текстов публикаций.

3.7. Обработка данных в медицине

Медицинские измерения и их интерпретация

Электрокардиограмма - результат регистрации и исследования электрических полей, образующихся при работе сердца. При интерпретации электрокардиограммы используется модель распространения электрического сигнала по проводящей системе сердца при его ритмичных

сокращениях. В норме, сигнал сокращения распространяется по всему сердцу, от участка, называемого *синусным узлом*, где возникает как согласованное возбуждение клеток ("типовидных клеток" *синусного узла*). Острые или хронические повреждения сердечной мышцы, и другие нарушения его работы, приводят к изменению в проводящей системе сердца, и, как следствие, к изменениям электрических полей, зарегистрированных на *электрокардиограмме*.

Электроэнцефалограмма - запись потенциалов суммарной электрической активности мозга, отводимой с поверхности кожи головы. На этой записи заметна ритмичность электрической активности мозга, и при интерпретации электроэнцефалограмм принято различать отдельные ритмы, называемые альфа-ритмами, бета-ритмами, и.т.п. По *электроэнцефалограмме* можно заметить изменения в характерной частоте и амплитуде ритмов. Явно заметные отклонения в ритмах могут быть использованы при диагностике больных эпилепсией.

Аллергологическая проба - результат теста, показывающего степень ответной реакции иммунной системы на определенный химический элемент или вещество, такое как бытовая пыль или пыльца цветущего растения.

Биохимический анализ крови - метод лабораторной диагностики, при котором проводится определение концентрации некоторых химических соединений в порции крови, полученной при проведении анализа. Среди используемых тестов, в анализе могут быть определены содержание химические элементы (кальций, калий и др.), продуктов метаболизма (мочевина, билирубин и др.), уровень холестерола.

Психодиагностика - методики для оценки и измерения индивидуально-психологических особенностей личности. Психодиагностическое тестирование может быть использовано в психотерапии и для диагностики заболеваний психики. Среди методик проведения тестирования, стандартизованная схема перевода ответов, полученных при заполнении опросного листа, в оценки риска, по направлениям, характеризующим возможные уязвимости в психическом состоянии пациента.

Подходы к получению доказательств в медицине

Проверка гипотез - основной из подходов в математической статистике. В его основе лежит построение модели, в рамках которой совокупность полученных в эксперименте данные является одной из возможностей. Выбор между возможностями, как это подразумевается, произошел случайно, но в при этом в модели определена вероятность каждого из допустимых исходов. Если исход, соответствующий экспериментальным данным, достаточно вероятен, то гипотезу можно называть проверенной.

Достоверность (значимость) - степень вероятности предложенного описания экспериментальных данных, рассматриваемого как гипотеза в рамках математической статистики.

Доверительный интервал - степень погрешности при оценке параметра модели. В этой схеме расчета погрешности предполагается, что точное значение параметра неизвестно, и выпало случайно; но при этом с достаточной достоверностью верна гипотеза, в которой утверждается, что это значение находится в границах доверительного интервала.

Нулевая гипотеза - предположение о том, что связь между изучаемыми в эксперименте явлениями отсутствует. Обычно рассматривается как дополнение к гипотезе о связи между явлениями, в соответствии с некоторой *вероятностной моделью*.

Клинические испытания (клинические исследования) - научное исследование (эксперимент) с участием людей (как объектов) для сбора информации, которую возможно использовать для оценки достоверности гипотезы об эффективности некоторой методики лечения

Рандомизированное контролируемое испытание (randomized control trial, RCT, РКИ) - схема проведения эксперимента, когда участники (объекты) случайным образом делятся на группы, в одной из которых объекты подвержены воздействию некоторого фактора, исключенного во второй (контрольной) группе.

Лекарственное средство - по формальному определению, "вещество или смесь веществ синтетического или природного происхождения в виде лекарственной формы (таблетки, капсулы, раствора, мази и т. п.), применяемые для профилактики, диагностики и лечения заболеваний". Эффективность лекарственного средства может являться объектом клинического исследования.

Плацебо - имитация лечения в контрольной группе, например при проведении клинических испытаний по схеме РКИ, для исключения вклада самовнушения "пациентов" в оценку эффективности методики лечения.

Доклинические испытания - научное исследование, которое проводится для оценки эффективности методики лечения, без участия людей как объектов эксперимента. Объектами при оценке эффективности в этом случае могут быть, например, лабораторные животные или культуры клеток.

Относительный риск (relative risk) - в приложениях статистики к медицине, оценка степени связи некоторого фактора с наступлением события. При расчете относительного риска предполагается разделение исследуемых объектов на две группы, подобное разделению в схеме РКИ. В каждой их групп, рассчитывается достоверность гипотезы о связи фактора с проявлением эффекта. Величиной относительного риска будет отношение достоверности при воздействии эффекта, пусть даже эта достоверность мала, и в отсутствии эффекта, когда по предположению эффект не должен проявляться.

Чувствительность и специфичность теста - меры для оценки эффективности теста, такого как диагностический тест, в котором возможно два варианта ответа. При проверке эффективности теста, зная правильный ответ, возможно четыре случая: положительный ответ, когда следует его дать, отрицательный ответ, когда следует ответить положительно, и оба варианта ответа, когда следует ответить отрицательно. *Чувствительность*, как определение меры - это доля правильно данных положительных ответов, а *специфичность* - доля правильно данных отрицательных ответов. При разработке теста возможно корректировать баланс между "чувствительностью", дополняющей долю ошибок "пропуск цели", и "специфичностью", дополняющей долю ошибок "ложная тревога". Выбор методики для сведения этих двух мер к единому показателю эффективности зависит от целей тестирования, в рамках проводимых исследований.

Многовариантный скорректированный коэффициент риска (multivariable adjusted risk ratio)

- оценка относительного риска, рассчитанная при необходимости свести воедино оценки, полученные на основании нескольких возможных методик расчета. Такая задача встает достаточно часто при обработке медицинских экспериментов, когда из-за сложности изучаемого явления нет возможности построить адекватную статистическую модель и выбрать наилучшую методику оценки относительного риска.

Некоторые из терминов, относящихся к экономическим отношениям в фармацевтике

Регистрация лекарственного средства - получение разрешения на употребление лекарственного средства в медицинской практике, выдаваемое уполномоченными органами государства, действующими на основании законов об обороте лекарственных средств.

Патент - право собственности на идею, как объект, регулируемый в рамках законов о распоряжении правом собственности. Патент на новую идею выдают уполномоченные государственные органы, после прохождения процедуры регистрации. Идеей при регистрации патента, может быть, например, новая методика лечения, или новая методика технологического процесса.

Дженерик (generic drug) — лекарственное средство, содержащее химическое вещество — активный фармацевтический ингредиент, идентичный запатентованному компанией - первоначальным разработчиком лекарства.

Стартап (startup) — коммерческий проект, инициированный для разработки нового продукта или услуги, с заведомо высокими шансами не компенсировать затраты и усилия. Характерными чертами стартапа также являются недолгое время существования, небольшой состав участников, и деятельность в области "высоких технологий", таких как фармацевтика. Среди возможных причин прекращения деятельности, неудача при попытках получить прибыль, и поглощение более крупной компанией.

Некоторые из гипотез, рассматриваемые в современной медицине

Гипотезы о причинах старения

- старение заложено в программе развития организма, поскольку, согласно теории естественного отбора, неминуемость смерти каждого отдельного представителя биологического вида способствует приспособлению вида как целого.

- при развитии отдельного организма, сопровождающегося делением клеток, в копиях генома новых клеток накапливаются мутации, приводящие к ошибкам в системах регуляции этих клеток.

- мутации, приводящие к сбоям, происходят случайно, и потому нет возможности предсказать, какие именно сбои произойдут при развитии организма, и какой из них приведет к смерти.

- при развитии отдельного организма, в составе метаболитов, в клетках и в межклеточном пространстве, накапливаются метаболиты, несущие вред дальнейшему развитию организма.

Гипотезы о причинах возникновения раковой опухоли

- случайно возникающие мутации могут привести к сбоям в системе регуляции клетки, так

что в клетке блокируется программа о необходимости ее смерти, которая в норме необходима для замены клеток при развитии организма.

- ослабление системы защиты организма приводят к сбоям в механизмах, обеспечивающим в норме уничтожение клеток организма, переродившихся в раковые из-за случайно возникших в них мутаций.

- заражение некоторыми вирусами, обладающими свойством встраивать свой генетический материал в геном зараженной клетки, способствует перерождению некоторых зараженных вирусом клеток в раковые.

- внешние факторы, такие как радиоактивное излучение или некоторые химические соединения, увеличивают вероятность мутаций при делении клеток организма, как, например, блокируя ферменты, необходимые для восстановления повреждений в ДНК.

Гипотезы о причинах сбоев в сердце и кровеносных сосудах

- закупорка сосудов происходит при накоплении на их стенках бляшек, образованных их молекул **холестерола**, химического соединения из класса жиров.

- холестерол в норме присутствует в организме, и жиры в норме присутствуют в пище. Но некоторые категории соединений из класса жиров, содержащиеся в пище определенного рода, способствуя смещению баланса в системе обмена веществ, повышают риск закупорки сосудов склеротическими бляшками и развития сердечно-сосудистых заболеваний.

Некоторые из терминов, относящихся к характеристикам лекарственных средств и продуктов питания

Антиоксидант - химическое соединение, приводящее к замедлению окислительных реакций в клетках организма, и, как следствие, замедлению накопления метаболитов, сопровождающему процессы старения в клетках организма.

Канцероген - химическое соединение, повышающее вероятность случайных мутаций в генетическом материале клеток, и, как следствие, увеличивающее риск возникновения раковой опухоли.

Насыщенные жирные кислоты - класс химических соединений из класса жиров, содержащиеся в большей степени в пище животного происхождения. Некоторые из проведенных исследований дали основание связать риск развития склеротических бляшек и содержание насыщенных жирных кислот в пище.

Поли-ненасыщенные жирные кислоты - класс химических соединений из класса жиров, в основном дополняющий насыщенные жирные кислоты в любой, по происхождению, жирной пище.

Антибиотик - вещество, подавляющее рост микроорганизмов, бактерий и иногда простейших эукариот. Антибиотики природного происхождения чаще всего вырабатываются *актиномицетами*, обычными бактериями, образующими тела подобные телу гриба. Некоторые антибиотики сильно подавляют рост и размножение бактерий и при этом почти не повреждают клетки организма-хозяина, и потому применяются в качестве лекарственных средств.

Антиsepтик - лекарственное средство, предназначенное для предотвращения процессов разложения на поверхности открытых ран, например в ранах, образующихся после больших операций или ушибов.

Иммуномодулятор - природные или синтетические вещества, способные оказывать регулирующее действие на иммунную систему.

Адаптоген - фармакологическая группа препаратов природного или искусственного происхождения, способных повышать неспецифическую сопротивляемость организма к широкому спектру вредных воздействий физической, химической и биологической природы.

Пробиотик - класс микроорганизмов и веществ микробного и иного происхождения, использующихся в терапевтических целях, а также пищевые продукты и биологически активные добавки, содержащие живые микрокультуры. Эффект при использовании пробиотиков наиболее явно заметен при их использовании для улучшения функционирования кишечника и стимулирования иммунной системы.

Стволовые клетки - клетки, составляющие зародыш (эмбрион) на ранних этапах роста, до стадии дифференцирования в клетки органов и тканей. Содержатся в значительном количестве в плаценте и околоплодных водах. Поскольку эти клетки не прошли стадий деления, то, по предположению и следуя принятым взглядам на причины старения, их трансплантация может в некоторых случаях способствовать частичному восстановлению организма.

Краткие комментарии к методам и терминологии в прикладных медицинских исследованиях

Медицина - наиболее прикладная из тем исследования, рассмотренных в книге, и для беспристрастности изложения при обсуждении этой темы следует ограничиться лишь несколькими замечаниями. Из этих замечаний, во-первых, подавляющее большинство исследований в этой области проводятся при попытках преодолеть болезни, которые являются причиной смерти большей части людей в мирное время. Во-вторых, объяснения причин этих болезней вводятся в контексте теории эволюции, где заведомо подразумевается неизбежность и необходимость старения и смерти. В-третьих, деньги, направляемые на проведение исследований в этой области, велики, и поступают в значительной степени из "благотворительных" фондов, чей бюджет пополняется "пожертвованиями" из семей, которых коснулись трагедии, сопряженные с упомянутыми заболеваниями. Но, как было показано в предыдущих разделах, на примере некоторых "передовых" технологий молекулярной биологии, достоверность выводов в таких исследованиях в целом следует считать невысокой.

Кратко также следует упомянуть о фактах споров и обсуждений, касающихся степени этичности при проведении испытаний на "добровольцах", при использовании послеродовых выделений женщин в технологии трансплантации стволовых клеток, о соответствии между оправданием необходимости понятия патента и использованием патентной защиты при установлении цен на лекарственные средства. Некоторые несообразности, замечаемые в современной медицине, в

темах подобных перечисленным выше, дают основание соотнести современные традиции с традициями древней медицины у разных рас и народов. Обобщения, которые можно вывести при таком сравнении, приведены в следующем разделе и в заключительной части книги.

Традиции медицины и их эволюция

Лекарственные растения всегда использовались для лечения заболеваний. Многие из химических соединений, которые применяют в современной медицине, были подобраны при исследовании состава лекарственных растений. Также, для многих из современных лекарственных средств, сырьем служат продукты переработки растений. И потому исследования лекарственных растений продолжают развиваться, и в базе научных публикаций возможно провести автоматический анализ текстов, с целью обобщить представленные в публикациях результаты по многим из прикладных задач.

В результате проведённой обработки публикаций, используя подход описанный в 3.6, были отобраны записи, содержащие родовое название растения, название химического соединения, и термин относящийся к хроническим заболеваниям. Полученную базу записей возможно представить в виде матрицы соответствий между названиями растений и заболеваниями. Для построенной матрицы был проведен анализ главных компонент, так что в итоге по парам главных компонент возможно построить диаграмму отношений между лекарственными растениями и спряженную диаграмму отношений между группами заболеваний, как это показано на рис. 3.28.

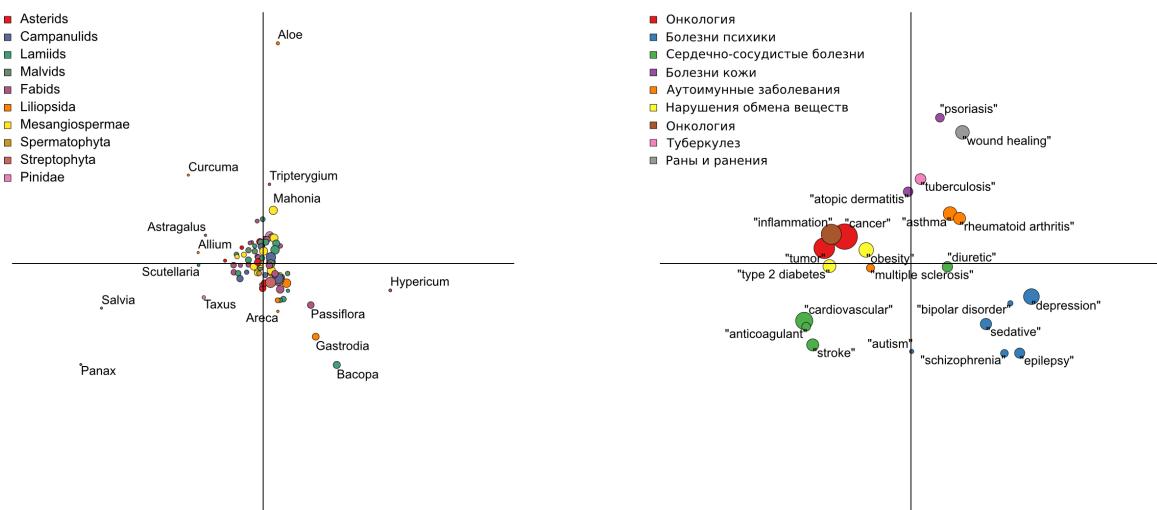


Рис. 3.28: Отношения между лекарственными растениями и между группами хронических заболеваний

в левой части, размер кругов обозначает относительное количество публикаций с упоминанием лекарственного растения, а цветом обозначена группа, которой отнесено растение по его ботаническому описанию.

в правой части, размер кругов обозначает относительное количество публикаций, относящихся к заболеванию.

Часто в публикациях при описании лекарственного растения упоминают, в какой из традиций медицины принято было его использовать для лечения. И потому, в проведенной обработке текстов выделялись также географических термины, относящиеся к традициям медицины. Результаты, показанные на рис. 3.29, основаны на разложении матрицы, составленной из связей между лекарственными растениями и традициями медицины. При обработке также были использованы термины "modern" и "official", чтобы охарактеризовать мета современной медицины среди традиций, выработанных при развитии цивилизации.

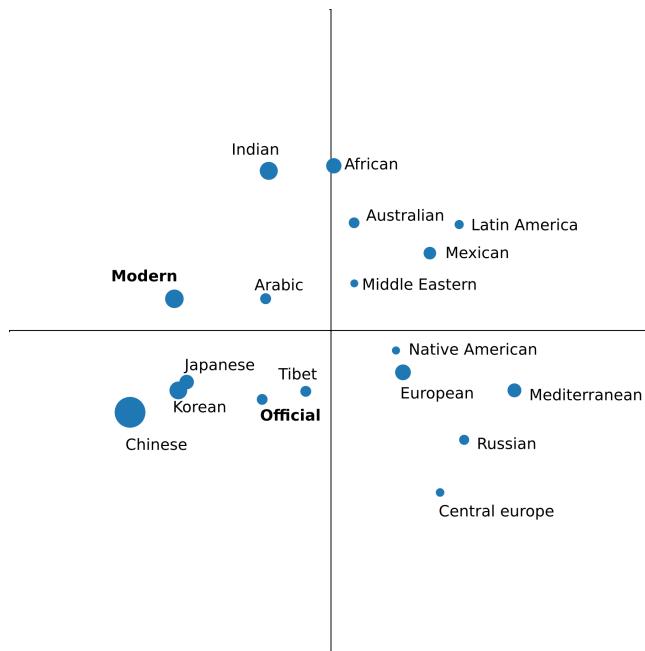


Рис. 3.29: Отношения между традициями медицины

размер кругов обозначает относительное количество публикаций, относящихся к традиции медицины.

В показанном способе расчета, традиционные "школы" медицины, с долей условности, разделены на культуры Востока, культуры Европы, и культуры народов Африки, Австралии и Америки. Подходы современной медицины, в этом представлении, подобны, в большей степени, культурам Востока. Но, говоря прямо, "переоцененные" научные идеи и термины, приводящие к накоплению систематических ошибок в современной науке, как это обсуждалось в разделах 3.5, 3.6, 3.7, следует сравнить с языческими богами, подобными языческим богам древних культур Востока. И потому обнаруженное сходство современной медицины с традициями Востока не следует считать неожиданным.

Другой подход к автоматической обработке текстов, описанный в 3.6, где названию растения сопоставлялись названия генов, также был использован для оценки отношений между заболеваниями, по предполагаемому эффекту лекарственных растений. Для анализа были также использованы серии экспериментов RNA-Seq, где исследовалось различие в экспрессии генов в тканях,

отобранных у пациентов, страдающих одним из заболеваний, и здоровых людей. Эффект лекарственного растения, оцененный, после обработки текстов, как направление регуляции некоторых генов, был обобщен до уровня, описывающего направление смещения профиля экспрессии генов. Это позволило сопоставить рассчитанные свойства растения со смещением профиля экспрессии, наблюдаемом в эксперименте, для каждого из заболеваний. Предварительная обработка серий экспериментов RNA-Seq, подобная описанной в разделе 3.5, была проведена, чтобы получить возможность использовать в основной серии расчетов обобщенные профили экспрессии.

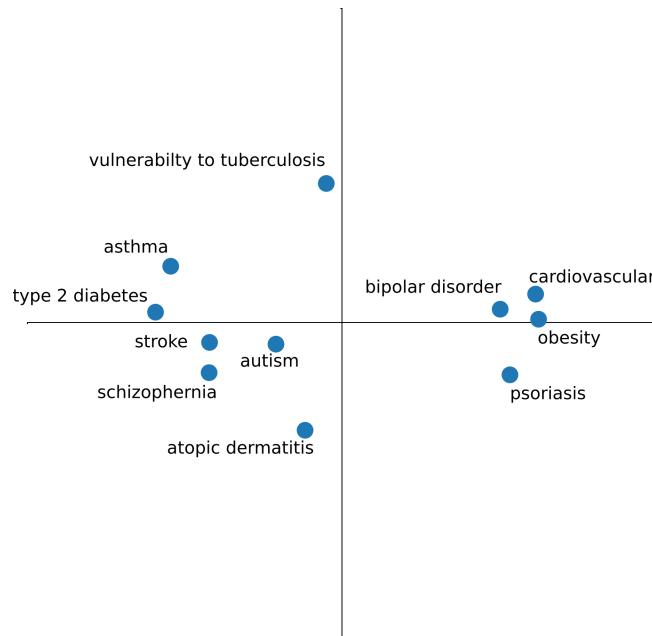


Рис. 3.30: Отношения между хроническими заболеваниями, рассчитанные по данным экспериментов RNA-Seq

Полученные при таком расчете соотношения между заболеваниями, показанные на рис. 3.30, не соответствуют соотношениям, полученным первым из подходов и показанные выше, на рис. 3.28. Но эти соотношения возможно согласовать, рассмотрев третью из главных компонент в представлении результатов РСА, оцененному при наиболее прямой обработке публикаций. При этом направление первой главной компоненты указывает, в основном, на систематические ошибки, вызванные смещениями акцентов в выборе тем исследования в обработанных публикациях. Полученное согласованное представление показано, с нескольких сторон, в заключительной части книги.

3.8. Математические модели в биологии

Понятия из теории дифференциальных уравнений

Динамическая система

Система обыкновенных дифференциальных уравнений вида $\frac{dx}{dt} = F(x)$ может представлять модель некоторого объекта, для которого состояние описывается набором параметром $x = (x_i, i = 1..N)$. Решение такой системы будет представлять эволюцию состояния объекта. В пространстве N измерений, решение системы можно представить как траекторию изменения состояния объекта; такое пространство называется *фазовым пространством*.

Описанная таким образом модель объекта может обозначаться термином *динамическая система*. Возможно также рассмотреть динамические системы с дискретным шагом по времени, в отличии от представления с помощью дифференциальных уравнений, где ось времени является непрерывной.

Устойчивость решения

Интуитивно, поведение динамической системы является устойчивым, если внесение малых отклонений в систему не приводят с течением времени к нарастающим отклонениям в траектории системы. Для системы дифференциальных уравнений, существуют более формальные определения устойчивости, в первую очередь - понятие "устойчивости по Ляпунову".

Особые точки уравнений

Систему из двух обыкновенных дифференциальных уравнений, без явной зависимости от времени, можно записать в виде $\frac{dx}{dt} = F_x(x, y), \frac{dy}{dt} = F_y(x, y)$. Решением этой системы будет пара функций $x_s(t), y_s(t)$, которую можно представить как кривую линию на плоскости, и можно интерпретировать как траекторию изменения системы. Скорость изменения системы при этом будет определяться значениями функций F_x, F_y . Особыми точками в таком двумерном пространстве будут точки, в которых система неподвижна, то есть $F_x = F_y = 0$. В окрестности особых точек значения функций малы, и можно ограничиться первым членом ряда Тейлора в разложении функций по обоим координатам. Упрощенные с помощью такого приближения уравнения несложно исследовать аналитически, и в результате возможно классифицировать возможные варианты поведения системы, разделив особые точки на три типа, как показано на рис. 3.31. В линейном приближении правая часть уравнений представляется как матрица 2x2, и тип особой точки определяется знаками собственных значений этой матрицы. Для вычисления собственных значений необходимо решить квадратное уравнение, и в некоторых случаях при вычислении решений необходимо использовать комплексные числа, поскольку значения решений будут выражаться через квадратный корень из отрицательного числа.

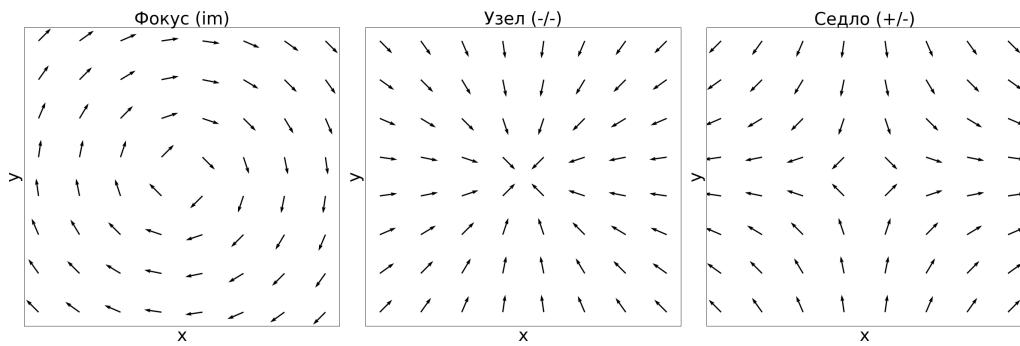


Рис. 3.31: Типы особых точек

*Фокус (слева) - собственные значения содержат минимую часть; узел (в центре) - собственные значения одного знака; седло (справа) - собственные значения разных знаков.
рисунок построен с помощью пакета matplotlib*

Для многих сложных многомерных дифференциальных уравнений, поиск положения и определение типа особых точек оказывается эффективным при аналитическом исследовании свойств уравнений.

Предельный цикл - обобщение понятия особой точки, для описания решения уравнений, в которых траектория системы по мере эволюции системы приближается к циклическому движению, или в которых для любых начальных условиях система будет двигаться по циклической траектории. Обобщением понятия предельного цикла является понятие *аттрактор*, когда устойчивой является траектория, не сводящаяся к циклической.

Бифуркация - изменение характерных свойств динамической системы, при малом изменении параметров системы. Так, например, в системе $\frac{dx}{dt} = rx - x^3$, при $r < 0$ устойчивым является решение $x = 0$, а при $r > 0$ - два решения $x = \pm\sqrt{r}$. В более широком смысле термин *бифуркация* используют, например, для описания развоения русла реки, и т.п.

Детерминированный хаос (динамический хаос) - поведение решения динамических систем, при котором внесение минимальных отклонений приводит к непредсказуемо большим последствиям при эволюции системы. В таком определении, понятие динамического хаоса обратно понятию устойчивости.

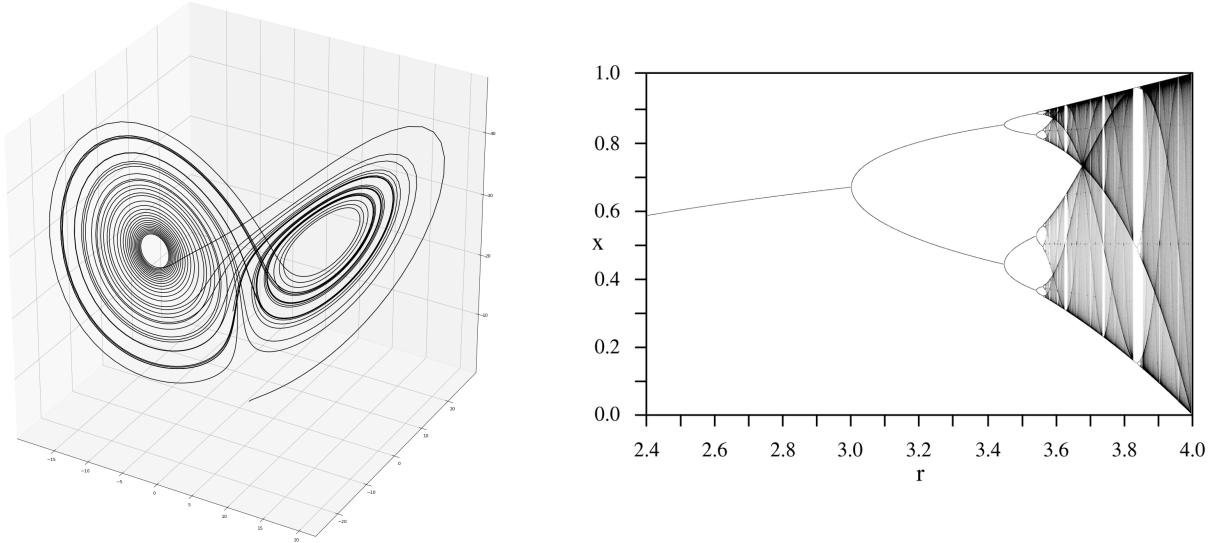


Рис. 3.32: Примеры систем с "детерминированным хаосом"

Слева: Аттрактор Лоренца, решение системы дифференциальных уравнений, известной как "система Лоренца".

Справа: Карта бифуркаций в дискретной системе $x_{n+1} = rx_n(1 - x_n)$, известной как "логистическое отображение". Горизонтальная ось: величина параметра r . Вертикальная ось: точки, составляющие предельный цикл системы.

рисунки по материалам проекта Wikipedia; credits: user PAR

Обзор и частные случаи прикладных моделей

Модель экспоненциального роста

С помощью простейшего дифференциального уравнения $\frac{dx}{dt} = ax$ оказывается возможным описать, на некотором этапе, эволюцию многих систем разных масштабов, в том числе в приложениях биологии. Так, например, численность микроорганизмов при условии неограниченных ресурсов будет удваиваться с каждым циклом деления клеток. Решением записанного уравнения будет экспоненциальная зависимость от времени ($x(t) = \exp at$).

Модель Лотка-Вольтерра ("хищник-жертва")

Система из двух уравнений $\frac{dx}{dt} = (\alpha - \beta y)x$; $\frac{dy}{dt} = (\delta x - \gamma)y$ позволяет качественно описать отношения между биологическими видами ("хищником" и "жертвой"), при этом решением системы являются периодическая зависимость от времени, для численности популяций обоих видов. В фазовом пространстве, решения уравнения Лотки-Вольтерра представляются как циклические кривые, и особая точка этого уравнения может быть классифицирована как фокус (рис. 3.31)

Обобщенная модель сосуществования двух видов

Для описания возможных отношений между двумя видами в математической экологии используются дифференциальные уравнения с двумя переменными, обобщающие модель Лотки-Вольтерра. При этом, классификацию отношений между видами можно свести к классификации особых точек дифференциальных уравнений. При этом используют следующие термины: "+/+"

(узел) - симбиоз или мутуализм; "+-" (фокус) = хищник - жертва или паразит - хозяин; "-/" (седло) - конкуренция.

Модель инфекционного заболевания

В иммунном ответе организма при инфекционном заболевании участвуют многие типы клеток, и потому модели, описывающие развитие инфекционного заболевания в организме, не сводятся к простейшей модели отношений между паразитом и хозяином. Для некоторых вариантов расширенных моделей, в которых учитывается несколько стадий иммунного ответа, возможно получить решения, соответствующие оструму и хроническому течению заболевания. С использованием таких моделей возможно качественно исследовать эффект так называемого "лечения через обострение", для перевода хронического заболевания в острое, с последующим полным выздоровлением. Однако степень сложности в организации иммунной системы, как правило, не позволяет довести соответствие моделей до уровня количественного описания заболевания.

* * *

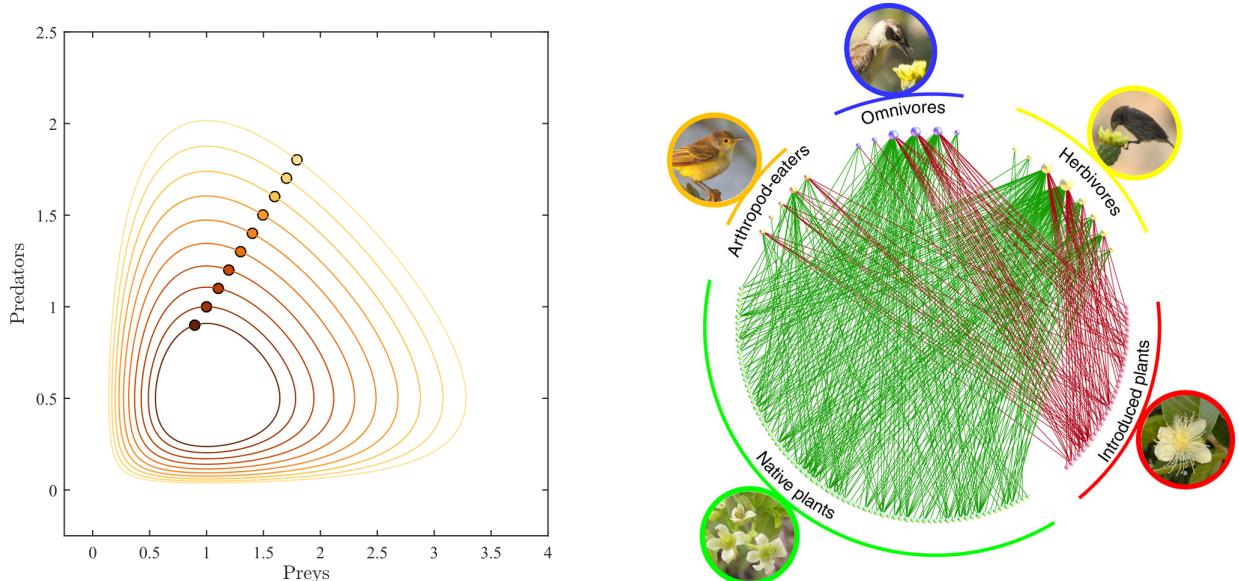


Рис. 3.33: Иллюстрации к некоторым темам математической экологии

Слева: Фазовая диаграмма решений уравнения Лотка-Вольтерра. Горизонтальная ось - условное количество "жертв"; вертикальная ось - условное количество "хищников". Траектории с разными начальными условиями показаны разным цветом.

Справа: Сводная схема связей между видами цветковых растений и видами птиц, которые их опыляют, по материалам полевых исследований на островах Галапагосского архипелага.

Рисунок слева: по материалам проекта Wikipedia; credits: user Brice2000. Рисунок справа: работа (Traveset и др. 2015).

Колебательные явления, которые возможно описать лишь с помощью системы нелинейных уравнениями, можно наблюдать даже в неорганической химии (так называемая "реакция Белоусова-Жаботинского"). И потому уместно применение методов из теории динамических систем для

описания биохимических процессов в клетке. Но описание некоторых явлений в многоклеточных организмах возможно только с использованием моделей динамических систем. Так, например, с использованием понятия бифуркации (так называемой "бифуркации Тьюринга") можно качественно описать появление полос на кожу у зебры, пятен у леопарда, и пр. Понятие бифуркации, в более широкой перспективе, описывает и само явление морфогенеза, когда соседние и однотипные клетки при делении дифференцируются в ткани различных органов.

Помимо упомянутой модели ответа иммунной системы на инфекционное заболевание, составлены модели для описания образования единого ритма в синусном узле сердца, в результате согласования активности клеток, называемых пэйсмэйкерами (pacemaker), а также модели распространения электрического импульса по тканям сердца. С помощью модели тока ионов натрия и калия в нейроне возможно описать распространение так называемого "потенциала действия" по протяженному аксону. Сложность устройства нервной системы не позволяет с достаточной адекватностью описать принципы, лежащие в основе мышления, однако исследования различных моделей сетей нейронов интенсивно развиваются, как будет подробно описано ниже.

Наконец, вся теория динамических систем, по своему происхождению и по степени применимости, тесно связана с описанием структуры и динамики экосистем. Для модельного описания отношений между видами в экосистеме, используют понятие *трофические сети*. При построении такой сети, связь между двумя биологическими видами устанавливается в терминах классификации отношений между видами, описанной выше. Также, на основании свойств графа трофических отношений, среди биологических видов выделяют так называемых "дженералистов" (*generalist*) - условно, "всеядные" виды, и "специалистов" (*specialist*), виды, адаптированные к определенному типу питания.

Динамика численности популяций при таком описании может быть в некоторых случаях сведена к отношениям между парами взаимодействующих видов, как это описано выше. В других случаях, в решениях уравнений проявляется "динамический хаос", как свидетельство ограниченности попыток описания экосистем с использование подобного класса методов. Также, существование эффекта, для обозначения которого используется понятие *виdeoобразование*, неявно ограничивает применимость описанных методов.

Нейробиология и модели сетей нейронов

Исследования, проводимые в области нейробиологии, науке изучающей работу нервной системы, сопоставимы по объему и интенсивности с исследованиями во всей молекулярной биологии, включая методы и подходы биоинформатики. И потому обсуждение методов и результатов, полученных в нейробиологии, выходит за рамки настоящего курса. Но модели сетей нейронов, построенные на основании сведений из нейробиологии, являются нетривиальными примерами моделей биологических систем. Обзор этих моделей, а также смежных понятий нейробиологии, приведен ниже. При построении моделей нейронных сетей вводится последовательный ряд упрощений и приближений; не все эти упрощения и не все исключения из правил оговариваются явно.

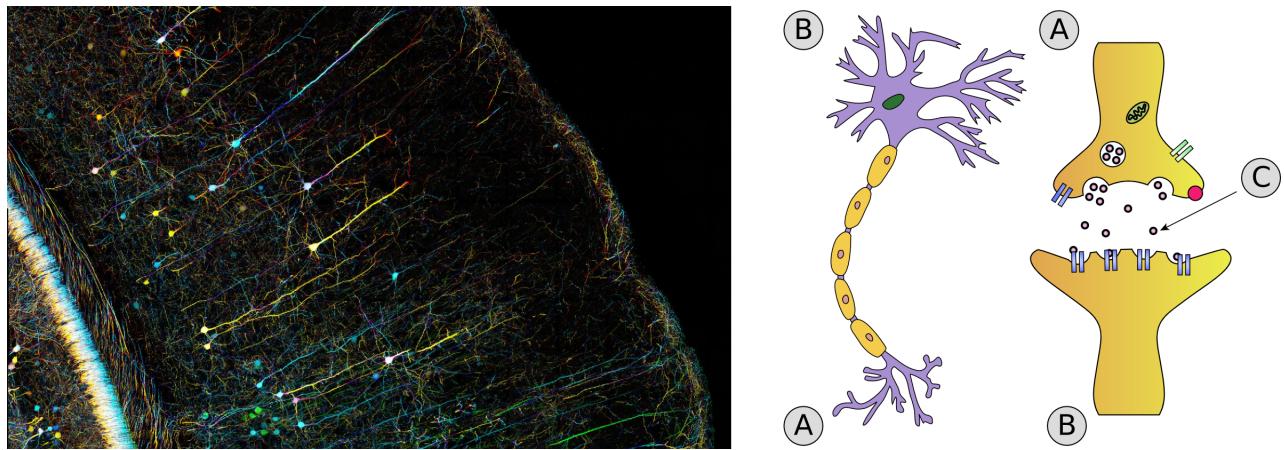


Рис. 3.34: Принципы устройства сети нейронов

Слева: кора головного мозга мыши, конфокальная лазерная сканирующая микроскопия.

Справа: схема устройства нейрона и синаптической щели. А - аксон, В - дендрит, С - нейромедиатор.

Credits: ZEISS Microscopy; Mouagip, Dake, Quasar Jarosz (Wikipedia).

Крупные и разветвленные клетки головного мозга, называемые *нейронами*, являются основой для происходящих в организме процессов хранения и переработки информации. Особенностью нейронов является возможность находиться в состоянии "возбуждения", что выражается в изменении электрического потенциала цитоплазмы по отношению к внешней среде. Разность потенциалов при этом поддерживается за счет переноса ионов натрия, калия и кальция через каналы в мембране нейрона. Состояние возбуждения может передаваться другим нейронам, через соединения между разветвленными частями двух нейронов, называемые *синапсами*.

Накопление возбуждения, передаваемого соседними нейронами через ответвления клетки, называемые *дендритами*, приводит к переходу в возбужденное состояние, и затем импульс возбуждения передается по ответвлению, называемому *аксоном*, следующим нейронам. При этом синапсы, через которые передается возбуждение, могут различаться по типам нейромедиатора и по интенсивности передаваемого сигнала. Для некоторых нейромедиаторов, сигнал, передаваемый через синапс, может приводить к торможению клетки, принимающей сигнал, а не к увеличению ее возбуждения.

Для описания разных категорий и свойств нейронных сетей предложено достаточно много подходов. Некоторые из моделей описывают развитие нейронной сети, выражющееся в изменении коэффициентов передачи сигнала в синапсах (так называемых *весов связей*). Другие модели описывают изменение интенсивности возбуждения клеток, при допущении, что веса связей не меняются при моделировании.

По топологии нейронной сети можно разделить модели, в которых граф сети не содержит циклов, и модели, где в графе допускаются циклы (рис. 3.35). В большей части моделей, для расчета возбуждения нейрона сигналы из входных нейронов суммируются с учетом весов связей. Но вид функции, связывающей итоговый входной сигнал и уровень возбуждения нейрона, зависит от выбора модели.

Таблица 3.3: Подходы к моделированию нейронных сетей

Наименование	наличие циклов	Комментарий
Обучение Хебба	-	<i>принцип ассоциативной памяти</i>
"Перцептрон"	нет	<i>модель распознавания образов при обучении "по образцу"</i>
Обратное распространение ошибки	нет	<i>алгоритм коррекции весов связей для сетей без циклов</i>
Сеть Кохонена	-	<i>используется для задач разделения на классы, без задания образцов</i>
Сеть Хопфилда	есть	<i>модель сети с циклами, как динамическая система</i>
"Победитель забирает все"	есть	<i>частный случай сети Хопфилда, каждый нейрон связан с остальными тормозящими связями.</i>
Ассоциативная сеть	есть	<i>модель сети с циклами, как ассоциативная память</i>
Глубинное обучение	нет	<i>уточненные методы настройки сети без циклов</i>

Модели сетей без циклов (рис. 3.35, слева) были предложены одними из первых и разработаны в наибольшей степени. Эти модели нашли приложения в решении задач классификации, и в уточненном виде, в технологиях, которые обозначаются термином *deep learning* ("глубинное обучение"). Настройка топологии, весов связей и формы преобразования входного сигнала в таких сетях производится на так называемой *стадии обучения*, за которой следует стадия использования сети для решения задач классификации.

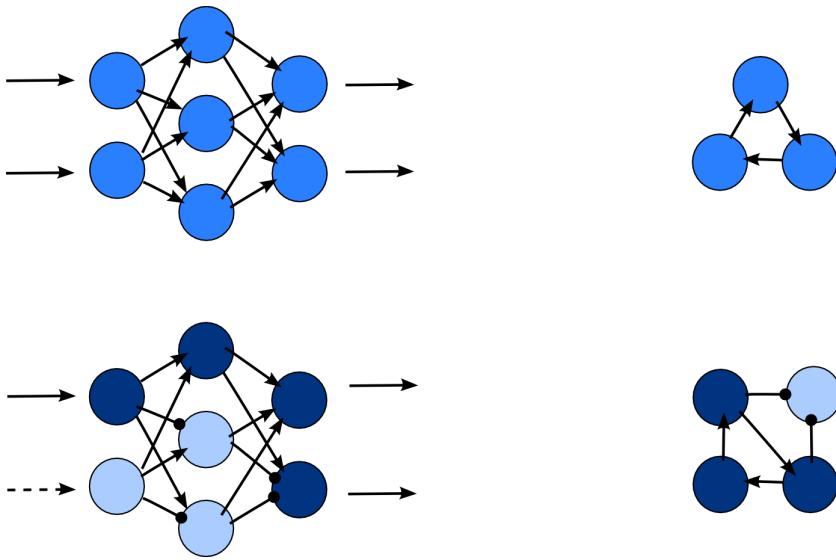


Рис. 3.35: Принципы построения моделей сетей нейронов

Слева - сети без циклов; справа - сети с циклами.

Вверху - структура сети, внизу - возбуждение в сети. Возбужденные нейроны помечены темно-синим. Тормозные синапсы обозначены кружками.

При моделировании сетей с циклами (рис. 3.35, справа), решением является устойчивое состояние сети, один из аттракторов, к которому сходится фазовое состояние системы, в зависимости от уровней возбуждения клеток в начале моделирования. Возможно использование такой сети ("сети Хопфилда") для решения некоторых комбинаторно сложных задач. Веса связей при таком подходе определяются условиями задачи, а состояние сети, полученное при моделировании, следует преобразовать в оптимальный выбор варианта при переборе.



Рис. 3.36: Пример восстановления повреждённого изображения с использованием ассоциативной сети

Слева: искаженный образ; справа: эталон.

По материалам проекта Wikipedia.

Более практичная и простая для восприятия модель приложения сетей с циклами - это возможность выбора одного из возможных вариантов состояния, заложенных на стадии обучения. Так, например, сопоставив клетки сети с точками изображения, можно использовать модель для восстановления искаженного изображения (рис. 3.36). Сеть может быть обучена для выбора одного из нескольких возможных изображений. Для выбора весов связей на стадии обучения возможно использовать так называемое правило Хебба (Hebb's rule). В общем случае, веса связей устанавливаются на основании ассоциаций между клетками в изображениях, используемых

для обучения. Ограничения применения сетей с циклами в такой постановке связаны со слишком малым (порядка $N / \log(N)$, где N - размер сети) количеством шаблонов, среди которых возможно выбрать решение.

Модель сети нейронов с двумя типами возбуждения

Простейшей моделью, состоящей из двух нейронов, является модель "winner takes all" ("победитель забирает все") (рис. 3.37, слева), частный случай сети Хопфилда. Эта модель, где два нейрона соединены между собой подавляющими связями, иллюстрирует, как мозг принимает решения, не попадая в ситуацию "Буриданова осла". Система имеет два устойчивых состояния, выбор между которыми может происходить до какой-то степени случайно, когда в начале оба нейрона не возбуждены. Когда один из нейронов становится возбужденным, в рамках модели система продолжает оставаться в устойчивом состоянии; вопрос о возврате в невозбужденное состояние можно поставить только в более широком контексте.

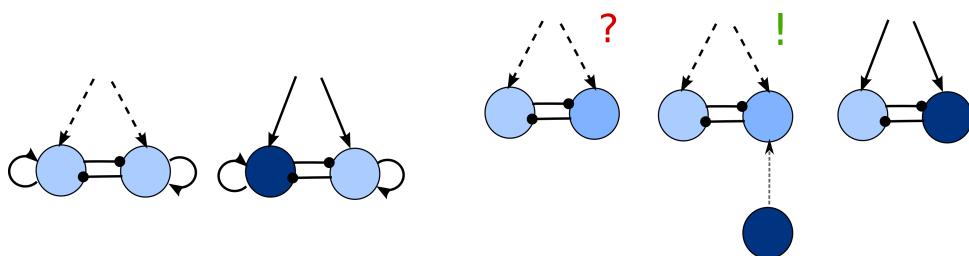


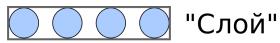
Рис. 3.37: Модель «Winner takes all»

Возбужденные нейроны помечены темно-синим. Связи показаны стрелками; неактивные связи показаны пунктиром. Справа показано возможное расширение модели за счет учета слабых уровней возбуждения (светло-синий).

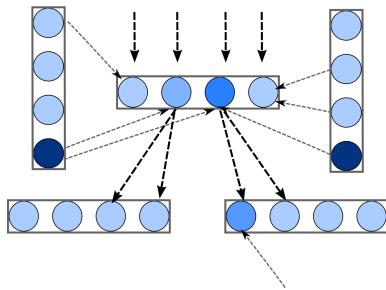
Более внимательный анализ существующих в мозге синапсов позволяет уточнить, что в большинстве синапсов медиатором служат глутамат и аспартат. Но в некоторых синапсах в ЦНС используется для передачи возбуждения медиатор ацетилхолин, основной медиатор в периферической нервной системе. При этом механизмы, через которые в ЦНС возбуждение передается через глутаматовые и ацетилхолиновые синапсы, принципиально отличаются тем, что:

- возбуждение клетки через рецепторы ацетилхолина приводит к высвобождению ионов кальция и интенсивному поглощению энергии возбужденной клеткой,
- возбуждение клетки через рецепторы глутамата регулирует в основном обмен ионов натрия и калия и не приводит к интенсивному поглощению энергии.

В предположении о двух типах возбуждающих синапсов, в модель "winner takes all" можно внести дополнение (рис. 3.37, справа). Если предположить, что возбуждение в этой модели передается ацетилхолиновыми связями, то предварительное малое возбуждение клетки через дополнительные рецепторы глутамата может обусловить выбор устойчивого состояния при появлении необходимости сделать такой выбор.



Шаг 1



Шаг 2

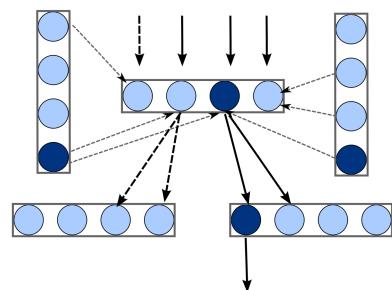


Рис. 3.38: Модель многослойной сети с двумя направлениями передачи информации
 Слой примерно соответствует ассоциативной сети. Шаги 1, 2 показывают этапы передачи информации. Возбужденные нейроны помечены темно-синим. Слабо возбужденные нейроны показаны светло-синим. Два типа связей показаны стрелками двух типов.

Эту модель можно обобщить и предположить, что в мозге существует два направления распространения информации (рис. 3.38). Первое направление служит для регулировки и подстройки и передается через рецепторы глутамата; тогда как, информация второго рода приводит в итоге к возбуждению мышц и передается через рецепторы ацетилхолина. Разделение сети на слои, схематично показанное на рис. 3.38, совместимо с предположением о двух направлениях передачи информации. Связи внутри каждого слоя при этом предполагаются тормозящими, как и между парой нейронов в модели "winner takes all". Допущение об отделении этого, третьего, типа связей, сигнал от которых приводит к безусловному подавлению активности клетки, также согласуется с данными нейробиологии. Медиатором в большей части тормозных синапсах является гамма-аминомасляная кислота (ГАМК), и при торможении через рецепторы ГАМК наблюдают эффект "шунтирования" (*shunting inhibition*), безусловного подавления остальных поступающих в клетку сигналов.

Разделение сети нейронов на слои в модели сети, хотя это и условное схематичное представление нервной системы, соответствует возможности разделения всего объема знаний об окружающем мире, сохраняющемся в ЦНС, на относительно независимые фрагменты. Язык человека отражает способ его мышления, и первичными фрагментами при передачи знания в языке являются слова, примерно соответствующие четко отделяемым объектам и устойчивым понятиям окружающего мира, так что значение каждого слова понятно как говорящему, так и слушающему. Каждый из объектов может быть устроен сколь угодно сложно, но, представив возможные и практически значимые состояния объекта как перечисляемое множество, модель такого объекта возможно составить из ограниченного числа нейронов, объединенных в общий слой через тормозящие связи.

В простейшем примере слоя, сети "winner takes all", количество нейронов определяет возможные состояния, а активность одного из нейронов в слое - выбор из возможных состояний.

В сети "winner takes all" все нейроны связаны между собой тормозящими связями, но, убрав из подобной сети некоторые из связей, количество возможных состояний следует скорректировать, и в некоторых из состояний могут быть активны сразу несколько нейронов. При "подстройке" ("обучении") сети, переход типа связи от тормозящей к нейтральной, и обратно, позволяет "погнать" систему связей в слое к объекту, который моделирует этот слой. Слой в описываемой сети следует рассматривать в контексте взаимодействия с другими слоями, и описанный процесс обучения, открывая возможность быть активными сразу нескольким нейронам в слое, позволяет настроить, среди свойств объекта, варианты его взаимодействия с другими объектами. С этим способом настройки модельной сети можно связать описанный в нейробиологии эффект переключения направления регуляции в рецепторах ГАМК при развитии мозга.

Разделение сети на слои в предложенной модели позволяет прояснить, как мозг способен принимать решения настолько быстро, при огромном объеме сохраняемых там знаний. А именно, в устройстве модели можно заметить сходство с моделью сворачивания белка (раздел 2.5), где черты принципа "разделяй и властвуй" совмещены с разделением задачи на согласованные части. В алгоритмах, относящихся к "динамическому программированию", где также проводятся расчеты по подобной схеме, с необходимостью вводятся два этапа, подобные двум направлениям передачи информации в описанной модели.

При моделировании некоторой системы по законам физики, возможность рассчитать, хотя бы приближенно, поведение системы, связана со степенью независимости частей, составляющих эту систему. Подобно этому, при настройке разделения сети на слои, в описываемой модели, переключение тормозящей связи до уровня нейтральной может отражать факт независимости отдельных объектов, и существенно повышать эффективность работы сети.

Парадоксальное отличие человека, и его поведения, от компьютерной системы, в том что, в отличии от человека, алгоритм для компьютера составляет разработчик, находящийся вне этой системы. Для животных, подобных человеку по анатомии и по устройству нервной системы, можно также примерно указать цели, определяющие поведение и необходимые для выживания особи и вида в целом. Поток информации, поступающий в мозг от органов чувств, подобен входным данным в компьютерной программе. Но цели поведения, как, например, потребность в пище, и даже потребность в пище определенного рода, можно рассматривать как поток информации другого рода, чем сигналы, посылаемые органами чувств.

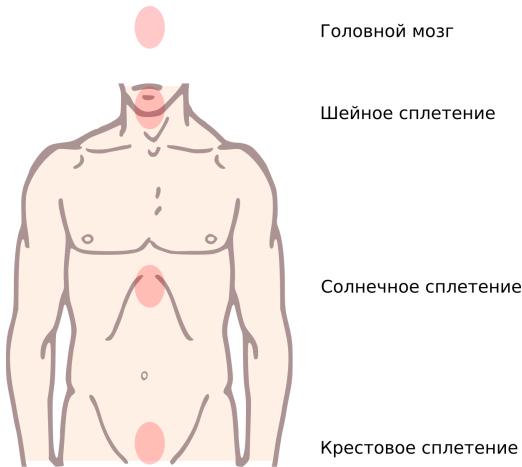


Рис. 3.39: Головной мозг и некоторые из "сплетений" в периферической нервной системе

credits: NASA

Модель, предполагающая два направления распространения информации, позволяет увязать меняющиеся и определяемые вне мозга цели поведения со "слабыми" сигналами, передающими через рецепторы глутамата. Этот род информации определят выбор из возможных альтернатив, когда от органов чувств приходит "сильный" сигнал, требующий немедленной реакции и распространяющийся по второму направлению, вплоть до возбуждения мышц.

Как все хорошо знают, поведение человека в полной мере не определяется его потребностями в насыщении и продолжении рода. Но, так же как цели поведения животного определяются вне мозга, цели и желания человека, любого рода, не следует искать внутри мозга. И потому, несмотря на то, что современные экспериментальные методы позволяют примерно локализовать в мозге зоны, относящиеся к владению языком или к восприятию языка искусства, категории информации такого рода не относятся непосредственно к целям поведения человека. Информация об их локализации и динамическом изменении не имеет смысла, хотя бы даже при намерениях прогнозировать и управлять поведением человека.

Прицельное воздействие на мозг, при примерном знании его устройства, может нанести травму психике. Эта травма может выражаться введением в состояние гипноза, но вопрос свободы воли человека при этом все же не имеет отношения к вопросу о его психическом здоровье. Регуляция дыхания и сердцебиения, необходимость в воздухе и циркуляции крови - наиболее глубокие и неотъемлемые из потребностей живого существа. И материя, в которой воплощена свобода воли человека, может быть увязана только с механизмами регуляции этих глубинных процессов. Так или иначе, но подлинные мотивы, руководящие человеческим разумом, следует искать за пределами человеческого разума.

Элементы теории фракталов

Термин *фрактал* был введен в 1960-1970 годы в серии работ Б. Мандельброта; в его статье «Какова длина побережья Британии» он применил понятие дробной размерности для описания так называемого «парадокса береговой линии», замеченного ранее Л.Ф. Ричардсоном. Парадокс береговой линии, который демонстрирует затруднения при измерении длины береговой линии, проиллюстрирован на рис. 3.40. Определения размерности Хаусдорфа или размерности Минковского, введенные в математику в начале XX века, позволяют рассчитать дробную (нецелую) размерность для некоторых специфических объектов геометрии. И подобные методы могут быть применены к описанию объектов, возникающих как явления природы. Цитируя Б. Мандельброта: «Облака - не сферы, горы - не конусы, береговые линии - не окружности, древесная кора не гладкая, и молния - далеко не прямая...»



Рис. 3.40: **Побережье Британии, измеренное в нескольких масштабах**

слева: единица длины - 200 км, длина побережья - около 2400 км; справа: единица длины - 50 км, длина - 3400 км

Credits: Avsa, Wapcaplet, Acadac (Wikipedia)

Объекты, которые можно описать как фракталы, возникают при различных обстоятельствах и их появление нельзя свести к наличию единого механизма развития. Свойства фрактальной структуры могут быть обнаружены в любом объекте, выделенном в окружающем мире, если в распределении его геометрических свойств, построенном в логарифмических координатах, обнаруживается линейная зависимость, как в примере, показанном на рис. 3.41.

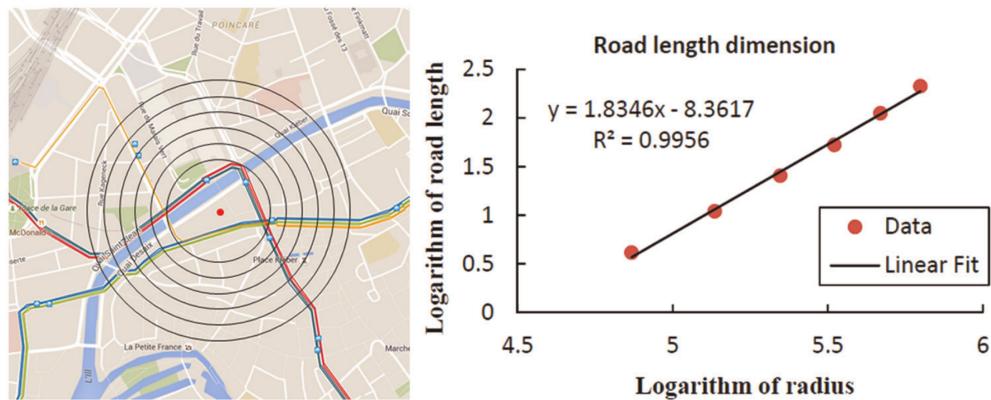


Рис. 3.41: **Фрактальной структура дорожной сети в Страсбурге**
по материалам (H. Wang и др. 2017)

В линейных координатах, эта зависимость будет выражаться так называемым *степенным законом* (power law), $y = A^D$; значение D , с отрицательным знаком, будет значением фрактальной размерности. Фрактальное представление сложных геометрических объектов позволяет охарактеризовать структуры с помощью одного параметра D , значения фрактальной размерности. Точное значение фрактальной размерности может быть разным для одного и того же объекта, в зависимости от методов, используемых для расчета размерности. Но, в рамках любого метода оценки, величина D может иметь точность, достаточную для использования этого значения при сравнении нескольких сопоставимых объектов, обеспечивая статистическую значимость выводов, полученных на основе такого сравнения.

Зависимость в форме степенного закона может быть замечена не только в свойствах геометрических объектов, но и в распределениях, относящимся к другим категориям и предметным областям. Распределение вероятностей, известное как *закон Ципфа*, описывающее частоту встречаемости слов в текстах на естественных языках, также является частным случаем степенного закона. В законе Ципфа, частота, с которой слово встречается в тексте, обратно пропорциональна его порядковому номеру в списке слов, отсортированном по количеству случаев использования каждого слова. Другим примером эмпирически обнаруженного степенного закона, относящимся к экономике, является *распределение Парето*. Это распределение было получено как обобщение наблюдения Вилфредо Парето о том, что 20% населения Италии владеет 80% ее земель. В общем случае, распределение Парето используют в экономике и социологии как модель распределения доходов или финансовых активов любого рода.

В некоторых случаях, когда степенного закона недостаточно, чтобы объяснить распределение полученные путем применения фрактальных подходов к изучаемой системе, используется так называемое *мультифрактальное* представление, где величина фрактальной размерности обобщается как непрерывный ряд показателей степени в степенном законе. Более простое по форме обобщение степенного закона используется в так называемой *эконофизике* (econophysics), для описания распределения доходов и активов в социологии.

Сам Б. Мандельброт указывал на принципиальное различие между вероятностными распределениями, описываемыми степенным законом, и вероятностными распределениями на основе

распределения Больцмана, используемыми в статистической физике. Методы статистической физики, на его взгляд, приводят к значительно более предсказуемым результатам, по сравнению с фрактальными распределениями. Однако в эконофизике, с использованием нескольких достаточно простых моделей перераспределения доходов, было показано, что оба типа распределений могут быть совмещены в рамках единого подхода (рис. 3.42).

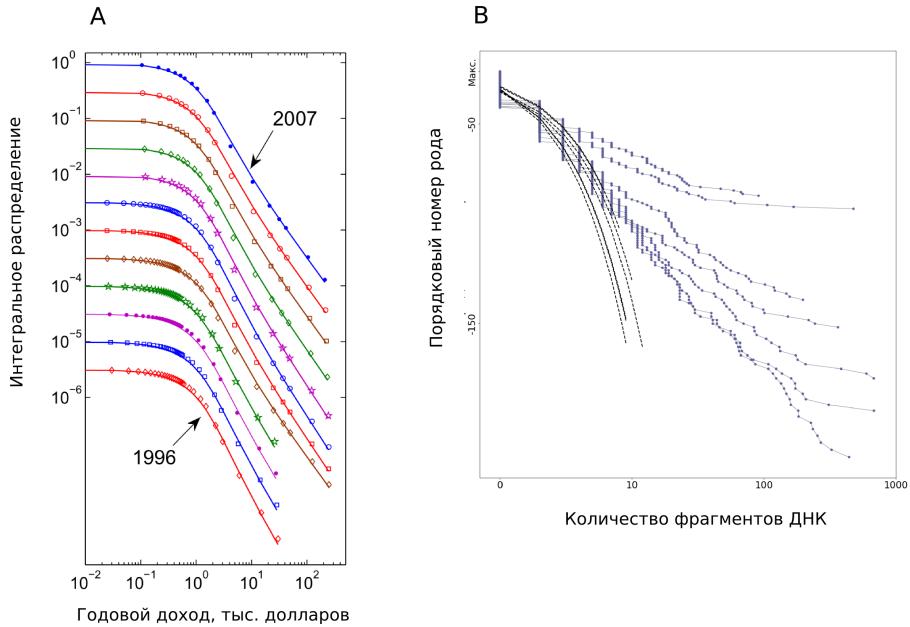


Рис. 3.42: Некоторые примеры описания данных с использованием совместно экспоненциального и степенного распределений вероятностей

Слева: распределение вероятностей описывающее доходы населения США, по результатам обработки налоговых деклараций. Теоретическое распределение показано линиями, данные федеральной резервной системы США показаны символами. Функции распределения показаны в логарифмическом масштабе, после перенормирования. Данные за разные годы сдвинуты по вертикали.

Справа: распределения относительной численности филотипов бактерий в микробиомах донных осадков, показанные в логарифмическом масштабе.
использованы материалы ([Banerjee и Yakovenko 2010](#)) и работы авторов курса.

В кратком пересказе, в моделях, используемых для совмещения двух подходов в социологии, распределение Парето использовано для описания доходов в высших слоях общества, когда доход пропорционален активам, как, например, через получение дивидентов. Распределение Больцмана выполняется, когда доход пропорционален потраченным на экономическую активность усилиям и времени. В первой категории, но не во второй, логарифмическое преобразование следует применять к размерам дохода. Полученная совмещенная модель позволяет адекватно описать данные из социологии и экономики (рис. 3.42 А).

Еще одно расширение теории фракталов касается моделей временных рядов, в которых, помимо фрактальной структуры, наблюдается периодичность, в которой период последовательно увеличивается или сокращается в геометрической прогрессии (рис. 3.43). Такое свойство

временных рядов можно, с долей условности, связать с наличием мнимой части у величины фрактальной размерности, как комплексного числа. Короткий промежуток времени, где период становится равным нулю, соответствует моменту кризиса, когда поведение системы нельзя достоверно описать на сколь угодно коротком промежутке времени.

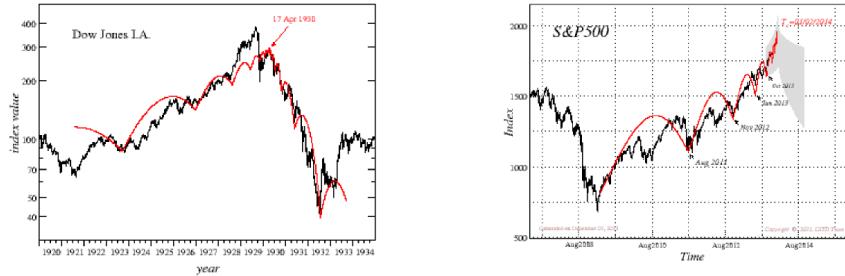


Рис. 3.43: Периодичность при анализе финансовых рынков

Слева: моделирование кризиса биржевых индексов во время Великой депрессии.

Справа: прогнозирование кризиса на основании анализа биржевых индексов в период с 2005 по 2014 гг.

использованы материалы с сайта института ядерной физики академии наук Польши (отдел теории сложных систем).

Ограничения теории фракталов следуют из того факта, что методы построенные на основе этой теории почти никогда не основаны на детальной модели исследуемой системы, и любые прогнозы, основанные на этих методах, являются рискованными. Как один из примеров, прогноз финансового кризиса в 2014 г. не подтвердился (рис. 3.43 справа). И более того, модель распределения доходов, в которой предполагается разделение общества на два класса, несложно связать с известной экономической теорией XIX века, которая послужила поводом и основанием для социальных потрясений, произошедших в России.

Биоразнообразие и модели распределения численности в экологии

Объектом моделирования при описании экосистем является, в первую очередь, распределение численности видов, полученное, как выборка, при полевых исследованиях экологов (рис. 3.44). Для сравнительного анализа, в рамках научного метода, возникает необходимость свести подобное распределение численности к одному или нескольким параметрам, характеризующим степень биоразнообразия. Первый и наиболее простой из этих параметров - количество биологических видов в экосистеме, называемый также *видовым богатством*. Другие индексы, используемые как характеристики биоразнообразия, приведены в таблице 3.4.

Количество используемых в экологии индексов неявно демонстрирует, что нет единого простого способа свести распределение численности к одному числовому параметру. И, в классификации, предложенной Уиттекером, подобный анализ распределения численности следует считать измерением *альфа-разнообразия*. Другие меры разнообразия, которые с долей условности принято называть *бета-разнообразием*, относятся к сравнению структуры сообществ, где

такое сравнение уместно. Так, например, на рис. 3.44 показано сравнение распределений численности видов моллюсков семейства Semelidae на разных глубинах, и в разные годы.

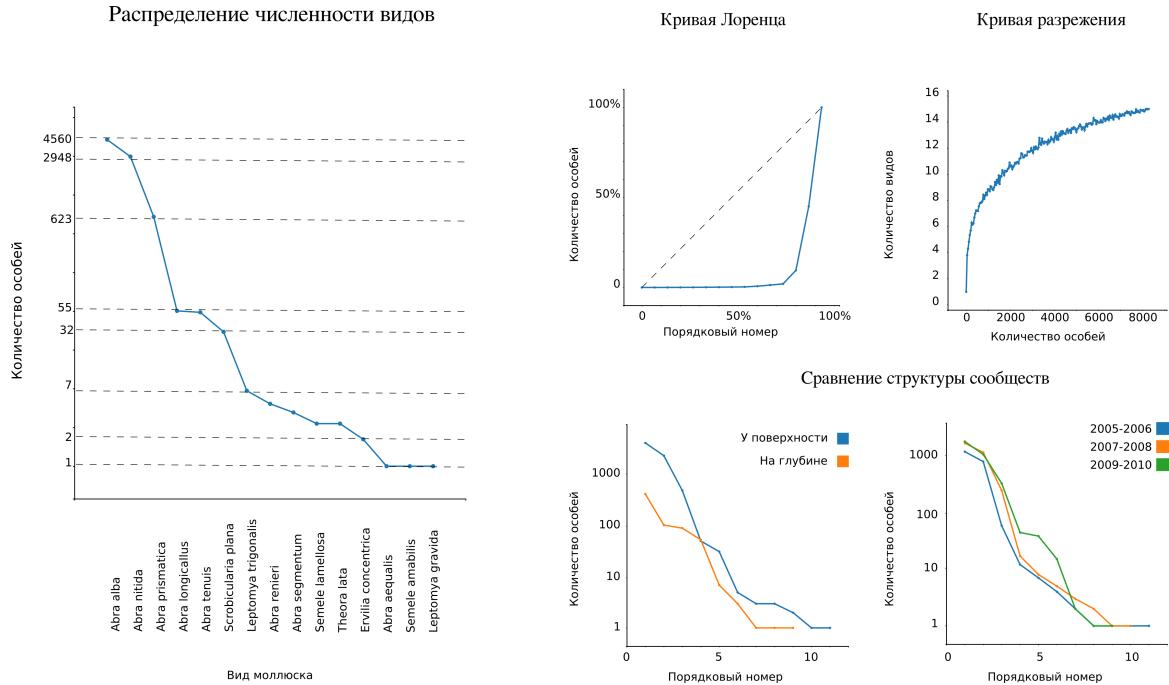


Рис. 3.44: Численность видов в экосистеме, в разных представлениях

Распределение численности моллюсков семейства Semelidae, на основании регистрации особей при подводных наблюдениях.

Слева: полная выборка, в представлении диаграммы численности видов ("представление Уиттекера").

Справа вверху: другие способы представления распределения численности: кривая Лоренца, для наглядного представления степени однородности сообщества, и "кривая разрежения", для оценки репрезентативности выборки.

Справа внизу: сравнение диаграмм численности сообществ, в разрезе глубины, где была зарегистрирована особь, и в разрезе времени регистрации. использованы базы данных проекта IOBIS ([Vandepitte и др. 2015](#)).

Диаграммы численности в различных представлениях, подобные показанным на рис. 3.44, обычно имеют вид (почти) непрерывных функций, привычных для математиков, знакомых с принципами моделирования. И создано много моделей, удовлетворительно описывающих распределения численности в некоторых частных случаях, и даже позволяющих сформулировать достоверные выводы применительно к задачам из практики экологов. Но никакая из моделей не является достаточно универсальной, чтобы рекомендовать эту модель для описания экосистем разного рода, с приемлемой степенью соответствия.

С учетом оговорок об "индивидуальности" каждой из изучаемых экосистем, наиболее эффективными из способов описания биоразнообразия и распределений численности являются

Таблица 3.4: Оценки индексов биоразнообразия для распределений, показанных на рис. 3.44

Индекс	полная выборка	у поверхности	на глубине	Комментарий
Количество видов	15	11	9	используется также термин "видовое богатство"
Индекс Chao1	16.5	11.5	12	асимптотическая оценка количества видов (*1)
Индекс Ace	17.03	12.2	14.4	асимптотическая оценка количества видов (*2)
Индекс Шеннона	1.45	1.35	1.65	мера Шеннона для количества информации
Индекс Симпсона	0.57	0.54	0.57	соответствует индексу Херфиндаля в экономике
Коэффициент Джини	0.91	0.90	0.82	в социологии, оценка степени расслоения общества
Коэффициент α Фишера	1.78	1.26	1.45	параметр модели Фишера $S_n = \alpha \frac{x^n}{n}$
Показатель степени D	3.57	4.06	3.12	параметр модели "степенного закона" $S_n = S_0 n^{-D}$

(*1,2) два подхода, используемые для учета недостаточной репрезентативности выборки

универсальные меры, используемые для оценки степени однородности распределений, относящимся к разным областям науки и разным объектам исследования. Так, например, наиболее популярный среди экологов индекс Шеннона выражает просто расчет количества информации, на основании распределения численности, используя формулу Шеннона, исходно выведенную при развитии кибернетики и теории кодирования. Мера информации, в интерпретации подходов кибернетики, совпадает с точностью до множителя с мерой энтропии, используемой в равновесной статистической физике. Коэффициент Джини, широко применяемый в социологии, также относится к эффективным мерам для оценки степени однородности в распределении численности видов. Но, впрочем, путь самого Корrado Джини, выступавшему за социальной равенство, с опорой на режим Муссолини, также может быть примером рискованности предсказаний, основанных на частных моделях, упрощающих интерпретацию данных.

Известное выражение "Сколько ангелов уместится на конце иглы" было использовано в свое время как пример бессмысленных споров средневековой схоластики, для оправдания веры в силу доводов разума. Но не стоит ли учитывать, что в масштабах, в которых проводятся исследования по экологии и эволюции, все же заняты своим делом слишком много ангелов, и легко ошибиться, сделав ставку лишь на одного из них.



Рис. 3.45: Виды северного Урала и другие
фотографии А. Чубинского

3.9. Молекулярная филогенетика и метагеномика

Анализ микробных сообществ

Возможность прочитывать последовательности нуклеотидов биоматериала любого происхождения связана с необходимостью интерпретации генетического материала, составленного из фрагментов геномов сразу многих организмов, как, например, сообщества микроорганизмов. В дополнении к задачам, которые можно поставить при изучении генетического материала из одного организма, как, например, асSEMBЛИРОВание генома или анализ дифференциальной экспрессии генов, в экспериментах по изучению сообщества возможные постановки задач включают:

- оценку биоразнообразия микроорганизмов в сообществе,
- определение видового состава микроорганизмов в биоматериале,
- оценку численности отдельных видов.

Среди понятий, используемых в этой области, термин *метагеном* обозначает совокупность ДНК организмов в сообществе, как, например, гены бактерий, присутствующих в биоматериале; термин *метатранскриптом* обозначает совокупность молекул РНК, как, например, матричных РНК, используемых для синтеза белков. Термины *микробиом* и *виром* обозначают, соответственно, совокупность бактерий и вирусов в биоматериале.

Секвенирование полного генетического материала сообщества - одна из возможных постановок экспериментов в микробиологии. В данных такого эксперимента содержится информация

о генах, присутствующих в совокупности микроорганизмов в сообществе, что позволяет интерпретировать эти данные в разрезе функциональных характеристик сообщества, как, например, особенностей биохимических процессов, протекающих в нем. Видовой состав и численность отдельных микроорганизмов также возможно оценить по данным секвенирования метагеномов, однако для оценки видового состава иногда используют другую технологию постановки экспериментов, когда перед проведением секвенирования проводят амплификацию фрагментов выбранных генов в биологическом материале, с помощью полимеразной цепной реакции (PCR).

Фрагмент генома бактерий, в котором кодируется последовательность рибосомной РНК, часто выбирается как мишень для амплификации, поскольку последовательность рибосомной РНК присутствует в геномах всех известных бактерий. Рибосомы, молекулярные комплексы в которых проходит трансляция белков, состоят из нескольких молекул РНК, имеющих устойчивую укладку за счет внутренних водородных связей, а также из вспомогательных белков-ферментов. Молекулы рибосомной РНК имеют относительно большой молекулярный вес, среди них принято выделять большую субъединицу (23S) и меньшую субъединицу - (16S). Подобно другим генам, обе субъединицы кодируются в бактериях как фрагменты геномной ДНК, и молекулы рибосомной РНК синтезируются в процессе транскрипции. И, подобно другим генам, в последовательностях генов рибосомной ДНК появляются мутации, и эти гены специфичны для каждого вида бактерий. Но, среди генов бактерий, гены рибосомной РНК (*rRNA*) являются наиболее консервативными, и сходными между собой у разных видов. И более того, в некоторых участках последовательности *rRNA* практически идентичны, что позволяет, с использованием специально подобранных праймеров, амплифицировать в процессе ПЦР фрагменты *rRNA* у большей части бактерий в биоматериале. И затем, при секвенировании ДНК после стадии амплификации, в прочтениях будут содержаться лишь фрагменты *rRNA*, относящиеся к разным бактериям из исходного биоматериала. В итоге, на основании специфических различий в прочитанных последовательностях *rRNA* между отдельными видами бактерий, возможно оценить видовой состав бактерий в исходном биоматериале и степень биоразнообразия в микробиоме.

Как правило, при анализе бактериальных сообществ используют ген малой субъединицы *rRNA*, 16S *rRNA*, но в подобных подходах со стадией амплификации консервативных генов, возможно использовать ген 23S *rRNA* бактерий, а также гены малой (18S) и большой (28S) субъединиц *rRNA* эукариот, для определения видового состава эукариотов в образце. Известны также более широкие и более специфичные методики постановки экспериментов с использованием амплификации, но их перечисление, также как и обсуждение возможных вариантов постановки экспериментов по секвенированию *rRNA*, выходит за рамки настоящего курса.

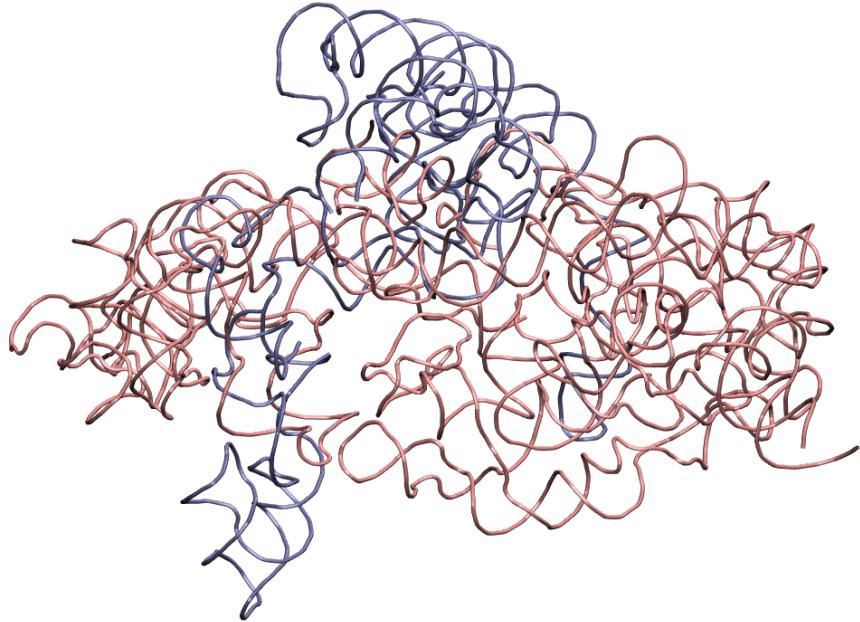


Рис. 3.46: Рибосомная РНК бактерий, малая субъединица

*Наличие вариабельных фрагментов V1-V7 в последовательности 16S rРНК бактерий определяют возможные наборы праймеров, используемых при амплификации. На рисунке, участок структуры, соответствующий вариабельным фрагментам V4 и V5, помечен голубым. по модели 1fka (бактерия *Thermus thermophilus*) с использованием пакета VMD*

Условные конвенции, принятые в подходах по обработке данных секвенирования сообществ микроорганизмов, в наибольшей степени детализированы в экспериментах с использованием амплификации генов. Ряд алгоритмов, приемов и пакетов программ разработан для использования при обработке данных, полученных при секвенировании метагеномов. Секвенирование метатранскриптомов, по отношению к первым двум группам подходов - наиболее "экспериментальная" область, хотя ряд пакетов программ может использоваться и в таких задачах.

В частности, в протоколах обработки данных при анализе микробиомов эффективным оказалось использование методов оценки биоразнообразия, разработанных в классической экологии. Но, использование подходов классической экологии в применении к анализу микробиомов, на основе данных секвенирования, привело к необходимости уточнения и расширения понятий, принятых в этих подходах. При исследовании различных сообществ микроорганизмов, количество возможных видов бактерий оказалось несопоставимо больше, чем это было принято в классической экологии, и даже определение биологического вида потребовало дополнительных оговорок. И потому в молекулярной микробиологии часто используют термин *OTE* (*операционная таксономическая единица*, OTU), как аналог понятия биологического вида. В отличии от

биологического вида, каждая ОТЕ остается определенной лишь в пределах отдельного исследования, и группировка ОТЕ зависит от метода расчетов, и проводится обычно на основании обработки серии экспериментов.

Филогенетические деревья

Белковые и нуклеотидные последовательности генов, имеющих сходные функции, у различных организмов часто оказываются близки между собой. Различие проявляется в заменах некоторых остатков, и в наличии вставок или выпадений в некоторых последовательностях. Понятие *эволюции*, в широком смысле, позволяет интерпретировать факты такого сходства, увязывая их с происхождением различных видов от общего предка, подобно тому как в истории династия ведет свой род от одного основателя. В такой интерпретации, факты замены остатков называют *мутациями*; предполагается, что мутации, происходящие в геномах отдельных индивидов, закрепляются со временем во всей популяции биологического вида.

Истории происхождения рассматриваемых видов, в рамках модели эволюции, может быть сведена к графу, имеющему структуру *дерева*. Такой граф, в широком смысле, называют *филогенетическим деревом*. Задача построения филогенетических деревьев, сформулированная на языке математики, не имеет однозначного решения, и для построения филогенетических деревьев используют эмпирические подходы различной степени сложности (рис. 3.47).

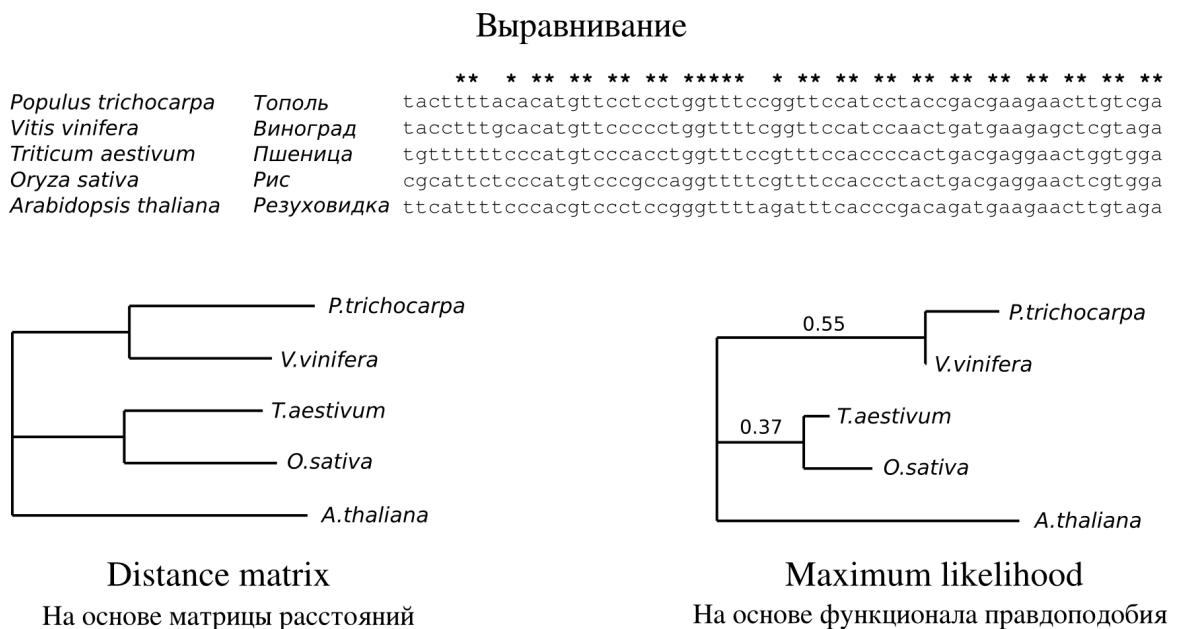


Рис. 3.47: Методы построения филогенетических деревьев

Выравнивание фрагмента генов семейства транскрипционных факторов NAC
расчеты с помощью программ *fastme* и *phym*

В наиболее простом из подходов, количество замен между каждой парой последовательностей записывается в виде матрицы, так называемой *матрицы расстояний* между последовательностями. Затем по матрице следует восстановить филогенетическое дерево, причем при этом разделяют понятия структуры графа дерева (*топологии дерева*), и длины ветвей дерева. Даже в

такой простой постановке можно отметить возможность неоднозначностей при восстановлении топологии дерева, как это показано на рис. 3.48. При формализации задачи построения дерева с помощью более сложных статистических моделей, и более сложных алгоритмов для нахождения оптимального решения в рамках модели, возможным шагом к уточнению и усложнению моделей является предположение о зависимости вероятности мутаций от типа нуклеотидных или аминокислотных остатков. Поиск решения построенных статистических моделей возможно проводить, следуя общим принципам подбора параметров в статистических моделях, на основе минимизации так называемого *функционала правдоподобия*.

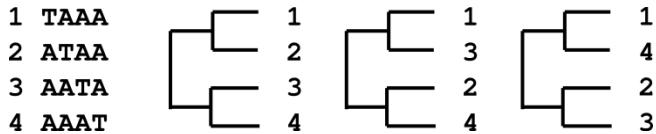


Рис. 3.48: Схема неоднозначностей при построении филогенетических деревьев

Частью таких статистических моделей будет так называемая *модель замен остатков*. В практике построения филогенетических деревьев используются более десятка моделей замен нуклеотидных остатков, предложенных на основании разного рода эмпирических подходов и обоснований. Но ни одну из предложенных моделей не следует считать безусловно оптимальной во всех ситуациях, где стоит задача построения деревьев.

В задаче восстановления времени, в которое происходило разделение видов, вводятся дополнительные требования к используемым статистическим моделям, и здесь для нахождения оптимального решения используют алгоритмы на основе "максимизации ожидания" (*expectation maximization*). Этот подход называют *Байесовым подходом* в филогении, по названию формулы ("формула Байеса") для записи условной вероятности события: $p(x) = p(y)p(x|y)$.

Моделирование МСМС

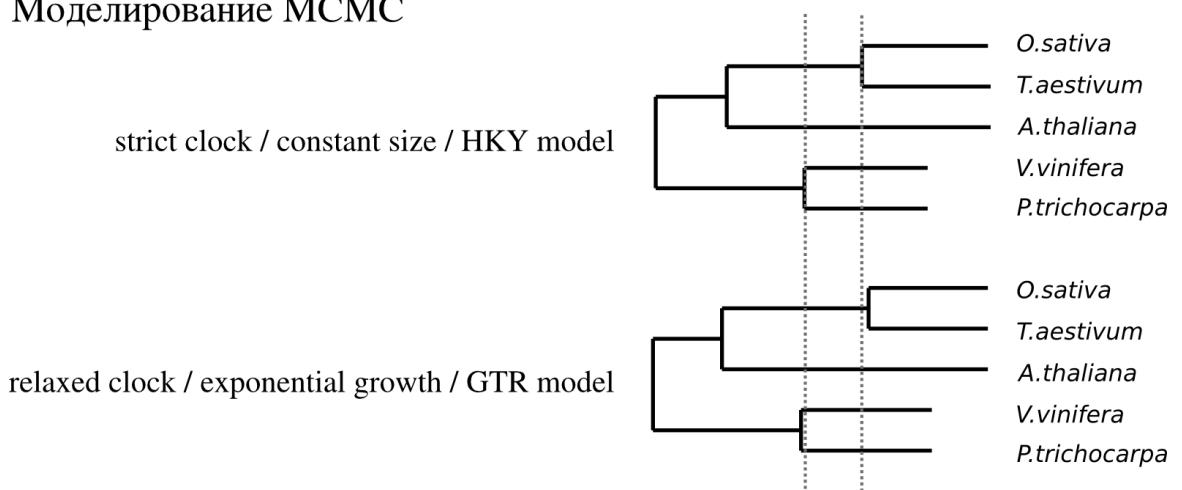


Рис. 3.49: Уточненные методы построения филогенетических деревьев
расчеты с помощью программ *beast-mcmc*

При построении статистических моделей в рамках Байесова подхода, помимо замен вводится

ряд других моделей, среди которых модель для описания размера популяции вида и так называемая *модель молекулярных часов*. Также, при реализации алгоритма максимизации ожидания в этой задаче, по причине вычислительной сложности, используют метод из класса методов Монте-Карло, так называемый метод Metropolis Chain Monte Carlo - МСМС. В простых задачах, введенные усложнения статистических моделей компенсируются достаточной степенью точности решения, независимо от выбора модели (рис. 3.49). Поддержкой и оправданием эффективности подобного подхода служит так называемая *теория нейтральности*, разрабатываемая в эволюционной биологии.

Среди возможных способов расширить и уточнить используемые вероятностные модели, следует упомянуть так называемую "*модель расслабленных молекулярных часов*" (relaxed molecular clock), когда скорость молекулярных часов на каждом из ветвей дерева представляется случайной величиной, описываемой общим распределением вероятностей. Но результаты расчетов с использованием такой уточненной модели, как и уточненных моделей другого рода, также не всегда являются адекватными; недостаточный объем исходной информации при восстановлении филогенетических деревьев не всегда удается компенсировать применением расчетных методов, как бы эти методы не были сконструированы.

Эволюция патогенов

В эволюции патогенов человека, болезнетворных вирусов, бактерий и эукариотических микрорганизмов, можно заметить некоторое сходство, трудно поддающееся формальным определениям. Гипотеза, проиллюстрированная на рис. 3.50, обосновывает сходство между патогенами, основанное на общих стратегиях при проникновении в организм. Для размножения, бактерии или вирусу необходимо преодолеть механизмы защиты хозяина. Можно предположить, что количество уязвимостей в механизме защиты хозяина, через которые может потенциально проникнуть чужеродный агент, ограничено. И может оказаться, что эволюционно не связанные бактерии или вирусы используют одну и ту же уязвимость в механизме защиты хозяина. В то же время для антител и для других белков при их взаимодействии существует эффект полиспецифичности, когда одно и то же антитело взаимодействует с сильно отличающимися по первичной структуре сайтами белков-антител.

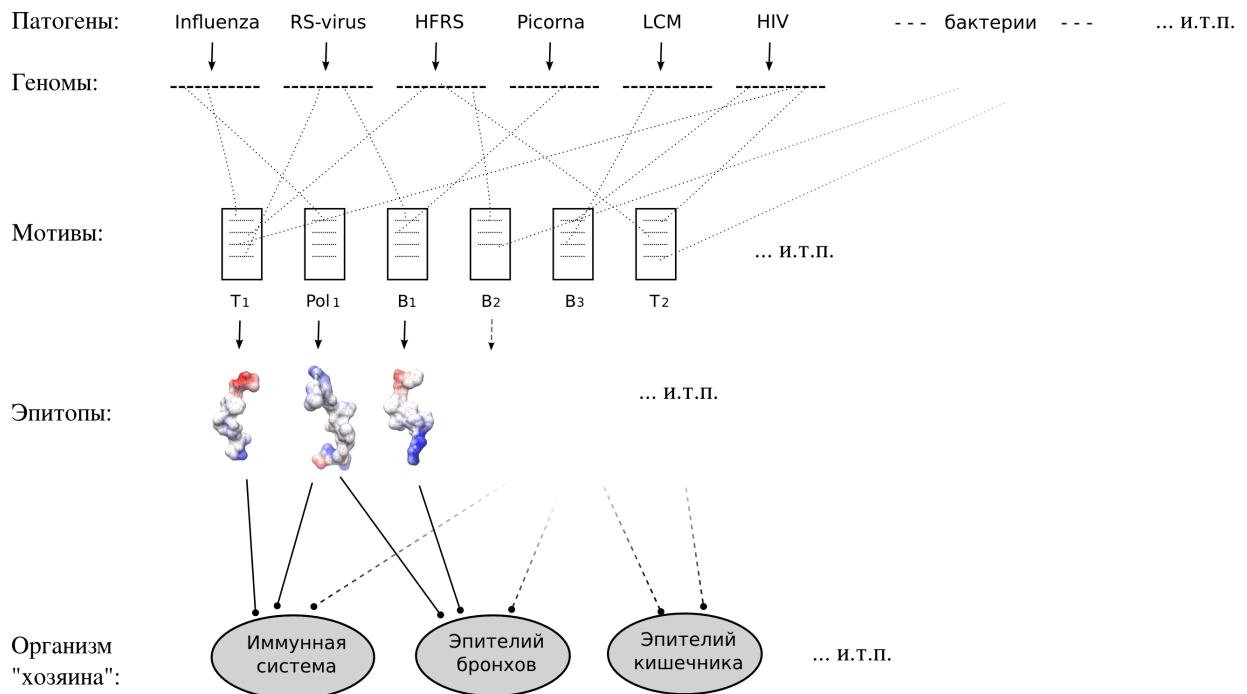


Рис. 3.50: Схема взаимодействия сайтов вирусов с организмом хозяина
по материалам ([Владыко и Петкевич 2001](#))

На рисунке 3.50 схематично показано, как сходные по структуре эпитопы, фрагменты белков в геномах патогенов, могли бы быть специфичны к одним и тем же белкам-мишеням в организме хозяина. Как другая модель эволюции патогенов, на рис. 3.51 приведено филогенетическое дерево, где реконструирована история распространения штаммов холеры. По предположениям, каждая из трех волн "пандемии" холеры начиналась с событий, когда патогенный штамм бактерии приобретал дополнительные свойства, такие как устойчивость к определенной группе антибиотиков, дававшие преимущество при дальнейшем распространении инфекции.

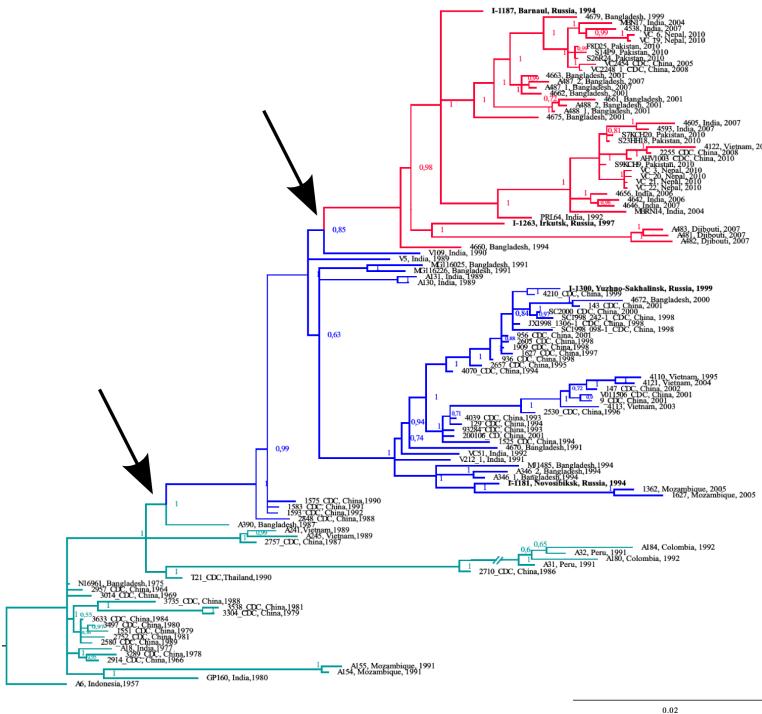


Рис. 3.51: Модель эволюции бактерии - возбудителя холеры

стрелками помечены этапы начала двух новых волн пандемии.
по материалам (Mironova и др. 2018)

Детальное сравнение геномов бактерий позволяет частично объяснить механизм передачи информации между штаммами и даже между разными видами. В геномах бактерий обнаруживают так называемые *геномные острова* или *мобильные генетические элементы*, фрагменты, имеющие сходство у бактерий разных штаммов и разных видов. Эволюция некоторых геномных островов, как и эволюция повторяющихся элементов в геноме, происходит отдельно от эволюции основного организма. По предположению, частично подтвержденному на молекулярном уровне, эти участки генома обладают свойством создавать свои копии, без непосредственной связи с процессом копирования всего генома при делении клетки.

На приведенном на рис. 3.52 фрагменте выравнивания белка NS5 flaviviruses показаны свойства смежных участков, относящихся к зоне, смежной между двумя доменами этого белка, структура которого была приведена на рис. 2.28. Аминокислотная последовательность одного из этих участков содержит вариабельные фрагменты, специфичные для каждого вида вирусов, и характеризующие тип вируса, по виду организма-переносчика (комары или клещи) и по симптомам заболевания. В частности, вирусы Денге (DV) и Западного Нила (WNV), переносимые комарами, сходны по последовательности междоменного интерфейса и по симптомам заболевания, то же можно отметить для вируса Зика и вируса желтой лихорадки (YFV), а так же для группы вирусов, переносимых клещами (TBEV, OHFV, KFDV)

TBEV	SRKLLAR FGDQ	RGP	TRVPELDL GVGTRCVVLAEDKVKEKD VQE RISALREOYGETWHMDREHPYRTWQYWGSYRTAP
	SRKLLAR FGDQ	RGP	TRVPELDL GVGTRCVVLAEDKVKEKD VQE RISALREOYGETWHMDREHPYRTWQYWGSYRTAP
	SRKLLAR FGDQ	RGP	TRVPELDL GVGTRCVVLAEDKVKEKD VQE RISALREOYGETWHMDREHPYRTWQYWGSYRTAP
OHFV	SRKLLAR FGDQ	RGP	IRVPEMD LGVGTRCVVLAEDKVKESD VQE RIKALQEYGD TWHVDREHPYRTWQYWGSYRTAP
	SRKLLAR FGDQ	RGP	IRVPEMD LGVGTRCVVLAEDKVKEHD VQE RIKALQEYOSDTWHVDREHPYRTWQYWGSYRTAP
KFDV	SRKLLAR FGDQ	RGP	TKVPEVD LGTGTRCVVLAEDKVREADVAE RIAALKTOYGSWSHVDKEH PYRTWQYWGSYKTEA
	SRKLLAR FGDQ	RGP	TKVPEVD LGTGTRCVVLAEDKVREADVAE RIAALKTOYGSWSHVDKEH PYRTWQYWGSYKTEA
YFV	SRLLMRRM	RRPTGKVTL EPDV IL PIGTRSVETDKGLD RGAIEERVERIKTEYAA TFHNDN PYRTWHYCGSYITKT	
	SRLLMRRM	RRPTGKVTL EPDV IL PIGTRSVETDKGLD RDAIEERVERIKTEYAA TWFYDND PYRTWHYCGSYITKT	
WNV	SOVLLGRMEKKTWKG P	QYEEDVN	LGSCTRAVGKP LLNSDTSKIRNIE RLKKEYSSTWHQD VNH PYRTWNYHGSYEV KP
	SOVLLGRMEKKTWKG P	QFEEDVN	LGSCTRAVGKP LLNSDTSKIRNIE RLKKEYSSTWHQD ANHPYRTWNYHGSYEV KP
USU	SOVLIGRMEKRTWHGP	KYEEDVN	LGSCTRAVGKP QPHTNQE IKARIQLKEEYAA THWDKDH PYRTWTYHGSYEV KP
	SOVLIGRMEKRTWHGP	KYEEDVN	LGSCTRAVGKP QPHTNQE IKARIQLKEEYAA THWDKDH PYRTWTYHGSYEV KP
JEV	SOVLLGRMDRVRWRGP	KYEEDVN	LGSCTRAVGKG EVHS DQGKIKKR IEKLKEEYAA TWHEDEP H PYRTWTYHGSYEV KA
	SOVLLGRMDRVRWRGP	KYEEDVN	LGSCTRAVGKG EVHS DQGKIKKR IEKLKEEYAA TWHEDEP H PYRTWTYHGSYEV KA
ZIKA	SQLLLGRMD	GPRRPV KYEE DVN	LGSCTRAVVSCAE APNM KIIGNR I EIRIRSEHAETWF FDE NHPYRTWAYHGSYEAPT
	SQLLLGRMD	GPRRPV KYEE DVN	LGSCTRAVVSCAE APNM KIIGNR I EIRIRSEHAETWF FDE NHPYRTWAYHGSYEAPT
DV3	SRLLLNRTMT H	TIEKD D V	LGAGTRHVNAE PETPNMD VIGERIKRIKEEHS STWHYDDENP YKTWAYHGSYEV KA
	SRLLLNRTMT H	TIEKD D V	LGAGTRHVNAE PETPNMD VIGERIKRIKEEHS STWHYDDENP YKTWAYHGSYEV KA
DV1	SRMLLNRT TM AH	RKP	TYERD D V LGAGTRHV AVEPEVANLD IQRIENI KEH KSTWHYD EDENP YKTWAYHGSYEV KP
	SRMLLNRT TM AH	RKP	TYERD D V LGAGTRHV AVEPEVANLD IQRIENI KEH KSTWHYD EDENP YKTWAYHGSYEV KP
DV2	SRMLLNRT TM RH	RKP	TYERD D V LGAGTRHV AVEPEVANLD IQRIENI KEH KSTWHYD EDENP YKTWAYHGSYEV KP
	SRMLLNRT TM RH	RKP	TYERD D V LGAGTRHV AVEPEVANLD IQRIENI KEH KSTWHYD EDENP YKTWAYHGSYEV KP
DV4	SKMLLNRT TT RH	RKP	TYERD D V LGAGTRSV ST TEKPD MT II GR RL Q LEEH K ETWHYD HN PYRTWAYHGSYEAPS
	SKMLLNRT TT RH	RKP	TYERD D V LGAGTRSV ST TEKPD MT II GR RL Q LEEH K ETWHYD HN PYRTWAYHGSYEAPS
	SKMLLNRT TT RH	RKP	TYERD D V LGAGTRSV ST TEKPD MT II GR RL Q LEEH K ETWHYD HN PYRTWAYHGSYEAPS

Рис. 3.52: Выравнивание фрагмента неструктурного белка NS5 у вирусов из рода *flavivirus*

Для представления выравнивания использованы по три последовательности от каждого представленного вида, из рода *flavivirus*. В показанном фрагменте междоменного интерфейса белка NS5, серым выделен наиболее вариабельный участок, бежевым наиболее консервативный участок, оранжевым обозначены консервативные позиции, со 100% идентичностью в использованной выборке флавивирусов.

Приведенное выравнивание иллюстрирует еще одну модель эволюции патогенов. В рамках этой модели, механизм перестройки доменов, осуществляемого при участии зоны междоменного интерфейса (рис. 2.28) может модифицироваться, для гибкой адаптации без критического влияния на жизненный цикл вируса. И, в этом случае, отмеченные мутации в зоне междоменного интерфейса белка NS5 являются следствием адаптации отдельных видов вируса к новым условиям окружения в процессе эволюции рода флавивирусов, и выражавшейся в накоплении мутаций в зоне междоменного интерфейса.

Наблюдая за развитием пандемии холеры или распространением устойчивых к антибиотикам штаммов стафилококка, возможно допустить факт ускорения эволюции этих патогенов. Расчет скорости эволюции возможен, среди перечисленных в разделе 3.9 методов, лишь с использованием Байесова подхода в филогении. Но ускорение эволюции при этом можно оценить лишь приближенно, по росту скорости молекулярных часов, рассчитанной разных фрагментах филогенетического дерева.

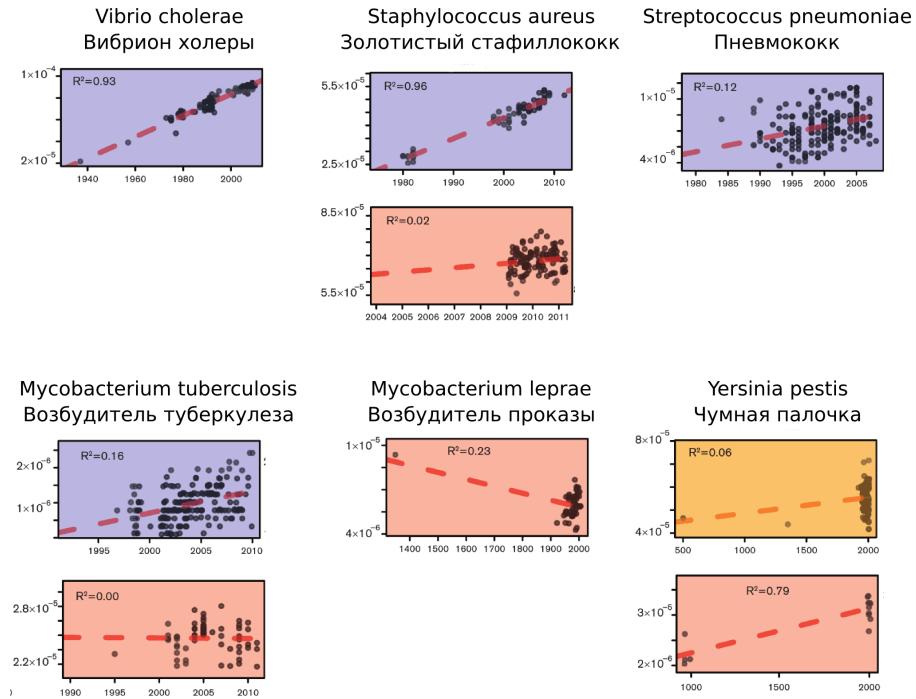


Рис. 3.53: Оценки изменения скорости эволюции некоторых патогенов

Для трех из шести бактерий, рассмотрено по две группы штаммов. Синим помечены группы, где с достоверностью замечен рост скорости эволюции, в рамках использованного метода расчетов.

по материалам ([Duchêne и др. 2016](#))

Такой анализ, приведенный на рис. 3.53, действительно позволяет заметить рост скорости эволюции для вибрион холеры и некоторых групп золотистого стафилококка. Хотя для других патогенов, таких как чумная палочка или возбудитель проказы, скорость эволюции примерно постоянна или имеет тенденцию к замедлению.

Применение Байесова подхода в филогении часто сопряжено с численной неустойчивостью решений и значительными флуктуациями в положениях узлов дерева (рис. 3.54). И потому возможно, без потери точности, объяснить волны в эволюции холеры, рассмотрев модель замедляющейся эволюции, упомянутую при изложении теории фракталов (раздел 3.8). При этом, для всех патогенных бактерий, в среднем, эволюция замедляется, и критическая точка в прошлом обозначает момент возникновению нового вида.

Задокументированная история пандемии холеры отражает зависимость развития патогена от непредсказуемых мелкомасштабных событий, таких как переселение некоторых народов во Второй мировой войне и после нее. Наличие волн может быть подтверждено и для других патогенов, таких как чумная палочка (*Y. pestis*) или золотистый стафилококк (*S. aureus*). Но для *Y. pestis*, страшные пандемии чумы остались в прошлом. И для быстро адаптирующегося *S. aureus*, приобретение устойчивости к следующим поколениям лекарств может и не отразиться на изменениях частоты мутаций.

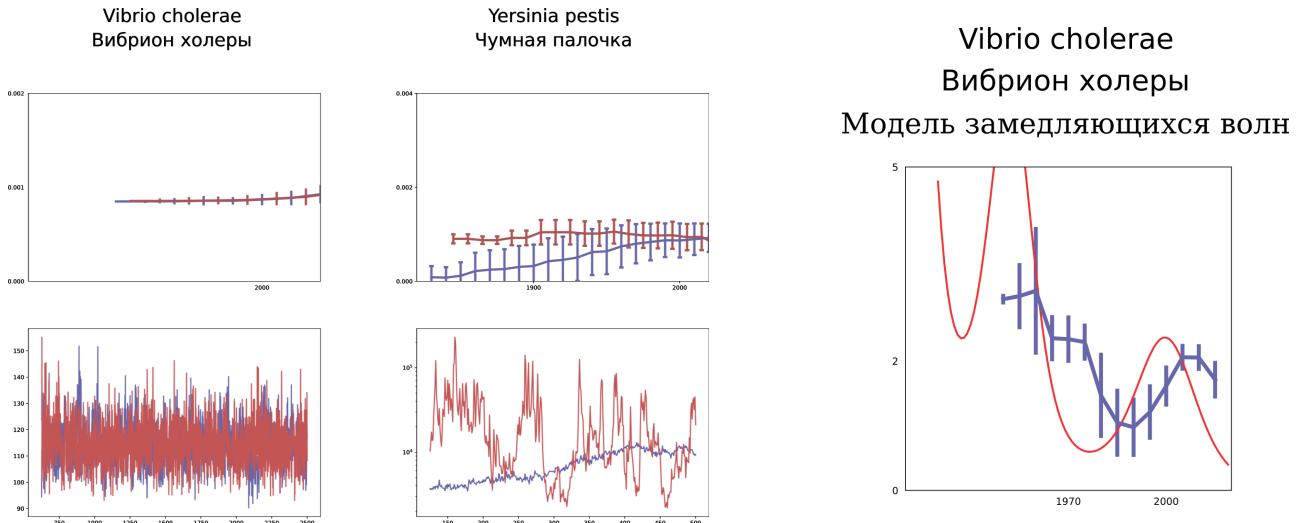


Рис. 3.54: Расчеты эволюции бактерий методом МСМС

Слева вверху: Изменение скорости эволюции двух патогенов, оцененное по оптимальному из деревьев, по результатам моделирования МСМС.

Слева внизу: Изменение оценки положения корня дерева в зависимости от шага при моделировании МСМС, для двух серий расчетов

Справа: Расчет эволюции штаммов холерного вибриона, с использованием модели волн с замедляющейся периодичностью

на всех графиках, по горизонтальной оси - годы, по вертикальной оси - оценки скорости мутаций расчеты с использованием пакета Beast, по материалом штаммов, перечисленных в (Cui и др. 2013; Didelot и др. 2015)

Бактерии, рассмотренные выше как патогенные, развиваются практически независимо друг от друга, за исключением того, что все штаммы происходят от общего предка, и потому критическая точка при эволюции каждого из видов находится в прошлом. Но, как пример обратной ситуации, озеро Байкал представляет собой измененную экосистему с сильной взаимозависимостью его компонентов. И недавний кризис на Байкале, по наблюдениям, развивался постепенно и "по нарастающей". Также и в эволюции человечества, такие признаки, как рост частоты экономических кризисов, могут обозначать наличие критической точки в будущем. Но и в человеческом обществе, в наши дни, отдельные слои и культуры все сильнее сильно зависят друг от друга. Но прогнозы с использованием модели ускорения эволюции следует сравнивать по точности с прогнозом погоды, и нет оснований говорить о возможности предсказать точные времена и сроки перехода в зону бифуркации.

Болезнь байкальской губки

В экспериментах по изучению микробиомов, во многих прикладных задачах микробиологии, было найдено достаточно свидетельств о богатом и сложном устройстве микробных сообществ, расширяя постановки вопросов, принятые в классической микробиологии. В одном из аспектов

интерпретации экспериментов, данные о количественном составе микробных сообществ оказалось возможным сопоставить с моделями, принятыми в экономике и социологии. Этот подход проиллюстрирован описанным ниже примером о составе симбиотических и патогенных микроорганизмов в байкальской губке (рис. 3.55).

Болезнь байкальской губки, как косвенный признак экологического кризиса в озере, начинала распространяться с 2011 г, и в настоящее время поразила большую часть губок, организмов, фильтрующих воду и обеспечивающих самоочищение Байкала. Один из аспектов различия микробиомов в здоровых и больных губках, показанный на рис. 3.55, состоит в значительной гетерогенности сообществ в губках с признаками заболевания. На языке социологии, сообщества в заболевших губках характеризуются значительно меньшей степенью "социального неравенства", как это проиллюстрировано на рис. 3.55 с помощью "кривых Лоренца".

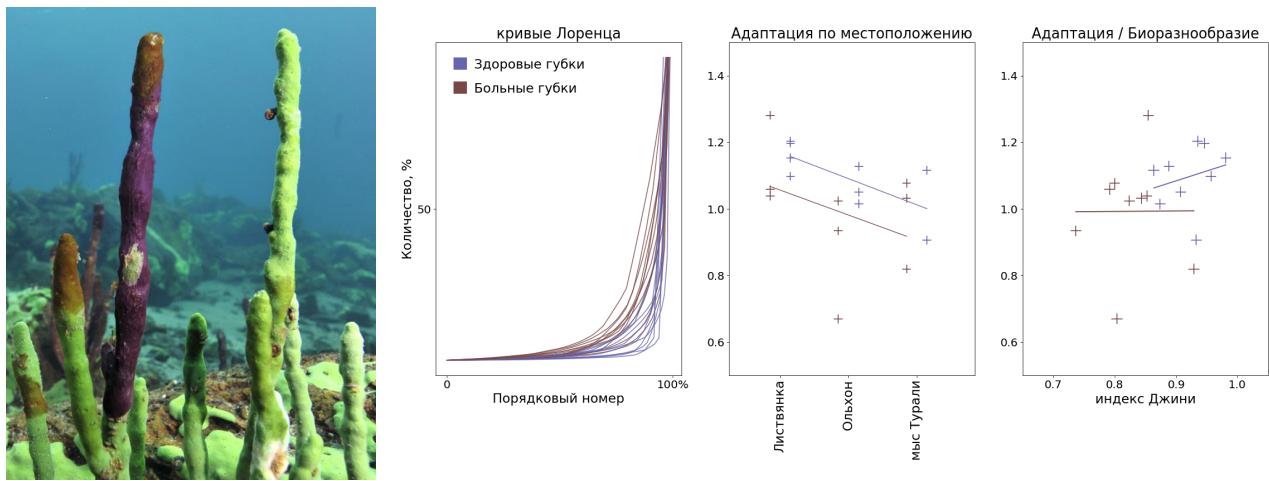


Рис. 3.55: Адаптация байкальской губки в условиях кризиса
Фото И.В.Ханаева. Расчеты по материалам (S.I. Feranchuk, Potapova и др. 2018)

Термином *оппортунистический патоген* обозначают те из симбиотических микроорганизмов, которые при ослаблении организма-хозяина переключают тип питания с симбиотического на паразитический (см. раздел 3.8). Именно за счет развития оппортунистических патогенов увеличивается биоразнообразие микроорганизмов в больных и гибнущих губках. Легко объяснить, что в условиях кризиса, в тех губках, которые остаются здоровыми, повышен уровень защиты от внешних микроорганизмов, и патогенных и симбиотических. Но малое биоразнообразие и жесткие отношения оставшихся симбионтов являются признаком хрупкости системы, делая здоровые губки более уязвимыми к атаке внешних патогенов, такой что сила атаки превосходит порог защиты.

При этом губка не может существовать без симбиотических организмов, и наиболее консервативные из симбионтов также могут существовать лишь совместно с губкой. И признаком по-настоящему здорового организма являлась бы степень биоразнообразия симбионтов, сравнимая со временами до наступления кризиса. Исследуя способность экосистемы озера к преодолению кризиса, было замечено, что в здоровых губках все же происходят изменения в геноме "добропачественных" симбионтов. Мутации там случаются относительно редко, но быстрее, чем в

отсутствии болезни хозяина, и тем быстрее, чем жестче симбиотическая система губки. Другими словами, жесткость регуляции симбионтов со стороны организма-хозяина следует считать не конечной целью, а этапом, необходимым при поиске возможностей для преодоления времен кризиса.

Если все же система защиты не выдержала атаки патогенов, болезнь, начав развиваться, неизбежно приводит к гибели губки. Но было замечено, что в "добропорядочных" симбионтах заболевшей губки происходит неожиданно много мутаций, так что это нельзя свести к обычным и известным механизмам изменений в геноме. Заболевший организм в итоге все же гибнет, но путем компенсации атак агрессивных патогенов при этом отчасти закрепляются в геномах его симбионтов. Горизонтальный перенос генов в экосистеме возможен, обычно он происходит в рамках формата *мобильных генетических элементов*, подобно тому как происходит обмен информацией среди людей через культуру. И, не игнорируя накопленные в погибших губках способы отражения атак патогенов, пусть даже они передаются через фрагменты геномов посторонних бактерий, здоровые организмы имеют больше шансов преодолеть кризис.

Эволюция человека

Методы молекулярной эволюции, с одной стороны, оказались эффективными в реконструкции некоторых сторон этой истории, и, в том числе, подтвердилась и уточнилась схема расселения племен с Африканского континента и Ближнего Востока, показанная на рисунке 3.56, в согласии с известными ранее данными антропологии. Также, в соответствии с правилами наследования генетического материала, было доказано на основании анализа геномов, что современные люди происходят от одного общего предка по мужской линии, и одной матери по женской линии.

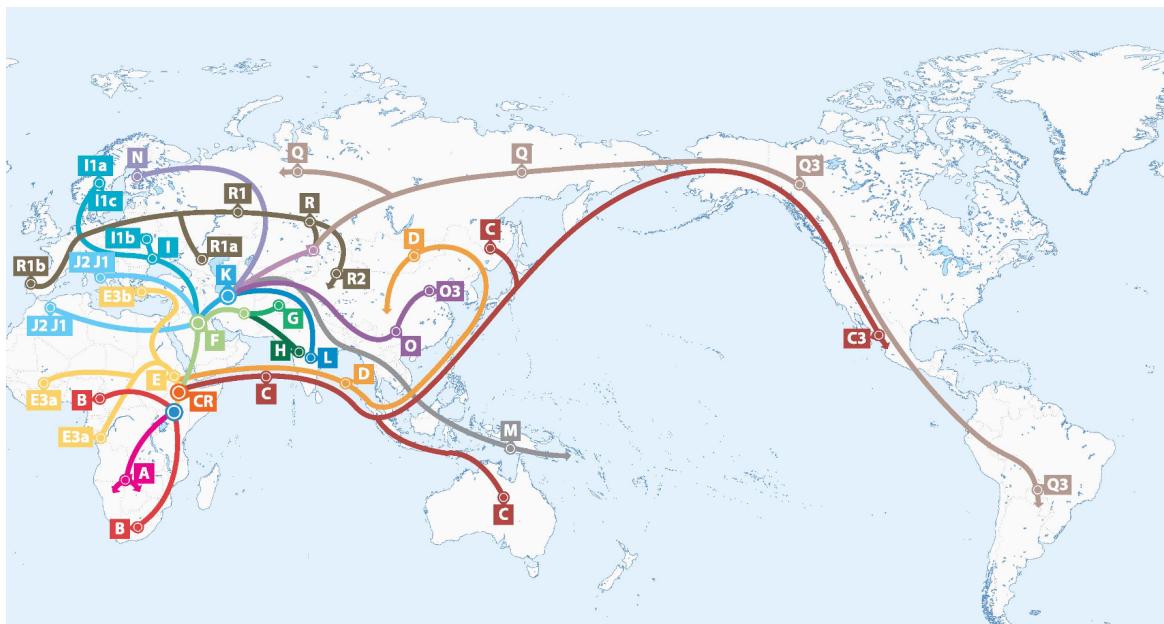


Рис. 3.56: Пути миграции современных людей при их расселении
по материалам проекта familytreedna.com

Однако при восстановлении некоторых деталей происхождения первых людей, результаты,

полученные на основании расчетов по молекулярной эволюции, оказались противоречивыми, и источник противоречий нельзя отыскать, оставаясь в рамках этих методов. В частности, так называемый "эффект основателя", наличие единого общего предка в группе первоходцев, не проявляется в расчетах по молекулярной филогенетике, хотя обнаруживается другими методами. И, что более важно, вопросы о времени происхождения обоих общих предков людей, и об истории людей до этапа расселения из Африки, остаются неразрешенными.

Ограниченнность, заложенную в методах молекулярной эволюции, при исследовании происхождения людей, можно прояснить на примере, приведенном на рис. 3.57. Несмотря на то, что для восстановления отношений между этническими группами на основании сравнения их полных геномов, проведенного в (Schiffels и Durbin 2014), были выбраны наиболее типичные представители этих групп (из Центральной Европы, Италии, Китая, Мексики и др.), оказалось, что при анализе мужской и женской линий, где генетический материал передается по другим правилам, происхождение выбранных людей оказалось часто не идентично происхождению их этнических групп. Результат, показанный на рис. 3.57 (вверху), демонстрирует недостатки формальных подходов филогенетики, по сравнению с историями жизни людей, историями их браков и их родов.

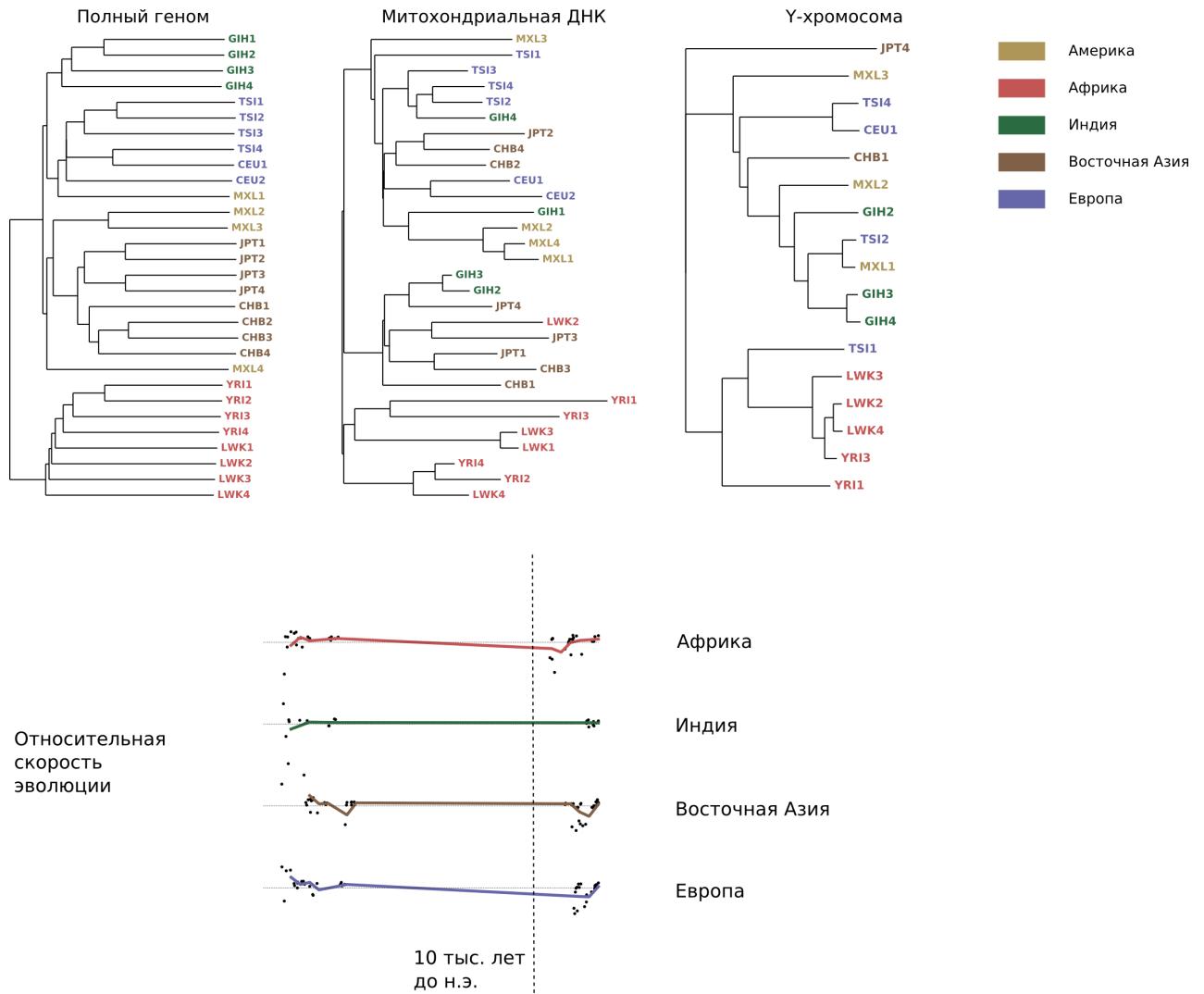


Рис. 3.57: Восстановление эволюции современных людей через филогенетические деревья

Вверху: отношения между представителями этнических групп, по разным линиям

Внизу: оценки относительного изменения скорости мутаций, для каждой из этнических групп расчеты по данным проекта 1000 Genomes

Предположение о постоянной скорости мутаций явно или неявно включено в вероятностные модели, на которых основаны расчеты по филогенетии. Но введя предположение о возможности изменения темпа молекулярных часов, возможно на качественном уровне оценить относительное изменение скорости мутаций, как это показано на рис. 3.57 (внизу), для эволюции современных людей. На этих графиках, рассчитанных по полиморфизмам в полном геноме, можно заметить рост скорости мутаций в течении последних нескольких тысяч лет. Это могло бы быть артефактом использованного подхода, но в модели, где допускается изменение скорости эволюции, с момента зарождения биологического вида, как и отдельного организма, скорость его эволюции постепенно снижается. Но возможно и ускорение эволюции, до коллапса или «критической точки», как это и ожидается в эволюции человечества. В этом случае, в какой-то момент скорость эволюции должна быть минимальной, и наблюдаемый рост скорости мутаций можно

объяснить переключением между этими двумя режимами.

3.10. Вместо заключения

Огромный поток информации, хлынувший в последние десятилетия, захлестнул, в том числе, и деятельность, которую принято называть научной работой. Несмотря на продолжающееся усовершенствование методов и накопление данных, результаты исследований не стали более точными и осмысленными. И, в этих условиях, занятие наукой все больше становится сходно с отправлением языческих ритуалов.

Сводная обработка публикаций по лекарственным растениям, про которую было сказано в разделе 3.7, при условии "фильтрации" искажений, внесенных смещением акцентов в публикациях, подходит для представления отношений между хроническими заболеваниями, согласованного с другим, независимым, подходом к расчетам (рис. 3.58, вверху). Но, полученные при сводном анализе результаты могут быть показаны с разных сторон, и, в одном из представлений, разделение традиций медицины выглядит как изображено на рис. 3.58 (внизу).

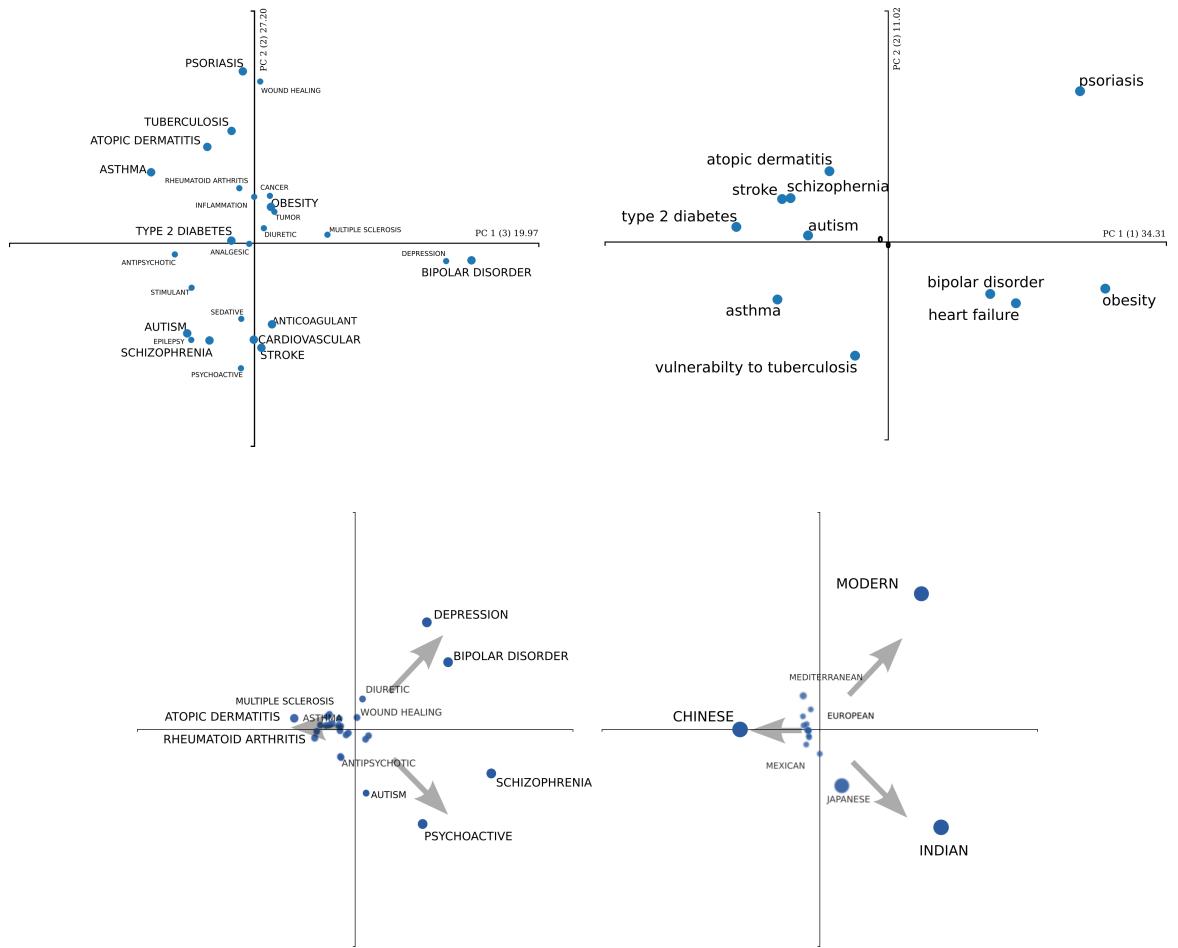


Рис. 3.58: Результаты сводной обработки биомедицинских публикаций

Вверху: Отношения между группами хронических заболеваний, по двух способам расчетов

Внизу: Отношения между заболеваниями и традициями медицины, в представлении сопряженных главных компонент

Разделения на Восток и Запад, и на центр и периферию, всегда существовали в цивилизации. Но поток информации, такой как происходит в наши дни, не с чем сравнить из событий истории. И сопряженное с ним язычество - другого рода, более жесткое и страшное, чем те культуры, которые были и продолжают оставаться в странах и народах.

То, что названо в книге "ускорением" развития цивилизации, нельзя не заметить, занимаюсь любыми из дел и глядя с любой из точек зрения. Мысли о приближающемся конце "витают в воздухе", и вместе с ними, то подспудно, то явно, "витает" безнадежность и отчаяние. Как лишь одна из точек зрения, в уточненных и расширенных расчетах ускорения эволюции геномов (рис. 3.59), остается заметным наблюдение о "точке перегиба" в развитии цивилизации несколько тысячелетий назад. Но, не говоря про времена и сроки больше чем это возможно, все же можно было бы сказать про события истории тех времен, которые до сих пор указывают нам на Путь, и Истину, и Жизнь.

Относительная скорость эволюции

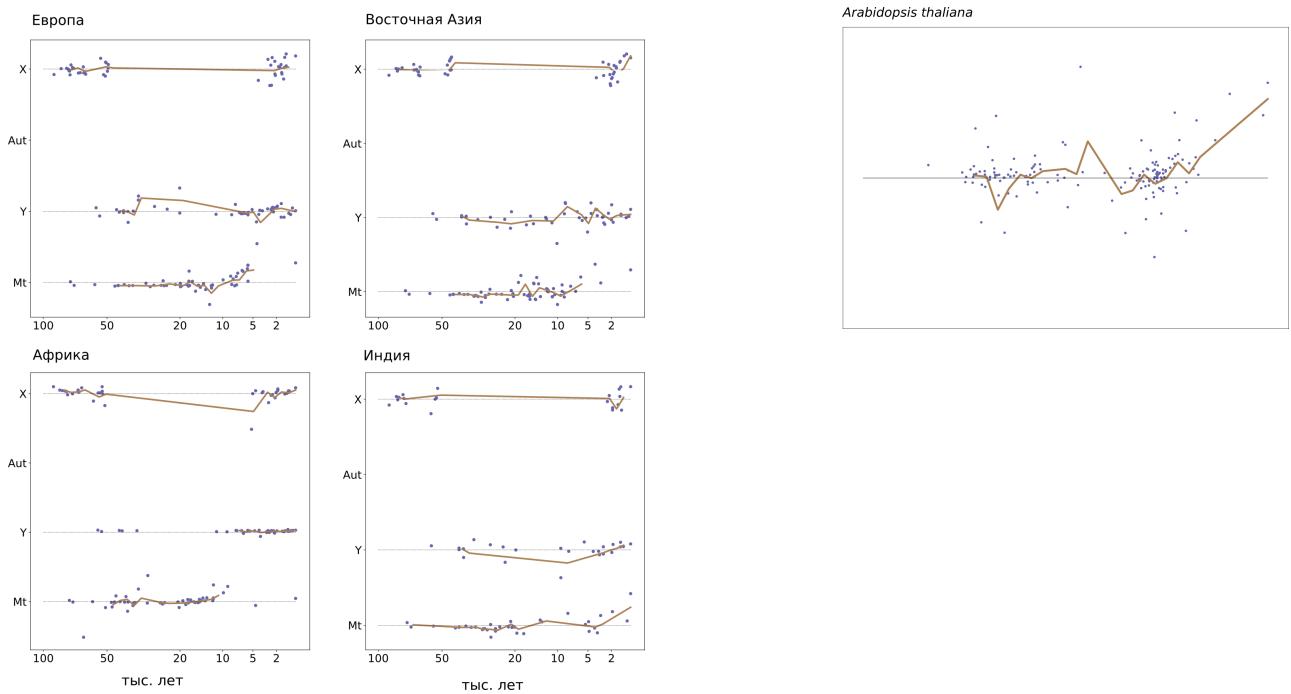


Рис. 3.59: Расширенный расчет изменений в скорости эволюции

И также, при работе в науке возможно заметить завораживающие детали, свидетельствующие о почерке Создателя. Вспоминая наших предшественников и учителей, надо вспомнить и Н.И. Вавилова, который, заметив сходство и параллели в развитии и изменчивости растений, назвал это "гомологичные ряды изменчивости". Некоторые другие из подобных "параллелей", и неожиданно замечаемой упорядоченности устройства мира, показаны на рис. 3.60. Но, впрочем, подобные признаки можно обнаружить и при разработке алгоритмов, моделей и приближенных описаний, как это было отмечено в нескольких из разделов книги.

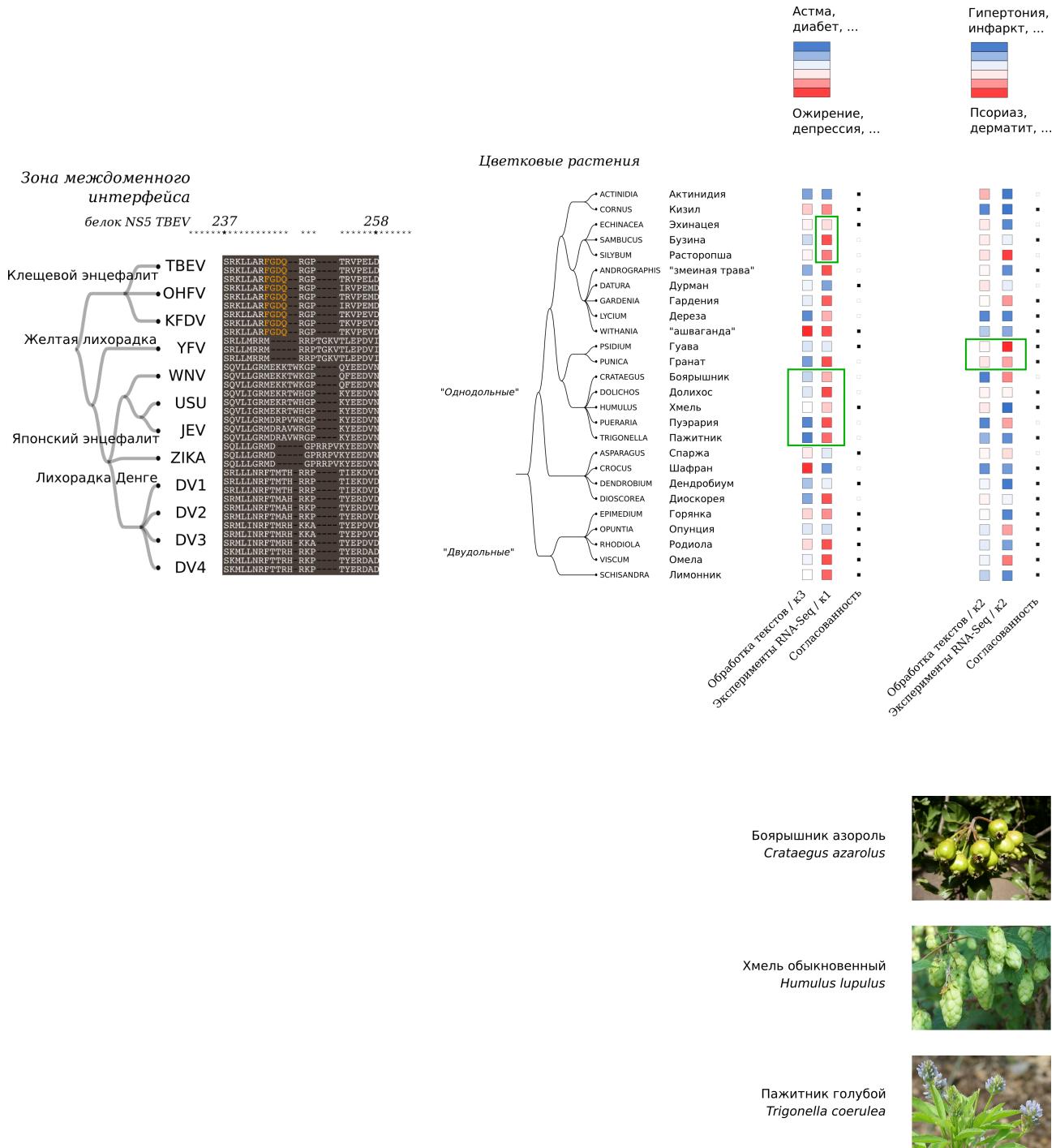


Рис. 3.60: Ряды сходства и изменчивости

Слева: Зона междоменного интерфейса flavivирусов. Сходство в свойствах выражаются в сходной "сигнатуре" вирусов, в выделенном участке последовательности белка NS5.

Справа: Оценки эффекта лекарственных растений, по двух способам расчетов. Для представления отобраны 25 растений, которые сравнимы по количеству упоминаний их родовых латинских названий в публикациях. Группы "гомологичных рядов изменчивости" помечены зелеными рамками.

выравнивание фрагментов белка flavивирусов построено по материалам (Потапова и др. 2018). фотографии растений: Cillas, Nejmlez, Flyout (Wikipedia)

Обнаружить подобные ряды изменчивости, занимаясь наукой - это знак благословения; но такое благословение нельзя разменять на деньги или почести, так уж заложено в устройстве мира. Замеченные сходства и параллели слишком трудно уловимы, чтобы их можно было использовать в тех занятиях, которые приносят "практическую пользу". "Пути Господни неисповедимы", или, как об этом сказано в псалме Асафа,

...

*Глас грома Твоего в круге небесном,
молнии освещали вселенную, земля содрогалась и трепетала.
Путь Твой в море, и стези Твои в водах великих, и пути Твои не познаются.
Как стадо, Ты вел народ свой, рукою Моисея и Аарона*

Но впрочем, ни деньги, ни почести не имеют отношения ни к началу, ни к концу жизни человека. В псалме Асафа, рассказчик вспоминает про исход из Египта; и, задумавшись о историях и притчах в Священных Писаниях в еще более широком контексте, несомненным становится, что "монеты" таких благословений пригодились бы как раз там, где содержится и начало и конец.

Как принято говорить про события во время кризиса при обсуждении математических моделей, те подходы, которые позволяют предсказать существование точки бифуркации, не подходят для описания событий в окрестности этого периода. И неопределенность, заведомо подразумеваемая в упомянутых выше наблюдениях, слишком неточных для "практического использования", могла бы быть компенсаций точности "общепринятых" подходов к планированию и предсказанию, когда за таким планированием с неотвратимостью следуют безнадежность и отчаяние. И именно некоторые из поэтичных и неточных параллелей и "аллюзий", те, которые завораживают ученых, которые их обнаруживают, могут оказаться более полезны, по мере того как времена меняются все быстрее и быстрее.



фото: pixel0201 (foto konkurs.ru), И.В. Ханаев

И все же, говоря про наши дни и наше время, прогнозы о скором "конце света" один за другим оказываются ложью. Но не стоит считать, что кризисных событий удается избежать из-за усилий экономистов и политиков. Дело в другом, среди тех прогнозов, которые не сбылись, были и точные и надежные. Но дело скорее в том, что "правила игры", даже те, которые считались непоколебимыми, понемногу подменяются, как это можно иногда даже заметить явно, в некоторых мелких и тонких деталях. И новые "правила", как это несложно признать, корректируют доселе незыблевые законы так, чтобы наиболее мягко преодолеть время кризиса.

4 Библиография

Список литературы, приведенный ниже, составлен без претензий на полноту представления тем, указанных в заголовках, и без претензий на полноту подкрепления материала, изложенного в основной части курса. В библиографии указаны некоторые из "центральных" работ и описаний методов, традиционно упоминаемых при обсуждении материалов в перечисленных темах. Но также, при составлении библиографии, в рамках некоторых тем, упомянуты материалы, относящиеся к теме, но при этом несколько расширяющие принятые там постановки вопросов. "Новое - это хорошо забытое старое", и часто для этого оказывалось достаточно вспомнить некоторые давно "вышедшие из моды" направления исследований.

4.1 Структурная биоинформатика

Теоретическая физика - учебники

Ландау и Лифшиц 1958; Ландау и Лифшиц 1960; Ландау и Лифшиц 1963; Ландау и Лифшиц 1964; Фейнман и др. 1965; Морс и Фешбах 1958; Морс и Фешбах 1960; Feynman et al. 1964; Morse and Feshbach 1958

Программирование - "классические" учебники

Вирт и Йенсен 1982; Керниган и Ритчи 1992; Jensen and Wirth 1975; Kernighan and Ritchie 1978; Stroustrup 1985.

Квантовые расчеты

Силовые поля и парциальные заряды: J. Gasteiger and Marsili 1980; Mulliken 1955; Jakalian et al. 2000; Hehre et al. 1969.

Пакеты программ: Gaussian (Frisch et al. n.d.), GAMESS (Schmidt et al. 1993), NWChem (Valiev et al. 2010).

Анализ нормальных мод : Skjærven et al. 2014; Hollup et al. 2005

Молекулярная динамика

Учебники и пособия на русском: Холмодуров и др. 2003; Хохлов и др. 2009.

Теория, классические работы, обоснование методов и подходов: Alder and Wainwright 1957; Verlet 1967; H. Berendsen et al. 1984; Cornell et al. 1995; Bultinck et al. 2002; Bonvin et al. 2010; Андрианов 2013.

Пакеты программ: Amber (Case et al. 2005), Charmm (Brooks et al. 2009).

Уравнение Пуассона-Больцмана и учет растворителя : J. Berendsen et al. 1981; Onufriev et al. 2002; Olsson et al. 2011; Hou et al. 2011

Моделирование структуры белков :

Учебный курс: Финкельштейн and Птицын 2012

Теория, классические работы, обоснование методов и подходов: Finkelstein and Badretdinov 1997; Karplus 1997; Galaktionov et al. 2001; McCammon et al. 1977; Liwo et al. 2005

Расчеты боковых цепей: Q. Wang et al. 2008; J. Xu and Berger 2006

Моделирование "по гомологии": Modeller (Sali and Blundell 1993), Jackal/Nest (Xiang 2006), Phyre (Kelley and Sternberg 2009)

Моделирование "de novo": Tasser (Yang and Zhang 2015), Rosetta (Das and Baker 2008)

"Эксперимент" CASP и сравнение методов: Moult et al. 2016

Пути сворачивания белка: Connaughton et al. 1981; Ngo and Marks 1992; Orengo and Thornton 1993; Efimov 1994; Tsai et al. 2000; Козырев, Козьмин, et al. 2010; Bellman 1952

Молекулярный докинг :

Докинг лигандов: Autodock (Goodsell and Olson 1990), Autodock Vina (Trott and Olson 2010), UCSF Dock (Brozell et al. 2012), Surflex (Jain 2009).

Докинг белков: Tovchigrechko and Vakser 2006

Пособие с описанием методов: Гуреев et al. 2018

Визуализация и компьютерная графика : UCSF Chimera (Pettersen et al. 2004), VMD (Humphrey et al. 1996).

Некоторые прикладные работы по структурной биоинформатике и смежным темам : Balevicius et al. 2010; De Marco et al. 2014; I. Feranchuk, Leonov, et al. 2016; I. Feranchuk, Komarov, et al. 1995; Skoromnik et al. 2017

публикации и издания к главе, на русском

Андрianов, А. (2013). Конформационный анализ белков: теория и приложения. Минск: Беларусь навука.

Вирт, Н. & Йенсен, К. (1982). Паскаль. руководство для пользователя и описание языка. М.: Финансы и статистика.

Гуреев, М., Кадочников, В. & Порозов, Ю. (2018). Молекулярный докинг и его верификация в контексте виртуального скрининга. Санкт-Петербург: Университет ИТМО.

- Керниган, Б. & Ритчи, Д. (1992). *Язык программирования Си*. М.: Финансы и статистика.
- Козырев, С., Козьмин, Ю., Богатова, О., Гарковенко, А. & Некрасов, А. (2010). Р-адический анализ первичной структуры белков и реализующий его веб сервис. *Гр. ун-т им. Я. Купалы*.
- Ландау, Л. Д. & Лифшиц, Е. М. (1958). *Механика*. М.: Физматгиз.
- Ландау, Л. Д. & Лифшиц, Е. М. (1960). *Теория поля*. М.: Физматгиз.
- Ландау, Л. Д. & Лифшиц, Е. М. (1963). *Квантовая механика (нерелятивистская теория)*. М.: Наука.
- Ландау, Л. Д. & Лифшиц, Е. М. (1964). *Статистическая физика. часть 1*. М.: Наука.
- Морс, Ф. & Фешбах, Г. (1958). *Методы теоретической физики. т.1*. Москва: ИЛ.
- Морс, Ф. & Фешбах, Г. (1960). *Методы теоретической физики. т.2*. М.: ИЛ.
- Фейнман, Р., Лейтон, Р. & Сэндс, М. (1965). *Фейнмановские лекции по физике*. М.: Наука.
- Финкельштейн, А. & Птицын, О. (2012). *Физика белка: курс лекций с цветными и стереоскопическими иллюстрациями и задачами* (3-е изд., испр. и доп.). Москва: КДУ.
- Холмодуров, Х., Алтайский, М., Пузынин, И., Дардин, Т. & Филатов, Ф. (2003). Методы молекулярной динамики для моделирования физических и биологических процессов. *Физика элементарных частиц и атомного ядра*, 34(2), 472—515.
- Хохлов, А., Рабинович, А. & Иванов, В. (Ред.). (2009). *Методы компьютерного моделирования для исследования полимеров и биополимеров*. М.: URSS.

публикации и издания к главе, на английском

- Alder, B. & Wainwright, T. E. (1957). Phase transition for a hard sphere system. *J. Chem. Phys.* 27, 1208–1209.
- Balevicius, V., Gdaniec, Z., Aidas, K., & Tamuliene, J. (2010). NMR and quantum chemistry study of mesoscopic effects in ionic liquids. *J Phys Chem A*. 114(16), 5365–5371.
- Bellman, R. (1952). On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA*, 38(8), 715–719.
- Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A., & Haak, J. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81, 3684.
- Brooks, B. R., Brooks III, C. L., Mackerell Jr., A. D., Nilsson, L., Petrella, R. J., Roux, B., ... Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10), 1545–1614.
- Brozell, S. R., Mukherjee, S., Balius, T. E., Roe, D. R., Case, D. A., & Rizzo, R. C. (2012). Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J. Comput-Aided Mol. Des.* 26, 749–773.
- Bultinck, P., Langenaeker, W., Lahorte, P., De Proft, G., Geerlings, P., Van Alsenoy, C., & Tollenaere, J. (2002). The electronegativity equalization method II: applicability of different atomic charge schemes. *J. Phys. Chem. A*, 106, 7895–7901.
- Case, D., Cheatham, T. I., Darden, T., Gohlke, H., Luo, R., Merz, K. J., ... Woods, R. (2005). The Amber biomolecular simulation programs. *J. Computat. Chem.* 25, 1668–1688.

- Connaughton, C., Rajesh, R., & Zaboronski, O. (1981). Stationary Kolmogorov solutions of the Smoluchowski aggregation equation with a source term. *Physical Review E*, 69(6), 061114.
- Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., ... Kollman, P. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117(19), 5179–5197.
- Das, R. & Baker, D. (2008). Macromolecular modeling with Rosetta. *Annual Review of Biochemistry*, 77, 363–382.
- De Marco, L., Thämer, M., Reppert, M., & Tokmakoff, A. (2014). Direct observation of intermolecular interactions mediated by hydrogen bonding. *J. Chem. Phys.* 141, 034502.
- Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., ... Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16), 1999–2012.
- Efimov, A. (1994). Common structural motifs in small proteins and domains. *FEBS Letters*, 355, 213–219.
- Feranchuk, I., Komarov, L., Nichipor, I., & Ulyanenkov, A. (1995). Operator method in the problem of quantum anharmonic oscillator. *Annals of Physics*, 238(2), 370–440.
- Feranchuk, I., Leonov, A., & Skoromnik, O. (2016). Physical background for parameters of the quantum rabi model. *J. Phys. A: Math. Theor.* 49(45), 454001.
- Feynman, R., Leighton, R., & Sands, M. (1964). *The Feynman lectures on physics*. Redwood City: Addison-Wesley.
- Finkelstein, A. & Badretdinov, A. (1997). Physical reasons for a rapid folding of stable protein structures: a solution of Levinthal's paradox. *Mol. Biol. (Russia, Eng. Edition)*, 31, 391–398.
- Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E. et al. (n.d.). Gaussian~09 Revision E.01. Gaussian Inc. Wallingford CT 2009.
- Galaktionov, S., Nikiforovich, G., & Marshall, G. (2001). Ab initio modeling of small, medium, and large loops in proteins. *Biopolymers*, 60(2), 153–168.
- Goodsell, D. & Olson, A. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins:Structure, Function and Genetics*, 8, 195–202.
- Hollup, S., Salensminde, G., & Reuter, N. (2005). WEBnm@: a web application for normal mode analyses of proteins. *BMC Bioinformatics*, 6, 52.
- Jain, A. (2009). Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J. Comput-Aided Mol. Des.* 23(6), 355–374.
- Jensen, K. & Wirth, N. (1975). *Pascal user manual and report, second edition*. Lecture Notes in Computer Science. Springer.
- Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Fold Des.* 2(4), S69–S75.
- Kelley, L. & Sternberg, M. (2009). Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols*, 4, 363–371.
- Kernighan, B. & Ritchie, D. (1978). *The C programming language*. Englewood Cliffs, NJ: Prentice Hall.

- Liwo, A., Khalili, M., & Scheraga, H. (2005). Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA*, 102(7), 2362–2367.
- MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., ... Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*. 102(18), 3586–3616.
- McCammon, A., Gelin, B., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267, 585–590.
- Morse, P. & Feshbach, H. (1958). *Methods of theoretical physics*. New York: McGraw-Hill.
- Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: progress and new directions in round xi. *Proteins: Structure, Function and Bioinformatics*, 84(S1), 4–14.
- Ngo, J. & Marks, J. (1992). Computational complexity of a problem in molecular structure prediction. *Protein Engineering*, 5(4), 313–321.
- Olsson, H., Søndergard, C., Rostkowski, M., & Jensen, J. (2011). PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J Chem Theory Comput.* 7, 525–537.
- Onufriev, A., Case, D., & Bashford, D. (2002). Effective Born radii in the generalized born approximation: the importance of being perfect. *J Comput Chem.* 23(14), 1297–1304.
- Orengo, C. & Thornton, J. (1993). Alpha plus beta folds revisited: some favoured motifs. *Structure*, 1(2), 105–120.
- Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., & Ferrin, T. (2004). UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem.* 25(13), 1605–1612.
- Sali, A. & Blundell, T. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 374, 461–491.
- Schmidt, M., Baldridge, K., Boatz, J., Elbert, S., Gordon, M., Jensen, J., ... Montgomery, J. (1993). General atomic and molecular electronic structure system. *J. Comput. Chem.* 14, 1347–1363.
- Skjærven, L., Yao, X., Scarabelli, G., & Grant, B. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics*, 15, 399.
- Skoromnik, O., Feranchuk, I., Leonau, A., & Keitel, C. (2017). Analytic model of a multi-electron atom. *Journal of Physics B: At. Mol. Opt.* 50(24), 245007.
- Stroustrup, B. (1985). *The C++ programming language*. Addison Wesley.
- Tovchigrechko, A. & Vakser, I. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* 34, W310–314.
- Trott, O. & Olson, A. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry* 55-461, 31, 455–461.
- Tsai, C., Maizel, J., & Nussinov, R. (2000). Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc. Natl. Acad. Sci. USA*, 97(22), 12038–12043.

- Valiev, M., Bylaska, E., Govind, N., Kowalski, K., Straatsma, T., van Dam, H., ... de Jong, W. (2010). NWChem: a comprehensive and scalable open-source solution for large scale molecular simulation. *Comput. Phys. Commun.* 181, 1477.
- Verlet, L. (1967). Computer "experiments" on classical fluids. I. thermodynamical properties of lennard-jones molecules. *Phys Rev*, 159(1), 98–103.
- Wang, Q., Canutescu, A., & Dunbrack, R. J. (2008). SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc.* 3(12), 1832–1847.
- Xiang, J. (2006). Advances in homology protein structure modeling. *Curr. Protein Pept. Sci.* 7, 217–227.
- Xu, J. & Berger, B. (2006). Fast and accurate algorithms for protein side-chain packing. *The Journal of the ACM*, 53(4), 533–557.
- Yang, J. & Zhang, Y. (2015). Protein structure and function prediction using I-TASSER. *Curr Protoc Bioinformatics*, 52, 5.8.1–5.8.15.

4.2 Системная биоинформатика



Figure 4.1: Некоторые из современных биоинформатиков

10 наиболее цитируемых современных специалистов по биоинформатике (по версии системы Google Scholar, декабрь 2018 г.)

Верхний ряд: W.Miller, G.Myers, P.Bork, S.Salzberg, D.G.Higgins; Нижний ряд: T.J.Gibson, R.Durbin, S.Altschul, S.Kumar, K.Tamura. Из них, W.Miller, G.Myers, S.Altschul - соавторы статьи с представлением программы Blast (1990), P.Bork, S.Salzberg, R.Durbin - участники проектов по определению генома человека (2001), D.G.Higgins, T.J.Gibson - соавторы статьи с описание пакета Clustal-W (1994), S.Kumar, K.Tamura - соавторы статьей с описанием пакета MEGA (версии 1-7, 1994-2007).

Выравнивание последовательностей

Алгоритмы и "классические" работы: Needleman and Wunsch 1970; Smith and Waterman 1981; Burrows and Wheeler 1994; Tatusov et al. 1994; Leontovich et al. 2002.

Пакеты программ: Blast (Altschul et al. 1990), Mafft (Katoh and Standley 2013), Muscle (Edgar 2004), Clustal (Larkin et al. 2007), Mummer (Kurtz et al. 2004), Bowtie (Langmead et al. 2009), Ublast (Edgar 2010).

Ассемблирование

Алгоритмы и "классические" работы: Lander and Waterman 1988; Idury and Waterman 1995.

Пакеты программ: Abyss (Simpson et al. 2009), SOAPdenovo (Luo et al. 2012), Spades (Bankevich et al. 2012).

Обработка экспериментов по масс-спектрометрии

Пакеты программ: Tandem (Craig and Beavis 2004), Andromeda (Cox et al. 2011), OpenMS (Röst et al. 2016), SearchGUI (Vaudel et al. 2011).

Некоторые прикладные работы: Phadke et al. 2017; Holland and Ohlendieck 2014

Обработка экспериментов RNA-Seq

Алгоритмы и программы с их реализацией: RSEM (B. Li and Dewey 2011), DeSeq (Anders and Huber 2010), edgeR (Robinson and McCarthy 2010)

Пакеты программ с "технологическими процессами": Tophat/Cufflinks (Trapnell et al. 2012), TrinityRNASeq (Grabherr et al. 2011).

Сравнение методов и прикладные работы: Anglicheau et al. 2008; Y. Chen et al. 2016; Won et al. 2014; Adamska et al. 2007; Anavy et al. 2014; Levin et al. 2016

Обработка данных в медицине

Обзорные работы: Rattan 2006; Wu et al. 2016; de Souza et al. 2015; Cuming 2009

Модели в вычислительной биологии

Издания и учебники на русском: Тимофеев-Ресовский et al. 1969; Марчук 1991; Ризниченко 2003

Некоторые из работ с обобщениями моделей: свойства детерминированного хаоса (Feigenbaum 1983), связи биологических моделей и статистической физики (Svirezhev 2000), универсальная модель квантового перехода (De Grandi and Polkovnikov 2010)

Математическая экология, классические работы, сравнение и обобщение методов: Preston 1948; Whittaker 1960; Hill 1973; Marrugan 2004; Jost 2006; McGill et al. 2007; Wittebolle et al. 2009; Eliazar and Sokolov 2012; Chiu and Chao 2014; Chao et al. 2014

Некоторые из прикладных работ по смежным темам: Остапеня 1985; Владыко и Петкевич 2001; Владыко, Счесленок и др. 2012; Belikov et al. 2005; Marchenkov et al. 2018

Нейронные сети

Классические работы: Hebb 1949; Rosenblatt 1957; Kohonen 1982; Hopfield and Tank 1985

К модели сети с двумя уровнями возбуждения: Lodish et al. 2000; Picciotto et al. 2012; Belousov et al. 2001; Irving et al. 1992; Paulus and Rothwell 2016; Ben-Ari 2014; Sum et al. 1999

Теория фракталов

Классические работы: Mandelbrot 1960; Mandelbrot 1967; Peng et al. 1994; Richardson 1961; Higuchi 1988

Обзорные работы и сравнение методов: Harte 2001; Jelinek and Fernandez 1998; Karbauskaitė and Dzemuda 2016; Seuront 2015

Некоторые из расширений и обобщений подходов: Yakovenko and Rosser 2009; Y. Xu et al. 2017; Sornette 2002; Rak et al. 2007

Эволюция и филогенетические деревья

Теория, классические работы, обоснование методов и подходов: Fitch and Margoliash 1967; Sanderson 2002; Felsenstein 1981; A. Drummond et al. 2006.

Пакеты программ: Beast (A. J. Drummond et al. 2012), RaxML (Stamatakis 2006), FastME (Lefort et al. 2015), PhyML (Guindon and Gascuel 2003), MrBayes (Huelsenbeck and Ronquist 2001), MEGA (Tamura et al. 2007).

Некоторые из прикладных работ: Cui et al. 2013; Didelot et al. 2015; Duchêne et al. 2016; Mironova et al. 2018; Потапова et al. 2018

Метагеномика

Обоснование и сравнение методов, подходов и алгоритмов: McDonald et al. 2012

Пакеты программ: QIIME (Caporaso et al. 2010), Mothur (Schloss et al. 2009), SortmeRNA (Kopylova et al. 2016), RDP (J. Cole et al. 2014), Vegan (Oksanen et al. 2007).

публикации и издания к главе, на русском

Владыко, А. & Петкевич, А. (2001). Новые и вновь появляющиеся инфекции: молекулярная эпидемиология. *Здравоохранение*, (11), 23—24.

Владыко, А., Счесленок, Е., Фомина, Е., Семижон, П., Игнатьев, Г., Школина, Т., ... Винокурова, Н. (2012). Получение и антигенная характеристика рекомбинантных нуклеокапсидных белков вирусов Ласса и Марбург. *Вопросы вирусологии*, (4), 41—44.

Марчук, Г. (1991). *Математические модели в иммунологии: вычислительные методы и эксперименты*. М.: Наука.

Остапеня, А. (1985). Органическое вещество в воде озёр и его трансформация. В *Экологическая система нарочинских озёр* (с. 229—245). Минск.

Ризниченко, Г. (2003). *Математические модели в биофизике и экологии*. Москва-Ижевск: Институт компьютерных исследований.

Тимофеев-Ресовский, Н. В., Воронцов, Н. & Яблоков, А. (1969). *Краткий очерк теории эволюции*. М.: Наука.

публикации и издания к главе, на английском

- Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11, R106.
- Anglicheau, D., Sharma, V., Ding, R., Hummel, A., Snopkowski, C., Dadhania, D., ... Suthanthiran, M. (2008). MicroRNA expression profiles predictive of human renal allograft status. *Proc Natl Acad Sci USA*, 106(13), 5330–5335.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19(5), 455–477.
- Belikov, S., Kaluzhnaya, O., Schöder, H., Krasko, A., Müller, I., & Müller, W. (2005). Expression of silicatein in spicules from the Baikalian sponge Lubomirskia baicalensis. *Cell Biol Int.* 29(11), 943–951.
- Belousov, A., O'Hara, B., & Denisova, J. (2001). Acetylcholine becomes the major excitatory neurotransmitter in the hypothalamus in vitro in the absence of glutamate excitation. *J Neurosci.* 21(6), 2015–2027.
- Burrows, M. & Wheeler, D. (1994). A block sorting lossless data compression algorithm. (Technical Report 124).
- Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335–336.
- Chao, A., Gotelli, N., Hsieh, T., Sander, E., Ma, K., Colwell, R. et al. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol Monogr*, 84, 45–67.
- Chen, Y., Lun, A., & Smyth, G. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, 5, 1438.
- Chiu, C. & Chao, A. (2014). Distance-based functional diversity measures and their decomposition: a framework based on hill numbers. *PLoS One*, 9, e100014.
- Cole, J., Wang, Q., Fish, J., Chai, B., McGarrell, D. et al. (2014). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucl. Acids Res.* 42, D633–D642.
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R., Olsen, J., & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res.* 10(4), 1794–1805.
- Craig, R. & Beavis, R. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20, 1466–1467.

- Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T. et al. (2013). Historical variations in mutation rate in an epidemic pathogen, *Yersinia pestis*. *Proc Natl Acad Sci USA*, *110*, 577–582.
- De Grandi, C. & Polkovnikov, A. (2010). Adiabatic perturbation theory: from Landau–Zener problem to quenching through a quantum critical point. In A. Chandra, A. Das, & B. Chakrabarti (Eds.), *Lecture notes in physics* (Vol. 802). Berlin, Heidelberg: Springer.
- Didelot, X., Pang, B., Zhou, Z., McCann, A., Ni, P., Li, D. et al. (2015). The role of China in the global spread of the current cholera pandemic. *PLoS Genetics*, *11*, e1005072.
- Drummond, A. J., Suchard, M., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUTi and the BEAST 1.7. *Mol. Biol. Evol.* *29*(8), 1969–1973.
- Drummond, A., Ho, S., Phillips, M., & A., R. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*(5), e88.
- Duchêne, S., Holt, K. E., Weill, F., Le Hello, S., Hawkey, J., Edwards, D. et al. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial genomics*, *2*(11), e000094.
- Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* *32*(5), 1792–1797.
- Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*, 2460–2461.
- Eliazar, I. & Sokolov, I. (2012). Measuring statistical evenness: a panoramic overview. *Physica A*, *391*, 1323–1353.
- Feigenbaum, M. (1983). Universal behavior in nonlinear systems. *Physica D Nonlinear Phenomena*, *7*(1), 16–39.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, *17*, 368–376.
- Fitch, W. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, *3760*(2), 279–284.
- Grabherr, M., Haas, B., Yassour, M. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, *29*(7), 644–652.
- Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* *52*(5), 696–704.
- Harte, D. (2001). *Multifractals: theory and applications* (4th edition.). London: Chapman and Hall.
- Hebb, D. (1949). *The organization of behavior*. New York: Wiley and Sons.
- Higuchi, T. (1988). Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, *31*, 277–283.
- Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, *54*, 427–432.
- Hopfield, J. & Tank, D. (1985). "Neural" computation of decisions in optimization problems. *Biol. Cybernet.* *52*, 141–152.
- Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: bayesian inference of phylogeny. *Bioinformatics*, *17*, 754–755.
- Idury, R. & Waterman, M. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, *2*(2), 291–306.

- Irving, J., Collingridge, G., & Schofield, J. (1992). L-glutamate and acetylcholine mobilise Ca²⁺ from the same intracellular pool in cerebellar granule cells using transduction mechanisms with different ca²⁺ sensitivities. *Cell Calcium*, 13(5), 293–301.
- Jelinek, H. & Fernandez, E. (1998). Neurons and fractals: how reliable and useful are calculations of fractal dimensions? *Journal of Neuroscience Methods*, 81, 9–18.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 114, 363–375.
- Karbauskaite, R. & Dzemuda, G. (2016). Fractal-based methods as a technique for estimating the intrinsic dimensionality of high-dimensional data: a survey. *Informatica*, 27, 257–283.
- Katoh, K. & Standley, D. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4), 772–780.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybernet.* 43(1), 59–69.
- Kopylova, E., Navas-Molina, J., Mercier, C., Xu, Z., Mahe, F., He, Y. et al. (2016). Open-source sequence clustering methods improve the state of the art. *mSystems*, 1, e00003–15.
- Kurtz, S., Phillippy, A., Delcher, A. et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5(2), R12.
- Lander, E. & Waterman, M. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3), 231–239.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.
- Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., ... Higgins, D. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), 2947–2948.
- Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10), 2798–2800.
- Leontovich, A., Brodsky, L., Drachev, V., & Nikolaev, V. (2002). Adaptive algorithm of automated annotation. *Bioinformatics*, 18(6), 838–844.
- Li, B. & Dewey, C. (2011). RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
- Lodish, H., Berk, A., Zipursky, S. et al. (2000). Molecular cell biology. (4th edition., Chap. Neurotransmitters, Synapses, and Impulse Transmission). New York: Freeman.
- Luo, R., Liu, B., Xie, Y. et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), 18.
- Mandelbrot, B. (1960). The Pareto-Levy law and the distribution of income. *International Economic Review*, 1, 76–106.
- Mandelbrot, B. (1967). How long is the coast of Britain? statistical self-similarity and fractional dimension. *Science*, 156, 636–638.
- Marchenkova, A., Petrova, D., Morozov, A., Zakharova, Y., Grachev, M., & Bondar, A. (2018). A family of silicon transporter structural genes in a pennate diatom *Synedra ulna* subsp. *danica* (Kütz.) Skabitsch. *PLoS One*, 13(8), e0203161.
- Marrugan, A. (2004). *Measuring biological diversity*. Oxford: Blackwell.

- McDonald, D., Clemente, J. C., Kuczynski, J., Rideout, J. R., Stombaugh, J. et al. (2012). The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, 1, 7.
- McGill, B., Etienne, R., Gray, J., Alonso, D., Anderson, M., Benecha, H. et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett*, 10, 995–1015.
- Naziroğlu, M., Çığ, B., & Özgül, C. (2014). Modulation of oxidative stress and ca(2+) mobilization through trpm2 channels in rat dorsal root ganglion neuron by hypericum perforatum. *Neuroscience*, 263, 27–35.
- Needleman, S. & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Oksanen, J., Kindt, P., Legendre, P., O'Hara, B., Stevens, M., & Oksanen, M. (2007). The vegan package. *Community ecology package*, 10, 631–637.
- Peng, C., Buldyrev, S., Havlin, S. et al. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, 49, 1685–1689.
- Picciotto, M., Higley, M., & Mineur, Y. (2012). Acetylcholine as a neuromodulator: cholinergic signaling shapes nervous system function and behavior. *Neuron*, 76(1), 116–129.
- Preston, F. (1948). The commonness, and rarity, of species. *Ecology*, 29, 254–283.
- Rak, R., Drożdż, S., & Kwapienie, J. (2007). Nonextensive statistical features of the Polish stock market fluctuations. *Physica A: Statistical Mechanics and its Applications*, 374, 315–324.
- Richardson, L. (1961). The problem of contiguity: an appendix to statistics of deadly quarrels. *General systems: Yearbook of the Society for the Advancement of General Systems Theory*, 61, 139–187.
- Robinson, M. & McCarthy, D. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.
- Rosenblatt, F. (1957). The perceptron - a perceiving and recognizing automaton. (Report 85-460-1).
- Röst, H., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H. et al. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods*. 13 (9): 741–8, 13(9), 741–748.
- Sanderson, M. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *mol. biol. evol.* 19, 101–109 (2002). *Mol. Biol. Evol.* 19, 101–109.
- Schiffels, S. & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 46(8), 919–925.
- Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E. et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, 75, 7537–7541.
- Seuront, L. (2015). On uses, misuses and potential abuses of fractal analysis in zooplankton behavioral studies: a review, a critique and a few recommendations. *Physica A: Statistical Mechanics and its Applications*, 432, 410–434.
- Simpson, J., Wong, K., Jackman, S., Schein, J., Jones, S., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123.

- Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1), 195–197.
- Sornette, D. (2002). *Why stock markets crash: critical events in complex financial systems*. Princeton University Press.
- Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21), 2688–2690.
- Sum, J., Leung, C., Tam, P., Young, G., Kan, W., & Chan, L. (1999). Analysis for a class of Winner-Take-All model. *IEEE Transactions On Neural Networks*, 10(1), 64–71.
- Svirezhev, Y. (2000). Thermodynamics and ecology. *Ecological Modelling*, 132(1), 11–22.
- Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24, 1569–1599.
- Tatusov, R., Altschul, S., & Koonin, E. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci.* 91, 12091–12095.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D., ... Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with TopHat and Cufflinks. *Nat Protoc.* 7(3), 562–578.
- Vaudel, M., Barsnes, H., Berven, F., Sickmann, A., & Martens, L. (2011). SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*, 11(5), 996–999.
- Whittaker, R. (1960). Vegetation of the siskiyou mountains, Oregon and California. *Ecol Mono*, 30, 279–338.
- Wittebolle, L., Marzorati, M., Clement, L., Ballo, A., Daffonchio, D., Heylen, K. et al. (2009). Initial community evenness favours functionality under selective stress. *Nature*, 2014(458), 623–626.
- Xu, Y., Wang, Y., Tao, X., & Lizbetinova, L. (2017). Evidence of Chinese income dynamics and its effects on income scaling law. *Physica A*, 147, 143–152.
- Yakovenko, V. & Rosser, J. (2009). Statistical mechanics of money, wealth, and income. *Rev Mod Phys*, 81, 1703–1725.
- Yao, Y., Kang, T., Jin, L., Liu, Z., Zhang, Z., Xing, H., ... Li, M. (2019). Temperature-dependent growth and hypericin biosynthesis in hypericum perforatum. *Plant Physiol Biochem*. 139, 613–619.

публикации и издания авторов, на русском и белорусском

- Потапов, В., Потапова, У., Феранчук, С., Приставка, А. & Беликов, С. (2011). *Решение задач биоинформатики при помощи веб- и интернет-сервисов (методическое пособие)*. Иркутск: ИГУ.
- Потапова, У., Феранчук, С., Беликов, С. & Леонова, Г. (2018). Сравнительный анализ белка ns5 штаммов трёх субтипов вируса клещевого энцефалита. *Acta Biomedica Scientifica*, 3(6), 36–47.
- Феранчук, С. (2001). Самаузгодненае аписанне дыфузнага рассейвання рэнтгенауских промня. *Веснік БГПУ, сер.1. Фізіка*, (1), 4—6.

публикации и издания авторов, на английском

- Brodsky, L., Vasiliev, A., Kalaidzidis, Y., Osipov, Y., Tatuzov, R., & Feranchuk, S. (1992). GeneBee: the program package for biopolymer structure analysis. *Dimacs*, 8, 127–139.
- Feranchuk, I. & Feranchuk, S. [S.I.]. (2007). Self-localized quasi-particle excitation in quantum electrodynamics and its physical interpretation. *Sigma*, 2, 117.
- Feranchuk, S. [S.I.]. (2010). De novo protein structure prediction by simulation of folding pathways. *Materials Physics and Mechanics*, 9, 162–166.
- Feranchuk, S. [S.I.], Belkova, N., Potapova, U., Kuzmin, D., & Belikov, S. (2018). Evaluating the use of diversity indices to distinguish between microbial communities with different traits. *Research in Microbiology*, 169, 245–261.
- Feranchuk, S. [S.I.], Potapova, U., Chernogor, L., Belkova, N., & Belikov, S. (2018). Microevolution processes are detected in symbiotic microbiomes of Baikal sponges by the methods of fractal theory. *Limnology and Freshwater Biology*, (2), 122–134.
- Kuzmin, D., Feranchuk, S., Sharov, V., Cybin, A., Makolov, S., Putintseva, Y., ... Krutovsky, K. (2019). Stepwise large genome assembly approach: a case of Siberian larch (*Larix sibirica* Ledeb.). *BMC Bioinformatics* 2019, 20(Suppl 1):37, 20(Suppl 1), 36–52.
- Mironova, L., Gladkikh, A., Ponomareva, A., Feranchuk, S., Bochalgin, N., Basov, E., ... Balakhonov, S. (2018). Comparative genomics of *Vibrio cholerae* El Tor strains isolated at epidemic complications in Siberia and at the Far East. *Infection, Genetics and Evolution*, 60, 80–88.
- Mukha, D., Feranchuk, S., Gilep, A., & Usanov, S. (2011). Molecular modeling of human lanosterol 14 α -demethylase complexes with substrates and their derivatives. *Biochemistry Mosc.* 76, 175–185.
- Potapova, U., Feranchuk, S. [S.I.], Leonova, G., & Belikov, S. (2018). The rearrangement of motif F in the flavivirus RNA-directed RNA polymerase. *International Journal of Biological Macromolecules*, 108, 990–998.
- Potapova, U., Feranchuk, S. [S.I.], Potapov, V., Kulakova, N., Kondratov, I., Leonova, G., & Belikov, S. (2012). NS2B/NS3 protease: allosteric effect of mutations associated with the pathogenicity of tick-borne encephalitis virus. *J. Biomol. Struct. Dyn.* 30(6), 638–651.