Национальная академия наук Украины
Институт кибернетики им. В.М. Глушкова

На правах рукописи

Белецкий Борис Александрович

УДК 519.217.2

ЭФФЕКТИВНОСТЬ БАЙЕСОВСКИХ МЕТОДОВ РАСПОЗНАВАНИЯ

01.05.01 – теоретические основы информатики и кибернетики

Диссертация на соискание научной степени
кандидата физико-математических наук

Научный руководитель,
доктор физ.-мат. наук, профессор
Гупал Анатолий Михайлович

Киев – 2007

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В последние годы возрос интерес к индуктивным методам в биологии и генетике. Этот фактор можно объяснить чрезвычайной сложностью объектов и процессов исследуемых в этих направлениях, что делает практически невозможным применение существующих детерминированных математических моделей. Например, геном простейшей бактерии состоит из порядка 500 тыс. пар нуклеотидов, которые полностью описывают строение, метаболизм, поведение этого микроорганизма. Функция белка в клетке зависит от его структуры, которая, в свою очередь, зависит от сотен, тысяч, десятков тысяч аминокислот, входящих в состав данного белка. Кроме того, процессы, протекающие на молекулярном уровне не достаточно хорошо изучены для построения надежных моделей.

С другой стороны, на сегодня существует множество открытых баз данных генетической информации, в которых в оцифрованном виде хранятся секвенированные геномы различных организмов, аминокислотные последовательности и структура белковых молекул. Несмотря на то, что эта информация получается с помощью чрезвычайно дорогостоящих экспериментов, количество записей в таких базах данных растет экспоненциально [1].

Информация из генетических баз данных используется в качестве обучающих выборок при построении всевозможных индуктивных процедур (machine learning approach) [1]. Индуктивный подход кардинально отличается от дедуктивного: он не требует знания уравнений причинно-следственных связей. В индуктивном подходе невозможно получить точные значения интересующих нас характеристик объекта или предсказать исход следующего эксперимента, поскольку известна информация только о конечном числе наблюдений и исходов предыдущих экспериментов. Индуктивные процедуры дают приближенное решение задачи, зато решение получается за один шаг вычислений.

Важным моментом при построении индуктивных процедур является исследование их эффективности. Ввиду того, что обучающие выборки (особенно в генетике) получаются в результате проведения множества дорогостоящих экспериментов, хотелось бы знать заранее, какой объем обучающей выборки необходим для эффективной работы конкретной процедуры. На сегодняшний день большинство методов в биоинформатике применяются без обоснования [1].

Актуальность темы

В настоящее время разработано много методов распознавания и прогнозирования, однако эффективность этих методов в достаточной мере не исследовалась. В работе исследуется эффективность байесовских процедур распознавания на однородных цепях Маркова в зависимости от размеров обучающих выборок и количества признаков исследуемых объектов. Полученные результаты актуальны в связи с тем, что, как оказывается, геномы и последовательности белков эффективно описываются моделью в виде однородной цепи Маркова.

Связь работы с научными программами, планами, темами

- ІП.235.07 Теоретические основы и принципы построения эффективных процедур индуктивного вывода з полиномиальными оценками погрешности 2003-2005 г., № 0103U003266;
- ВФ.235.08 Разработка математических моделей для поиска структурных закономерностей записи генетической информации на основе статистического анализа геномов 2006-2010 г., № 0106U000548.
- ВК.245.21 Обеспечение повышения надежности функционирования аппаратно-программных средств суперкомпьютерных систем СКИТ-1 и СКИТ-2, разработка интеллектуальных информационных технологий и прикладного программного обеспечения для решения сложных задач в экономике, нац. Безопасности, обороне Украины а

также в других отраслях народного хозяйства, 2005 г., №0105U005692.

- ВК.245.19 Разработка и создание ряда высокопродуктивных интеллектуальных ЭВМ на основе кластерных архитектур для решения традиционных вычислительных задач и задач искусственного интеллекта, 2006р., №0104U007390.

Цель и задачи исследования.

Цель работы – построить эффективные процедуры распознавания вторичной структуры белка и исследовать их сложность в зависимости от входа задачи: размеров обучающей выборки и количества признаков.

В основе исследований лежат идеи и положения теории сложности процедур распознавания, построенной сотрудниками Института кибернетики имени В.М. Глушкова [2,3], а также классические результаты Т. Андерсона и Л. Гудмена в исследовании цепей Маркова [4].

Объект исследования – аминокислотные последовательности белков. Длина отдельных белков достигает нескольких десятков тысяч. Белки строятся из 20 различных аминокислот, каждая из которых кодируется триплетом (или несколькими триплетами) нуклеотидов. Основной особенностью исследования аминокислотных последовательностей является то, что частоты соседних остатков не являются независимыми.

Принимая во внимание специфику аминокислотных последовательностей, необходимо было обобщить оценки сложности байесовской процедуры распознавания, полученные в [2] на дискретный случай, а также использовать модель описания белков, учитывающую зависимость соседних аминокислотных остатков.

Отдельно рассматривался вопрос выбора модели описания аминокислотных последовательностей. Поскольку белки синтезируются путем трансляции информации записанной в ДНК, необходимо было провести статистический анализ записи белков в геномах. Важным моментом

при обосновании выбора модели для описания белков является закон комплементарность по одной цепочке ДНК (второе правило Чаргаффа).

Кроме того, необходимо было решить вопрос выбора параметров модели таких как стационарность и порядок цепи. Показано, что проблема выбора параметров модели сводится к решению серии задач распознавания гипотез с применением критерия χ^2.

Научная новизна полученных результатов

В работе рассмотрен случай использования байесовской процедуры распознавания на объектах с зависимыми признаками $x_1, x_2, ..., x_n$. Т.е. $P(x_1, x_2, ..., x_n) \neq P(x_1)P(x_2), ..., P(x_n)$. В качестве модели описания объектов используются цепи Маркова. Отдельно рассматривается вопрос соответствия описаний объектов выбранной модели. Показано, что генетическая информация, а следовательно, и аминокислотные последовательности белков, эффективно описываются моделями цепей Маркова.

Для того чтобы исследовать эффективность процедур распознавания на цепях Маркова (случай зависимых испытаний) и обосновать применение этих процедур на практике, необходимо было изучить поведение оценок переходных вероятностей. Как и в схеме Бернулли для независимых признаков переходные вероятности заменяются частотами. В отличии от схемы Бернулли математические ожидания оценок переходных вероятностей не совпадают с их точными значениями (оценки являются смещенными). Доказано, что оценки переходных вероятностей асимптотически нормальны и получены дисперсии и ковариации этого предельного распределения.

Показано, что верхняя оценка погрешности байесовской процедуры распознавания на однородных цепях Маркова совпадает с оценкой для независимых признаков. Принципиально, что этот результат получен для произвольных распределений переходных вероятностей.

В процессе построения процедур распознавания вторичной структуры белка проводился статистический анализ геномов бактерий и растений.

Исследовались соотношения комплементарности по одной цепочке ДНК. Показано, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова, совпадают. Показано, что для стационарных цепей Меркова вероятность нуклеотидной последовательности совпадает с вероятностью комлементарной последовательности.

Относительно простая структура геномов бактерий позволила провести статистический анализ записи белок-кодирующих генов. Для стационарных цепей показано, что переходные вероятности подсчитанные по белок-кодирующим участкам совпадают для комплементарных нитей ДНК.

Практическое значение полученных результатов

Проведенные исследования эффективности байесовских процедур распознавания на однородных цепях Маркова и полученные полиномиальные оценки погрешности позволяют сделать вывод о том, что эти процедуры могут быть широко использованы для решения прикладных задач распознавания и прогнозирования.

Полученные результаты применялись, в частности, для решения задачи предсказания вторичной структуры белка. Задача ставиться следующим образом: необходимо по поступившей на вход аминокислотной последовательности и имеющимся последовательностям с уже известной вторичной структурой (т.е. с помощью обучающей выборки) определить вторичную структуру заданного белка. В качестве обучающей выборки использовались последовательности с экспериментально установленной вторичной структурой, хранящиеся в свободном доступе на интернет-ресурсе NCBI[5]. Следует отметить, что экспериментальное определение вторичной структуры отдельного белка – дорогостоящая процедура, основанная на применении методов магнитно-ядерного резонанса и рентгено-структурного анализа.

По состоянию на апрель 2006 г. в базе данных NCBI находилось более 80000 белков с известной вторичной структурой. Основным недостатком в

организации подобных ресурсов является дублирование данных. После соответствующей обработки удалось сформировать множество из порядка 20000 уникальных белков с известной вторичной структурой, которые использовались в качестве обучающей выборки.

Применение кластера Института кибернетики им. В.М. Глушкова позволило сравнить предсказанную в результате применения метода вторичную структуру с экспериментально установленной для всех 20000 белков из обучающей выборки.

В процессе выбора модели описания аминокислотных последовательностей белков проводился статистический анализ геномов бактерий и растений. Исследования проводились на фактических данных с интернет-ресурса NCBI. Было проанализировано геномы более 50 бактерий и двух растений.

Публикации

Результаты работы опубликованы в шести международных научных журналах, в одном сборнике научных трудов Института кибернетики имени В.М. Глушкова НАН Украины и одних тезисах международной научно-практической конференции.

Личный вклад соискателя

В работе [6] автором реализовано алгоритм для подсчета частоты переходных вероятностей последовательностей ДНК бактерий, произведены численные расчеты (проанализировано более 50 геномов), выведены соотношения комплементарности основ для одной цепочки ДНК. В работе [7] проведен статистический анализ геномов двух растений. Разработано программное обеспечение для анализа больших геномов. В работе [8] проводился анализ записи белок-кодирующих участков в геномах бактерий. Выведены новые соотношения комплементарности для белок-кодирующих генов. Разработано соответствующее программное обеспечение. В работе [9] получена нижняя оценка байесовской процедуры распознавания для случая конечного числа значений признаков исследуемого объекта. В работе [10]

байесовская процедура распознавания применяется для распознавания вторичной структуры белков. В работе [11] байесовская процедура распознавания применяется для определения вторичной структуры пар аминокислотных остатков. В работах [12, 13] проведен общий анализ методов распознавания вторичной структуры белков. Обосновано использование цепей Маркова в качестве модели описания аминокислотных последовательностей белков.

РАЗДЕЛ 1
ОБОЗРЕНИЕ ЛИТЕРАТУРЫ И ВЫБОР НАПРАВЛЕНИЙ ИССЛЕДОВАНИЙ

Наиболее полный обзор по методам распознавания и обучения содержится в монографии К.В. Воронцова [14]. Поскольку диссертационная работа посвящена исследованию эффективности методов распознавания, то в обзоре затрагиваются вопросы современной теории обобщающей способности методов распознавания, при изложении будем придерживаться обозначений работы [14, 15].

1.1. Задачи обучения

Основные понятия и определения. Пусть имеются множество объектов X, множество ответов Y, и существует целевая функция $y^*: X \to Y$, значения которой $y_i = y^*(x^i)$ известны только на конечном подмножестве объектов $\{x^i,...,x^l\} \subset X$. Пары «объект–ответ» (x^i, y_i) называются прецедентами. Совокупность пар $X^l = (x^i, y_i)_{i=1}^l$ называется обучающей выборкой.

Задача обучения заключается в том, чтобы восстановить функциональную зависимость между объектами и ответами, то есть построить отображение $a: X \to Y$, удовлетворяющее следующей совокупности требований:

- отображение a должно допускать эффективную компьютерную реализацию. По этой причине будем называть его алгоритмом;
- алгоритм $a(x)$ должен воспроизводить на объектах выборки заданные ответы: $a(x^i) = y_i, i = 1,...,l$. Равенство здесь может пониматься как точное или как приближённое, в зависимости от конкретной задачи;

- на алгоритм $a(x)$ могут накладываться разного рода априорные ограничения, например, требования непрерывности, гладкости, монотонности, и т. д., или сочетание нескольких требований. В некоторых случаях может задаваться модель алгоритма – функциональный вид отображения $a(x)$, определённый с точностью до параметров;
- алгоритм a должен обладать обобщающей способностью, то есть достаточно точно приближать целевую функцию $y^*(x_i)$ не только на объектах обучающей выборки, но и на всём множестве X.

Вероятностная постановка задачи. Постановка задачи приводилась в упрощенном виде, упуская из виду, что элементы множества X – это не реальные объекты, а лишь их описания, содержащие доступную часть информации об объектах. Точные описания могут достигаться при достаточно больших количествах признаков, определяющих объект. Невозможно исчерпывающим образом охарактеризовать, скажем, человека, геологический район, производственное предприятие или экономику страны. Поэтому одному и тому же описанию x могут соответствовать различные объекты, а, значит, и целое облако ответов $y^*(x)$. Для формализации этих соображений и вводится вероятностная постановка задачи.

Вместо существования неизвестной целевой функции $y^*(x)$ предполагается, что существует неизвестное вероятностное распределение $p(x, y)$ на множестве $X \times Y$, согласно которому сгенерирована выборка пар $X^l = (x^i, y_i)_{i=1}^l$, x – непрерывный вектор $x = (x_1, x_2, ..., x_n)$ размерности n, y принимает два значения нуль и единица.

Следующий шаг в постановке наиболее важный. Он направлен на то, чтобы придать точный смысл тому, как выбираются наблюдения, по которым строится правило для классификации и определяется качество построенного правила. Принято считать, что на пространстве векторов X существует неизвестная нам вероятностная мера $P(x)$. В соответствии с $P(x)$ случайно и

независимо появляются ситуации *x*, которые классифицируются с помощью правила $p(y \mid x)$, т.е. строится обучающая последовательность $X^l = (x^i, y_i)_{i=1}^l$. Для всякого решающего правила $a(x)$ определяет качество обучения как вероятность различной классификации с помощью правила $a(x)$ и правила $p(y \mid x)$. Чем меньше эта вероятность, тем выше качество обучения. Формально качество решающего правила можно записать в виде

$$I(a(x)) = \int (a(x) - y)^2 P(x, y) dxdy \ . \qquad (1.1)$$

Объект *x* является вектором размерности *n*, а *y* принимает два значения. Минимизация среднего риска (1.1) является обобщением классических задач, решаемых на основе метода наименьших квадратов, т.е. когда наблюдению $x = (x_1, x_2, ..., x_n)$ соответствует не одно, а несколько состояний объектов (исходов экспериментов). В.Н. Вапник был одним из первых, кто придал задачам распознавания строгую математическую трактовку. Многие исследователи считают, что задачи распознавания сводятся к минимизации среднего риска (1.1) в специальном классе решающих правил [19].

С другой стороны, существует и другая точка зрения: специалисты видят проблему в том, чтобы найти такие описания объектов, при которых можно построить эффективные и даже оптимальные процедуры распознавания.

Существуют и другие подходы, в частности, теория возможности Пытьева [17] и теоретико-множественный подход Трауба, Васильковского и Вожьняковского [18].

1.2. Теория минимизации эмпирического риска

Если *x* – непрерывный вектор $x = (x_1, x_2, ..., x_n)$ размерности *n*, а *y* принимает два значения, то мощность множества решающих правил составляет величину

$$2^{\aleph^n} = 2^{\aleph} > \aleph,$$

где $\aleph$ - мощность континуума, т.е. эта мощность бесконечна и превосходит мощность континуума. Поэтому чтобы получить содержательные результаты, рассматривается ситуацию, когда мера разнообразия класса решающих функций мала (например, конечна) по сравнению с объемом выборки. Задается параметрическое множество функциональных зависимостей $F(x,\alpha)$ (класс решающих правил). Все функции класса $F(x,\alpha)$ – характеристические, т.е. принимают только два значения нуль и единица.

Рассматривается задача минимизации среднего риска

$$I(\alpha) = P(\alpha) = \int (y - F(x,\alpha))^2 P(x,y)dxdy \qquad (1.2)$$

по эмпирическим данным

$$x^1, y_1, ..., x^l, y_l \ . \qquad (1.3)$$

Функционал (1.2) для каждого решающего правила определяет вероятность ошибочной классификации. Изучается принцип минимизации эмпирического риска, согласно которому за точку минимума (1.2) принимается точка минимума эмпирического функционала

$$I_э(\alpha) = \nu(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - F(x^i,\alpha))^2, \qquad (1.4)$$

который вычисляется по случайной независимой выборке (1.3). Пусть минимум функционала (1.4) достигается на функции $F(x,\alpha_э)$. Необходимо

установить, в каких случаях функция $F(x,\alpha_э)$ близка к функции $F(x,\alpha_0)$, которая минимизирует (1.2) в классе $F(x,\alpha)$.

В работах В.Н. Вапника эта проблема связывается с проблемой существования равномерной сходимости средних к математическим ожиданиям, т.е. с ситуацией, когда для любой заданной величины уклонения может быть указано неравенство

$$P\left\{\sup_{\alpha}\left|I(\alpha)-I_э(\alpha)\right|>\varepsilon\right\}<\eta. \qquad (1.5)$$

Из (1.5) вытекает неравенство

$$P\{I(\alpha_э)-I(\alpha_0)>2\varepsilon\}<\eta. \qquad (1.6)$$

Согласно классическим теоремам теории вероятностей частота появления любого события сходится к вероятности этого события при неограниченном увеличении числа испытаний. Формально это означает, что для любых фиксированных α и ε имеет место соотношение

$$\lim_{l\to\infty} P\{|P(\alpha)-\nu(\alpha)|>\varepsilon\}=0. \qquad (1.7)$$

Однако из условия (1.7) не следует, что правило, которое минимизирует (1.4), будет доставлять функционалу (1.2) значение, близкое к минимальному. Для достаточно больших l близость найденного решения к наилучшему следует из более сильного условия, когда для любого ε выполняется равенство

$$\lim_{l\to\infty} P\left\{\sup_{\alpha}\left|P(\alpha)-\nu(\alpha)\right|>\varepsilon\right\}=0. \qquad (1.8)$$

Соотношение (1.8) – определение равномерной сходимости частот появления событий к их вероятностям по классу событий $S(\alpha)$. Каждое событие $S(\alpha^*)$ в классе $S(\alpha)$ задается решающим правилом $F(x,\alpha^*)$ как множество пар *x*, *y*, на которых выполняется равенство $\left(y - F(x,\alpha^*)\right)^2 = 1$.

Определение равномерной сходимости в виде (1.8) является неклассическим, поскольку в нем присутствует предельный переход $l \to \infty$. С практической точки зрения такой предельный переход выполнить нельзя, так как невозможно произвольно увеличивать объем выборки. В булевском случае объем выборки ограничен величиной 2^{n+1}, поскольку по построению обучающая выборка – множество пар наблюдений описаний объектов *x* и их состояний и повторы в ней исключены. Более того, в реальных задачах распознавания обучающие выборки фиксированы.

В.Н. Вапник показал, что равномерная сходимость частот появления событий к их вероятностям вытекает из условия $l \to \infty$. В начале они были выведены для простого случая, когда множество функций $F(x,\alpha)$ конечно и состоит из *N* правил:

$$F(x,\alpha_1),\ldots,F(x,\alpha_N).$$

Каждому решающему правилу $F(x,\alpha_i)$ может быть поставлено в соответствие событие A_i, состоящее из тех пар *x*, *y*, на которых $\left(y - F(x,\alpha_i)\right)^2 = 1$, т.е. определено конечное число *N* событий $A_1,\ldots,A_N$.

Для каждого фиксированного события справедлив закон больших чисел (частота сходится к вероятности при неограниченном увеличении числа испытаний). Одним из конкретных выражений этого закона является оценка (неравенство Берштейна)

$$P\left\{\left|P(\alpha_i) - \nu(\alpha_i)\right| > \varepsilon\right\} < \exp\left\{-\varepsilon^2 l\right\}, \qquad (1.9)$$

поскольку дисперсия отдельного члена в эмпирическом риске (1.4) не превосходит 1/4.

Из неравенства (1.9) следует соотношение

$$P\left\{\sup_i \left|P(\alpha_i) - \nu(\alpha_i)\right| > \varepsilon\right\} < N \exp\left\{-\varepsilon^2 l\right\}. \qquad (1.10)$$

В булевом случае число функций составляет величину 2^{2^n}, поэтому неравенство (1.10) при фиксированном n выполняется при значениях l, которые превосходят величину 2^{n+1}, что противоречит определению обучающей выборки. Следовательно, число функций N должно быть существенно меньше величины 2^{2^n}.

Из неравенства (1.10) вытекает, что для конечного числа событий выполняется равномерная сходимость частот появления событий к их вероятностям, т.е. справедливо

$$\lim_{l\to\infty} P\left\{\sup_i \left|P(\alpha_i) - \nu(\alpha_i)\right| > \varepsilon\right\} = 0.$$

Теорема 1.1. Пусть множество решающих правил состоит из N элементов, и пусть для решающих правил $F(x,\alpha_i)$ частоты ошибок на обучающей последовательности длины l равны $\nu(\alpha_i)$. Тогда с вероятностью $1-\eta$ можно утверждать, что одновременно для всех решающих правил выполняются неравенства

$$\nu(\alpha_i) - \sqrt{\frac{\ln N - \ln\eta}{l}} < P(\alpha_i) < \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln\eta}{l}}. \qquad (1.11)$$

Так как неравенства справедливы для всех N правил, то теорема 1.1 устанавливает доверительный интервал для качества решающего правила

$F(x,\alpha_э)$, которое минимизирует среди N правил эмпирический риск. Он равен

$$\nu(\alpha_э)-\sqrt{\frac{\ln N-\ln\eta}{l}}<P(\alpha_э)<\nu(\alpha_э)+\sqrt{\frac{\ln N-\ln\eta}{l}}.$$

По мнению В.Н. Вапника, дальнейшее развитие теории минимизации эмпирического риска состоит в обобщении этих теорем на случай бесконечного числа решающих правил.

Пусть задано множество S решающих правил $F(x,\alpha)$ и дана выборка $x^1,...,x^l$. Эта выборка, вообще говоря, может быть разделена на два класса 2^l способами. (С помощью правила $F(x,\alpha)$ множество $x^1,...,x^l$ делится на два подмножества – подмножество, на котором $F(x,\alpha)=1$, и подмножество, на котором $F(x,\alpha)=0$.) Число таких способов разделения зависит как от класса решающих правил $F(x,\alpha)$, так и от состава выборки. Обозначим это число величиной $\Delta^s(x^1,...,x^l)$.

Функция $m^S(l)=\max\limits_{x^1,...,x^l}\Delta^s(x^1,...,x^l)$ называется функцией роста системы событий, образованной решающими правилами $F(x,\alpha)$, где максимум берется по всем возможным выборкам длины l. Функция роста имеет простой смысл: она вычисляет максимальное число способов разделения l точек на два класса с помощью решающих правил.

Функция роста либо тождественно равна 2^l, либо при $l>h$ она мажорируется функцией

$$m^S(l)<1{,}5\frac{l^h}{h!},$$

где $h+1$ – минимальный объем выборки, при котором нарушается условие $m^S(l)=2^l$. Число h служит мерой разнообразия класса решающих правил.

Класс характеристических функций имеет емкость h, если справедливо неравенство

$$m^S(l) < 1{,}5\frac{l^h}{h!},\ l > h.$$

В случае выполнения равенства $m^S(l) = 2^l$ говорят, что емкость класса характеристических функций $F(x,\alpha)$ бесконечна. Если емкость класса характеристических функций конечна, то имеет место равномерная сходимость частот к вероятностям.

Нетрудно найти функцию роста для класса событий, заданных линейными решающими правилами:

$$F(x,\alpha) = \theta\left(\sum_{i=1}^{n}\alpha_i\varphi_i(x)\right); \qquad \theta(z) = \begin{cases}1, & \text{если } z \ge 0,\\ 0, & \text{если } z < 0\end{cases}. \tag{1.12}$$

Для этого достаточно определить максимальное число точек h в пространстве размерности n, которые можно с помощью гиперплоскости разбить на два класса всеми 2^h способами. Известно, что это число равно n. Поэтому для класса линейных решающих правил (1.12) функция роста оценивается величиной $m^s(l) < 1{,}5\frac{l^n}{n!}$ $(l > n)$.

В.Н. Вапник получил оценку скорости равномерной сходимости частот к вероятности по классу событий $S(\alpha)$. Показано, что имеет место неравенство

$$P\left\{\sup_{\alpha}\left|P(\alpha) - \nu(\alpha)\right| > \varepsilon\right\} < 6m^s(2l)\exp\left\{-\frac{\varepsilon^2 l}{4}\right\}. \tag{1.13}$$

Оценка (1.13) становится содержательной, когда емкость класса решающих правил конечна

$$m^S(l) < 1{,}5\frac{l^h}{h!}.$$

В таком случае имеет место следующая теорема.

Теорема 1.2. Пусть $F(x,\alpha)$ – класс решающих правил ограниченной емкости h, и пусть $\nu(\alpha)$ – частота ошибок, вычисленная по обучающей последовательности для правила $F(x,\alpha)$. Тогда с вероятностью $1-\eta$ для всех правил $F(x,\alpha)$ вероятность ошибочной классификации заключена в пределах

$$\nu(\alpha) - \sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{9}}{l}} < P(\alpha) < \nu(\alpha) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{9}}{l}} \qquad (1.14)$$

Из теоремы 1.2 следует, что для правила $F(x,\alpha_э)$, которое минимизирует эмпирический риск, с вероятностью $1-\eta$ справедлива оценка сверху

$$P(\alpha_э) < \nu(\alpha_э) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{9}}{l}}. \qquad (1.15)$$

В вероятностных оценках (1.14), (1.15) присутствует параметр η. Необходимо задать этот параметр, подставить в формулы и вычислить оценку погрешности. Для достижения высокой надежности параметр η следует положить близким к нулю, и вследствие этого оценки принимают завышенный характер.

В работе [15] указан метод, который, казалось бы, минимизирует эмпирический риск до нуля, но при этом абсолютно не способен обучаться.

Получив обучающую выборку X^l, он запоминает ее и строит алгоритм, который сравнивает предъявляемый объект x с обучающими объектами x^i из X^l. В случае совпадения $x = x^i$ алгоритм выдает ответ y_i. Эмпирический риск принимает наименьшее возможное значение, равное нулю. Отмечается, что этот алгоритм не способен восстановить зависимость вне объектов обучения. Заметим, что этот пример для вероятностной постановки задачи лишен смысла, поскольку нам неизвестны вероятности $p(x,0)$ и $p(x,1)$. Не исключено, что в обучающей выборке могут присутствовать объекты x, имеющие одновременно оба состояния нуль и единица. В таком случае повторить ответ нельзя, так как не будет выполнено требование относительно того, что метод обучения является функцией.

Первые оценки обобщающей способности были получены в конце 60-х годов Вапником и Червоненкисом [15, 19]. Их численные значения были сильно завышены и позволяли лишь на качественном уровне обосновывать применимость некоторых методов обучения. На протяжении последующих десятилетий статистическая теория обучения развивалась активно. Современные оценки уже позволяют делать достаточно точные количественные предсказания для отдельных частных случаев [16, 31, 32], см. также обзор [33]. Однако для многих методов, успешно применяемых на практике, строгие обоснования до сих пор не получены. Вывод точных количественных оценок обобщающей способности пока остаётся открытой проблемой.

Замечания относительно изложенной теории.

Если алгоритм *a* доставляет минимум функционалу (1.4) на заданной обучающей выборке X^l, то это еще не гарантирует, что он будет хорошо приближать целевую зависимость на произвольной *контрольной выборке* $X^k = (x'^i, y'_i)_{i=1}^k$. Специфика методов минимизации эмпирического риска состоит в том, что выборка X^L разбивается на обучающую подвыборку X^l

длины l и контрольную X^k длины k, $L = l + k$. На этапе обучения на основе выборки X^l строится алгоритм $a = \mu(X^l)$ и затем его работа проверяется на контрольной выборке X^k. Естественно, что при таком сокращении выборки эффективность алгоритма уменьшается, поскольку в приведенных выше оценках фигурирует длина выборки. Для байесовских процедур распознавания, исследуемых в диссертации, разбиения выборки на обучающую и контрольную проводить не надо, поскольку при подсчете погрешности учитывается работа процедуры на всем множестве обучающих выборок.

Заметим, что все объекты (и их состояния) в обучающей выборке представляют большую ценность для процесса распознавания, они получаются, как правило, с помощью дорогостоящих процедур. Например, вторичная структура белка определяется на основе измерений с помощью рентгеноструктурного анализа и ядерного магнитного резонанса.

Теория минимизации эмпирического риска построена для непрерывных объектов $x = (x_1, x_2, ..., x_n)$ размерности n. В приведенных выше теоремах утверждается следующий факт: если мера разнообразия класса решающих правил мала по сравнению с объемом выборки, то метод минимизации эмпирического риска позволяет выбрать правило, близкое к наилучшему в ограниченном классе решающих правил. «Характерной особенностью изложенной теории минимизации эмпирического риска является полное отсутствие, каких бы то ни было, указаний на конструктивную возможность построения алгоритмов» [15]. Таким образом, теория минимизации эмпирического риска является некоторой методологической схемой рассуждений без построения конкретных алгоритмов. Поэтому окончательных выводов относительно эффективности предложенного подхода сделать нельзя.

В теории Вапника-Червоненкиса верхние оценки погрешности метода минимизации эмпирического риска получены для узкого класса

параметрических функций. Поэтому функция $a(x)$, которая минимизирует средний риск (1.1), может не принадлежать классу параметрических функций.

Наиболее важной проблемой в теории распознавания образов является оценка сложности решения задачи минимизации среднего риска (1.1) (например, для задач в дискретной постановке). Для этого необходимо получить более сильные нижние оценки погрешности процедур (а не верхние, как в теории минимизации эмпирического риска). Поэтому данный вопрос остается за рамками теории минимизации эмпирического риска.

Верхние оценки (1.14), (1.15) в теории минимизации эмпирического риска были получены на основе эмпирических данных

$$x^1, y_1, ..., x^l, y_l \ ,$$

которые являются некоторой случайной последовательностью описаний объектов и их состояний. Поэтому они носят вероятностный характер. В эту оценки входит длина выборки, емкость h и вероятностный параметр η.

Следуя канонам теории сложности алгоритмов, задача распознавания в силу своей специфики в дискретном случае определяется следующими входными параметрами: размерами классов объектов в обучающей выборке, количеством признаков и количеством значений признаков. В оценки (1.14), (1.15) размеры классов не входят. Этот факт, на наш взгляд, является ошибкой. Представим себе, что медицинская экспертная система строится только на классе больных или здоровых пациентов (или размеры этих классов отличаются друг от друга в сотни раз), понятно, что эффективных процедур распознавания, в таком случае, построить нельзя.

Для булевских векторов $x = (x_1, x_2, ..., x_n)$ число различных функций $a(x) \in \{0,1\}$ равно экспоненте 2^{2^n}. Средний риск (1.1) в булевом случае записывается в виде

$$I(a(x)) = \sum_{x \in X} \sum_{y=0}^{1} (a(x) - y)^2 P(x, y), \qquad (1.16)$$

где усреднение проводится по всем булевским векторам $x = (x_1, x_2, ..., x_n)$.

Если обучающая выборка содержит только один класс объектов, то можно построить пример, когда эмпирический риск окажется нулевым, а средний риск (1.16) будет максимальным. Пусть $n = 1$

$$P(x = 0, y = 0) = 0{,}1; \quad P(x = 0, y = 1) = 0{,}3$$
$$P(x = 1, y = 0) = 0{,}2; \quad P(x = 1, y = 1) = 0{,}4.$$

Тогда

$$a_1(x = 0) = 1, \quad a_1(x = 1) = 1;$$
$$a_2(x = 0) = 1, \quad a_2(x = 1) = 0;$$
$$a_3(x = 0) = 0, \quad a_3(x = 1) = 1;$$
$$a_4(x = 0) = 0, \quad a_4(x = 1) = 0;$$

Функция a_1 минимизирует риск (1.16), он равен 0,3. Если обучающая выборка содержит объекты только одного класса 0, и в выборке присутствуют объекты $x = 0$ или $x = 1,$ то функция a_4 дает нулевой эмпирический риск, однако у этой функции наблюдается максимальный риск (1.16), равный 0,7.

По смыслу вероятностных оценок построенные контр примеры, вообще говоря, не опровергают оценки (1.14), (1.15), поскольку для любого фиксированного l может существовать ненулевая вероятность η того, что они не выполняются. Другими словами, в диапазон вероятности η может попадать такое семейство распределений или задач, для которых размеры классов могут значительно отличаться друг от друга.

Выясним, почему в оценки В.Н. Вапника и его последователей входит длина выборки. Ясно, что эмпирический риск, как случайная величина, имеет биномиальное распределение с параметром l и вероятностью $P(a(x) \neq y)$, т.к. единица в эмпирическом риске появляется с вероятностью $P(a(x) \neq y)$ [33]. Поэтому полученные оценки очевидным образом вытекают из обобщенных неравенств Чебышева.

Чтобы исключить рассмотренные выше примеры, нужно в обучающей выборке зафиксировать размеры классов. Пусть l_0 – число объектов в обучающей выборке

$$x^1, y_1, ..., x^l, y_l \ ,$$

у которых состояние $y = 0$, а l_1 – число объектов, имеющих состояние $y = 1$. Тогда вместо эмпирического риска

$$I_э(\alpha) = \nu(\alpha) = \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2$$

вычисляются два эмпирических риска

$$\nu_{l_0}(\alpha) = \frac{1}{l_0} \sum_{l=1}^{l_0} (0 - F(x_i, \alpha))^2 \quad \text{и} \quad \nu_{l_1}(\alpha) = \frac{1}{l_1} \sum_{l=1}^{l_1} (1 - F(x_i, \alpha))^2 \ .$$

Поэтому, как легко заметить, в оценки (1.14), (1.15) вместо длины выборки l будет входить минимальный размер классов $l' = \min(l_0, l_1)$. Если в обучающей выборке отсутствует один из классов, т.е. $l' = 0$, то нужно доказать (как это сделано в диссертации), что любой алгоритм работает непредсказуемо плохо и его оценка погрешности строго положительна.

В 1995 году в работах Гупала А.М., Пашко С.В., Сергиенко И.В. построена теория статистического оценивания процедур распознавания [2].

Рассматривается задача минимизации среднего риска

$$I(a(x)) = \sum_{x \in X} \sum_{y=0}^{1} (a(x) - y)^2 P(x, y). \qquad (1.17)$$

Показано, что в булевом случае байесовская полиномиальная процедура для объектов с независимыми признаками является субоптимальной, построена верхняя и нижняя оценки погрешности процедуры в зависимости от размеров обучающей выборки, которые совпадают с точностью до абсолютной константы

Постановка задачи. Наблюдаемый объект описывается вектором $x_1, x_2, ..., x_n$, f, где $x_1, x_2, ..., x_n$ – признаки (измерения) объекта, f – состояние объекта. Пусть проведено m наблюдений над объектами и в каждом случае было зафиксировано состояние объекта. Имеем обучающую выборку $V = (V_0, V_1, V_2)$ вида (рис 1.1).

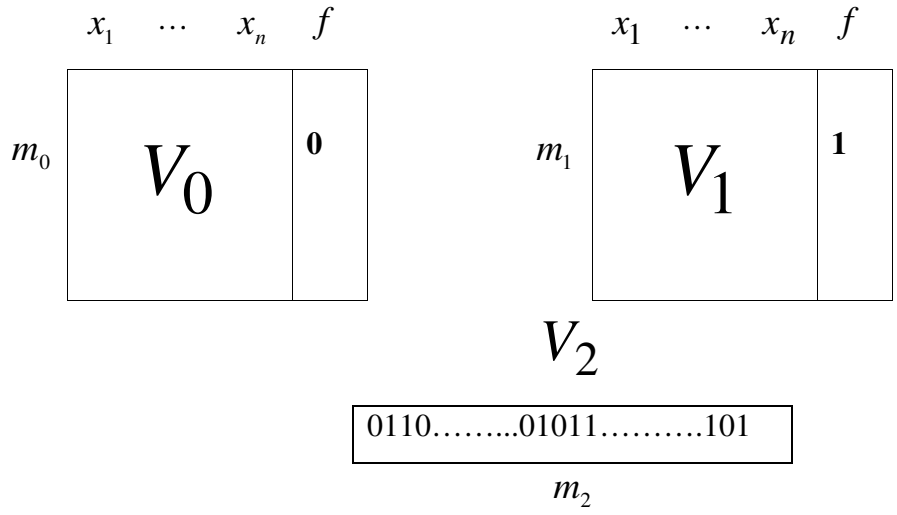

Рис.1.1 Обучающая выборка $V = (V_0, V_1, ..., V_{h-1}, V_h)$

На множестве B описаний объектов $(x_1, x_2, ..., x_n, f)$ задано некоторое распределение вероятностей P, которое заранее неизвестно. Первая часть V_0 – булева матрица размерности $m_0 \times n$, где m_0– число строк. Каждая строка представляет собой вектор ы, описывающий объект класса 0 (имеющий

состояние 0); он выбран из множества описаний в соответствии с распределением P при условии $f = 0$. Вторая часть V_1 – булева матрица размерности $m_1 \times n$, где $m_1 -$ число строк. Каждая строка матрицы представляет собой вектор x, описывающий объект класса 1; он выбран в соответствии с распределением P при условии $f = 1$. Последняя часть V_2 – булев вектор размерности m_2. Каждая компонента этого вектора – наблюдаемое значение состояния f, она выбирается в соответствии с распределением P, т.е. вероятность того, что i-я компонента вектора V_2 принимает значение 0, равна $P(f = 0)$, а значение $1 - P(f = 1)$. Поскольку наблюдения проводятся также как и в теории минимизации эмпирического, можно считать, что $m_2 = m = m_0 + m_1$.

Индуктивный шаг. Требуется построить такую процедуру индуктивного вывода, которая по измерениям $x_1, x_2, ..., x_n$ любого следующего объекта и обучающей выборке $V = (V_0, V_1, V_2)$ определяет состояние f объекта.

Поступивший на вход обучающей системы вектор $x = (x_1, x_2, ..., x_n)$ может присутствовать в выборке V_0 или в выборке V_1 или в обеих выборках вместе. Однако указать к какому классу принадлежит $x = (x_1, x_2, ..., x_n)$ нельзя, так как не известны вероятности $P(x,0)$ и $P(x,1)$. Этот вопрос решает байесовская процедура распознавания.

Изучение эффективности байесовского подхода к решению описанной задачи, а также сложности подобных задач требует формализации таких понятий, как класс задач, наилучшая функция распознавания, процедура распознавания, погрешность процедуры распознавания и т.д.

Погрешность процедуры. Сложность класса задач. Полагаем, что процесс определения состояния f объекта по известным значениям вектора входа $x = (x_1, x_2, ..., x_n)$ проводится с помощью функции $a(x)$, т.е. $f = a(x)$. Так как число функций $a(x)$ конечно, то среди них существует наилучшая

функция $a^*(x)$, такая, что $P(x,a^*(x)) = \max(P(x,0),P(x,1))$. Легко заметить, что функция $a^*(x)$ минимизирует средний риск (1.17).

Погрешность функции $a(x)$ – усредненная величина

$$\upsilon(a,P) = \sum_x P(x,a^*(x)) - P(x,a(x)), \qquad (1.18)$$

поэтому $\upsilon(a^*) \equiv 0$. Погрешность (1.18) в отличие от методов минимизации эмпирического риска указывает отклонение функции $a(x)$ от минимума среднего риска, поэтому для практических расчетов с ней удобнее работать.

Индуктивная процедура распознавания Q строит по данной выборке $V = (V_0, V_1, V_2)$ и вектору x функцию $A(x) = Q(V, x)$.

Погрешность процедуры Q на распределении P – усредненная величина

$$\upsilon(Q,P) = \sum_{V \in W} \upsilon(A,P) P_1(V). \qquad (1.19)$$

Усреднение в (1.19) проводится по всем обучающим выборкам V из некоторого множества W, $P_1(V)$ – вероятность получения выборки V. Для независимых признаков вероятность $P_1(V)$ полностью определяется распределением P. Усреднение (1.19) проводится для того, чтобы получить детерминированные оценки погрешности байесовской процедуры распознавания. Операция усреднения по своей сути выполняет функцию контроля: формула (1.19) оценивает качество работы процедуры на всевозможных новых объектах, не входящих в состав обучающей выборки. При этом у всех обучающих выборок размеры классов одинаковы.

Класс задач $C \equiv C(m_0, m_1, m_2, n)$ – совокупность всевозможных распределений вероятностей P, детерминированные числа m_0, m_1, m_2, n – вход задачи, они определяют размеры выборки.

Погрешностью процедуры распознавания Q на классе C называется число

$$\upsilon(Q, C) = \sup_{P \in C} \upsilon(Q, P).$$

Сложность класса C определяется величиной

$$\mu(C) = \inf_{Q} \upsilon(Q, C) = \inf_{Q} \sup_{p \in C} \upsilon(Q, P).$$

Нужно построить такую процедуру распознавания Q, для которой число $\upsilon(Q, C)$ мало отличается от числа $\mu(C)$.

Байесовская процедура распознавания. Пусть $d = (d_1, d_2, ..., d_n)$ – булев вектор. Считаем, что распределения P из класса C при каждом d удовлетворяют условию

$$P(x_1 = d_1, x_2 = d_2, ..., x_n = d_n \mid f = i) = \prod_{j=1}^{n} P(x_j = d_j \mid f = i),\ i = 0,1,$$

что означает независимость признаков x_j для каждого класса объектов.

Рассмотрим случайные величины $\xi(d, i)$, которые зависят от d и i как от параметров:

$$\xi(d, i) = \prod_{j=1}^{n} (k(d_j, i) / m_i) k_i / m_2, i = 0,1; \qquad (1.20)$$

Здесь $k(d_j, i)$ – количество значений, равных d_j, j-го признака в j-м столбце матрицы V_i; k_i – количество значений целевого признака, равных i, в векторе V_2. Тогда функция распознавания определяется формулой

$$A(d)=\begin{cases}0, \text{ если } \xi(d,0)\geq\xi(d,1),\\ 1, \text{ если } \xi(d,0)<\xi(d,1).\end{cases} \qquad (1.21)$$

Процедуру распознавания, определяемую соотношениями (1.20), (1.21), обозначим Q_B. Заметим, что величины $\xi(d,i)/(\xi(d,0)+\xi(d,1))$ представляют собой приближенные значения вероятностей $P(f=i|x_1=d_1,x_2=d_2,...,x_n=d_n)$, вычисленных по теореме Байеса. При подсчете оценок этих вероятностей существенным образом используется информация относительно размеров классов в обучающей выборке, поэтому эти значения входят в приводимые ниже оценки погрешности байесовской процедуры. Для байесовской процедуры распознавания проводить разбиение на обучающую и контрольную выборки не надо, поскольку при подсчете погрешности (1.19) учитывается работа процедуры на всем множестве обучающих выборок. Байесовская процедура (1.20), (1.21) строится программным образом, найти ее аналитический вид невозможно. В отличие от этого методы минимизации эмпирического риска работают с известными функциями.

В работе [2] получена оценка сверху погрешности байесовской процедуры на классе C.

Теорема 1.3. Существует абсолютная константа $\alpha<\infty$, такая, что справедливо неравенство

$$\upsilon(Q_B,C)\leq\min\left(1,a\sqrt{\frac{n}{\min(m_0,m_1)}+\frac{1}{m_2}}\right). \qquad (1.22)$$

При условии $m=\min(m_0,m_1)\geq 2n$: оценка сверху погрешности байесовской процедуры задается квадратным корнем, в противном случае (для малых выборок) она не превосходит единицы. Из доказательства теоремы вытекает, что абсолютная константа $a=4\sqrt{2}$.

В [34] доказана теорема о том, что если в обучающей выборке отсутствует один из классов, т.е. $\min(m_0, m_1) = 0$, то любая процедура, в том числе и байесовская, работает плохо и ее погрешность строго положительна.

Теорема 1.4. Для любой индуктивной процедуры Q существует такое распределение P, что ее погрешность удовлетворяет соотношению

$$\upsilon(Q, P) \geq C > 0.$$

Наиболее сложно получить оценку снизу погрешности байесовской процедуры на классе C [2, 34]. Справедлива следующая теорема.

Теорема 1.5. Существует абсолютная константа $a_1 > 0$ такая, что справедливо следующее: каковы бы ни были целые числа m_0, m_1, m_2, удовлетворяющие неравенствам $m_1 \geq m_0 \geq 0$, $m_2 \geq 0$, натуральное число n и процедура распознавания Q, существует такое распределение вероятностей P из класса C, что выполняется неравенство

$$\upsilon(Q, P) \geq a_1 \min\left(1, \sqrt{\frac{n}{m_0} + \frac{1}{m_2}}\right). \qquad (1.23)$$

Из теоремы 1.5 следует, что погрешность $\upsilon(Q_B, C)$ отличается от сложности $\mu(C)$ класса задач C не более чем в константу раз. В этом смысле байесовская процедура распознавания Q_B является субоптимальной. Таким образом, оценка (1.22) определяет сложность класса C.

Аналогичные теоремы для дискретного случая были получены в [12] .

В отличие от вероятностных оценок (1.14), (1.15), полученных в теории минимизации эмпирического риска, оценки погрешности байесовской процедуры распознавания детерминированы, и их удобно использовать для практических расчетов.

Таким образом, с учетом результатов, полученных в диссертации, проблема минимизации среднего риска (1.17) полностью решена для дискретных объектов с независимыми признаками.

В работе В.Б. Берикова и Г.С. Лбова [35] исследуется байесовская процедура распознавания на дискретном множестве событий. Рассмотрен частный случай объектов с одной переменной, при этом авторы используют определение погрешности функций распознавания из работ Вапника В.Н., Червоненкиса А.Я. [15,19]. Приводится байесовская оценка сверху вероятности ошибки

$$\frac{8M^2 + 5MN + N^2 - 4M - N}{4(N + 2M - 1)(N + 2M)}, \qquad (1.24)$$

которая имеет асимптотический характер; здесь M - число значений единственного признака, N - объем выборки. Легко заметить, что при $N \to \infty$ оценка (1.24) стремиться к $\frac{1}{4}$.

РАЗДЕЛ 2.
ЭФФЕКТИВНОСТЬ БАЙЕСОВСКОЙ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ. ДИСКРЕТНЫЙ СЛУЧАЙ

В основе дедуктивных вычислений лежит последовательный принцип организации вычислений. При доказательстве конкретной теоремы образуется цепочка истинных утверждений, которая начинается аксиомой и заканчивается выводимой теоремой. Промежуточные утверждения получаются на основе дедуктивных правил по типу «modus ponens» или метода резолюций [36]. При доказательстве конкретной теоремы нельзя заранее оценить длину этой цепочки. Скажем, знаменитая теорема Ферма будоражила умы математиков на протяжении трех столетий и, наконец, была решена Э.Уайлсом в 1995 г. Дедуктивные правила вывода имеют высокую сложность и низкую эффективность. Так, алгоритм Британского музея, который последовательно порождает все возможные выводы, может быть эффективнее известного метода резолюций [36].

В дедуктивной математике объекты и явления природы изучаются с помощью математических моделей. Эти модели формируются на основе известных уравнений о причинно-следственных связях между переменными или признаками объекта. Процесс наблюдений или экспериментов определяет конкретные значения переменных. Если не удается получить аналитическое решение задачи, то поиск точного решения строится в виде последовательной цепочки приближений.

Индуктивный подход кардинально отличается от дедуктивного: он не требует знания уравнений причинно-следственных связей. В индуктивном подходе невозможно получить точные значения интересующих нас характеристик объекта или предсказать исход следующего эксперимента, поскольку известна информация только о конечном числе наблюдений и исходов предыдущих экспериментов. В этом смысле не выполняется основная догма дедуктивной математики – получение точного решения.

Индуктивные процедуры дают приближенное решение задачи, зато решение получается за один шаг вычислений. Как и в квантовой механике нужно привлекать вероятностный аппарат. Индуктивные процедуры, как и квантовые вычисления по своей природе являются вероятностными. В квантовой механике невозможно заранее прогнозировать исход отдельного эксперимента (он объективно является случайным), остается только прогнозировать среднее число результатов большого количества экспериментов.

Общая методика исследований в диссертации основана на теории сложности процедур индуктивного вывода, разработанной в [2,3]. Развиваемый подход проводится с позиций решения задач распознавания или обучения.

2.1. Постановка задачи

Изучение природных объектов проводится на основе измерений с помощью различных измерительных устройств или приборов. Для наблюдаемого объекта строится его описание в виде вектора

x_1	x_2	$\cdots$	x_n	f

Здесь $x_1, x_2, ..., x_n$ – измерения объекта, а f – исход эксперимента или значение интересующей нас характеристики. Например, в медицине $x_i, i = 1,...,n$ – результаты анализов, а f – состояние пациента (болен или здоров). В задачах распознавания эти величины называют признаками. Рассмотрим дискретный случай, когда каждый признак и исход эксперимента принимают конечное число значений.

Пусть имеется конечное множество B описаний объектов $b = (x_1, x_2, ..., x_n, f)$; здесь n – натуральное число, $x_j \in \{0,1,..., g-1\}$, $j = 1,2,...,n$; $f \in \{0,1,..., h-1\}$; g, h – натуральные числа, $g \geq 2$, $h \geq 2$.

Предположим, на множестве B задано распределение вероятностей P, которое нам заранее не известно. Из этого множества выделена обучающая выборка V, структура и способ получения которой описаны ниже. Пусть некоторое описание объекта получено из множества B независимо от выборки V в соответствии с распределением P, причем известны только значения признаков $x_1, x_2, ..., x_n$. Требуется по этим значениям и по обучающей выборке V определить значение целевого признака f.

Индуктивный шаг. Требуется построить такую процедуру индуктивного вывода, которая по измерениям $x = (x_1, x_2, ..., x_n)$ любого следующего объекта и обучающей выборке V определяет значение целевого признака f.

В индуктивном подходе невозможно получить точное значение f, т.к. известна информация только о конечном числе наблюдений над объектами.

Основная проблема в естественных науках – уметь предсказывать исходы наблюдений, для любых измерений объекта $x_1, x_2, ..., x_n$, эксперименты в них недешевы и часто проводятся в реальном масштабе времени. Для решения этой проблемы нужны эффективные процедуры индуктивного вывода, обладающие полиномиальными оценками погрешности.

Отметим следующий основной момент: нужно отличать объекты от их описаний, на самом деле это разные понятия. Легко заметить, что для точного описания объектов может понадобиться большое число измерений, не исключено, что и бесконечное. При изучении объектов на основе конечного числа измерений, вектору $x = (x_1, ..., x_n)$ может соответствовать несколько значений f, например, один и тот же вектор x может соответствовать как больному, так и здоровому пациенту.

Чтобы различать такие описания объектов, нужно на множестве всех описаний объектов ввести некоторое распределение вероятностей P, которое нам заранее неизвестно.

2.2. Необходимые понятия: класс задач, процедура обучения, погрешность процедуры, обучающая выборка

Для того чтобы изучить вопрос об эффективности индуктивного подхода к решению описанной задачи, а также сложность подобных задач, необходимо формализовать такие понятия, как класс задач, наилучшая функция распознавания, процедура распознавания, погрешность процедуры, обучающие выборки и их вероятностные распределения.

Число булевых векторов $x = (x_1, x_2, ..., x_n)$, где $x_i \in \{0,1\}, i = 1,2,...n$, равно g^n. Они образуют множество описаний объектов M. Назовем функцией распознавания функцию $A \equiv A(x)$, определенную на множестве булевых векторов $x \in M$ и принимающую значения во множестве $\{0,1,...,h-1\}$. Известно, что число (мощность) таких функций $A(x) \in \{0,1,...,h-1\}$ составляет h^{g^n} [37].

Считаем, что процесс распознавания целевого признака f объекта по известным значениям вектора входа x проводится с помощью функции $A(x)$, т.е. полагаем $f = A(x)$. Функцию $A(x)$ целесообразно выбрать так, чтобы как можно большим было значение

$$\sum_{d \in M} P(x = d, f = A(d)) = P(f = A(x)).$$

Известно, что среди функций $A(x)$ существует в определенном смысле наилучшая функция $A^*(x)$, такая, что среди всех $A(x)$ и x выполняется неравенство $P(x, A^*(x)) \geq P(x, A(x))$. Ее можно получить лексикографическим упорядочиванием функций $A(x)$ по значению вероятности $P(x, A(x))$ [37].

Каждому булевому вектору $b=(x,f)$ приписана вероятность $P(b)$, такая что $\sum P(b)=1$. Назовем *погрешностью* функции $A(x)$ на распределении P усредненную величину

$$\upsilon(A,P)=\sum_{x\in M}P(x,A^{*}(x))-P(x,A(x)). \qquad (2.1)$$

В усреднении (2.1) присутствуют все 2^n векторов x множества M, поскольку в процессе распознавания может участвовать любой вектор из M.

Процедурой распознавания Q назовем однозначную функцию, которая определена на некотором множестве обучающих выборок $W=\{V\}$ и принимает значения во множестве функций распознавания $A(x)$; процедура Q строит функцию $A(x)$ по *выборке* V, т.е. $A(x)=Q(V,x)$. Таким образом, процедура Q – это некоторый вполне определенный принцип построения по данной выборке V функции $A=Q(V,x)$, один и тот же для всех выборок.

Погрешностью процедуры Q на распределении P назовем усредненную величину

$$\upsilon(Q,P)=\sum_{V\in W}\upsilon(A,P)P_1(V). \qquad (2.2)$$

В усреднении участвуют все выборки V из множества W.

Здесь $A=Q(V,x)$, $P_1(V)$ – вероятность получения выборки V. Этот момент – самый трудный в теории обучения, поскольку неясно, как оценить вероятность выборки $P_1(V)$. Позже мы покажем, что $P_1(V)$ полностью определяется распределением P.

Структура обучающей выборки. Подмножество объектов из B, у которых целевой признак равен i, называется i-м классом объектов. В обучающей выборке V количества объектов различных классов заданы;

более того, на практике они часто определяются заранее. Поэтому считаем, что выборка V состоит из $h+1$ части, $V=(V_0,V_1,...,V_h)$ рис.2.1. Пусть $m_0,m_1,...,m_h$ – целые неотрицательные детерминированные числа; обозначим $m^{(h)}=(m_0,m_1,...,m_h)$. В случае когда $0<m_i<\infty$ и $i<h$, часть V_i представляет собой целочисленную матрицу размерностью $m_i \times n$. Каждая строка этой матрицы является наблюдаемым значением вектора $x=(x_1,x_2,...,x_n)$, описывающего объект класса i, который выбран из множества B в соответствии с распределением вероятностей P при условии $f=i$. Используя матрицу V_i, можно вычислить частоты событий $\{x_j=s\}$, $s=0,1,...,g-1$, при условии $f=i$.

Если $0<m_h<\infty$, то последняя часть V_h представляет собой целочисленный вектор размерностью m_h. Каждая компонента этого вектора есть наблюдаемое значение целевого признака f, которое выбирается в соответствии с распределением P. Иными словами, вероятность того, что j-я компонента вектора V_h принимает значение i, равна $P(f=i)$. Все строки матрицы V_i, $i=0,1,...,h-1$ и компоненты вектора V_h являются независимыми случайными элементами.

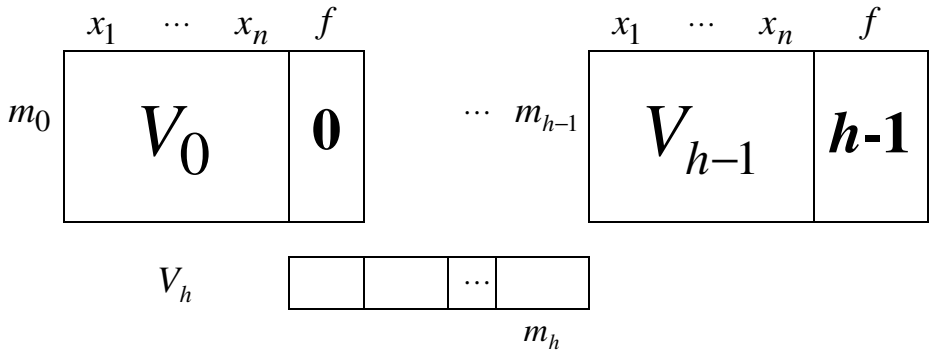

Рис.2.1 Обучающая выборка $V=(V_0,V_1,...,V_{h-1},V_h)$

Как уже отмечалось, поступивший на вход обучающей системы вектор $x = (x_1, x_2, ..., x_n)$ может присутствовать в любой выборке V_i , $i = 1,2,...,h-1$ или в нескольких выборках одновременно (например в V_0 и в V_1).

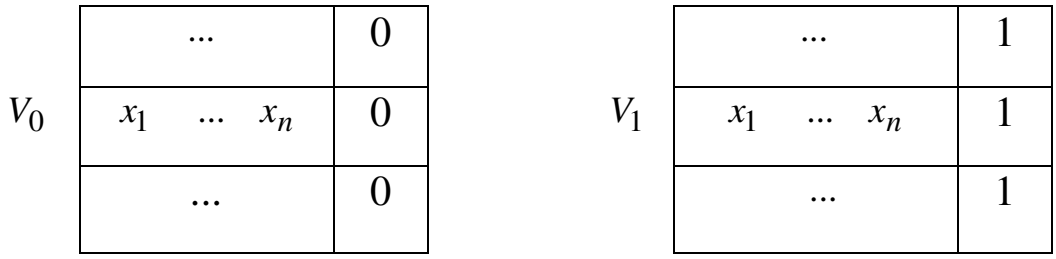

Рис. 2.2. Присутствие вектора x в V_0 и V_1

В этом нет ничего удивительного, так как в таком случае он входит в описания разных наблюдаемых объектов. Однако сказать к какому классу принадлежит $x = (x_1, x_2, ..., x_n)$ мы не можем, поскольку не известны вероятности $P(x_1, x_2, ..., x_n, 0)$ и $P(x_1, x_2, ..., x_n, 1)$. Этот вопрос решает байесовская процедура распознавания.

Сложность класса задач. Класс задач $C \equiv C(g, h, m^{(h)}, n)$ – совокупность всевозможных распределений вероятности P на множестве B вместе с величинами $g, h, m^{(h)}, n$. Множество задач и множество распределений данного класса находятся во взаимно однозначном соответствии. Погрешностью процедуры распознавания Q на классе C – число

$$\nu(Q, C) = \sup_{P \in C} \nu(Q, P),$$

а сложностью класса C – число

$$\mu(C) = \inf_{Q} \upsilon(Q, C) = \inf_{Q} \sup_{p \in C} \upsilon(Q, P).$$

Ясно, что $0 \le \mu(C) \le \nu(Q,C) \le 1$, и следует пользоваться такими процедурами распознавания Q, для которых число $\nu(Q,C)$ мало отличается от числа $\mu(C)$.

Пусть $d = (d_1, d_2, ..., d_n)$ – целочисленный вектор. Считаем, что распределения P из класса C при каждом d удовлетворяют условию

$$P(x_1 = d_1, x_2 = d_2, ...x_n = d_n \,|\, f = i) = \prod_{j=1}^{n} P(x_j = d_j \,|\, f = i),\ i = 0,1,..., h-1,$$

что означает независимость признаков x_j для каждого класса объектов. Считаем, что выполняются неравенства $h \le 2g^n$, $P(f = i) > 0$, $i = 0,1,..., h-1$, а также для простоты изложения неравенство $\frac{gn}{m'} + \frac{h}{m_h} < 1$, где $m' = \min\limits_{0 \le i \le h-1} m_i$.

2.3. Байесовские индуктивные процедуры. Независимые признаки.

Понятия независимости двух или нескольких событий, а также условной вероятности занимают центральное место в теории вероятностей. Напомним, что если $P(A) > 0$, то частное

$$P(B|A) = \frac{P(AB)}{P(A)}$$

называют условной вероятностью события B при условии A. Если произошло событие A, то могут иметь место три случая: а) условная вероятность $P(B|A)$ равна $P(B)$, т.е. события A и B независимы; б) вероятность $P(B|A)$ больше чем $P(B)$ и с) вероятность $P(B|A)$ меньше чем $P(B)$. Таким образом, наступление некоторого события A может в принципе изменить всю последующую цепочку событий.

Для независимых признаков совместное распределение величин $x_1, x_2, ..., x_n$ задается произведением:

$$P(x_1, ...x_n) = \prod_{j=1}^{n} P(x_j).$$

Байесовская индуктивная процедура. Определим случайные величины $\xi(d,i)$, зависящие от d и i как от параметров, с помощью формулы

$$\xi(d,i) = \frac{k_i}{m_h} \prod_{j=1}^{n} \frac{k(d_j,i)}{m_i},\ i \in \{0,1,...,h-1\}, \qquad (2.3)$$

здесь $k(d_j,i)$ – количество значений, равных d_j, j-го признака в j-м столбце матрицы V_i; k_i – количество значений целевого признака f, равных i, в векторе V_h. Выберем $A(d)$ равным минимальному числу s из множества $\{0,1,...,h-1\}$ такому, что

$$\xi(d,s) \geq \xi(d,i),\ i \in \{0,1,...,h-1\}. \qquad (2.4)$$

Процедуру распознавания, определяемую соотношениями (2.3), (2.4), обозначим Q_B. Заметим, что величины $\xi(d,i) / \sum_{j=0}^{h-1} \xi(d,j)$ представляют собой приближенные значения вероятностей $P(f = i | x_1 = d_1, x_2 = d_2, ..., x_n = d_n)$, вычисленных по формуле Байеса, поэтому процедуру распознавания Q_B будем называть байесовской.

Сама формула Байеса связывает указанные условные вероятности зависимостью

$$P(f = i \mid x = d) = \frac{\prod_{j=1}^{n} P(x_j = d_j \mid f = i) P(f = i)}{\sum_{i=0}^{h-1} \prod_{j=1}^{n} P(x_j = d_j \mid f = i) P(f = i)},\ i = 0,1,...,h-1.$$

Рассмотрим теперь вопрос распределений случайных выборок $V=(V_0,V_1,...,V_h)$. Пускай пространство результатов получения вектора V_h

$$\Omega=\{\omega:\omega=(a_1,a_2,...,a_{m_h}),a_j\in\{0,1,...,h-1\}\}$$

и $p(\omega)=p_0^{k_0(\omega)}p_1^{k_1(\omega)}...p_{h-1}^{k_{h-1}(\omega)}$, де $k_i(\omega)$ – количество элементов i в последовательности ω, $p_i=P(f=i)$, $\sum_{\omega\in\Omega}p(\omega)=1$. Анализ схемы Бернулли примененный к случайным величинам

$$\xi_j(\omega)=\begin{cases}1, a_j=i\\ 0, a_j\neq i\end{cases},\quad j=1,2,...,m_h$$

показывает, что математическое ожидание $k_i(\omega)$ равно $m_h p_i$. Таким образом, вектор V_h обучающей выборки отождествляется с вектором ω $V_h=(a_1,a_2,...,a_{m_h})$, $P_1(V_h)=p(\omega)$, $\sum P_1(V_h)=1$, количество векторов V_h равно h^{m_h} и распределение $P_1(V_h)$ определяется распределением $P(f=i)$.

Несложно заметить, что множество обучающих выборок $W=\{V\}$ имеет структуру прямого произведения $hn+1$ вероятностных пространств, полученных на основе бернулливских распределений, поэтому $\sum_{V\in W}P_1(V)=1$

2.4. Верхняя оценка погрешности байесовской процедуры распознавания

Дадим оценку сверху погрешности процедуры Q_B на классе C. Имеет место следующая теорема [3].

Теорема 2.1 Пусть выполняется условие $gn/m' + h/m_h < 1$. Существует абсолютная константа $a_0 < \infty$ такая, что справедливо неравенство $\upsilon(Q_Б, C) \le \min(1, a_0\sqrt{gn/m' + h/m_h})$, где $m' = \min\limits_{0 \le i \le h-1} m_i$.

Доказательство. Если $m' = 0$ или $m_h = 0$, то справедливость теоремы очевидна. Пусть $m' > 0$, $m_h > 0$, $P_1(V)$ – вероятность получения выборки V; M_1, D_1 – соответственно математическое ожидание и дисперсия по вероятностному пространству, множество элементарных исходов которого состоит из всевозможных значений V, а вероятность есть $P_1(V)$. Тогда для каждого целочисленного вектора $d = (d_1, d_2, ..., d_n)$ справедливы соотношения

$$M_1 \frac{k_i}{m_h} = p_i, \quad M_1 \frac{k_{jid_j}}{m_i} = p_{jid_j},$$

$$D_1 \frac{k_i}{m_h} = p_i(1 - p_i)/m_h, \quad D_1 \frac{k_{jid_j}}{m_i} = p_{jid_j}(1 - p_{jid_j})/m_i, \quad j = 1,2,...,n,$$

где +/.

Из этих соотношений и выражений (2.3) получаем

$$M_1\xi(d,i) = \prod_{j=1}^{n} p_{jid_j} p_i = P(x = d, f = i). \qquad (2.5)$$

Для дисперсий случайных величин $\xi(d,i)$ легко получить соотношения

$$D_1\xi(d,i) = M_1\xi^2(d,i) - (M_1\xi(d,i))^2 =$$

$$= \prod_{j=1}^{n} (M_1 k_{jid_j}^2 / m_i^2) M_1 k_i^2 m_h^2 - \prod_{j=1}^{n} p_{jid_j}^2 p_i^2 =$$

$$= \prod_{j=1}^{n} [(m_i p_{jid_j} (1 - p_{jid_j}) m_i^2 p_{jid_j}^2) m_i^2] \cdot$$

$$\cdot (m_h p_i (1 - p_i) + m_h^2 p_i^2) / m_h^2 - \prod_{j=1}^{n} p_{jid_j}^2 p_i^2 =$$

$$= P(x = d, f = i) \lambda(d, i), \; i \in \{0,1,...,h-1\}, \qquad (2.6)$$

где

$$\lambda(d,i) = \prod_{j=1}^{n} [(1 - p_{jid_j}) / m_i + p_{jid_j}][(1 - p_{jid_j}) / m_h + p_i] - \prod_{j=1}^{n} p_{jid_j} p_i .$$

Из определений погрешности процедуры для любого распределения P следует

$$\upsilon(Q_Б, P) = M_1[P(A^*) - P(A)] =$$

$$= \sum_V [\sum_d P(x = d, A^*(d)) - P(x = d, A(d))] P_1(V) =$$

$$= \sum_d [\sum_V P(f = A^*(d) | x = d) P_1(V) - \sum_V P(f = A(d) | x = d) P_1(V)] P(x = d) =$$

$$= \sum_d [P(f = A^*(d) | x = d) - \sum_{i=0}^{h-1} P(f = i | x = d) P_1(A(d) = i)] P(x = d) =$$

$$= \sum_d \sum_{i=0}^{h-1} \delta_i(d) P_1(A(d) = i) P(x = d) . \qquad (2.7)$$

Здесь $A = Q_Б(d)$, $\delta_i(d) = P(f = A^*(d) | x = d) - P(f = i | x = d)$, $A^*(d)$ – наилучшая функция распознавания, которая не зависит от выборки V.

Для каждого d обозначим $\alpha \equiv \alpha(d) = P(f = A^*(d) | x = d)$ и рассмотрим два множества $K = \{i : i \in I_0, \delta_i(d) > 3\alpha / 4\}$ и $L = \{i : i \in I_0, \varepsilon < \delta_i(d) \le 3\alpha / 4\}$; здесь $I_0 = \{0,1,...,h-1\}$, $\varepsilon \in (0,1]$, число ε определяется ниже. Таким образом,

$K=\{i: i\in I_0, P(f=i\mid x=d)<\alpha/4\}$, $L=\{i: i\in I_0, \alpha/4\le P(f=i\mid x=d)<\alpha-\varepsilon\}$.

Из (2.7) следует

$$\nu(Q_Б,P)\le\varepsilon+\sum_d[\sum_{i\in K}\delta_i(d)P_1(A(d)=i)+$$

$$+\sum_{i\in L}\delta_i(d)P_1(A(d)=i)]P(x=d). \qquad (2.8)$$

Рассмотрим сумму $S_0\equiv\sum_{i\in K}\delta_i(d)P_1(A(d)=i)$. Учитывая, что $\delta_i(d)\le\alpha$, получаем

$$S_0\le\alpha\sum_{i\in K}P_1(A(d)=i)=\alpha P_1(\sum_{i\in K}\{A(d)=i\}). \qquad (2.9)$$

Из определения $A(d)$ следует

$$\sum_{i\in K}\{A(d)=i\}\subseteq\bigcup_{i\in K}\{\xi(d,i)\ge\alpha P(x=d)/2\}\bigcup\bigcup_{i\in K}\{\xi(d,i)\le\alpha P(x=d)/2\}$$

$$\sum_{i\in K}\{A(d)=i\}\subseteq\left\{\bigcup_{i\in K}\{\xi(d,i)\ge\alpha P(x=d)/2\}\bigcup\{\xi(d,A^*(d))\le\alpha P(x=d)/2\}\right\}.$$

Из последнего соотношения и (2.9) вытекает

$$S_0\le\alpha\left[\sum_{i\in K}P_1\{\xi(d,i)\ge\alpha P(x=d)/2\}+P_1\{\xi(d,A^*(d))\le\alpha P(x=d)/2\}\right]. \qquad (2.10)$$

Используя неравенство Чебышева и соотношения (2.5), (2.6), при $i\in K$ получаем

$$P_1\{(\xi(d,i)\ge\alpha P(x=d)/2)\}=P_1\{(\xi(d,i)-M_1(\xi(d,i)\ge\alpha P(x=d)/2-$$

$$-M_1(\xi(d,i))\} \le P_1\{|\,\xi(d,i) - M_1\xi(d,i)\,| \ge \alpha P(x=d)/2 -$$

$$-P(f=i\,|\,x=d)P(x=d)\} \le P_1\{|\,\xi(d,i) - M_1\xi(d,i)\,| \ge \alpha P(x=d)/4\} \le$$

$$\le D_1(\xi(d,i)/[\alpha P(x=d)/4]^2 =$$

$$= 16P(x=d, f=i)\lambda(d,i)/[\alpha P(x=d)]^2 \le$$

$$\le 4\lambda(d,i)/[\alpha P(x=d)]; \qquad (2.11)$$

Здесь использовалось неравенство $P(f=i\,|\,x=d) < \alpha/4$, $i \in K$. Вторая вероятность в (2.10) оценивается следующим образом

$$P_1\{\xi(d,A^*(d)) \le \alpha P(x=d)/2\} = P_1\{\xi(d,A^*(d)) - M_1\xi(d,A^*(d)) \le$$

$$\le \alpha P(x=d)/2 - P(f=A^*(d)\,|\,x=d)P(x=d)\} \le P_1\{|\,\xi(d,A^*(d) - M_1\xi(d,A^*(d))\,| \ge$$

$$\ge \alpha P(x=d)/2\} \le D_1\xi(d,A^*(d)/[\alpha P(x=d)/2]^2 = 4P(x=d,A^*(d)) \times$$

$$\times \lambda(d,A^*(d))/[\alpha P(x=d)]^2 \le 4\lambda(d,A^*(d))/[\alpha P(x=d)]. \qquad (2.12)$$

Из соотношений (2.10)-(2.12) выводим

$$S_0 \le 4\sum_{i\in I_0} \lambda(d,i)/P(x=d). \qquad (2.13)$$

Оценим сумму $S_1 \equiv \sum_{i\in L} \delta_i(d) P_1(A(d)=i))$ при $0 < \varepsilon < 3\alpha/4$. Рассмотрим множество интервалов $J_j = [\alpha - 2^j\varepsilon, \alpha - 2^{j-1}\varepsilon)$, $j = 1,2,...,k-1$, $J_k = [\alpha/4, \alpha - 2^{k-1}\varepsilon)$; здесь k натуральное число, удовлетворяющее соотношениям $\alpha - 2^k\varepsilon \le \alpha/4$, $\alpha - 2^{k-1}\varepsilon > \alpha/4$.

Пусть $G = \{j_1, j_2,..., j_t\}$ - множество индексов тех интервалов J_j, которые содержат не менее одного числа вида $P(x=d, f=i)$, $i \in L$; $t \le k$. Тогда

$$S_1 = \sum_{j\in G} S_{1j}, \qquad (2.14)$$

где

$$S_{1j} = \sum_{j \in L_j} \delta_i(d) P_1(A(d) = i),\ L_j = \{i : i \in L, P(f = i \mid x = d) \in J_j\}.$$

Рассмотрим S_{1j}, $j \in G$. Определим целое число $r = r(j)$ следующим образом. Если $j = 1$ или $j = 2$, то выберем $r = A^*(d)$. Если $j = u$, $u \geq 3$, то в качестве r выберем произвольное число из множества $L_{j_{u-2}}$, если $L_{j_{u-2}} \neq \varnothing$. Поскольку $\delta_i(\alpha) < 2^j \varepsilon$ и $(\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon) > 0$, то аналогично (2.10) имеем

$$S_{1j} \leq 2^j \varepsilon \sum_{i \in L_j} P_1(A(d) = i) = 2^j \varepsilon P_1\{\sum_{i \in L_j} \{A(d) = i\}\} \leq$$

$$\leq 2^j \varepsilon \{\sum_{i \in L_j} P_1\{\xi(d,i) \geq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon) P(x = d)\} +$$

$$+ P_1\{\xi(d,r) \leq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon) P(x = d)\}\}. \qquad (2.15)$$

Используя неравенство Чебышева, при $i \in L_j$ получаем

$$P_1\{\xi(d,i) \geq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon) P(x = d)\} =$$

$$= P_1\{\xi(d,i) - M_1\xi(d,i) \geq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon - P(f = i \mid x = d) P(x = d)\} \leq$$

$$\leq P_1\{|\xi(d,i) - M_1\xi(d,i)| \geq 2^{j-3}\varepsilon P(x = d)\} \leq D_1\xi(d,i) / [2^{j-3}\varepsilon P(x = d)]^2 =$$

$$= P(x = d, f = i)\lambda(d,i) / [2^{j-3}\varepsilon P(x = d)]^2 \leq \alpha\lambda(d,i) / [4^{j-3}\varepsilon^2 P(x = d)];$$

поскольку для $i \in L_j$ $P(f = i \mid x = d) < \alpha - 2^{j-1}\varepsilon$

Вторая вероятность в (2.15) оценивается следующим образом:

$$P_1\{\xi(d,r) \leq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon) P(x = d)\} = P_1\{\xi(d,r) - M_1\xi(d,r) \leq$$

$$\leq (\alpha - 2^{j-1}\varepsilon + 2^{j-3}\varepsilon - P(f = r \mid x = d)) P(x = d)\} \leq$$

$$\le P_1\{|\,\xi(d,r)-M_1\xi(d,r)\,|\ge (2^{j-1}-2^{j-3}-2^{j-2})\varepsilon P(x=d)\}\le$$

$$\le D_1\xi(d,r)/[2^{j-3}\varepsilon P(x=d)]^2=$$

$$=P(x=d,\,f=r)\lambda(d,r)/[2^{j-3}\varepsilon P(x=d)]^2\le\alpha\lambda(d,r)/[4^{j-3}\varepsilon^2 P(x=d)]$$

Здесь использовались соотношения

$$\alpha-2^{j-2}\le f(f=r\,|\,x=d),\ 2^{j-1}-2^{j-3}-2^{j-2}=2^{j-3}(4-1-2)=2^{j-3},$$
$$f(f=r\,|\,x=d)<\alpha$$

Из последних оценок и соотношений (2.14), (2.15) выводим

$$S_1=\sum_{j\in G}S_{1j}\le\sum_{j\in G}\{2^j\varepsilon\alpha/[4^{j-3}\varepsilon^2P(x=d)]\}\{\sum_{i\in L}\lambda(d,i)+\lambda(d,r)\}\le;$$

$$\le 64\alpha\sum_{i\in L}\lambda(d,i)/\varepsilon P(x=d)$$

здесь $L'=L\cup\{A^*(d)\}$.

Если $i\in L'$, то $\alpha\le 4P(d,f=i)$. Отсюда получаем

$$S_1\le 256\sum_{i\in L'}\lambda(d,i)/[\varepsilon P(x=d)]. \qquad (2.16)$$

Заметим, что если $\varepsilon\ge 3\alpha/4$, то $L=\varnothing$ и $S_1=0$, поэтому соотношение (2.16) справедливо при всех положительных значениях ε.

Пусть D – множество всех целочисленных векторов $d=(d_1,d_2,...,d_n)$, таких, что $d_j\in\{0,1,...,g-1\}$, $j=1,2,...,n$. Из соотношений (2.8), (2.13), (2.16) следует

$$v(Q_Б,P)\le\varepsilon+4\sum_{d\in D}\sum_{i\in I_0}\lambda(d,i)+256\sum_{d\in D}\sum_{i\in I_0}\lambda(d,i)/\varepsilon.$$

Выбирая $\varepsilon=16\left[\sum_{d\in D}\sum_{i\in I_0}\lambda(d,i)\right]^{1/2}$, получаем неравенство

$$\upsilon(Q_Б,P)\le 4\sum_{d\in D}\sum_{i\in I_0}\lambda(d,i)+32\left[\sum_{d\in D}\sum_{i\in I_0}\lambda(d,i)\right]^{1/2}. \qquad (2.17)$$

Используем известный результат [], который легко доказывается по индукции.

Пусть θ_{ji} - вещественные числа, $j=1,2,...,n$; $i\in I_0$. Очевидно,

$$\sum_{d\in D}\prod_{j=1}^{n}\theta_{jd_j}=\prod_{j=1}^{n}(\theta_{j0}+\theta_{j1}+...+\theta_{j,g-1}).$$

Используя это тождество, получаем

$$\sum_{d\in D}\lambda(d,i)=\sum_{d\in D}\{\prod_{j=1}^{n}[(1-p_{jid_j})/m_i+p_{jid_j}][(1-p_i)/m_h+p_i]-$$

$$-\prod_{j=1}^{n}p_{jid_j}p_i\}=(1+(g-1)/m_i)^n(p_i+(1-p_i)/m_h)-p_i. \qquad (2.18)$$

Если $gn/m_i\le 1$, $i\in I_0$, то при $n\ge 2$ получаем

$$[1+(g-1)/m_i]^n\le[(1+g/m_i)^{n/2}]^2=[(1-g/(m_i+g))^{n/2}]^{-2}\le$$

$$\le[1-(gn/2)/(m_i+g)]^{-2}=[1+(gn/2)/(m_i+g-gn/2)]^2\le$$

$$\le(1+gn/m_i)^2\le 1+3gn/m_i.$$

Очевидно, если $n=1$, то неравенство $[1+(g-1)/m_i]^n\le 1+3gn/m_i$, $i\in I_0$, также справедливо. Используя эту оценку и соотношение (2.18), получаем

$$\sum_{i\in I_0}\sum_{d\in D}\lambda(d,i)\le\sum_{i\in I_0}[(1+3gn/m_i)(p_i+1/m_h)-p_i]\le$$

$$\le\sum_{i\in I_0}(3p_i gn/m_i+4/m_h)\le 4(\max_{i\in I_0} gn/m_i+h/m_h). \qquad (2.19)$$

Если $\max\limits_{i\in I_0} gn/m_i+h/m_h\le 1$, то

$$(\max_{i\in I_0} gn/m_i+h/m_h)\le[\max_{i\in I_0} gn/m_i+h/m_h]^{1/2}.$$

Поэтому из соотношений (2.17) –(2.19) выводим

$$\upsilon(Q_Б,P)\le 81[\max_{i\in I_0} gn/m_i+h/m_h]^{1/2}.$$

Из последней оценки и неравенства $\upsilon(Q_Б,P)\le 1$ следует утверждение теоремы.

2.5 Вспомогательные утверждения

Приведем утверждения, которое используется при доказательстве теоремы о нижней оценке сложности класса задач, с другой стороны, эти утверждения имеют самостоятельное значение.

Лемма 2.1. Если $m_0=0$, то для любой индуктивной процедуры Q существует такое распределение P, что ее погрешность

$$v(Q,P)\ge C>0.$$

Доказательство. Рассмотрим два вероятностных распределения (две задачи). Пусть для обеих из них имеет место

$$P(f=0)=P(f=1)=\frac{1}{2}-\varepsilon,\ P(f=i)=\frac{2\varepsilon}{h-2},i\ge 2,$$

$$P(x_j = 0 \mid f = i) = 1,\ j \geq 2,\ i = 0,1,...,h-1,$$

$$P(x_1 = 0 \mid f = i) = P(x_1 = 1 \mid f = i) = \frac{1}{2},\ i \geq 1,\ \varepsilon = \frac{1}{256}, h > 2. \qquad (2.20)$$

Из соотношений (2.20) видно, что априорные вероятности классов известны. Поэтому ответ процедуры Q есть 0 или 1, т.е. $A(x) \in \{0,1\}$.

В задаче 0 выберем распределение $P^0(x_1 = 0 \mid f = 0) = 1$. В задаче 1 – распределение $P^1(x_1 = 1 \mid f = 0) = 1$, т.е. $P^s(x_1 = s \mid f = 0) = 1$, $s = 0,1$. Таким образом, оба эти распределения различаются на классе 0 по первой переменной x_1.

Ясно, что процедура Q строит ответ 0 или 1 по переменной x_1, а из (2.20) вытекает, что $A_s^*(x_1) = s \oplus x_1$, где $\oplus$ - операция сложения по модулю 2. Действительно

$$P^s(f = 0 \mid x_1 = 0) - P^s(f = 1 \mid x_1 = 0) =$$

$$= \frac{P^s(x_1 = 0 \mid f = 0)P(f = 0) - P^s(x_1 = 0 \mid f = 1)P(f = 1)}{P^s(x_1 = 0 \mid f = 0)P(f = 0) + P^s(x_1 = 0 \mid f \geq 1)P^s(f \geq 1)} =$$

$$= \frac{\begin{Bmatrix} 1, s = 0 \\ 0, s = 1 \end{Bmatrix} - \frac{1}{2}}{\begin{Bmatrix} 1, s = 0 \\ 0, s = 1 \end{Bmatrix}\left(\frac{1}{2} - \varepsilon\right) + \frac{1}{2}\left(\frac{1}{2} + \varepsilon\right)}\left(\frac{1}{2} - \varepsilon\right) =$$

$$= \frac{1 - s \oplus x_1 - \frac{1}{2}}{(\cdot)}\left(\frac{1}{2} - \varepsilon\right) = \frac{\frac{1}{2} - (s \oplus x_1)}{(\cdot)}\left(\frac{1}{2} - \varepsilon\right).$$

Аналогично имеем

$$P^s(f=0|x_1=d)-P^s(f=1|x_1=d)=$$

$$=\frac{\frac{1}{2}-(s\oplus d)}{P^s(x_1=d|f=0)P(f=0)+P^s(x_1=d|f\geq 1)P(f\geq 1)}\left(\frac{1}{2}-\varepsilon\right).$$

Поэтому

$$A_s^*(x_1=d)=\begin{cases}0, \text{если } s\oplus d=0\\ 1, \text{если } s\oplus d=1\end{cases}=s\oplus d\,.$$

Пусть процедура Q такова, что при входе $x_1=0$ выполняется неравенство

$$P_1(V: A(x_1=0)=0)\geq\frac{1}{2}. \qquad (2.21)$$

Заметим, что по определению вероятности

$$P_1(V: A(x_1=0)=0)+P_1(V: A(x_1=0)=1)=1.$$

Погрешность процедуры Q определяется соотношением [3].

$$v(Q,P)=\sum_d\sum_{i=0}^{h-1}\delta_i(d)P_1(A(d)=i)P(x=d), \qquad (2.22)$$

где $\delta_i(d)=P(A^*(d)|x=d)-P(f=i|x=d)$.

В случае (2.21) используем задачу 1 $(s=1)$, для которой $P^1(x_1=0)=\frac{1}{2}\left(\frac{1}{2}+\varepsilon\right)>\frac{1}{4}$ и $A_1{}^*(x_1=0)=1$.

Из соотношения (2.22) вытекает, что

$$\delta_{i=0}(x_1 = 0) = P^1(f = 1 | x_1 = 0) - P^1(f = 0 | x_1 = 0) \geq \frac{1}{2},$$

так как $P^1(f = 1 | x_1 = 0) = \dfrac{\frac{1}{2}\left(\frac{1}{2} - \varepsilon\right)}{\frac{1}{2}\left(\frac{1}{2} + \varepsilon\right)} \geq \frac{1}{2}$, $P^1(f = 0 | x_1 = 0) = 0$.

Поэтому

$$v(Q, P^1) \geq \frac{1}{2}\frac{1}{2}\frac{1}{4} = \frac{1}{16}.$$

Если у процедуры Q при входе $x_1 = 0$ выполняется условие

$$P_1(V : A(x_1 = 0) = 1) \geq \frac{1}{2},$$

то переходим к задаче 0 $(s = 0)$, для которой $A_0^*(x_1 = 0) = 0$.

В таком случае

$$\delta_{i=1}(x_1 = 0) = P^0(f = 0 | x_1 = 0) - P^0(f = 1 | x_1 = 0) \geq \frac{1}{8}.$$

$$P^0(x_1 = 0) = \frac{1}{2} - \varepsilon + \frac{1}{2}\left(\frac{1}{2} + \varepsilon\right) \geq \frac{1}{2}.$$

Поэтому из (2.22) вытекает

$$v(Q, P^0) \geq \frac{1}{2}\frac{1}{8}\frac{1}{2} \geq \frac{1}{32}.$$

Заметим, что переход от одной задачи к другой возможен, так как вероятность выборки $P_1(V)$ для двух распределений одна и та же.

Следующий результат леммы 2.2 показывает, что при отсутствии в выборке V вектора V_h, любая процедура также работает неудовлетворительно.

Лемма 2.2. Если $m_h = 0$, то для любой индуктивной процедуры существует такое распределение P, что

$$v(Q, P) \geq C > 0.$$

Доказательство. Рассмотрим два распределения (2 задачи):

Задача 0: при $s = 0$ полагаем

$$\left.\begin{aligned} P(x_1 = 0,..., x_n = 0, f = 0) &= \frac{1}{4} - \varepsilon, \\ P(x_1 = 0,..., x_n = 0, f = 1) &= \frac{3}{4} - \varepsilon, \\ P(x_1 = 0,..., x_n = 0, f = i) &= \frac{2\varepsilon}{h-2}, i \geq 2. \end{aligned}\right\} P^0.$$

Задача 1: при $s = 1$

$$\left.\begin{aligned} P(x_1 = 0,..., x_n = 0, f = 0) &= \frac{3}{4} - \varepsilon, \\ P(x_1 = 0,..., x_n = 0, f = 1) &= \frac{1}{4} - \varepsilon, \\ P(x_1 = 0,..., x_n = 0, f = i) &= \frac{2\varepsilon}{h-2}, i \geq 2. \end{aligned}\right\} P^1.$$

Из определений вероятностных распределений видно, что на вход процедуры Q поступает всегда один вектор $x \equiv 0$, поэтому ответ процедуры на этот вход тоже один, так как $A(x)$ однозначная функция. Поскольку в

выборке V отсутствует вектор V_h, ответ процедуры Q может указывать на любой из классов, т. е. $A(x) = A(x=0) \in \{0,1,...,h-1\}$, при этом $A_s^*(0) = \bar{s}$.

Если ответ процедуры Q не совпадает с 1, то рассматривается задача 0, для которой $A_0^*(x=0) = 1$. Тогда

$$\delta_i(x=0) = P^0(f=1|x=0) - P^0(f=i|x=0) \geq \frac{1}{2}, i \neq 1,$$

$$P_1^0\big(A(x=0)=i\big) = 1, \;\; P^0(x=0) = 1.$$

Поэтому

$$v(Q, P^0) \geq \frac{1}{2}.$$

Если $A(0) \neq 0$, то переходим к задаче 1, для которой $A_1^*(x=0) = 0$. В этом случае выполняются соотношения

$$\delta_0(x=0) = P^1(f=0|x=0) - P^1(f=i|x=0) \geq \frac{1}{2}, i \neq 0,$$

$$P_1^1\big(A(x=0)=i\big) = 1, \; P^1(x=0) = 1.$$

Следовательно, погрешность процедуры удовлетворяет условию

$$v(Q, P^1) \geq \frac{1}{2}.$$

Лемма доказана.

Лемма 2.3. Пусть m – натуральное число, s – булева случайная величина, $u = (u_1, u_2, ..., u_m)$ – троичный случайный вектор, $u_i \in \{0,1,2\}$, $\theta \in (0,1]$, $\beta \in (0,1/2]$, $t \in (0, \beta)$, $\{w\}$ – множество всех троичных векторов $w = (w_1, w_2, ..., w_m)$, $w_i \in \{0,1,2\}$, $i = 1,...,m$. Пусть вероятность P,

определенная на декартовом произведении $\{0,1\}\times\{w\}$, удовлетворяет соотношениям

$$P(s=0)=P(s=1)=1/2,\ P(u=w|s=k)=\prod_{i=1}^{m}P(u_i=w_i|s=k),$$

$$P(u_i=w_i|s=k)=\begin{cases}\theta(\beta+(-1)^k t, w_i=0,\\ \theta(1-\beta-(-1)^k t, w_i=1,\\ 1-\theta, w_i=2,\end{cases}\quad i=1,2,...,m;\ k=0,1.$$

Тогда существует абсолютная константа $c_0<\infty$ такая, что справедливо неравенство

$$I(s,u)\le c_0 m\theta t^2/\beta,$$

где $I(s,u)$ – шенноновское количество информации относительно случайной величины s, содержащееся в случайном элементе u.

Доказательство. Имеем

$$I(s,u)=H(u)-MH(u|s)=H(u)-[H(u|s=0)+H(u|s=1)]/2\le$$
$$\le m\{H(u_1)-[H(u_1|s=0)+H(u_1|s=1)]/2\};$$

здесь символом H обозначена энтропия, символом M – математическое ожидание.

Так как

$$P(u_1=0)=[P(u_1=r|s=0)+P(u_1=r|s=1)]/2=\theta\beta,$$
$$P(u_1=1)=\theta(1-\beta),$$

то

$$I(s,u)\le m\{-\theta\beta\log_2(\theta\beta)-\theta(1-\beta)\log_2(1-\beta)-(1-\theta)\log_2(1-\theta)+$$
$$+[(\theta(\beta+t))\log_2(\theta(\beta+t))+(\theta(1-\beta-t))\log_2(\theta(1-\beta-t))+$$

$$+(1-\theta)\log_2(1-\theta)+(\theta(\beta-t)\log_2(\theta(\beta-t))+$$

$$+(\theta(1-\beta+t))\log_2(\theta(1-\beta+t))+(1-\theta)\log_2(1-\theta)]\}=$$

$$=(\log_2 e)(m\theta/2)\{-2\beta\ln(\theta\beta)-2(1-\beta)\log_2(\theta(1-\beta))+$$

$$+(\beta+t)(\ln(\theta\beta)+\ln(1+t/\beta))+(1-\beta-t)(\ln(\theta(1-\beta))+$$

$$+\ln(1-t/(1-\beta))+(\beta-t)(\ln(\theta\beta)+\ln(1-t/\beta))+$$

$$+(1-\beta+t)(\ln(\theta(1-\beta))+\ln(1+t/(1-\beta))\}=$$

$$=(\log_2 e)(m\theta/2)\{(\beta+t)\ln(1+t/\beta)+(\beta-t)\ln(1-t/\beta)+(1-\beta-t)\times$$

$$\times\ln(1-t/(1-\beta))+(1-\beta+t)\ln(1+t/(1-\beta))\ \}\le$$

$$\le(\log_2 e)(m\theta/2)\{(\beta+t)t/\beta+(\beta-t)(-t/\beta)+(1-\beta-t)(-t/(1-\beta))+$$

$$+(1-\beta+t)t/(1-\beta)=(\log_2 e)m\theta\left\{\frac{t^2}{\beta}+\frac{t^2}{1-\beta}\right\}\le 2(\log_2 e)m\theta\frac{t^2}{\beta}.$$

Здесь использовалось неравенство $\ln(1+x)<x$, $x\neq 0$. Лемма доказана.

В теории сложности задач оптимизации важную роль играет одно неравенство относительно шенноновской информации между случайными величинами [57]. Количество информации между двумя случайными величинами z_0 и z_1 определяется соотношением

$$I(z_1,z_0)=H(z_1)-MH(z_1\mid z_0)$$

где $H(z_1)$ – энтропия случайной величины z_1 [57], $H(z_1\mid z_0)$м – условная энтропия случайной величины z_1 при условии z_0. Величина $I(z_1,z_0)$ означает убыль энтропии случайной величины z_1 при наблюдении случайной величины z_0 и наоборот, поскольку $I(z_1,z_0)=I(z_0,z_1)$. Величину убыли понимают как количество информации о случайной величине z_1, полученной при наблюдении случайной величины z_0. Равенство информации $I(z_1,z_0)=I(z_0,z_1)$ подчеркивается следующей простой формулой, которая во многих случаях оказывается весьма удобной

$$I(z_1, z_0) = H(z_1) + H(z_0) - H(z_1, z_0) \tag{2.23}$$

Лемма 2.4. Пусть s и z — булевы случайные величины, для которых выполняются неравенства

$$P(z \neq s) \leq \frac{1}{4}$$

$$P(s = 0) = P(s = 1) = \frac{1}{2} \tag{2.24}$$

Тогда выполняется соотношение

$$I(s, z) \geq c_0 > 0, \tag{2.25}$$

где c_0 — абсолютная константа.

Доказательство. Из соотношений (2.24) вытекают неравенства

$$P(z = s) \geq \frac{3}{4},$$

$$P(z = 0, s = 0) + P(z = 1, s = 1) \geq \frac{3}{4},$$

$$P(z = 0 \mid s = 0) + P(z = 1 \mid s = 1) \geq \frac{3}{2},$$

$$P(z = 0 \mid s = 0) \geq \frac{1}{2},\ P(z = 1 \mid s = 1) \geq \frac{1}{2}, \tag{2.26}$$

так как

$$P(z = 0 \mid s = 0) = \frac{P(z = 0, s = 0)}{P(s = 0)},$$

$$P(z = 1 \mid s = 1) = \frac{P(z = 1, s = 1)}{P(s = 1)},$$

$$P(z = 0 \mid s = 0) \leq 1,\ P(z = 1 \mid s = 1) \leq 1.$$

Далее очевидным образом получаем цепочку неравенств

$$P(z=0,s=0)\geq\frac{1}{4},$$

$$P(z=1,s=1)\geq\frac{1}{4},$$

$$P(z=0)\geq\frac{1}{4},\ P(z=1)\geq\frac{1}{4}. \qquad (2.24)$$

Информация $I(s,z)$ вычисляется следующим образом

$$I(s,z)=1-P(z=0)\log P(z=0)-P(z=1)\log P(z=1)+$$
$$+P(z=0,s=0)\log P(z=0,s=0)+P(z=0,s=1)\log P(z=0,s=1)+$$
$$+P(z=1,s=0)\log P(z=1,s=0)+P(z=1,s=1)\log P(z=1,s=1), \qquad (2.25)$$

здесь через $\log$ обозначается двоичный логарифм,

$$H(s)=-P(s=0)\log(s=0)-P(s=1)\log P(s=1)=1.$$

Обозначим

$$P(z=0)=a,\quad P(z=0,s=0)=b.$$

Из (2.25), учитывая соотношения

$$P(z=0,s=0)+P(z=1,s=0)=P(s=0)=\frac{1}{2},$$

$$P(z=0,s=1)+P(z=1,s=1)=P(s=1)=\frac{1}{2},$$

$$P(z=0,s=1)+P(z=0,s=1)=P(z=0)=a,$$

$$P(z=0)+P(z=1)=1, \qquad (2.26)$$

Получаем

$$I(z,s)=\varphi(a,b)=1-a\log a-(1-a)\log(1-a)+b\log(b)+(a-b)\log(a-b)+$$
$$+\left(\frac{1}{2}-b\right)\log\left(\frac{1}{2}-b\right)+\left(\frac{1}{2}-a+b\right)\log\left(\frac{1}{2}-a+b\right).$$

Из (2.24), (2.26) заключаем, что величины a и b должны удовлетворять условиям

$$b\geq\frac{1}{4},\ \ a\geq\frac{1}{4},\ \ b\leq\frac{1}{2},\ \ b\leq a,\ \ b\geq\frac{a}{2}+\frac{1}{8},$$

последнее следует из второго в (2.26) и соотношений

$$P(z=0,s=1)+P(z=1,s=1)=\frac{1}{2},$$
$$P(z=0,s=0)+P(z=0,s=1)=a.$$

Множество допустимых значений a и b лежит в треугольнике ABC (Рис. 2.1).

Имеем

$$\varphi'(a,b)=\log e\left[\ln b-\ln(a-b)-\ln\left(\frac{1}{2}-b\right)+\ln\left(\frac{1}{2}-a+b\right)\right]=\log\frac{b\left(\frac{1}{2}-a+b\right)}{(a-b)\left(\frac{1}{2}-b\right)}.$$

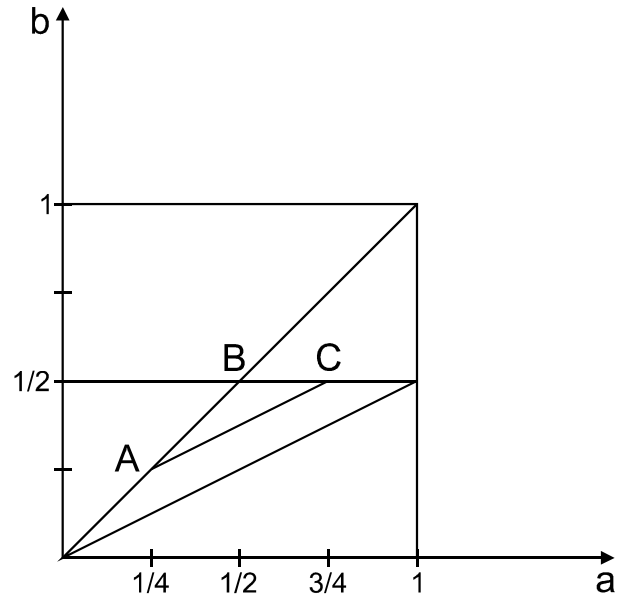

Рис. 2.1

Так как $b > \frac{a}{2}$, то

$$\varphi'(a,b) > \log \frac{\frac{a}{2}\left(\frac{1}{2}-\frac{a}{2}\right)}{\frac{a}{2}\left(\frac{1}{2}-\frac{a}{2}\right)} = 0.$$

при каждом a из допустимой области.

Поэтому минимальное значение функции $\varphi(a,b)$ лежит на отрезке АС, т.е. при $b = \frac{a}{2} + \frac{1}{8}$.

Таким образом

$$I(z,s) \geq 1 - a\log a - (1-a)\log(1-a) + \frac{4a+1}{8}\log\frac{4a+1}{8} + \frac{4a-1}{8}\log\frac{4a-1}{8} +$$

$$+ \frac{3-4a}{8}\log\frac{3-4a}{8} + \frac{5-4a}{8}\log\frac{5-4a}{8} \equiv \varphi(a),$$

где $\frac{1}{4} \leq a \leq \frac{3}{4}$. Отсюда следуют, что

$$\varphi'(a)=\log e\left[-\ln a+\ln(1-a)+\frac{1}{2}\ln\frac{4a+1}{8}+\frac{1}{2}\ln\frac{4a-1}{8}-\frac{1}{2}\ln\frac{3-4a}{8}-\frac{1}{2}\ln\frac{5-4a}{8}\right]=$$

$$=\frac{1}{2}\log\frac{(1-a)^2(4a+1)(4a-1)}{a^2(3-4a)(5-4a)}.$$

Поэтому, если $a\to\frac{1}{4}+$, то $\varphi'(a)\to-\infty$, и если $a\to\frac{3}{4}-$, то $\varphi'(a)\to\infty$.

Таким образом то $\varphi'(a)$ принимает значение 0 только в одной точке то $a=\frac{1}{2}$. Действительно,

$$(1-2a+a^2)(1ba^2-1)=a^2(15-32a+16a^2),$$

$$16a^2-1-32a^3+2a+16a^4-a^2=15a^2-32a^3+16a^4,$$

$$2a=1.$$

Значит функция $\varphi(a)$ достигает минимума в точке $a=\frac{1}{2}$, т.е.

$$I(z,s)\ge 2\left(1+\frac{3}{8}\log\frac{3}{8}+\frac{1}{8}\log\frac{1}{8}\right)=2\left(\frac{3}{8}\log 3+\frac{1}{8}\log 1-\frac{1}{2}\right)=\frac{1}{4}(3\log 3-4)=$$

$$=\frac{1}{4}\log\frac{3^3}{2^4}=\frac{1}{4}\log\frac{27}{16}>0.$$

Неравенство (2.25) доказано.

Результат леммы показывает, что взаимная шенноновская информация двух случайных величин z и s, связанных соотношениями (2.24), оказывается не слишком малой, т.е. при проведении опытов над этими случайными величинами приобретается конечная информация.

2.6 Нижняя оценка погрешности

Дадим оценку снизу сложности класса задач. Справедлива следующая теорема.

Теорема 2.2. Существует абсолютная константа $a_1 > 0$ такая, что справедливо следующее. Каковы бы ни были натуральные числа g, h, n, удовлетворяющие неравенствам $g \geq 2$, $2 \leq h \leq 2g^n$ целые неотрицательные числа $m_0, m_1, ..., m_h$ и процедура распознавания Q, существует такое распределение вероятностей P из класса C, что выполняется неравенство

$$\upsilon(Q,P) \geq a_1 \min\left(1, \sqrt{\frac{gn}{m'} + \frac{h}{m_h}}\right),$$

где $m' = \min\limits_{0 \leq i \leq h-1} m_i$.

Доказательство. Ситуация, когда $m' = 0$ или $m_h = 0$ уже рассматривалась. Предположим, погрешность некоторой процедуры Q на классе C не превосходит числа $\upsilon \in (0,1)$. При условии $m_h < \infty$ докажем неравенство

$$\upsilon \geq c_1 \sqrt{h / m_h}\,, \tag{2.27}$$

где $c_1 > 0$ – абсолютная константа. Если $\upsilon \geq 1/16$, то неравенство (2.27) справедливо. Пусть $\upsilon < 1/16$.

Обозначим $2\eta - 1$ наибольшее нечетное число из множества $I = \{0,1,...,h-1\}$. Имеем $\eta \geq 1$, $2\eta \leq h \leq 2g^n$, $\eta \leq g^n$. Из класса C выделим подкласс задач следующим образом. Пусть $s = (s_0, s_1, ..., s_{\eta-1})$ – булев вектор; t – вещественное число, $t \in (0,1/2)$; θ – вещественное число, такое что

$$\theta = \begin{cases} 1/2\eta, & 2\eta - 1 < h - 1, \\ 1/n, & 2\eta - 1 = h - 1. \end{cases} \tag{2.28}$$

Выберем η различных векторов $d^{(0)}, d^{(1)}, ..., d^{(\eta-1)}$, все компоненты которых принадлежат множеству $\{0,1,\ldots,g\text{-}1\}$. Если $2\eta < h$, то $\eta < g^n$; поэтому существует n – мерный вектор $d^{(\eta)}$, отличный от векторов $d^{(0)}, d^{(1)}, ..., d^{(\eta-1)}$, каждая компонента $d_j^{(\eta)}$ которого принадлежит множеству $\{0,1,\ldots,g\text{-}1\}$. Определим на множестве B распределение вероятностей P^s:

$$P^s(f = 2i) = \theta(1/2 + (-1)^s t),$$

$$P^s(f = 2i+1) = \theta(1/2 - (-1)^s t),$$

$$P^s(x = d^{(i)} \mid f = 2i) = 1,\ P^s(x = d^{(i)} \mid f = 2i+1) = 1,$$

$$i = 0,1,...,\eta - 1.$$

В случае $2\eta - 1 < h - 1$ дополнительно определим $P^s(f = h-1) = 1/2$, $P^s(x = d^{(\eta)} \mid f = h-1) = 1$. Нетрудно видеть, что множество распределений вероятностей P^s при каждом t принадлежит классу C.

Считаем, что компоненты s_j вектора s являются независимыми случайными величинами и принимают значения 0 и 1 с вероятностями

$$P^{(0)}(s_j = 0) = P^{(0)}(s_j = 1) = 1/2,\ j = 0,1,...,\eta - 1.$$

Сказанное определяет на декартовом произведении $\{x\}\times\{f\}\times\{V\}\times\{s\}$ распределение вероятностей P, которое для всех $d \in \{x\}, i \in \{f\}, W \in \{V\}, k \in \{s\}$ удовлетворяет соотношению

$$P(x = d, f = i, V = W, s = k) = P^k(x = d \,|\, f = i) P^k(f = i) P_1^k(V = W) \prod_{j=0}^{\eta-1} P^{(0)}(s_j = k_j) \tag{2.29}$$

вероятность P_1^k определяется по вероятности P^k и способу построения выборки V.

Используя определение (2.29) вероятности P, для всех $k \in \{s\}$ выводим соотношение

$$\upsilon(Q, P^k) \geq \sum_{i=0}^{\eta-1} \sum_{j=0}^{h-1} P(A(d^{(i)}) = j \mid s = k) \delta_{jk}(d^{(i)}) P(x = d^{(i)} \mid s = k); \quad (2.30)$$

здесь $\delta_{jk}(d^{(i)}) = P(f = A_k^*(d^{(i)}) \mid s = k, x = d^{(i)}) - P(f = j \mid s = k, x = d^{(i)})$;

$A_k^*(d)$ есть определенная выше наилучшая функция распознавания $A^*(d)$, соответствующая распределению P^k.

Аналогично неравенству (7), получаем

$$\upsilon(Q, P^k) \geq \sum_{i=0}^{\eta-1} \sum_{j=0}^{h-1} P(A(d^{(i)}) = j \mid s = k) \delta_{jk}(d^{(i)}) P(x = d^{(i)}) \quad (2.31)$$

где $A = Q(V), A(d^{(i)}) = Q(d^{(i)}, V)$,

$\delta_{jk}(d^{(i)}) = P^k(f = A_k^*(d^{(i)}) \mid x = d^{(i)}) - P(f = j \mid x = d^{(i)})$, $j \in I = \{0,1,...,h-1\}$.

Имеем

$$P_1^k(A(d^{(i)}) = j) = P_1^k\{W : W \in \{V\}, A(d^{(i)}) = j\}. \quad (2.32)$$

Заметим, что

$$P_1^k(V = W) = \frac{P_1^k(V = W) P^{(0)}(s = k)}{P^{(0)}(s = k)} =$$

$$\frac{\sum_{d^{(j)}}\sum_{j} P^k(x=d^{(i)} \mid f=j)P^k(f=j)P_1^k(V=W)P^{(0)}(s=k)}{P^{(0)}(s=k)}$$

$$=\frac{\sum_{d^{(i)}}\sum_{j} P(x=d^{(i)}, f=j, V=W, s=k)}{\sum_{d^{(i)}}\sum_{j}\sum_{W} P(x=d^{(i)}, f=j, V=W, s=k)}=\frac{P(V=W, s=k)}{P(s=k)}=P(V=W, s=k).$$

Из последнего соотношения и из (2.32) следует $P_1^k(A(d^{(i)})=j)=P(A(d^{(i)})=j \mid s=k)$.

Аналогично булевому случаю [2] имеем $P(x=d^{(i)} \mid s=k)=P^{(k)}(x=d^{(i)})$.

Из формул

$$P^k(f=2i \mid x=d^{(i)})=\frac{P^k(x=d^{(i)} \mid f=2i)P^k(f=2i)}{P^k(x=d^{(i)})},$$

$$P^k(f=2i+1 \mid x=d^{(i)})=\frac{P^k(x=d^{(i)} \mid f=2i+1)P^k(f=2i+1)}{P^k(x=d^{(i)})},$$

$$P^k(x=d^{(i)})=P^k(x=d^{(i)} \mid f=2i)P(f=2i)+$$

$$+P^k(x=d^{(i)} \mid f=2i+1)P(f=2i+1)=\theta,$$

получаем соотношения

$$P^k(f=2i \mid s=k, x=d^{(i)})=1/2+(-1)^{k_i}t,$$

$$P^k(f=2i+1 \mid s=k, x=d^{(i)})=1/2-(-1)^{k_i}t.$$

Отсюда при условии $j \neq A_k^*(d^{(i)})$ следуют неравенства

$$\delta_{jk}(d^{(i)}) \geq 2t, i=0,1,...,\eta-1 \qquad (2.33)$$

и соотношения

$$A_k^*(d^{(i)}) = 2i + k_i, i = 0,1,...,\eta - 1. \qquad (2.34)$$

Из (2.30), (2.33) и (2.34) получаем

$$\sum_{i=0}^{\eta-1} P(A(d^{(i)}) \neq 2i + s_i) 2t\theta \leq v. \qquad (2.35)$$

Пусть q- целое число, такое, что $0 \leq q \leq \eta - 1$ и выполняются неравенства $P(A(d^{(q)}) \neq 2q + s_q) \leq P(A(d^{(i)}) \neq 2i + s_i), i = 0,1,...,\eta - 1$. Из соотношений (2.28) вытекает, что

$$\eta\theta = \begin{cases} 1/2, 2\eta < h, \\ 1, \quad 2\eta = h. \end{cases}$$

Выбирая в (2.35) $t = 4v$, если $2\eta < h$, и $t = 2v$, если $2\eta = h$, получаем $P(A(d^{(q)}) - 2q \neq s_q) \leq \frac{1}{4}$. Обозначим $z = A(d^{(q)}) - 2q$, тогда

$$P(z \neq s_q) \leq \frac{1}{4}. \qquad (2.36)$$

Из последнего неравенства и соотношений $P(s_q = 0) = P(s_q = 1) = \frac{1}{2}$ для шенноновской взаимной информации случайных величин s_q, z следует оценка $I(s_q, z) \geq c_2$, где $c_2 > 0$ z абсолютная константа [20, с. 205]. Величина z является однозначной функцией случайных элементов $V_0, V_1, ..., V_h$, поэтому

$I(s_q, z) \le I(s_q, V_0 V_1 ... V_h)$. Из свойств распределения вероятностей (2.29) следует $I(s_q, V_0 V_1 ... V_h) = I(s_{q,} V_h)$.

Таким образом, $I(s_q, V_h) \ge c_2$. Пусть $V_h^{'}$ - случайный вектор размерностью m_h, компоненты $\upsilon_{hi}^{'}$ которого определяются по компонентам υ_{hi} вектора V_h по формулам

$$\upsilon_{hi}^{'} = \begin{cases} \upsilon_{hi} - 2q, \upsilon_{hi} \in \{2q, 2q+1\}, \\ 2, \upsilon_{hi} \notin \{2q, 2q+1\}. \end{cases}$$

Из определения вероятности $P^s(f)$ вытекает, что компонента s_q вектора s задает вероятности только двух значений целевого признака $f : P^s(f = 2q)$ и $P^s(f = 2q+1)$. Поэтому $I(s_q, V_h) = I(s_q, V_h^{'})$. Применяя лемму 2.3, получаем

$$I(s_q, V_h^{'}) \le \frac{32 c_0 m h \upsilon^2}{h}$$

а так как $I(s_q, V_h^{'}) \ge c_2$, то из последней оценки вытекает неравенство (2.27).

Докажем теперь соотношения

$$\upsilon \ge c_3 \min\left(1, \sqrt{\frac{gn}{m_u}}\right), u = 0,1,...,h-1, \qquad (2.37)$$

где $c_3 > 0$ - абсолютная константа.

Не ограничивая общности, число g считаем четным, $g = 2\eta, \eta \ge 1$. Пусть S - булева матрица, состоящая из элементов

$s_{ij}, i = 0,1,...,\eta - 1;\ j = 1,2,...,n; t \in \left(0, \frac{1}{2}\right)$. Доказательство проводим для произвольного $u \in \{0,1,..., h-1\}$. Полагаем без ограничения общности

$$w = \begin{cases} 0, & u \geq 1, \\ 1, & u = 0. \end{cases}$$

Определим на множестве B распределение вероятностей P^S. Пусть $P^S(s = d \mid f) = \prod_{j=1}^{n} P^S(x_j = d_j \mid f), d \in \{d\}$; $\{d\}$ - множество всех векторов $d = (d_1, d_2, ...d_n)$, таких, что $d_j \in \{0,1,..., g-1), j = 1,2,..., n$; здесь и ниже соотношения выполняются с вероятностью 1. В случае $h = 2$ выберем $P^S(f = 0) = P^S(f = 1) = \frac{1}{2}$; в случае $h > 2$ выберем

$$P^S(f = u) = P_s^1(f = w) = \frac{3}{8}, P^S(f = i) = \frac{1}{4(h-2)}, i \in \{0,1,..., h-1\} \setminus \{u, w\}.$$

Пусть также для $h \geq 2$

$$P^S(x_j = 2i \mid f = u) = (1 + (-1)^{s_{ij}} t)\left(\frac{1}{2gn}\right)$$

$$P^S(x_j = 2i+1 \mid f = u) = \frac{2}{g} - (1 + (-1)^{s_{ij}} t)\left(\frac{1}{2gn}\right) = \frac{2}{g}\left[1 - (1 + (-1)^{s_{ij}} t)\left(\frac{1}{4n}\right)\right]$$

$$P^S(x_j = 2i \mid f = k) = \frac{1}{2gn},$$

$$P^S(x_j = 2i+1 \mid f = k) = \frac{2}{g} - \frac{1}{2gn} = \frac{2}{g}\left(1 - \frac{1}{4n}\right),$$

$$j = 1,2,..., n; i = 0,1,..., \eta - 1; k = 0,1,..., h-1, k \neq u$$

Рассматриваемый способ задания вероятностей является корректным. Из приведенных соотношений видно, что компоненты s_{ij} матрицы S определяют вероятностное распределение признаков x_j, $j = 1,...,n$ только в классе $f = u$, в остальных классах они участия не принимают. Кроме того, каждая компонента s_{ij} при фиксированных i, j задает вероятности двух значений признака x_j, а именно $x_j = 2i$ и $x_j = 2i+1$. Множество распределений вероятностей P^S при фиксированном параметре t образуют подкласс задач, принадлежащий классу C.

Считаем, что компоненты s_{ij} матрицы S являются независимыми случайными величинами и принимают значения 0 и 1 с вероятностями $P^{(0)}(s_{ij} = 0) = P^{(0)}(s_{ij} = 0) = \frac{1}{2}, i = 0,1,...,\eta - 1;\ j = 1,2,...,n$. Таким образом, на декартовом произведении $\{x\}\times\{f\}\times\{V\}\times\{S\}$ выбрано распределение вероятностей P, которое для всех $d \in \{x\}, i \in \{f\}, W \in \{V\}, K \in \{S\}$ удовлетворяет соотношению

$$P(x = d, f = i, V = W, S = K) =$$

$$= P^K(x = d \mid f = i) P^K(f = i) P_1^K(V = W) \prod_{i=1}^{\eta-1} \prod_{j=1}^{n} P^{(0)}(s_{ij} = k_{ij}), \qquad (2.38)$$

где k_{ij} - элементы матрицы K.

Рассмотрим множество $D(j)$ всех векторов $d = (d_1, d_2,..., d_n)$ таких, что $d_i \in \{0,1,..., g-1\}, i = 1,2,...,n$, причем компонента d_j является четным числом, а все остальные компоненты – нечетные.

С помощью (2.38) для всех $K \in \{S\}$ получим соотношение

$$\sum_{j=1}^{n} \sum_{d \in D(j)} \sum_{i=o}^{h-1} P(A(d) = i \mid s = k) \delta_i^K(d) P(x = d \mid S = K) \le v; \qquad (2.39)$$

здесь $A = Q(V)$; $\delta_i^K(d) = P(f = A_k^*(d) \mid S = K, x = d) - P(f = i \mid S = K, x = d)$.

Рассмотрим случай $h = 2$ (для удобства используем те же обозначения для целевого признака f, что и для $h > 2$), в таком случае $u \in \{0,1\}$. Согласно определению (2.38) вероятности P для векторов $d \in D(j)$, имеем

$$P(f = u \mid S = K, x = d) = P(x = d \mid S = K, f = u)P(f = u \mid S = K) /$$
$$/[P(x = d \mid S = K, f = u)P(f = u \mid S = K) +$$
$$+ P(x = d \mid S = K, f = w)P(f = w \mid S = K)] = \frac{\alpha_u}{\alpha_u + \alpha_w}, \qquad (2.40)$$

$$P(f = w \mid S = K, x = d) = \frac{\alpha_w}{\alpha_u + \alpha_w}, \qquad (2.41)$$

где $\alpha_u = \left[1 + (-1)^{k_{d_j/2,j}} t\right] \prod_{\substack{i=1 \\ i \neq j}}^{n} \left[1 - \frac{1 + (-1)^{k_{(d_i-1)/2,i}} t}{4n}\right]$, $\alpha_w = \left[1 - \frac{1}{4n}\right]^{n-1}$.

В случае $h > 2, d \in D(j)$ точно также имеем

$$P(f = u \mid S = K, x = d) = \frac{\frac{3}{8}\alpha_u}{\frac{3}{8}\alpha_u + \frac{5}{8}\alpha_w} = \frac{3\alpha_u}{3\alpha_u + 5\alpha_w}, \qquad (2.42)$$

$$P(f = w \mid S = K, x = d) = \frac{3\alpha_w}{3\alpha_u + 5\alpha_w}, \qquad (2.43)$$

$$P(f = i \mid S = K, x = d) = \frac{2\alpha_w}{(h-2)(3\alpha_u + 5\alpha_w)}, \quad i = 0,1,\ldots,h-1, i \neq u, i \neq w. \qquad (2.44)$$

Из соотношений (2.40) – (2.44) следует, что если $i \neq A_K^*(d), h \geq 2, d \in D(j)$, то

$$\delta_i^K(d) \geq \frac{\min\{3|\alpha_u - \alpha_w|, \alpha_w\}}{3\alpha_u + 5\alpha_w}, \qquad (2.45)$$

кроме этого, $A_K^*(d) \in \{u, w\}$ при всех $h \geq 2$.

Если $k_{d_j/2,j} = 0$, то, учитывая, что $t \in \left(0, \frac{1}{2}\right)$, получаем

$$\alpha_u - \alpha_w \geq (1+t)\left[1 - \frac{1}{4n} - \frac{t}{4n}\right]^{n-1} - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= (1+t)\left[1 - \frac{1}{4n}\right]^{n-1}\left[1 - \frac{t}{(1-1/4n)4n}\right]^{n-1} - \left[1 - \frac{1}{4n}\right]^{n-1} \geq$$

$$\geq \left[1 - \frac{1}{4n}\right]^{n-1}(1+t)\left[1 - \frac{t(n-1)}{4n-1}\right] - \left[1 - \frac{1}{4n}\right]^{n-1} \geq$$

$$\geq \left[1 - \frac{1}{4n}\right]^{n-1}(1+t)\left[1 - \frac{t(n-1)}{4(n-1)}\right] - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= \left[1 - \frac{1}{4n}\right]^{n-1}(1+t)(1-t/4) - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= \left[1 - \frac{1}{4n}\right]^{n-1}\left(t - \frac{t}{4} - \frac{t^2}{4}\right) > \frac{3t}{8}.$$

Здесь использовалось неравенство $(1+a)^n \geq 1 + na, a > -1$. В этом случае $A_K^*(d) = u$. Если $k_{d_j/2,j} = 1$, то аналогично выводим

$$\alpha_u - \alpha_w \leq (1-t)\left[1 - \frac{1}{4n} + \frac{t}{4n}\right]^{n-1} - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= (1-t)\left[1 - \frac{1}{4n}\right]^{n-1}\left[1 + \frac{t}{(1-1/4n)4n}\right]^{n-1} - \left[1 - \frac{1}{4n}\right]^{n-1} \leq$$

$$\leq \left[1 - \frac{1}{4n}\right]^{n-1}(1-t)\left[1 + \frac{2t(n-1)}{4n-1}\right] - \left[1 - \frac{1}{4n}\right]^{n-1} \leq$$

$$\leq \left[1 - \frac{1}{4n}\right]^{n-1}(1-t)\left[1 + \frac{2t(n-1)}{4(n-1)}\right] - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= \left[1 - \frac{1}{4n}\right]^{n-1} (1-t)(1+t/2) - \left[1 - \frac{1}{4n}\right]^{n-1} =$$

$$= \left[1 - \frac{1}{4n}\right]^{n-1} \left(\frac{t}{2} - t - \frac{t^2}{2}\right) \le -\frac{t}{2}\left[1 - \frac{1}{4}\right]^{n-1} \le -\frac{t}{2}\left(1 - \frac{1}{4}\right) = -\frac{3t}{8},$$

здесь использовалось неравенство $(1+a)^n \le 1 + 2na, na \le 1/2, a \ge 0$. В этом случае $A_K^*(d) = w$. Из этих соображений, а также из (2.45) при $i \ne A_K^*(d), d \in D(j)$ следует, что

$$A_K^*(d) = \begin{cases} u, \ k_{d_j/2,j} = 0, \\ w, k_{d_j/2,j} = 1 \end{cases} \tag{2.46}$$

и

$$\delta_i^K(d) \ge \frac{\min\{9/8t, \alpha_w\}}{3\alpha_u + 5\alpha_w} \tag{2.47}$$

Таким образом, значение наилучшей функции распознавания на векторе определяется только лишь одним значением компоненты матрицы K. Для получения окончательного результата этот момент является важным.

Так как $\alpha_w = \left[1 - \frac{1}{4n}\right]^{n-1} \ge \frac{3}{4}, 1 \ge 2t, \alpha_u \le 1 + t \le \frac{3}{2}, \alpha_w \le 1$, то из (2.47)

Следует

$$\delta_i^K(d) \ge \frac{\min\left\{\frac{9}{8}t, \frac{3}{2}t\right\}}{3\frac{3}{2} + 5} = \frac{9t}{8\left(\frac{9}{2} + \frac{40}{8}\right)} \ge \frac{9t}{81} = \frac{t}{9}.$$

Из последнего неравенства и (2.39) заключаем, что для всех $K \in S$ справедливы соотношения

$$\frac{t}{9}\sum_{j=1}^{n}\sum_{d\in D(j)} P(A(d) \neq A_k^*(d) \mid S = K)P(x = d \mid S = K) \leq v \qquad (2.48)$$

Учитывая условие $t \leq \frac{1}{8}$, при $d \in D(j)$ получаем

$$P(x = d \mid S = K) = P(x = d \mid S = K, f = u)P(f = u \mid S = K) +$$
$$+ P(x = d \mid S = K, f \neq u)P(f \neq u \mid S = K) \geq$$
$$\geq \frac{1}{4n}\left(\frac{2}{g}\right)^n\left(\frac{3}{8}\right)\left\{(1-t)\left[1-\frac{(1+t)}{4n}\right]^{n-1} + \left[1-\frac{1}{4n}\right]^{n-1}\right\} \geq$$
$$\geq \frac{1}{n}\left(\frac{2}{g}\right)^n\left(\frac{3}{32}\right)\left\{\frac{1}{2}\left(1-\frac{3}{8}\right)+\frac{3}{4}\right\} \geq \frac{1}{11n}\left(\frac{2}{g}\right)^n.$$

Из этой оценки и из (2.48) для всех $K \in \{S\}$ получаем неравенства

$$\frac{t}{100n}\left(\frac{2}{g}\right)^n \sum_{j=1}^{n}\sum_{d\in D(j)} P(A(d) \neq A_K^*(d) \mid S = K) \leq \upsilon,$$

из которых следует соотношение

$$\frac{t}{100n}\left(\frac{2}{g}\right)^n \sum_{j=1}^{n}\sum_{d\in D(j)} P(A(d) \neq A_S^*(d)) \leq \upsilon. \qquad (2.49)$$

Обозначим $D' = \bigcup_{j=1}^{n} D(j)$. Пусть d' - такой вектор, что $d' \in D'$ и для всех $d \in D'$ выполняется неравенство $P(A(d') \neq A_S^*(d')) \leq P(A(d) \neq A_S^*(d))$. Из (2.49) следует

$$\frac{t}{100}P(A(d')\neq A_S^*(d'))\le \upsilon, \qquad (2.50)$$

поскольку мощность множества $D(j)$ равна $\left(\frac{g}{2}\right)^n$.

Если $\upsilon \ge \frac{1}{800}$, то очевидно $\upsilon \ge \frac{1}{800}\min\left(1,\sqrt{\frac{gn}{m_u}}\right)$, т.е. неравенство (2.37) справедливо.

Пусть $\upsilon < \frac{1}{800}$. Выберем $t = 400\upsilon$. Условие $t < \frac{1}{2}$ выполнено; из (2.50) выводим

$$P(A(d')\neq A_S^*(d'))\le \frac{1}{4}. \qquad (2.51)$$

Из определения случайной матрицы S и из (2.46) заключаем, что $P(A_S^*(d')=u)=P(A_S^*(d')=w)=\frac{1}{2}$; и этих соотношений и (2.51) следует оценка $I(A_S^*(d'),A(d'))\ge c_4$, где $c_4>0$ - абсолютная константа. Так как $A(d')$ - однозначная функция от случайного элемента V, то справедливо неравенство

$$I(A_S^*(d'),A(d'))\le I(s_{d'_{q/2,q}},V);$$

здесь q таково, что d_q' - четное число. Поскольку каждая компонента s_{ij} матрицы S одинаковым образом определяет вероятности значения признаков j, $j=1,...,n$ и выполняются соотношения $P^{(0)}(s_{ij}=0)=P^{(0)}(s_{ij}=0)=\frac{1}{2}$, то отсюда следует цепочка равенств

$I(s_{d'_{q/2,q}},V)=I(s_{d'_{q/2,q}},V_u)=I(s_{01},V_u)=I(s_{01},\upsilon_u^{(1)})$, где $\upsilon_u^{(1)}$ - первый столбец матрицы V_u.

Таким образом, $I(s_{01},\upsilon_u^{(1)})\ge c_4$. Пусть υ'_u - случайный вектор размерности m_u, компоненты υ'_{ui} которого определяются по компонентам $\upsilon_{ui}^{(1)}$ вектора $\upsilon_u^{(1)}$ по формулам

$$\upsilon'_{ui}=\begin{cases}\upsilon_u^{(1)}, & \upsilon_u^{(1)}\le 1,\\ 2, & \upsilon_u^{(1)}\ge 2\end{cases}$$

Из (2.38) следует $I(s_{01},\upsilon_u^{(1)})=I(s_{01},\upsilon'_u)$. Последнее утверждение справедливо, так как компонента s_{01} определяет вероятности значений $x_1=0$ и $x_1=1$ и не влияет на остальные значения x_1.Используя лемму 2.3 и соотношение $t=1400\nu$, получаем

$$c_4\le I(s_{01},\upsilon'_u)\le c_0 mu\left(\frac{2}{g}\right)\left(\frac{t}{4n}\right)^2\Big/\left(\frac{1}{4n}\right)\le\frac{c_5\upsilon^2 m_u}{gn},$$

где $c_5<\infty$ – абсолютная константа.

Из этих неравенств вытекают соотношения (2.37), а из неравенств (2.27) и (2.37) - утверждение теоремы.

Выводы. Из теорем 2.1 и 2.2 и следует, что

$$\frac{\upsilon(Q_B,C)}{\mu(C)}\le\frac{\max(1,a_0)}{a_1},$$

т.е. отношение погрешности $\upsilon(Q_B,C)$ к сложности $\mu(C)$ класса задач C не превосходит абсолютную константу. В этом смысле байесовская процедура распознавания Q_B является субоптимальной. Таким образом, установлена (с

точностью до абсолютной мультипликативной константы) сложность класса задач C.

РАЗДЕЛ 3.
БАЙЕСОВСКИЕ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ НА НЕСТАЦИОНАРНЫХ ЦЕПЯХ МАРКОВА

Процедуры распознавания для независимых признаков имеют ограниченную область практических приложений. Большой интерес представляет развитие байесовских процедур распознавания на случай зависимых испытаний, связанных в цепь Маркова. Для того чтобы исследовать эффективность процедур распознавания на цепях Маркова и обосновать применение этих процедур на практике, необходимо изучить поведение оценок переходных вероятностей. Для этой цели в работе используются классические результаты работы [4].

Опишем схему рассуждений А.А. Маркова, которая использовалась при исследовании литературных текстов. В диссертации она применяется при анализе генетических последовательностей ДНК и белков.

3.1. Цепи Маркова

В начале 20 века А.А. Марков опубликовал ряд работ, которые привели к общей схеме « испытаний, связанных в цепь » [39, 40]. На этой схеме Марков установил ряд замечательных закономерностей, положивших начало всей современной теории марковских процессов.

Рассматривается неограниченный ряд последовательных испытаний, которые отмечаются по порядку номерами

$$1, 2, 3, \ldots, k, \ k+1, \ldots .$$

В результате испытаний наступает некоторое событие E или противоположное ему событие F (т.е. вся последовательность испытаний записывается в двухбуквенном алфавите). Если в результате одного из

испытаний установлено, что появилось событие E или противоположное ему F, то все последующие испытания зависимы от этого испытания, но совершенно не зависят от результатов всех предшествующих испытаний. Марков назвал такой ряд испытаний *простой цепью*.

Такая ситуация соответствует схеме Бернулли для независимых испытаний: для всей цепи установлены два числа p_1, p_2, вместо одного числа p для случая Бернулли.

Число p_1 означает вероятность события E при $k+1$ - м испытании, если дано, что E появилось при k - м испытании. Число p_2 означает также вероятность события E при $k+1$-м испытании, но только при задании, что k - е испытание привело к появлению противоположного события F. Согласно вышеприведенному объяснению связи испытаний в *простую цепь,* указанные вероятности события E при $k+1$ - м испытании, устанавливаются совершенно независимо от результатов предыдущих $k-1$ испытаний.

Чтобы сообщить выводам полную определенность (т.е. уметь вычислять вероятности событий E или F при любом испытании), следует ввести еще число p', представляющее вероятность события E при первом испытании.

Вместе с числами p_1, p_2, p' для сокращения записи и для симметрии формул вводятся их дополнения до единицы

$$q_1 = 1 - p_1,\ q_2 = 1 - p_2,\ q' = 1 - p',$$

представляющие вероятности события F, противоположного E.

Марков первым начал изучать эргодические свойства цепей: он исследовал ряд чисел

$$p', p'', p''', ..., p^{(k)}, p^{(k+1)}, ...,$$

соответственно представляющих вероятности события E при каждом из испытаний $1,2,3,\ldots,k,k+1,\ldots$.

На основании аксиом сложения и умножения вероятностей выполняется соотношение

$$p^{(k+1)} = p_1 p^{(k)} + p_2(1 - p^{(k)}) \quad .$$

Этому уравнению можно придать вид

$$p^{(k+1)} - p = \delta(p^{(k)} - p),$$

где δ и p определяются равенствами

$$\delta = p_1 - p_2,\ p_2 = p(1-\delta).$$

При этом исключаются случаи $\delta = \pm 1$, не представляющие интереса, т.е.

$$-1 < \delta < 1. \qquad (3.1)$$

Из предыдущего уравнения имеет место общая формула

$$p^{(k)} = p + (p' - p)\delta^{k-1}, \qquad (3.2)$$

откуда видно, что $p = \dfrac{p_2}{1 + p_2 - p_1}$ служит пределом, к которому стремится $p^{(k)}$, когда $k \to \infty$. Заметим, что скорость сходимости в (3.2) – геометрическая и предел p не зависит от начальной вероятности p'.

Анализ литературных текстов.

Марков рассмотрел поучительный пример связанных испытаний, совокупность которых, с некоторым приближением, можно рассматривать как простую цепь. Этот пример выясняет, что суммы многих связанных величин могут образовать (почти) независимые величины. Он взял последовательность 20 000 букв в романе Пушкина “Евгений Онегин”, не считая ъ и ь; эта последовательность занимает всю первую главу и шестнадцать строф второй. Она составляет 20 000 зависимых испытаний, каждое из которых дает гласную или согласную букву.

Рассмотрим небольшой фрагмент начала поэмы:

Мой дядя самых честных правил когда не в шутку занемог он уважат себя заставил и лучше выдумат не мог его пример другим наука.

Литературный текст поэмы А.С. Пушкина – осмысленная и зависимая последовательность 32 букв русского алфавита. Двухбуквенная последовательность гласных и согласных букв, выделенная из этого фрагмента текста, как мы видим, никакого явного смысла не имеет:

сгг сгсг сгсгс сгсссгс ссгсгс сгссг сг с сгссг сгсгсгс гс гсгсгс сгсг сгссгсгс г сгссг сгсгсгг сг сгс гсс ссгсгс ссгсгс сггсг.

Соответственно этому Марков допускает существование неизвестной постоянной вероятности p – быть букве гласной и приближенную величину числа p он ищет из наблюдений, считая число появившихся гласных и согласных букв. Кроме числа p Марков нашел, также из наблюдений, приближенные величины двух других чисел p_1 и p_2, представляющих вероятности:

первое, p_1 – гласной букве следовать за гласной,

второе, p_2 – гласной букве следовать за согласной.

Разыскивая число p, Марков сначала нашел 200 приближенных величин, из которых выводится среднее арифметическое. А именно вся последовательность 20 000 букв разбивается на 200 последовательностей по 100 букв; подсчитывается сколько гласных в каждой сотне букв; получаются 200 чисел, которые при делении на 100 дают двести приближенных величин p. Полученное таким способом значение p оказалось равным 0,432.

Вычисление вероятностей p_1 и p_2 проводится следующим образом: просматривается весь текст из 20 000 букв, подсчитывается, сколько в нем встречается пар гласная, гласная; получается число 1104, которое при делении на число всех гласных в тексте дает для p_1 приближенную величину

$$\frac{1104}{8638} \approx 0{,}128.$$

Подобным образом для p_2 получается приближенная величина

$$\frac{7534}{11362} \approx 0{,}663.$$

Подставив, полученные значения в формулу $p = \frac{p_2}{1 + p_2 - p_1}$, находим число 0,4319, близкое к уже полученному 0,432. Заметим, что поделив число гласных в тексте, равное 8638 на 20000, получим величину 0,4319.

В таком совпадении нет ничего удивительного, позже из теоремы эргодичности вытекает, что если переходные вероятности p_1 и p_2 оценивать частотами (как это делал А.А. Марков), то предел p совпадает с частотами

гласной буквы в тексте.

Отсюда видно, что вероятность букве быть гласной значительно изменяется в зависимости от того, предшествует ей гласная или согласная.

Такое же исследование Марков выполнил над произведением другого автора (С.Т.Аксаков, “Детские годы Багрова-внука”), рассматривая совокупность 100000 букв. Полученные результаты показывают, что если рассматривать буквы какого-нибудь текста, то вероятность гласной и согласной изменяется в зависимости от характера одной или двух предыдущих букв.

Конечное число состояний цепи.

Рассмотрим теперь случай цепи Маркова с конечным числом состояний.

Будем предполагать, что

$$\Omega = \{\omega : \omega = (x_0, x_1, ..., x_n), x_i \in X\},$$

где X – некоторое конечное множество состояний. Пусть заданы также неотрицательные функции $p_0(x)$, $p_1(x, y)$,…, $p_n(x, y)$ такие, что

$$\sum_{x \in X} p_0(x) = 1, \tag{3.3}$$

$$\sum_{y \in X} p_k(x, y) = 1 \ , \ k = 1, ..., n \ , \ x \in X \ . \tag{3.4}$$

Для каждого исхода $\omega = (x_0, x_1, ..., x_n)$ положим

$$p(\omega) = p_0(x_0) p_1(x_0, x_1) ... p_n(x_{n-1}, x_n) \, . \tag{3.5}$$

Нетрудно проверить, что $\sum_{\omega\in\Omega} p(\omega)=1$ и, следовательно, набор этих чисел $p(\omega)$ вместе с пространством Ω определяют некоторую вероятностную модель, которую принято называть *моделью испытаний, связанных в цепь Маркова.* Множество X называется *пространством состояний* цепи. Набор вероятностей $p_0(x)$, $x\in X$ называют *начальным* распределением, а матрицу $\|p_k(x,y)\|$, $x,y\in X$ где $p_k(x,y)=p(x_k=y|x_{k-1}=x)$ – *матрицей переходных вероятностей* из состояний x в состояния y в момент $k=1,...,n$. В том случае, когда переходные вероятности $p_k(x,y)$ не зависят от k, $p_k(x,y)=p(x,y)$, последовательность $x_0,...,x_n$ называется *однородной* марковской цепью с матрицей переходных вероятностей $\|p(x,y)\|$. Заметим, что матрица $\|p(x,y)\|$ является стохастической: ее элементы неотрицательны и сумма элементов любой ее строки равна единице, $\sum_y p(x,y)=1,\ x\in X$.

Будем обозначать через $x=(x_k,\Pi,\mathrm{P})$ однородную марковскую цепь с вектором (строкой) начальных вероятностей $\Pi=\|p_i\|$ и матрицей переходных вероятностей $\mathrm{P}=\|p_{ij}\|$. Ясно, что

$$p_{ij}=\mathrm{P}(x_1=j|x_0=i)=...=\mathrm{P}(x_n=j|x_{n-1}=i).$$

Обозначим $p_{ij}^{(k)}=\mathrm{P}(x_k=j|x_0=i)$ – вероятность перехода за k шагов из состояния i в состояние j, и $p_j^k=P(x_k=j)$ – вероятность нахождения частицы в момент времени k в точке j. Пусть также

$$\Pi^{(k)}=\|p_i^{(k)}\|,\ \mathrm{P}^{(k)}=\|p_{ij}^{(k)}\|.$$

Переходные вероятности $p_{ij}^{(k)}$ удовлетворяют *уравнению Колмогорова-Чепмена* [40]

$$p_{ij}^{(k+l)} = \sum_{\alpha} p_{i\alpha}^{(k)} p_{\alpha j}^{(l)}, \tag{3.6}$$

или в матричной форме

$$\mathrm{P}^{(k+l)} = \mathrm{P}^{(k)} \cdot \mathrm{P}^{(l)}. \tag{3.7}$$

Особо важны следующие два частных случая уравнения (3.6):
обратное уравнение

$$p_{ij}^{(l+1)} = \sum_{\alpha} p_{i\alpha} p_{\alpha j}^{(l)} \tag{3.8}$$

и *прямое уравнение*

$$p_{ij}^{(k+1)} = \sum_{\alpha} p_{i\alpha}^{(k)} p_{\alpha j} \cdot \tag{3.9}$$

В матричной форме прямые и обратные уравнения записываются соответственно следующим образом:

$$\mathrm{P}^{(k+1)} = \mathrm{P}^{(k)} \cdot \mathrm{P}, \tag{3.10}$$

$$\mathrm{P}^{(k+1)} = \mathrm{P} \cdot \mathrm{P}^{(k)}. \tag{3.11}$$

Аналогично для (безусловных) вероятностей $p_j^{(k)}$ получаем, что

$$p_j^{(k+l)} = \sum_{\alpha} p_{\alpha}^{(k)} p_{\alpha j}^{(l)}, \qquad (3.12)$$

или в матричной форме

$$\Pi^{(k+l)} = \Pi^{(k)} \cdot \mathrm{P}^{(l)}.$$

В частности,

$$\Pi^{(k+1)} = \Pi^{(k)} \cdot \mathrm{P}^{(1)} \qquad (\textit{прямое уравнение})$$

и

$$\Pi^{(k+1)} = \Pi^{(1)} \cdot \mathrm{P}^{(k)} \qquad (\textit{обратное уравнение}).$$

Поскольку

$$\mathrm{P}^{(1)} = \mathrm{P},\ \Pi^{(0)} = \Pi$$

то

$$\mathrm{P}^{(k)} = \mathrm{P}^k,\ \Pi^{(k)} = \Pi\mathrm{P}^k. \qquad (3.13)$$

Тем самым для однородных марковских цепей вероятности перехода за k шагов $p_{ij}^{(k)}$ являются элементами k- х степеней матриц P, в связи с чем многие свойства этих цепей можно изучать методами матричного анализа. В частности, подтвердить результат Маркова, вытекающий из соотношения (3.2).

Рассмотрим однородную марковскую цепь с двумя состояниями 0 и 1 и матрицей

$$\mathrm{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

У Маркова состояние 0 обозначается буквой E, а состояние 1 - F. Нетрудно подсчитать, по индукции, что

$$\mathrm{P}^n = \frac{1}{2-p_{00}-p_{11}}\begin{pmatrix}1-p_{11} & 1-p_{00}\\ 1-p_{11} & 1-p_{00}\end{pmatrix}+$$

$$+\frac{(p_{00}+p_{11}-1)^n}{2-p_{00}-p_{11}}\begin{pmatrix}1-p_{00} & -(1-p_{00})\\ -(1-p_{11}) & 1-p_{11}\end{pmatrix}.$$

Отсюда видно, что если элементы матрицы P таковы, что $|p_{00}+p_{11}-1|<1$, (т.е. выполняются неравенства (3.1)), то при $n\to\infty$ скорость сходимости к пределу

$$\mathrm{P}^n \to \frac{1}{2-p_{00}-p_{11}}\begin{pmatrix}1-p_{11} & 1-p_{00}\\ 1-p_{11} & 1-p_{00}\end{pmatrix},$$

является геометрической. В обозначениях Маркова это означает, что

$$\lim_n p_{i0}^{(n)} = \frac{1-p_{11}}{2-p_{00}-p_{11}} = \frac{p_2}{1+p_2-p_1} = p,$$

$$\lim_n p_{i1}^{(n)} = \frac{1-p_{00}}{2-p_{00}-p_{11}} = \frac{1-p_1}{1+p_2-p_1}. \qquad (3.14)$$

Поэтому результат Маркова легко следует из соотношений (3.13) и (3.14).

Таким образом, если $|p_{00}+p_{11}-1|<1$, то поведение рассматриваемой марковской цепи подчиняется следующей закономерности: влияние начального состояния на вероятность нахождения частицы в том или ином состоянии (т.е. $\Pi^{(k)}$) исчезает с ростом времени: $p_{ij}^{(n)}$ сходится к

предельным значениям π_j, не зависящим от i и образующим распределение вероятностей

$$\pi_0 \geq 0,\ \pi_1 \geq 0,\ \pi_0 + \pi_1 = 1.$$

То же самое можно сказать и о $\Pi^{(k)}$. Если к тому же все элементы $p_{ij} > 0$, то тогда все предельные значения $\pi_0 > 0$, $\pi_1 > 0$.

Эргодическая теорема.

Следующая известная теорема описывает широкий класс марковских цепей, обладающих так называемым свойством *эргодичности*: пределы

$$\pi_j = \lim_n p_{ij}^{(n)}$$

не только существуют, не зависят от i, образуют распределение вероятностей

$$\pi_j \geq 0,\ \sum_j \pi_j = 1,$$

но и таковы, что $\pi_j \geq 0$ при всех j. Такие распределения π_j называются *эргодическими*. Из этой теоремы результат Маркова (3.2) следует очевидным образом.

Эргодическая теорема [40]. Пусть $\mathrm{P} = \left\| p_{ij} \right\|$ – матрица переходных вероятностей марковской цепи с конечным множеством состояний $\mathrm{X} = \{1,2,...,N\}$.

а) Если найдется n_0 такое, что

$$\min_{i,j} p_{ij}^{(n_0)} > 0, \tag{3.15}$$

то существуют числа $\pi_1,...,\pi_N$ такие, что

$$\pi_j > 0, \ \sum_j \pi_j = 1 \tag{3.16}$$

и для любого $i \in \mathrm{X}$

$$p_{ij}^{(n)} \to \pi_j, \ n \to \infty. \tag{3.17}$$

b) Обратно, если существуют числа $\pi_1,...,\pi_N$, удовлетворяющие условиям (3.16) и (3.17), то найдется n_0 такое, что выполнено условие (3.15).

c) Числа $\pi_1,...,\pi_N$ удовлетворяют системе уравнений

$$\pi_j = \sum_{\alpha=1}^{N} \pi_\alpha p_{\alpha j}, \ j = 1,..., N. \tag{3.18}$$

Из теоремы эргодичности вытекает, что вероятность нахождения частицы в определенном состоянии стабилизируется с течением времени и в дальнейшем почти не меняется по длине цепочки.

Сходимость $p_{ij}^{(n)}$ к предельным значениям π_j происходит с геометрической скоростью, знаменатель прогрессии равен $(1 - \min_{i,j} p_{ij})$.

Результат Маркова (3.2) очевидным образом вытекает из соотношений (3.18). Решая систему уравнений

$$\pi_0 = \pi_0 p_{00} + \pi_1 p_{10}, \ \pi_0 + \pi_1 = 1,$$

получаем (в обозначениях Маркова), что

$$\pi_0 = \frac{p_2}{1 + p_2 - p_1} = p\,.$$

Предложение 3.1. Если в соотношениях (3.18) переходные вероятности $p_{\alpha j}$ заменить частотами

$$\hat{p}_{ij} = \frac{m(ij)}{m(i)}$$

то системе уравнений (3.18) удовлетворяют частоты

$$\pi_j = \frac{m(j)}{m},$$

где $m(ij)$ – число пар состояний (ij), $m(i)$ $(m(j))$ – число состояний $i(j)$, m – длина цепи.

Доказательство проводится прямой подстановкой в (3.18).

3.2. Статистическое оценивание переходных вероятностей в цепях Маркова

Чтобы запустить в работу процедуры обучения необходимо построить оценки переходных вероятностей и воспользоваться соотношением (3.5). Как и в схеме Бернулли для независимых признаков переходные вероятности заменяются частотами. Однако исследование этих оценок достаточно сложно, и для их изучения привлекаем асимптотическую теорию, разработанную в [4].

Предположим, существует m объектов, которые описываются цепью Маркова. Обозначим времена наблюдений $t = 0,1,...,T$, состояния $i = 1,...,k$, $p_{ij}(t)$ $(i, j = 1,...,k,\ t = 1,...,T)$ – вероятность состояния j в момент времени t при заданном состоянии i в момент времени $t-1$. Далее рассматриваются как стационарные переходные вероятности такие, что $p_{ij}(t) = p_{ij}$ для всех $t = 0,1,...,\mathrm{T}$, так и нестационарные, которые меняются со временем. Пусть $m_i(0)$ – число объектов, находящихся в состоянии i в начальный момент 0, $i \in \{1,...,k\}$. Вначале считаем, что $m_i(0)$ – неслучайные величины, т.е. фиксированные числа.

Пусть $i(0), i(1),..., i(T)$ – последовательность состояний объекта в моменты времени $t = 0,...,T$. При заданном начальном состоянии $i(0)$ всего есть k^{T} возможных последовательностей, которые представляют взаимно исключающие события с вероятностями

$$p_{i(0)i(1)}\ p_{i(1)i(2)} \cdots p_{i(T-1)i(T)} \tag{3.21}$$

в ситуации, когда переходные вероятности стационарны. В случае нестационарных переходных вероятностей $p_{i(t-1)i(t)}$ заменяются на $p_{i(t-1)i(t)}(t)$.

Обозначим $m_{ij}(t)$ число объектов, находящихся в состоянии i в момент времени $t-1$ и в состоянии j в момент времени t, $(i, j = 1,...,k$, $t = 1,...,T)$.

Пусть $m_{i(0)i(1)...i(T)}$ – число объектов, у которых последовательность состояний есть $i(0)i(1)...i(T)$. Тогда

$$m_{gj}(t) = \sum m_{i(0)i(1)...i(T)}\ , \tag{3.22}$$

где суммирование проводится для всех таких величин, у которых $i(t-1)=g$ и $i(t)=j$. Пусть

$$m_i(t-1)=\sum_{j=1}^{k} m_{ij}(t),$$

$$m_{ij}=\sum_{t=1}^{T} m_{ij}(t). \qquad (3.23)$$

Известно, что для стационарных переходных вероятностей p_{ij} оценками максимального правдоподобия являются величины

$$\hat{p}_{ij}=\frac{m_{ij}}{\sum_{j=1}^{k} m_{ij}}=\frac{\sum_{t=1}^{T} m_{ij}(t)}{\sum_{j=1}^{k}\sum_{t=1}^{T} m_{ij}(t)}=\frac{\sum_{t=1}^{T} m_{ij}(t)}{\sum_{t=0}^{T-1} m_i(t)}. \qquad (3.24)$$

Для нестационарных переходных вероятностей такие оценки имеют вид

$$\hat{p}_{ij}(t)=\frac{m_{ij}(t)}{m_i(t-1)}=\frac{m_{ij}(t)}{\sum_{j=1}^{k} m_{ij}(t)}. \qquad (3.25)$$

Основная трудность исследования этих оценок состоит в том, что знаменатель в (3.25) есть случайная величина, а не фиксированная, как в схеме Бернулли. Например, в булевом случае при переходе по времени от $t-1$ к t «разыгрываются» две переходные вероятности (всего их четыре, но есть два связывающих уравнения), определяемые соотношениями

$$p_{00}(t) + p_{01}(t) = 1 \text{ и } p_{10}(t) + p_{11}(t) = 1.$$

Детерминированной величиной является сумма

$$m_{00}(t) + m_{01}(t) + m_{10}(t) + m_{11}(t) = m.$$

В отличие от схемы Бернулли математические ожидания оценок (3.24), (3.25) не совпадают с их точными значениями p_{ij} и $p_{ij}(t)$, т.е. эти оценки являются смещенными. Поэтому нужно исследовать их асимптотическое поведение.

Для этого рассмотрим величины $m_{ij}(t)$. Считаем, что

$$\frac{m_s(0)}{\sum m_j(0)} \to \eta_s \quad \text{при} \quad \sum m_j(0) \to \infty,\ \eta_s > 0,\ \sum \eta_s = 1.$$

Для каждого фиксированного начального состояния $i(0)$ множество объектов $n_{i(0)i(1)\ldots i(T)}$ является мультиномиальной случайной величиной с объемом выборки $m_{i(0)}(0)$ и вероятностями

$$p_{i(0)i(1)} p_{i(1)i(2)} \cdots p_{i(T-1)i(T)} \cdot$$

Поэтому асимптотическая нормальность при увеличении объема выборки вытекает из предельных теорем теории вероятностей, а с другой стороны, $m_{ij}(t)$ являются линейными комбинациями мультиномиальных величин и отсюда также вытекает их асимптотическая нормальность. Проведем обоснование этих выводов.

Пусть $\mathrm{P} = \left\| p_{ij} \right\|$, а $p_{ij}^{(t)}$ - элементы матрицы P^t. Тогда $p_{ij}^{(t)}$ есть вероятность состояния j в момент времени t при заданном i в момент времени 0. Обозначим через $m_{s;ij}(t)$ число последовательностей объектов, имеющих состояния s в момент времени 0, i в - $t-1$ и j в - t. Рассмотрим моменты низкого порядка величин

$$m_{ij}(t) = \sum_{s=1}^{k} m_{s;ij}(t) \tag{3.26}$$

Вероятность, соответствующая $m_{s;ij}(t)$, есть $p_{si}^{(t-1)} p_{ij}$ с объемом выборки, равным $m_s(0)$. Поэтому

$$\mathrm{M} m_{s;ij}(t) = m_s(0)\, p_{si}^{(t-1)} p_{ij}, \tag{3.27}$$

$$\mathrm{D} m_{s;ij}(t) = m_s(0)\, p_{si}^{(t-1)} p_{ij} [1 - p_{si}^{(t-1)} p_{ij}], \tag{3.28}$$

$$\operatorname{cov}(m_{s;ij}(t), m_{s;eh}(t)) = -m_s(0)\, p_{si}^{(t-1)} p_{ij}\, p_{se}^{(t-1)} p_{eh},$$
$$(i, j) \neq (e, h), \tag{3.29}$$

так как величины $m_{s;ij}(t)$ имеют мультиномиальное распределение.

Вычислим теперь моменты величин

$$m_{s;ij}(t) - m_{s;i}(t-1)\, p_{ij}, \quad \text{где } m_{s;i}(t-1) = \sum\nolimits_j m_{s;ij}(t).$$

Очевидно, что условное распределение $m_{s;ij}(t)$ при заданном $m_{s;i}(t-1)$ мультиномиально с вероятностью p_{ij}. Поэтому

$$\mathrm{M}\{m_{s;ij}(t) \mid m_{s;i}(t-1)\} = p_{ij} m_{s;i}(t-1), \tag{3.30}$$

$$\mathrm{M}(m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}) =$$

$$= \mathrm{MM}\{[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}] \mid m_{s;i}(t-1)\} = 0. \tag{3.31}$$

Дисперсия этой величины определяется выражением

$$\mathrm{M}[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}]^2 =$$

$$= \mathrm{MM}\{[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}]^2 \mid m_{s;i}(t-1)\} =$$

$$= \mathrm{M} m_{s;i}(t-1) p_{ij}(1 - p_{ij}) =$$

$$= m_s(0) p_{si}^{(t-1)} p_{ij}(1 - p_{ij}). \tag{3.32}$$

Ковариации таких пар величин определяются соотношениями

$$\mathrm{M}[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}][m_{s;ih}(t) - m_{s;i}(t-1) p_{ih}] =$$

$$= \mathrm{MM}\{[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}] \times$$

$$\times [m_{s;ih}(t) - m_{s;i}(t-1) p_{ih}] \mid m_{s;i}(t-1)\} =$$

$$= \mathrm{M}[-m_{s;i}(t-1) p_{ij} p_{ih}] = -m_s(0) p_{si}^{(t-1)} p_{ij} p_{ih}, \; j \neq h, \tag{3.33}$$

$$\mathrm{M}[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}][m_{s;eh}(t) - m_{s;e}(t-1) p_{eh}] =$$

$$= \mathrm{MM}\{[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}] \times$$

$$\times [m_{s;eh}(t) - m_{s;e}(t-1) p_{eh}] \mid m_{s;i}(t-1), m_{s;e}(t-1)\} = 0,$$

$$i \neq e. \tag{3.34}$$

$$\mathrm{M}[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}][m_{s;eh}(t+r) - m_{s;e}(t+r-1) p_{eh}] =$$

$$= \mathrm{MM}\{[m_{s;ij}(t) - m_{s;i}(t-1) p_{ij}][m_{s;eh}(t+r) -$$

$$- m_{s;e}(t+r-1)\,p_{eh}\,]\,p_{eh}\,]\,|\,m_{s;e}(t+r-1), m_{s;i}(t-1), m_{s;ij}(t)\} = 0\,,$$

$$r > 0\,. \qquad (3.35)$$

Итак, показано, что случайные величины $m_{s;ij}(t) - m_{s;i}(t-1)\,p_{ij}$ для $j = 1,...,k$ имеют среднее 0, а дисперсии и ковариации аналогичны таковым для мультиномиальных величин с вероятностями p_{ij} и объемом выборки $n_s(0)\,p_{si}^{(t-1)}$. Случайные величины $m_{s;ij}(t) - m_{s;i}(t-1)\,p_{ij}$ и $m_{s;eh}(g) - m_{s;e}(g-1)\,p_{eh}$ некореллированы, если $t \neq g$ или $i \neq e$.

Поскольку $m_s(0)$ фиксировано, то величины $m_{s;ij}(t)$ и $m_{l;eh}(t)$ независимы, если $s \neq l$. Поэтому

$$\mathrm{M}[\,m_{ij}(t) - m_i(t-1)\,p_{ij}\,] = 0\,, \qquad (3.36)$$

$$\mathrm{M}[\,m_{ij}(t) - m_i(t-1)\,p_{ij}\,]^2 = \sum_{s=1}^{k} m_s(0)\,p_{si}^{(t-1)}\,p_{ij}(1-p_{ij})\,, \qquad (3.37)$$

$$\mathrm{M}[\,m_{ij}(t) - m_i(t-1)\,p_{ij}\,][\,m_{ih}(t) - m_i(t-1)\,p_{ih}\,] =$$

$$= -\sum_{s=0}^{k} m_s(0)\,p_{si}^{(t-1)}\,p_{ij}\,p_{ih}\;,\; j \neq h\,, \qquad (3.38)$$

$$\mathrm{M}[\,m_{ij}(t) - m_i(t-1)\,p_{ij}\,][\,m_{eh}(g) - m_e(g-1)\,p_{eh}\,] = 0\,,$$

$$t \neq g \text{ или } i \neq e\,. \qquad (3.39)$$

Асимптотическое распределение оценок. Покажем, что при $m \to \infty$ величина

$$\sqrt{m}(\hat{p}_{ij} - p_{ij}) = \sqrt{m}\left[\frac{\sum_{t=1}^{T} m_{ij}(t)}{\sum_{t=1}^{T} m_i(t-1)} - p_{ij}\right] =$$

$$= \sqrt{m}\left[\frac{\sum_{t=1}^{T} m_{ij}(t) - p_{ij} m_i(t-1)}{\sum_{t=1}^{T} m_i(t-1)}\right] =$$

$$= \sqrt{m}\left[\frac{\sum_{s=1}^{k}\sum_{t=1}^{T} m_{s;ij}(t) - p_{ij} m_{s;i}(t-1)}{\sum_{t=1}^{T} m_i(t-1)}\right] \qquad (3.40)$$

имеет предельное нормальное распределение, а также определим математическое ожидание, дисперсию и ковариацию этого предельного распределения. Поскольку $m_{s;ij}(t)$ является мультиномиальной величиной, известно, что

$$\frac{m_{s;ij}(t)}{m} \approx \frac{m_{s;ij}(t)}{m_s(0)} \eta_s \qquad (3.41)$$

сходится по вероятности к математическому ожиданию при $\frac{m_s(0)}{m} \to \eta_s$. Поэтому

$$p \lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{T} m_i\,(t-1) = \lim_{m\to\infty} \frac{1}{m} \mathrm{M} \sum_{t=1}^{T} m_i\,(t-1) =$$

$$= \sum_{s=1}^{k} \eta_s \sum_{t=1}^{T} p_{si}^{(t-1)} . \qquad (3.42)$$

Следовательно, как вытекает из [41], величина $\sqrt{m}(\hat{p}_{ij} - p_{ij})$ имеет то же предельное распределение, что и

$$\frac{\sum_{t=1}^{T} [m_{ij}(t) - p_{ij} m_i(t-1)] / \sqrt{m}}{\sum_{s=1}^{k} \sum_{t=1}^{T} \eta_s p_{si}^{(t-1)}} . \qquad (3.43)$$

Из предыдущих результатов вытекает, что числитель в (3.41) имеет среднее 0 и дисперсию

$$\mathrm{M}[\sum_{t=1}^{T} m_{ij}(t) - p_{ij} m_i(t-1)]^2 / m =$$

$$= \Big(\sum_{s=1}^{k} \sum_{t=1}^{T} m_s(0) p_{si}^{(t-1)} p_{ij}(1 - p_{ij}) \Big) / m . \qquad (3.44)$$

Ковариация двух разных числителей в (3.43) определяется выражением

$$\mathrm{M}\Big([\sum_{t=1}^{T} m_{ij}(t) - p_{ij} m_i(t-1)][\sum_{t=1}^{T} m_{eh}(t) - p_{eh} m_e(t-1)] \Big) / m =$$

$$= -\delta_{ie} \Big(\sum_{s=1}^{k} \sum_{t=1}^{T} m_s(0) p_{si}^{(t-1)} p_{ij} p_{eh} \Big) / m , \qquad (3.45)$$

где $\delta_{ie} = 0$, если $i \neq e$ и $\delta_{ii} = 1$.

Обозначим

$$\sum_{s=1}^{k}\sum_{t=1}^{T}\eta_s p_{si}^{(t-1)} = \phi_i \,. \qquad (3.46)$$

Тогда предельной дисперсией числителя в (3.43) является величина $\phi_i p_{ij}(1-p_{ij})$, которая с точностью до ϕ_i совпадает с дисперсией бернуллиевской величины. Предельной ковариацией двух различных числителей является – $\delta_{ie}\phi_i p_{ij} p_{eh}$. Так как числители в (3.43) есть линейные комбинации нормированных мультиномиальных величин с постоянными вероятностями и возрастающим объемом выборки, то они асимптотически нормальны, а дисперсии и ковариации этого предельного распределения являются пределами соответствующих дисперсий и ковариаций. Этот факт вытекает из теоремы 2 [42].

Поскольку $\sqrt{m}(\hat{p}_{ij} - p_{ij})$ имеют такое же предельное распределение, что и выражение (3.43), случайные величины $\sqrt{m}(\hat{p}_{ij} - p_{ij})$ имеют предельное нормальное распределение со средним 0, дисперсиями

$$\frac{p_{ij}(1-p_{ij})}{\phi_i} \qquad (3.47)$$

и ковариациями

$$-\frac{\delta_{ie} p_{ij} p_{eh}}{\phi_i}\,. \qquad (3.48)$$

Соответственно величины $\sqrt{m\phi_i}(\hat{p}_{ij} - p_{ij})$ имеют предельное нормальное распределение со средним 0, дисперсиями $p_{ij}(1-p_{ij})$ и

ковариациями $-\delta_{ie} p_{ij} p_{eh}$. Аналогично получаем, что совокупность величин $\sqrt{m_i^*}(\hat{p}_{ij} - p_{ij})$ имеет предельное нормальное распределение со средним 0, дисперсиями $p_{ij}(1-p_{ij})$ и ковариациями $-\delta_{ie} p_{ij} p_{eh}$, где $m_i^* = \sum_{t=0}^{T-1} m_i(t)$.

Другими словами, совокупность случайных величин $\sqrt{m\phi_i}(\hat{p}_{ij} - p_{ij})$ для заданного состояния i имеет такое же предельное распределение, что и оценки мультиномиальных вероятностей p_{ij} с объемом выборки $m\phi_i$. Величины $\sqrt{m\phi_i}(\hat{p}_{ij} - p_{ij})$ для k различных значений i $(i = 1,2,...,k)$ асимптотически независимы и, следовательно, имеют предельное распределение, аналогичное тому, которое можно было бы получить, проводя подобные действия над оценками мультиномиальных вероятностей p_{ij} по k независимым выборкам объемом $m\phi_i$ $(i = 1,2,...,k)$.

Подытожим полученные основные результаты, которые понадобятся в дальнейшем, в виде следующей теоремы [4].

Теорема 3.1. Величины $\sqrt{m}(\hat{p}_{ij} - p_{ij})$ имеют предельное нормальное распределение со средним 0, дисперсиями (3.47) и для k различных значений i $(i = 1,2,...,k)$ асимптотически независимы.

Нестационарные переходные вероятности. Оценки

$$\hat{p}_{ij}(t) = \frac{m_{ij}(t)}{m_i(t-1)} \tag{3.49}$$

для заданного состояния i и времени t имеют такое же асимптотическое распределение, что и оценки мультиномиальных вероятностей с объемом

выборки $\mathrm{M}m_i(t-1)$, и величины $\hat{p}_{ij}(t)$ для двух различных значений состояний i $(i=1,2,...,k)$ или времени t асимптотически независимы.

Ранее предполагалось, что $m_i(0)$ – неслучайные величины. Рассмотрим теперь ситуацию, когда $m_i(0)$ – мультиномиальная случайная величина с вероятностью η_i и объемом выборки m. Оценка максимального правдоподобия η_i есть

$$\hat{\eta}_i = \frac{m_i(0)}{m}. \tag{3.50}$$

Средние, дисперсии и ковариации величин $m_{ij}(t) - m_i(t-1)p_{ij}$ определяются формулами (3.36) – (3.39), в которых $m_s(0)$ заменяются на $m\eta_s$. В таком случае $m_{ij}(t) - m_i(t-1)p_{ij}$ некоррелированы с $m_i(0)$. Асимптотические свойства величин $\sqrt{m}(\hat{p}_{ij} - p_{ij})$ остаются прежними.

Асимптотические выражения дисперсий и ковариаций упрощаются, если цепь Маркова стартует из стационарного начального состояния, т.е. такого, что выполняется система уравнений (3.18)

$$\sum_{s=1}^{k} \eta_s p_{si} = \eta_i \ . \tag{3.51}$$

Тогда

$$p_i^{(1)} = \sum_{\alpha} \eta_\alpha p_{\alpha i} = \eta_i$$

и вообще $p_i^{(t)} = \eta_i$. Иначе говоря, если в качестве начального распределения взять $(\eta_1,...,\eta_k)$, то это распределение не будет изменяться со временем, т.е. для любого t

$$P(x_t = i) = P(x_0 = i),\ i = 1,2,...,k.$$

Более того, с таким начальным распределением совместное распределение вектора $(x_k, x_{k+1},..., x_{k+l})$ не зависит от k для любого l. Поэтому $\sum_\alpha \eta_\alpha p_{\alpha i}^{(t-1)} = \eta_i$ и из соотношения (3.46) получаем явное выражение для ϕ_i:

$$\phi_i = T\eta_i \tag{3.52}$$

Таким образом, для стационарных цепей дисперсия оценок выражается соотношением

$$\frac{p_{ij}(1-p_{ij})}{mT\eta_i}. \tag{3.53}$$

В геноме дисперсия (3.53) исчезающе мала, поскольку m – порядка десятка тысяч, а T (длина аминокислотных цепочек белков) – порядка тысяч.

3.3. Байесовские процедуры распознавания на нестационарных цепях Маркова.

Рассмотрим случай, когда $x_1, x_2,..., x_n$ образуют последовательность зависимых случайных величин, связанных в цепь Маркова с конечным числом состояний.

Для модели цепи Маркова первого порядка вероятность цепочки $x_1, x_2, ..., x_n$ задается соотношением

$$P(x_1, x_2, ..., x_n | f) = P(x_1 | f) \times P(x_2 | x_1, f) \times ... \times P(x_n | x_{n-1}, f), \quad (3.54)$$

где $P(x_k | x_{k-1}, f)$, $k = 2, ..., n$ – нестационарные переходные вероятности; как и в дискретном случае полагаем, что признаки $x_j \in \{0,1,..., g-1\}$, $j = 1,2,..., n$; $f \in \{0,1,..., h-1\}$; g, h – натуральные числа. В численных расчетах используются оценки переходных вероятностей, построенные в виде частот

$$\hat{p}(x_k = j | x_{k-1} = i, f) = \frac{m(x_{k-1} = i, x_k = j, f)}{m(x_{k-1} = i, f)}, \quad (3.55)$$

где $m(x_{k-1} = i,\ x_k = j,\ f)$ – число объектов $x_1, x_2, ..., x_n$, принадлежащих заданному классу f в обучающей выборке, у которых признак x_{k-1} принимает значение i, и признак x_k – значение j; $m(x_{k-1} = i, f)$ – число объектов, для которых признак x_{k-1} принимает значение i.

В теореме 3.1 исследовались свойства оценок переходных вероятностей (3.55). Разность величин $\hat{p}(x_k | x_{k-1}, f) - p(x_k | x_{k-1}, f)$ имеет асимптотическое нормальное распределение со средним 0 и дисперсией порядка $\frac{1}{m}$, где m – объем обучающей выборки, при этом оценки переходных вероятностей (3.55) асимптотически независимы. Поэтому для больших выборок оценки погрешностей байесовских процедур распознавания на нестационпрных цепях Маркова аналогичны оценкам, полученным для независимых признаков в дискретном случае, но они уже носят асимптотический характер.

Выводы. В разделе 3.1 исследуются эргодические свойства цепей Маркова и излагается статистический анализ литературных текстов, проведенный А.А. Марковым. Аналогичные вопросы для четырехбуквенного алфавита ДНК и двадцатибуквенного алфавита белков рассматриваются в разделах 4, 5.

Исследование эффективности байесовских процедур распознавания основано на статистическом анализе оценок переходных вероятностей. В отличие от независимых бернуллиевских величин математическое ожидание частот смещено и не совпадает с точными значениями вероятностей. В разделе 3.2 показано, что оценки переходных вероятностей, построенных в виде частот, асимптотически нормальны, получены дисперсии и ковариации этого предельного распределения. Оценки погрешности байесовской процедуры для нестационарных цепей Маркова аналогичны оценкам, полученным для независимых признаков.

РАЗДЕЛ 4
СТАТИСТИЧЕСКИЙ АНАЛИЗ ГЕНОМОВ

Для эффективной работы процедур распознавания на белках, необходимо учитывать специфику этих объектов и использовать наиболее подходящие модели для их описания. Поскольку инструкции по синтезу белков записаны в геномах, то целесообразно вначале провести статистический анализ геномов, а потом использовать полученные результаты при построении процедур распознавания на аминокислотных последовательностях белков.

В текущем разделе будет показано, что цепь Маркова является удобной и экономичной моделью описания геномов. Отдельное внимание будет уделено проблеме выбора параметров модели (таких как стационарность и порядок цепи). Также будут обозначены некоторые статистические закономерности свойственные для ДНК живых организмов. Полученные результаты, будут использованы для обоснования применения моделей цепей Маркова для описания аминокислотных последовательностей белков.

4.1 Структура ДНК и механизм реализации генетической информации

Напомним, что генетическая информация клетки хранится в хромосомах, представляющих собой двойную цепочку ДНК. Каждая цепочка состоит из нуклеотидных звеньев (нуклеотидов, оснований) четырех типов: A, T, C, G. Две цепочки спариваются по закону комплементарности (A соединяется с T, а C – с G) и образуют хромосому. Таким образом, одна цепочка ДНК однозначно определяет себе комплементарную и хромосому в целом.

Атомная структура нуклеотидов позволяет ввести ориентацию (направление) цепочки ДНК: один конец цепочки обозначается 5’, а другой – 3’. Цепочки в хромосоме расположены антипараллельно (направлены в

противоположные стороны), т.е. 3’-концу одной цепочки соответствует 5’-конец комплементарной цепочки и наоборот (рис. 4.1). Таким образом, хромосома в целом направления не имеет и состоит из комплементарных пар нуклеотидов, в которых измеряется величина генома.

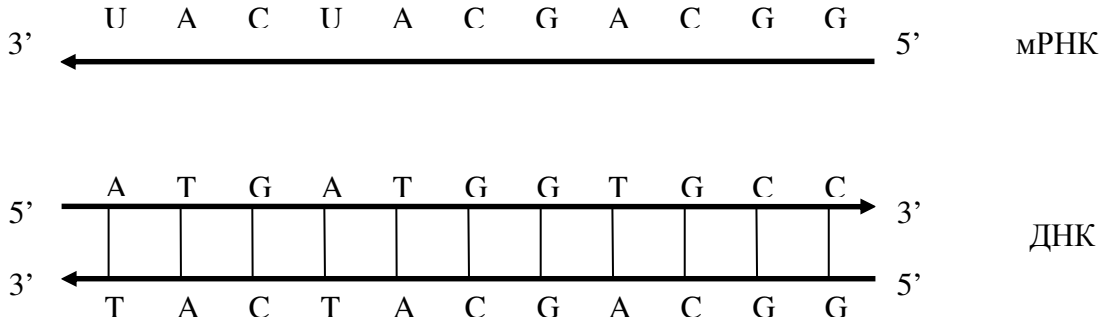

Рис. 4.1 Схематическое изображение молекулы ДНК и соответствующей ей мРНК.

Три нуклеотида (триплет) кодируют одну из 20 аминокислот, из которых строятся белки. Полимераза – фермент, считывающий информацию с ДНК для последующего синтеза белка, работает в направлении 5’ → 3’. Таким образом, в силу антипараллельности комплементарных цепочек ДНК, один и тот же участок хромосомы можно считывать по каждой из нитей (в противоположных направлениях), в результате чего, в общем случае, получаются 2 разные нуклеотидные последовательности. Действительно, если на одной из нитей имеется нуклеотидная последовательность ATG, кодирующая аминокислоту метионин, то на комплементарной нити этому же участку будет соответствовать последовательность CAT (берем последовательность, коплементарную ATG – TAC и читаем в обратном направлении, получаем CAT), кодирующая уже другую аминокислоту – гистидин (Таблица 4.1).

Процесс реализации генетической информации называется экспрессией генов и заключается в координированном синтезе множества различных

белков и РНК, выполняющих специфические функции в клетке. В геномах самых простых бактерий закодировано около 500 генов. Сотрудники лаборатории TIGR с помощью точечных отключений единичных генов пытаются определить наименьшее число генов, достаточных для жизни. Для поддержания жизнедеятельности в лабораторных условиях число генов удалось сократить до двухсот.

Синтез белка происходит в 2 шага по следующей схеме:

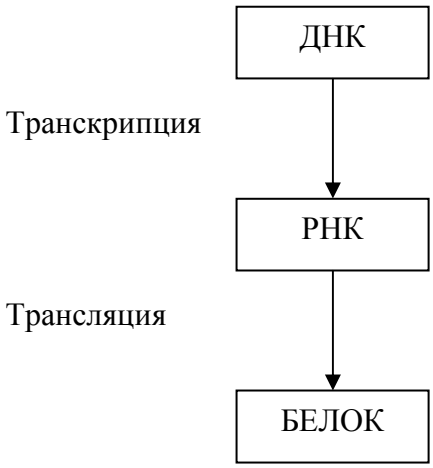

Рис. 4.2 Схема синтеза белка

Транскрипция. В процессе транскрипции белок-кодирующая последовательность копируется с ДНК на РНК. Фермент полимераза находит начало гена и связывается с ДНК, разрывая водородные связи между комплементпрными нитями. Смещаясь в направлении 5’ → 3’, полимераза копирует информацию с одной из нитей ДНК на молекулу РНК. РНК, в отличие от ДНК представляет собой цепочку, состоящую из единичных нуклеотидов (а не комплементарных пар, как в случае ДНК), кроме того РНК отличается от ДНК тем, что вместо основания тимин в РНК входит урацил. Полученная в процессе транскрипции молекула РНК называется мРНК (матричная РНК).

Полимераза узнает местонахождение начала белок-кодирующего участка на ДНК по специфичным сигнальным последовательностям (промоторам). Это последовательности «TATTGACA», находящаяся на расстоянии 35 нуклеотидов до начала белок-кодирующего участка, и «TATAAT» на расстоянии 10 нуклеотидов до начала белок-кодирующего участка. Кроме того, все белки начинаются с аминокислоты метионин: следовательно любой белок-кодирующий участок начинается с триплета, кодирующего эту аминокислоту, т.е. с «AGT».

Трансляция. В процессе трансляции аминокислотная последовательность белка синтезируется по молекуле мРНК. мРНК попадает на рибосому – белковый комплекс, ответственный за синтез беков. Рибосома ставит в соответствие каждому следующему триплету нуклеотидов мРНК одну из 20 аминокислот и смещается по мРНК на три позиции в направлении 3’ → 5’. Этот процесс продолжается, пока рибосома не достигнет стоп-кодона – триплета, который определяет конец белок-кодирующей последовательности. Таким образом аминокислотная последовательность постепенно наращивается, пока все триплеты не будут переведены в аминокислоты.

Принцип, согласно которому рибосома переводит триплет нуклеотидов в аминокислоту называется генетическим кодом. Генетический код универсален, т.е. одинаков для все живых организмов, его можно изобразить в виде таблицы (таблица 4.1).

Генетический код является вырожденным, общее количество всех триплетов, которые можно составить из оснований четырех типов (A, C, T, G) – 64, а количество аминокислотных остатков – 20. Таким образом, некоторые аминокислоты кодируются несколькими триплетами, и, следовательно, однозначного преобразования обратного генетическому коду не существует.

Таблица 4.1

Генетический код

Первое основание	Второе основание				Третье основание
	T(U)	C	A	G	
T(U)	Phe F	Ser S	Tyr Y	Cys C	T(U)
	Phe F	Ser S	Tyr Y	Cys C	C
	Leu L	Ser S	STOP	STOP	A
	Leu L	Ser S	STOP	Trp W	G
C	Leu L	Pro P	His H	Arg R	T(U)
	Leu L	Pro P	His H	Arg R	C
	Leu L	Pro P	Gln Q	Arg R	A
	Leu L	Pro P	Gln Q	Arg R	G
A	Ile I	Tre T	Asn N	Ser S	T(U)
	Ile I	Tre T	Asn N	Ser S	C
	Ile I	Tre T	Lys K	Arg R	A
	Met M	Tre T	Lys K	Arg R	G
G	Val V	Ala A	Asp D	Gly G	T(U)
	Val V	Ala A	Asp D	Gly G	C
	Val V	Ala A	Glu E	Gly G	A
	Val V	Ala A	Glu E	Gly G	G

По мере синтеза аминокислотной последовательности белок сворачивается в трехмерную структуру от которой и будет зависеть его функция в клетке [43]. Трехмерная структура белка определяется его аминокислотной последовательностью. Белки имеющие одинаковую аминокислотную последовательность образуют одинаковые пространственные структуры (с точность до 5-12%). Активные центры белков, т.е. подпоследовательности аминокислотной цепочки, играющие определяющую роль в функционировании белка в целом, всегда имеют одинаковую структуру. В противном случае белок не будет выполнять своей функции.

4.2 Статистический анализ геномов

Одной из основных особенностей в анализе аминокислотных (нуклеотидных) последовательностей является то, что частоты встречаемости соседних остатков (оснований) не являются независимыми. В частности, частоты пар соседних остатков обычно отличаются от произведений частот самих остатков. Иными словами, если p_u частота остатка типа u в последовательности, и p_{uv} – частота, с которой соседние остатки принадлежат к типам u и v, то $p_{uv} \neq p_u p_v$.

Эта особенность частот была проанализирована в [44], где также изучался вопрос соответствия между марковской цепью определенного порядка и отдельными последовательностями ДНК. Отмечалось, что анализ марковских цепей следует проводить на уровне *всего генома*, а не на уровне отдельного гена.

Поведение однородной связанной цепи первого порядка применительно к геному определяется начальным распределением вероятностей четырех букв: $p(A)$, $p(C)$, $p(G)$, $p(T)$, составляющих в сумме 1, и переходных вероятностей, записанных в виде матрицы,

$$\mathrm{P} = \begin{Vmatrix} p(AA) & p(AC) & p(AG) & p(AT) \\ p(CA) & p(CC) & p(CG) & p(CT) \\ p(GA) & p(GC) & p(GG) & p(GT) \\ p(TA) & p(TC) & p(TG) & p(TT) \end{Vmatrix},$$

где суммы вероятностей по строкам также равны единице. Переходные вероятности $p(ij)$, $i, j \in \{A, C, G, T\}$ имеют такой же смысл, как и у А.А Маркова.

Марковская цепь порядка k предполагает, что нахождение основания в определенном месте последовательности зависит только от оснований, находящихся в предшествующих k положениях. В цепи порядка 1

вероятность нахождения какого-либо основания в позиции i зависит от вероятности присутствия одного из четырех оснований в позиции i-1.

Последовательность, состоящая из независимых оснований, соответствует марковской цепи 0-го порядка. Порядок цепи может быть установлен методами правдоподобия (на основе вероятностей нуклеотидной цепочки) путем решения серии задач распознавания гипотез.

В работе [4] показано, что для больших выборок величина

$$-2\ln\frac{L(k)}{L(k+1)}$$

подчиняется известному распределению χ^2, где $L(k)$ – правдоподобие цепи порядка k. Для цепи порядка k число независимых параметров – $3\cdot 4^k$. Число степеней свободы статистики χ^2 равно разности между числами степеней свободы для каждого из двух порядков.

Проведенные расчеты на последовательностях генома человека показали, что существенное различие (на 5% -ном уровне) имеется только между цепями порядка 0 и 1.

При оценке порядка цепи k, при котором достигается наилучшее соответствие, следует иметь в виду, что цепи более высокого порядка имеют большое число степеней свободы. Это проявляется при анализе последовательностей конечной длинны (таких как ДНК и белки), которой может быть недостаточно для построения необходимых статистик.

Оценки переходных вероятностей $p(ij)$ вычисляются по формулам

$$\hat{p}(ij)=\frac{m(ij)}{m(i)}, \qquad (4.1)$$

здесь $m(ij)$ – число пар (ij), $m(i)$ – число букв i в хромосоме. Хромосомы имеют длину порядка 10^6-10^8 букв, поэтому дисперсии этих оценок малы.

В работах [45, 46] анализировались геномы высших организмов. В данной работе мы остановимся на геномах растений и бактерий. На момент написания работы в базе данных NCBI [5] имелось 286 секвенированных геномов бактерий и всего 2 секвенированных генома растений.

При анализе геномов отдельное внимание уделяется свойству комплементарности по одной цепочке ДНК. Стандартный закон комплементарности хорошо известен, он утверждает, что комплементарные нити ДНК спариваются вследствие образования водородных связей между комплементарными основаниями A-T и C-G на противоположных нитях (Рис. 4.1). Менее известен закон комплементарности по одной нити, который утверждает, что количество оснований A приблизительно совпадает с количеством оснований T по одной нити (аналогично для C-G) [45, 46]. Это свойство нетривиально и не является следствием стандартного закона комплементарности. Можно привести пример двойной цепочки ДНК, где стандартный закон комплементарности выполняется, а комплементарность по одной цепочке места не имеет. Например, пусть имеется нить ДНК на которой записано 10 – A, 20 – C, 30 – T и 40 – G. Следовательно, на противоположной нити будет записано 10 – T, 20 – G, 30 – A и 40 – C. Оказывается в ДНК живых организмов количество A (C) практически совпадает с количеством T (G) по одной цепочке, чего в приведенном примере не наблюдается.

4.2.1 Статистический анализ геномов растений

Геномы Oryza sativa (рис) и Arabidopsis thaliana (сорняк) состоят из 340 млн. п.н. (12 хромосом) и 119 млн. п.н. (5 хромосом) соответственно.

В геномах исследованных растений отчетливо наблюдается комплементарность по одной цепочке ДНК. В таблицах 4.2 и 4.3 приведены частоты нуклеотидных оснований для хромосом Arabidopsis thaliana и Oryza sativa. Частоты комплементарных оснований (A и T, C и G), подсчитаные по

одной цепочке, практически совпадают. Более того, частоты оснований для разных хромосом растения очень близки. В случае бактерий этот факт не так сильно выражен [6].

Таблица 4.2

Частоты нуклеотидных оснований в хромосомах Arabidopsis thaliana

Основание	Хромосома 1	Хромосома 2	Хромосома 3	Хромосома 4	Хромосома 5
A	0,319104835	0,320594514	0,31902276	0,319635813	0,319582
C	0,178642134	0,179878174	0,181553253	0,181409168	0,179112686
T	0,318690805	0,320643486	0,317460437	0,318210688	0,320698153
G	0,178174379	0,178756297	0,181729856	0,180581136	0,180095061

Таблица 4.3

Частоты нуклеотидных оснований в хромосомах Oryza sativa

Осно-вание	Хромосома 1	Хромосома 2	Хромосома 3	Хромосома 4	Хромосома 5	Хромосома 6
A	0,281344	0,283508	0,281362	0,279717	0,279774	0,282720
C	0,218993	0,216641	0,218547	0,220705	0,219709	0,217689
T	0,280952	0,283197	0,281806	0,278753	0,280920	0,282158
G	0,218709	0,216652	0,21825	0,220619	0,219595	0,21743
Осно-вание	Хромосома 7	Хромосома 8	Хромосома 9	Хромосома 10	Хромосома 11	Хромосома 12
A	0,282301	0,28295	0,28319	0,282057	0,286488	0,284995
C	0,216896	0,216981	0,217171	0,218773	0,21359	0,214933
T	0,282847	0,283185	0,282808	0,282258	0,285973	0,28456
G	0,217954	0,216874	0,216825	0,216688	0,213944	0,215504

Абсолютные частоты комплементарных последовательностей также практически не отличаются.

Напомним, что последовательности $x = \{x_1, x_2, ..., x_n\}$ и $y = \{y_1, y_2, ..., y_n\}$, $x_i,\ y_i \in \{\text{A, T, C, G}\}$ – комплементарны, если $x = \{\overline{y_n}, \overline{y_{n-1}}, ..., \overline{y_1}\}$,

где

$$\underline{y_i} = \begin{cases} \text{A}, & \text{если } y_i = \text{T} \\ \text{T}, & \text{если } y_i = \text{A} \\ \text{C}, & \text{если } y_i = \text{G} \\ \text{G}, & \text{если } y_i = \text{C} \end{cases} \qquad i = 1,...,n\,.$$

В таблице 4.4 приведены количества пар оснований на примере первой хромосомы каждого из растений. Отметим, что для четных n некоторые последовательности x_n будут самокомплементарны (например “CG” или “ACGT”).

Таблица 4.4

Абсолютные частоты пар оснований

Oryza sativa, хромосома 1				Arabidopsis thaliana, хромосома 1			
Пара	Кол-во	Компл. пара	Кол-во	Пара	Кол-во	Компл. пара	Кол-во
AA	3847082	TT	3834480	AA	3520106	TT	3512669
AC	2279028	GT	2274604	AC	1588532	GT	1580093
AG	2535663	CT	2535842	AG	1795172	CT	1798448
CA	2816956	TG	2809631	CA	1924186	TG	1920441
CC	2247409	GG	2242670	CC	1016264	GG	1009062
TC	2591476	GA	2588396	TC	1929073	GA	1930497
TG	2809631	CA	2816956	TG	1920441	CA	1924186
GC	2298328			GC	902633		
AT	3435406			AT	2807304		
CG	1816034			CG	697611		
TA	2844744			TA	2336338		

Подобные свойства наблюдаются для более длинных последовательностей. Конкретные результаты мы не приводим ввиду громоздкости таблиц.

Еще одним интересным свойством ДНК является комплементарность изолированных последовательностей из N одинаковых оснований, т.е. таких, которые не являются подпоследовательностями более длинных последовательностей. В таблице 4.5 представлены абсолютные частоты изолированных последовательностей одинаковых оснований длинной от 1 до 25 в третьей хромосоме каждого из растений.

Таблица 4.5

Абсолютные частоты изолированных N-ок

Oryza sativa, хромосома 3					Arabidopsis thalina, хромосома 3				
N	A	T	C	G	N	A	T	C	G
1	4814420	4820898	4505779	4504391	1	3115350	3114904	2785052	2787076
2	1320746	1322565	1140093	1138868	2	1045590	1039758	558711	560452
3	443468	443970	223147	222059	3	391242	389399	94588	93882
4	168630	168560	55470	54897	4	153450	151344	14938	15043
5	55220	55488	15707	15567	5	51715	51291	2413	2594
6	18373	18546	3494	3426	6	15862	15662	375	301
7	9996	10116	789	876	7	7350	7067	28	43
8	6396	6389	116	139	8	3406	3269	11	10
9	3309	3397	59	44	9	2274	2087	10	2
10	1858	1888	76	66	10	1407	1343	5	5
11	543	551	75	83	11	672	595	8	8
12	259	273	65	55	12	363	349	7	5
13	130	131	42	52	13	239	229	5	1
14	74	76	35	35	14	160	157	6	0
15	32	47	31	38	15	124	115	0	2
16	26	14	36	24	16	95	94	4	0
17	8	6	34	32	17	80	70	3	2
18	6	3	26	21	18	48	55	2	2
19	3	5	27	18	19	45	36	0	0
20	0	5	14	9	20	37	35	0	1
21	3	1	10	12	21	20	17	0	2
22	1	0	7	7	22	18	20	0	0
23	0	2	2	1	23	11	11	0	0
24	0	1	4	2	24	10	6	0	0
25	0	0	0	0	25	7	11	0	0

Как и у многих бактерий, в хромосомах растений наблюдается локальное расхождение частот комплементарных оснований. Это явление можно продемонстрировать на следующем примере: в то время как в одной половине хромосомы основание A преобладает над T, а в другой – наоборот, T преобладает над A, количества оснований A и T по всей длине хромосомы равны.

4.2.2 Статистический анализ геномов бактерий

Нами было исследовано ДНК 53 представителей 16 разных типов эубактерий и архебактерий. По тонкому строению клетки, выявляемому с помощью электронного микроскопа, архебактерии принципиально не отличаются от эубактерий и ближе к грамположительной (Firmicutes) их ветви. Одно из существенных отличий архебактерий связано с химическим составом клеточных стенок, в которых не обнаружен характерный для эубактерий пептидогликан. Особенность генома архебактерий — наличие многократно повторяющихся нуклеотидных последовательностей, а в генах, кодирующих белки, тРНК и рРНК, — интронов, что характерно для организации генетического материала эукариот. У некоторых архебактерий обнаружены основные гистоноподобные белки, связанные с ДНК. Функция их предположительно заключается в обеспечении определенной упаковки ДНК в нуклеоиде [47].

Бактерии были первыми живыми организмами, появившимися на Земле приблизительно 3,6 млрд. лет назад. Их развитию природа уделила больше времени, чем всем остальным организмам вместе взятым (Рис. 4.3), расселив их в различных условиях окружающей среды, к которым те с блестящим успехом приспособились, выработав все необходимые гены для поддержания процессов жизнедеятельности. С появлением эукариот и полового размножения темпы эволюции резко ускорились, так как новые организмы, видимо, “складывались” из ранее выработанных генов как из готовых деталей конструктора.

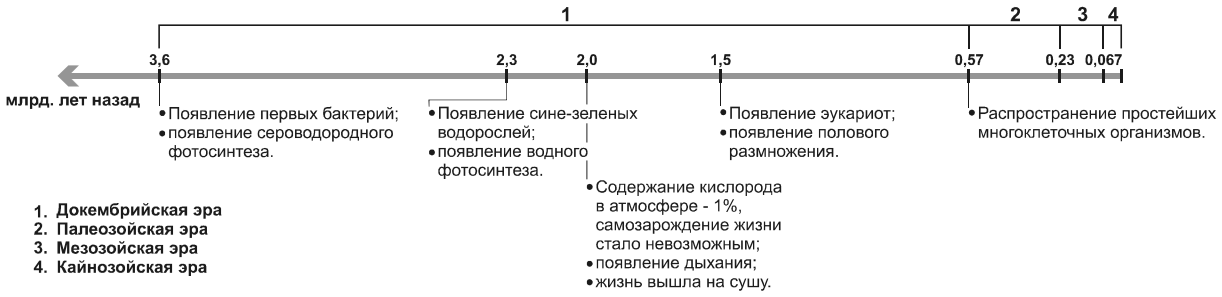

Рис. 4.3 Некоторые этапы эволюции живой природы

Следует обратить внимание на важность подобных исследований, т.к. бактерии представляют собой простейшие функционирующие клетки, что отражается на их ДНК, размеры которой колеблются от 0,5 до 10 млн. пар нуклеотидов. Самый маленький геном обнаружен на сегодняшний день у архебактерии Nanoarchaeum equitans, составляющий 490 тыс. пар нуклеотидов, кодирующих 582 гена (для сравнения: геном человека состоит из 3,2 млрд. пар нуклеотидов, кодирующих 20 – 25 тыс. генов). Относительная простота организации геномов (по сравнению с высшими организмами) в сочетании с большим количеством секвенированных геномов бактерий делает их наиболее удобными объектами для анализа.

У бактерий наблюдаетя комплементарность по одной цепочке ДНК, как и у высших эукариот [45, 46]. Вдоль одной нити ДНК выполняются соотношения

$$n(A) \approx n(T), \quad n(C) \approx n(G),$$

где $n(j)$ – число оснований j, $j \in \{A, C, G, T\}$.

Таблица 4.6

Частоты оснований по одной цепочке ДНК некоторых бактерий

Бактерия	A	T	C	G	|A-T|/n	|C-G|/n
Streptomyces avermitilis	0,1471169	0,1457176	0,3537280	0,3534376	0,0013992	0,0002904
Thermus thermophilus	0,1533830	0,1522442	0,3465650	0,3478078	0,0011389	0,0012428
Mycobacterium avium	0,1539113	0,1531026	0,3469663	0,3460198	0,0008087	0,0009464
Bordetella parapertussis	0,1591042	0,1598963	0,3378518	0,3431477	0,0007921	0,0052958
Halobacterium	0,1605246	0,1603449	0,3400778	0,3390526	0,0001797	0,0010252
Mycobacterium tuberculosis	0,1719213	0,1719813	0,3287840	0,3273034	0,0000600	0,0014806
Bradyrhizobium japonicum	0,1795740	0,1798322	0,3201739	0,3204200	0,0002582	0,0002461
Mesorhizobium loti	0,1863620	0,1861674	0,3162045	0,3112662	0,0001946	0,0049383
Gloeobacter violaceus	0,1905854	0,1894360	0,3100539	0,3099247	0,0011494	0,0001292
Bifidobacterium longum	0,1997465	0,1990552	0,3005650	0,3006332	0,0006913	0,0000682
Aeropyrum pernix	0,2156214	0,2212656	0,2835117	0,2796014	0,0056441	0,0039103
Rhodopirellula baltica	0,2243119	0,2216552	0,2791587	0,2748741	0,0026568	0,0042846
Treponema pallidum	0,2354200	0,2368278	0,2620672	0,2656842	0,0014078	0,0036170
Escherichia coli	0,2461871	0,2459159	0,2542320	0,2536650	0,0002711	0,0005671
Methanothermobacter thermautotrophicus	0,2508529	0,2537078	0,2473083	0,2481310	0,0028549	0,0008228
Dehalococcoides ethenogenes	0,2545049	0,2569925	0,2419706	0,2465320	0,0024876	0,0045614
Archaeoglobus fulgidus	0,2580316	0,2561518	0,2420584	0,2437583	0,0018798	0,0016999
Synechocystis	0,2609125	0,2618850	0,2382729	0,2389297	0,0009724	0,0006568
Thermotoga maritima	0,2696981	0,2678246	0,2276478	0,2348294	0,0018735	0,0071816
Bacillus subtilis	0,2818074	0,2830144	0,2180801	0,2170981	0,0012070	0,0009821
Bacteroides fragilis	0,2838898	0,2842045	0,2156974	0,2162084	0,0003147	0,0005110
Aquifex aeolicus	0,2841288	0,2811095	0,2168204	0,2179413	0,0030193	0,0011210
Nostoc	0,2928839	0,2936407	0,2064018	0,2070737	0,0007568	0,0006718
Chlamydophila pneumoniae	0,2957341	0,2985175	0,2025957	0,2031528	0,0027834	0,0005570
Thermoplasma volcanium	0,3016215	0,2991935	0,1990675	0,2001175	0,0024281	0,0010500
Mycoplasma pneumoniae	0,3052583	0,2946617	0,1995605	0,2005196	0,0105966	0,0009591
Haemophilus influenzae	0,3101648	0,3083168	0,1916447	0,1898486	0,0018480	0,0017961
Sulfolobus solfataricus	0,3193525	0,3227757	0,1785051	0,1793667	0,0034232	0,0008616
Parachlamydia	0,3256639	0,3271553	0,1731498	0,1740228	0,0014914	0,0008731
Staphylococcus aureus	0,3358153	0,3360093	0,1630207	0,1651546	0,0001940	0,0021339
Nanoarchaeum equitans	0,3422003	0,3422044	0,1575950	0,1580004	0,0000041	0,0004054
Mycoplasma genitalium	0,3457197	0,3373621	0,1577799	0,1591383	0,0083576	0,0013584
Campylobacter jejuni	0,3496075	0,3473311	0,1521916	0,1508698	0,0022764	0,0013218
Candidatus Pelagibacter ubique	0,3530826	0,3500866	0,1477232	0,1491077	0,0029960	0,0013845
Fusobacterium nucleatum	0,3584249	0,3700589	0,1399513	0,1315645	0,0116339	0,0083868

В табл. 4.6 *n* обозначает число оснований в одной цепочке ДНК.

Расчеты показывают, что для пар букв выполняются следующие соотношения комплементарности

$$n(AC) \approx n(GT), \quad n(AG) \approx n(CT)$$

$$n(TC) \approx n(GA), \quad n(TG) \approx n(CA)$$

$$n(AA) \approx n(TT), \quad n(CC) \approx n(GG).$$

Заметим, что пары AT, TA, CG, и GC не присутствуют в (2), поскольку они сами себе антикомплементарны (табл. 4.7).

Таблица 4.7

Количества пар оснований в геноме Escherichia coli K12

Пара	Количество	Комплементарная пара	Количество
AA	337870	TT	339482
AC	256662	GT	255608
AG	237877	CT	236061
CA	325149	TG	322239
CC	271673	GG	270137
TC	267288	GA	267247
TA	211961		
AT	309819		
GC	383931		
CG	346670		

Отдельные *n*-ки оснований связаны следующими соотношениями комплементарности

$$n(ij...k) \approx n(\bar{k}...\bar{j}\,\bar{i}),$$

где $i, j, k \in \{A, C, G, T\}$, $\overline{A} = T$, $\overline{C} = G$, $\overline{T} = A$, $\overline{G} = C$, *n*-ка $(\bar{k} \ldots \bar{j}\,\bar{i})$ – антикомплементарна *n*-ке (*ij*…*k*). Для нечетных последовательностей каждая *n*-ка имеет антикомплементарную *n*-ку, исключений в этом случае нет. Для 64 триплетов получаем 32 соотношения: кодон – антикодон (табл. 4.8).

Таблица 4.8

Количества кодонов в геноме бактерии Escherichia coli

Кодон	кол-во	Кодон	кол-во	Кодон	кол-во	Кодон	кол-во
AAA	108924	TTT	109831	CAG	104799	CTG	102909
AAC	82582	GTT	82598	CCA	86436	TGG	85141
AAG	63369	CTT	63655	CCC	47775	GGG	47495
AAT	82995	ATT	83398	CCG	87036	CGG	86877
ACA	58637	TGT	58375	CGA	70938	TCG	71739
ACC	74897	GGT	74301	CGC	115695	GCG	114632
ACG	73263	CGT	73160	CTA	26764	TAG	27243
ACT	49865	AGT	49772	CTC	42733	GAG	42465
AGA	56621	TCT	55472	GAA	83494	TTC	83848
AGC	80860	GCT	80298	GAC	54737	GTC	54221
AGG	50624	CCT	50426	GCA	96028	TGC	95232
ATA	63697	TAT	63288	GCC	92973	GGC	92144
ATC	86486	GAT	86551	GGA	56197	TCC	56028
ATG	76238	CAT	76985	GTA	52672	TAC	52592
CAA	76614	TTG	76975	TAA	68838	TTA	68828
CAC	66751	GTG	66117	TCA	84048	TGA	83491

Каждая четверка оснований имеет антикомплементраную ей четверку кроме 16 исключений, которые получаются из пар AT, TA, CG, и GC вставками в середину каждой из этих пар. Вычисления показали, что аналогичные соотношения комплементарности выполняются для возрастающих *n*-ок, расчеты не приводятся из-за больших размеров таблиц.

В геномах бактерий обнаружены другие интересные регулярности относительно повторов одинаковых комплементарных букв. Компьютер подсчитывал число изолированных последовательностей, состоящих из

одинаковых букв A, T, C, G. Изолированная буква A не входит в состав пар AA, троек AAA и т.д., пара AA не входит в состав троек AAA, четверок AAAA и т.д. Таким образом, последовательности, состоящие из одинаковых букв A, не пересекаются и в сумме дают общее число букв A в хромосоме. То же самое относится к последовательностям, состоящим из одинаковых букв T, C, G. В табл. 4.9 приведены данные о количествах последовательностей, состоящих из одинаковых букв A, T,C, G. Отсюда можно сделать вывод о том, что выполняются следующие соотношения

$$n(A...A) \approx n(T...T), \;\; n(C...C) \approx n(G...G).$$

Приведенные соотношения были подтверждены для всех исследуемых геномов бактерий.

Таблица 4.9

Количества изолированных последовательностей из n одинаковых оснований

Haemophilus influenzae					Thermoplasma acidophilum				
N	A	T	C	G	N	A	T	C	G
1	210775	211283	226132	225368	1	517201	515470	714598	714614
2	84886	83604	47375	46869	2	111010	111003	234462	234084
3	31958	32090	7331	7070	3	29300	28763	59087	59207
4	12602	12388	1570	1363	4	8595	8395	13546	13515
5	4977	4911	252	257	5	3572	3433	4331	4315
6	1949	1855	39	46	6	1107	1027	1000	983
7	513	513	6	13	7	230	230	208	211
8	69	70	4	2	8	33	30	47	55
9	8	8	0	0	9	6	6	7	8
10	0	2	0	0	10	1	1	3	2

Генетическая информация большинства прокариот, в отличие от эукариот, содержится в одной молекуле ДНК. Среди исключений, например,

бактерия Vibrio cholerae – возбудитель болезни холера, имеющая две хромосомы. Интересно, что

Так, например, хромосомы все той же Vibrio cholerae O1 biovar eltor str. N16961 имеют длины 2961118 и 1072311 пар оснований (табл. 4.10).

Таблица 4.10

Частоты оснований по хромосомам бактерий

Бактерия	Хромосома 1				Хромосома 2			
	A	T	C	G	A	T	C	G
Leptospira interrogans serovar Copenhageni	0,3249677	0,3245370	0,1757862	0,1747091	0,3242095	0,3259800	0,1738901	0,1759205
Vibrio fischeri	0,3180030	0,3117471	0,1860337	0,1842154	0,3046419	0,3057307	0,1950888	0,1945352
Photobacterium profundum	0,2894233	0,2908856	0,2092672	0,2104216	0,2936844	0,2940812	0,2071380	0,2050892
Vibrio parahaemolyticus	0,2716546	0,2744890	0,2270339	0,2268225	0,2724604	0,2740245	0,2273116	0,2262035
Vibrio vulnificus	0,2680685	0,2678181	0,2323598	0,2317537	0,2626650	0,2651937	0,2365050	0,2356364
Vibrio vulnificus	0,2666992	0,2688019	0,2313119	0,2331870	0,2640151	0,2647929	0,2357722	0,2354198
Vibrio cholerae	0,2597782	0,2632678	0,2375400	0,2394133	0,2648523	0,2660105	0,2326545	0,2364827
Brucella abortus	0,2144865	0,2139559	0,2866285	0,2849291	0,2122760	0,2143686	0,2845017	0,2888538
Brucella suis	0,2140816	0,2137780	0,2870105	0,2851299	0,2129560	0,2138803	0,2874677	0,2856961
Brucella melitensis	0,2138948	0,2145215	0,2848725	0,2867075	0,2137865	0,2128331	0,2857911	0,2875877
Agrobacterium tumefaciens	0,2045392	0,2017090	0,2997459	0,2940060	0,2034039	0,2038351	0,2971410	0,2956200
Haloarcula marismortui	0,1883828	0,1879786	0,3118934	0,3117452	0,2102795	0,2174310	0,2872314	0,2850581
Deinococcus radiodurans	0,1648171	0,1650622	0,3354105	0,3347053	0,1695502	0,1635625	0,3333018	0,3335855
Burkholderia mallei	0,1593326	0,1591249	0,3436969	0,3378456	0,1545873	0,1555570	0,3448182	0,3450375

Как уже отмечалось выше, геномы бактерий имеют сравнительно простую структуру: белок-кодирующие участки не прерываются некодирующими вставками – интронами. Эта особенность бактериальных геномов позволяет выделять и отдельно анализировать белок-кодирующие участки.

Напомним, три нуклеотида (триплет) кодируют одну из 20 аминокислот, из которых строятся белки. Полимераза – фермент, считывающий информацию с ДНК для последующего синтеза белка, работает в направлении 5’ → 3’ (рис 4.1).

Известно, что белок-кодирующие последовательности содержатся на обеих цепочках хромосомы. Как показывают наблюдения, белок-кодирующие участки в геномах бактерий распределены примерно одинаково по комплементарным нитям. Более того, белок-кодирующие

последовательности, записанные в одноименном направлении, локализованы в определенных участках хромосомы.

В геномах бактерий белок-кодирующие последовательности, записанные в противоположных направлениях, сконцентрированы у противоположных концов хромосомы. Их концентрация падает по мере движения в направлении к середине хромосомы.

На рис. 4.4 показано изменение количества белок-кодирующих последовательностей вдоль длины хромосомы бактерии *Bacillus subtilis subsp. subtilis str. 168*, состоящей из 4214630 п.н. Каждая кривая соответствует участкам, записанным по одной из цепочек хромосомы.

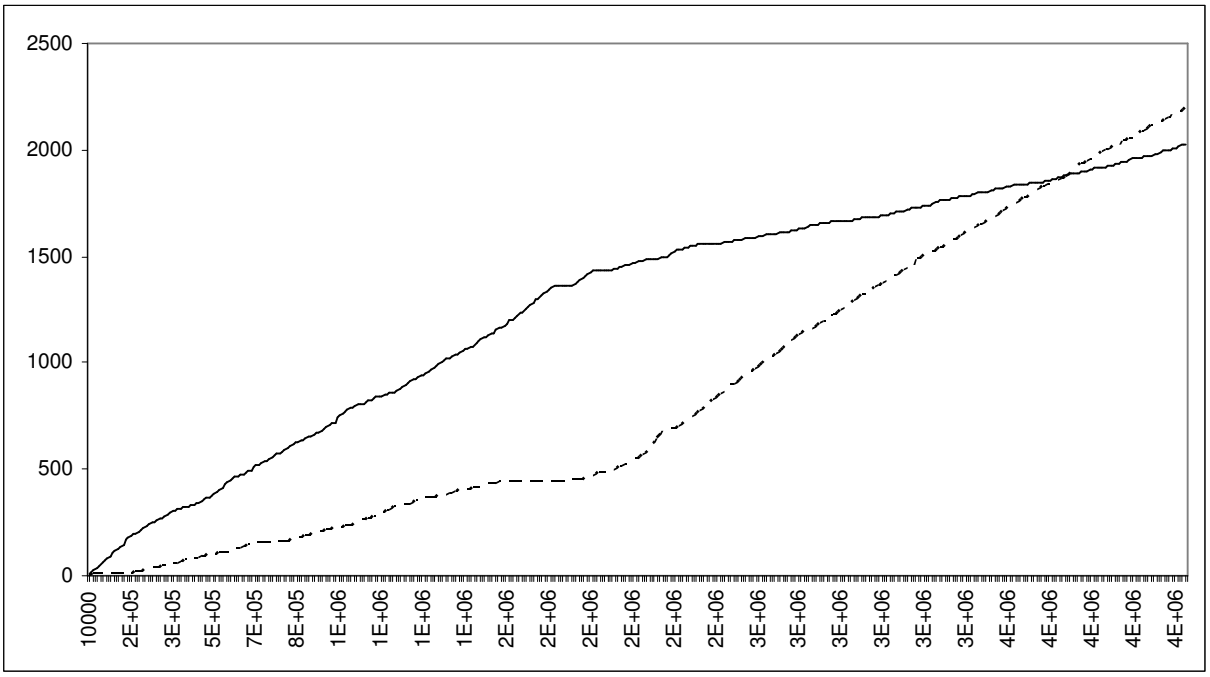

Рис. 4.4

Белки, в общем случае, состоят из различного количества аминокислот. Количества аминокислот для отдельных белков может отличаться на порядок, следовательно, длины соответствующих белок-кодирующих участков также отличаются. Тем не менее, вышеописанные особенности остаются в силе, если подсчитывать количество не белок-кодирующих участков, а нуклеотидов, из которых они состоят.

На рис. 4.5 показано изменение количества нуклеотидов, входящих в состав белок-кодирующих участков бактерии *Bacillus subtilis subsp. subtilis str. 168*. Как и в предыдущем случае, каждая из кривых отвечает определенному направлению записи.

Такие особенности напоминают распределение комплементарных оснований по длине цепочки ДНК, о чем уже сообщалось раньше [48, 6]. Подтвердилось, что количество комплементарных оснований, подсчитанных по одной цепочке ДНК, приблизительно одинаковы; иными словами, количество оснований T практически совпадает с количеством A, а количество C практически совпадает с количеством G. Подчеркиваем, что речь идет не о комплементарных цепочках хромосомы, где это свойство очевидно – в данном случае рассматривается только одна цепочка ДНК.

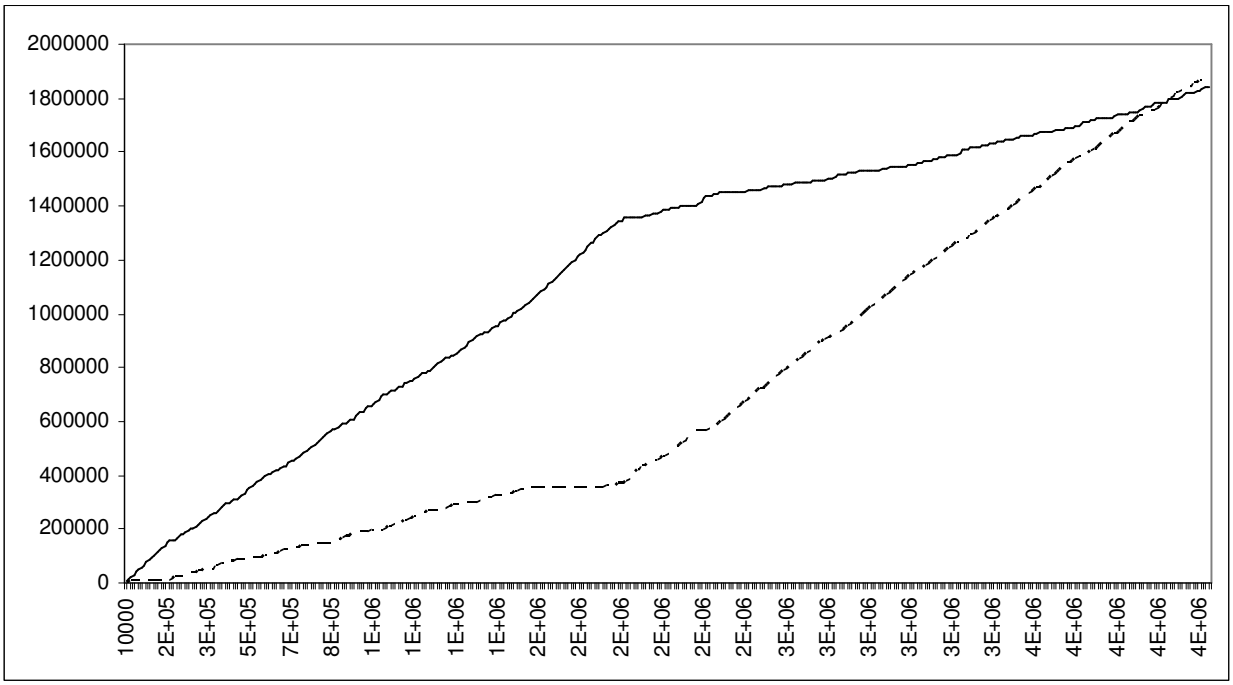

Рис. 4.5

Более того, в геномах бактерий комплементарные основания также концентрируются у противоположных концов цепочки ДНК. Например, у одного конца цепочки преобладает нуклеотид A, у другого – T, тем не менее, их количество по всей цепочке в целом совпадает. То же самое касается оснований C и G.

Для примера рассмотрим цепочку ДНК бактерии *Bacillus subtilis subsp. subtilis str. 168.* Вместо количеств A, T, C и G для удобства используются соответствующие относительные частоты: $\nu(\mathrm{A}),\ \nu(\mathrm{T}),\ \nu(\mathrm{C}),\ \nu(\mathrm{G})$.

$$\nu(x) = \frac{n(x)}{m},\ x \in \{\mathrm{A},\ \mathrm{T},\ \mathrm{C},\ \mathrm{G}\},$$

где $n(x)$ – количество нуклеотидных оснований x, а m – длина участка ДНК, на котором осуществляется подсчет. Очевидно, что

$$n(\mathrm{A}) + n(\mathrm{T}) + n(\mathrm{C}) + n(\mathrm{G}) = m,$$
$$\nu(\mathrm{A}) + \nu(\mathrm{T}) + \nu(\mathrm{C}) + \nu(\mathrm{G}) = 1.$$

На графиках (рис. 4.6, 4.7) отображается сходимость частот комплементарных нуклеотидов по длине цепочки ДНК. На первом графике отчетливо видно, что в начале цепочки основание A преобладает над T. Начиная примерно с середины цепочки количество A падает, а количество T, наоборот, возрастает.

Более наглядно этот факт проявляется, если рассматривать разности частот комплементарных оснований $\nu(\mathrm{A}) - \nu(\mathrm{T})$ и $\nu(\mathrm{C}) - \nu(\mathrm{G})$ локально, т.е. не по всей цепочке, а по сравнительно небольшим (но достаточно длинным для построения соответствующих статистик) участкам. Таким образом, избавляемся от эффекта накопления.

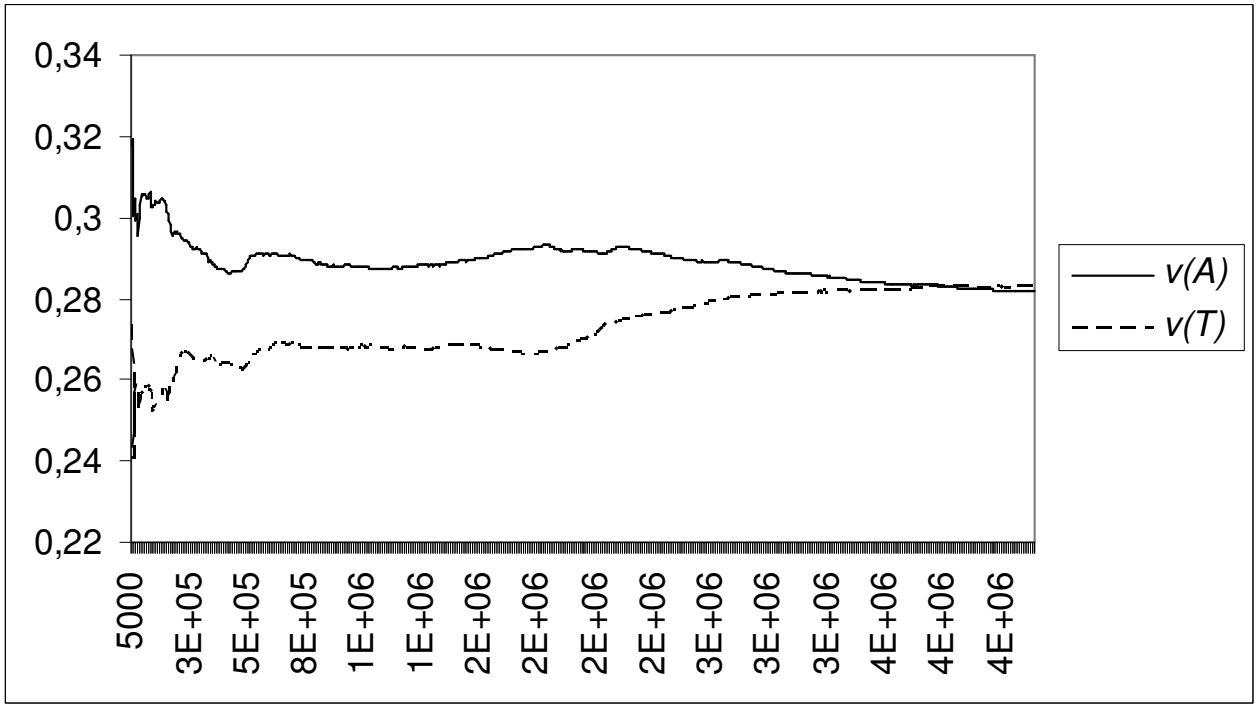

Рис 4.6

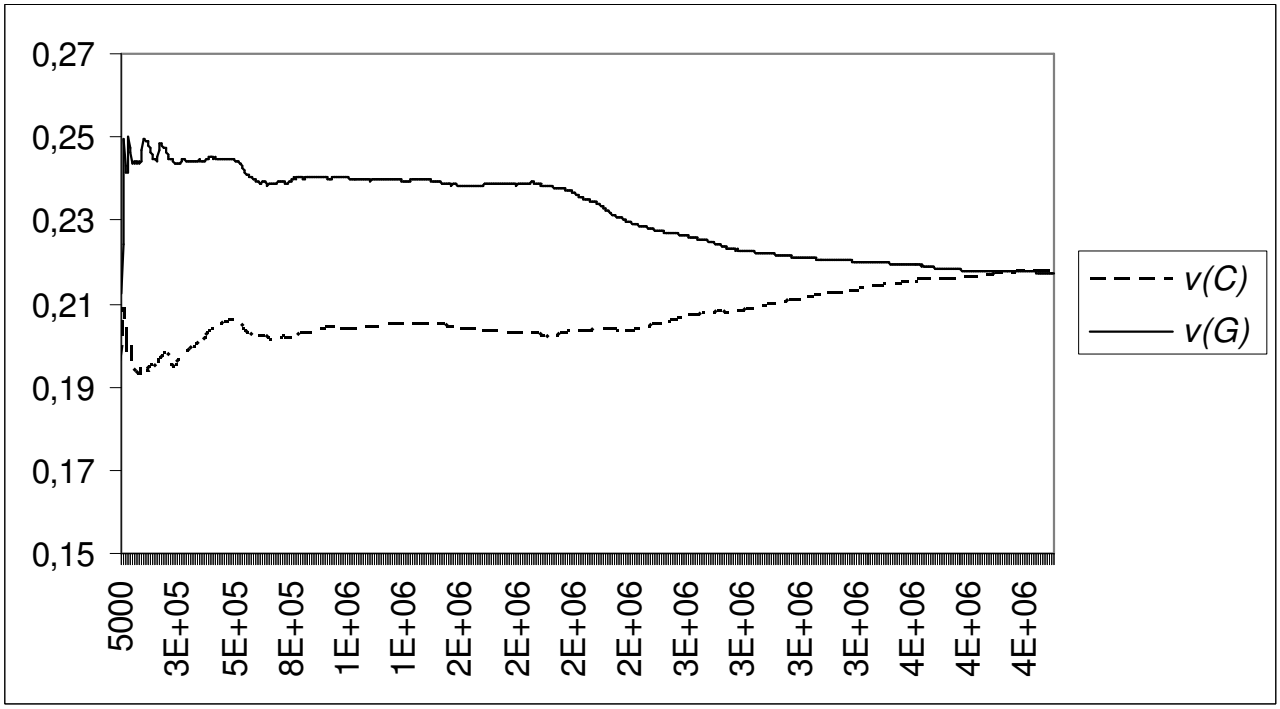

Рис. 4.7

Компьютер разбивал цепочку ДНК бактерии *Bacillus subtilis subsp. subtilis str. 168* на отрезки по 5000 нуклеотидов. На каждом из этих отрезков

подсчитывались разности относительных частот комплементарных нуклеотидов и отображались на графиках (рис. 4.8, 4.9). При подсчете по всей длине цепочки относительные частоты комплементарных оснований практически сравниваются, принимая значения

$$\nu(\mathrm{A}) = 0,2818,\ \nu(\mathrm{T}) = 0,2830;$$
$$\nu(\mathrm{C}) = 0,2181,\ \nu(\mathrm{G}) = 0,2171.$$

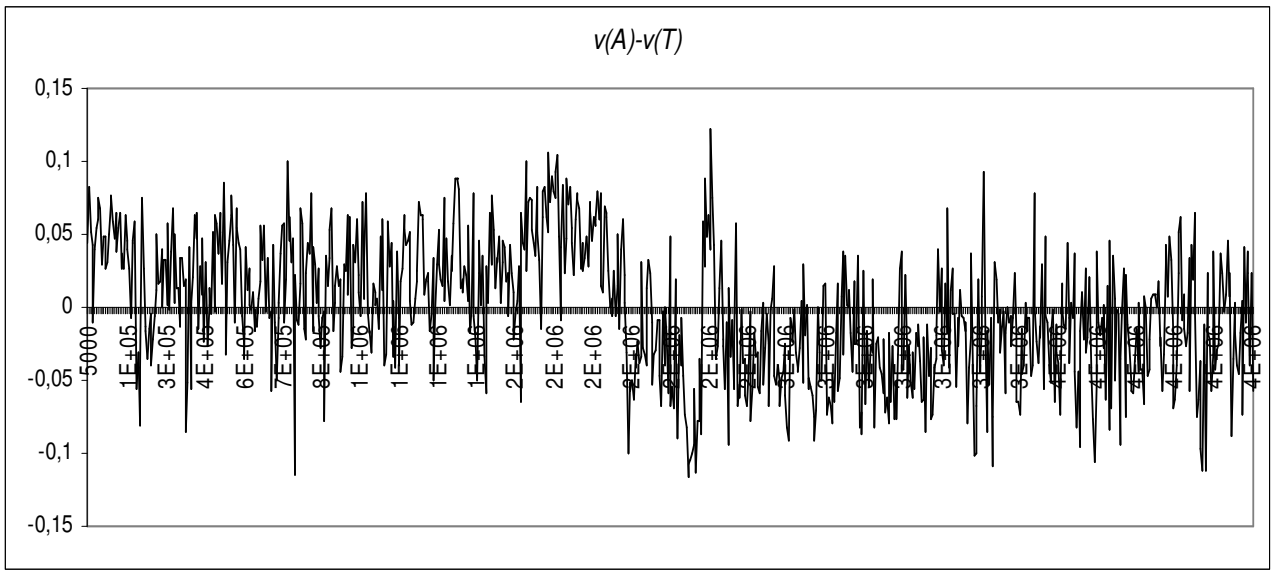

Рис. 4.8

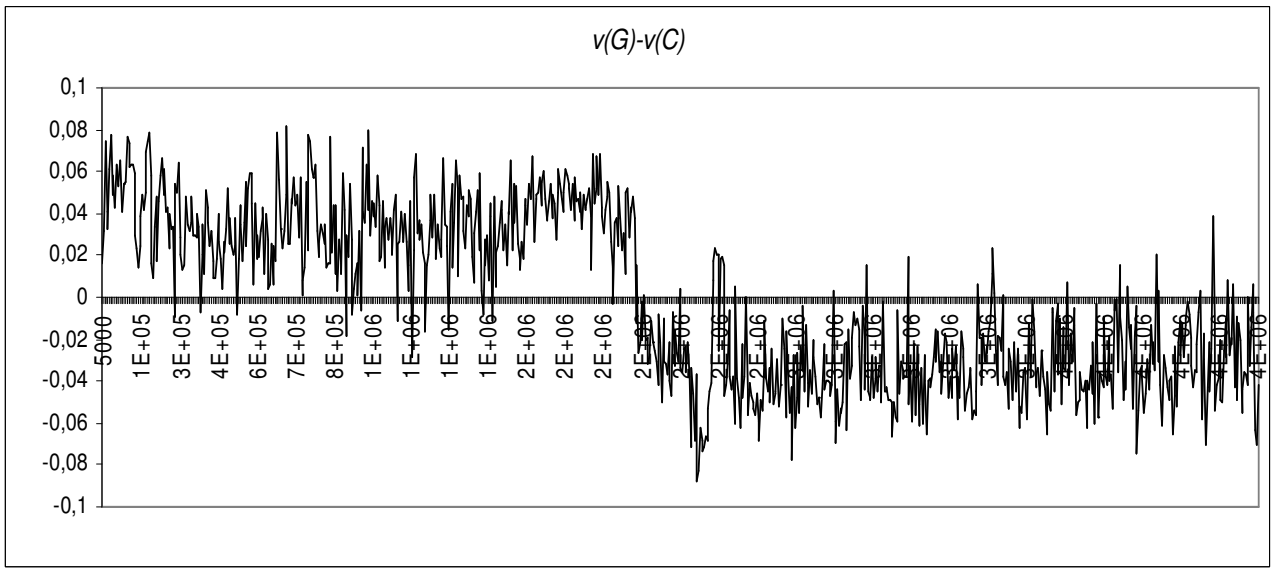

Рис. 4.9

Сравнивая распределения белок-кодирующих участков, записанных в разных направлениях (по разным цепочкам ДНК), и распределения комплементарных нуклеотидных оснований, можно предположить, что они взаимосвязаны. Для подтверждения этого по отдельности подсчитывались относительные частоты нуклеотидных оснований для белок-кодирующих участков, записанных в разных направлениях.

В таблице 4.11 приведены количества и относительные частоты оснований, подсчитанные отдельно по белок-кодирующим участкам, записанным в разных направлениях, по всем белок-кодирующим участкам и по цепочке в целом. Участкам, записанным в противоположных направлениях, отвечают симметричные отклонения частот комплементарных оснований, что приводит к их взаимной компенсации при подсчете по всем белок-кодирующим участкам или по всей цепочке ДНК.

Таблица 4.11

Bacillus subtilis subsp. subtilis str. 168

Основание	3'→5'		5'→3'	
	Количество	Частота	Количество	Частота
A	485855	0,257436	550371	0,298345
C	454257	0,240693	375345	0,203467
T	566608	0,300224	471419	0,255547
G	380566	0,201647	447611	0,242641
Основание	Белок-кодирующие участки		Вся цепочка	
	Количество	Частота	Количество	Частота
A	1035550	0,277476	1187714	0,281807
C	829885	0,222368	919127	0,21808
T	1038333	0,278222	1192801	0,283014
G	828264	0,221934	914988	0,217098

Частоты, полученные при подсчете по всем белок-кодирующим участкам, мало отличаются от полученных при подсчете по всей цепочке ДНК. Это связано с тем, что белок-кодирующие участки составляют 80-90 %

геномов бактерий. Для высших организмов, где белок-кодирующие участки иногда занимают всего несколько процентов от длины всей хромосомы, ситуация может быть иной.

Таким образом, подтвердилось, что в геномах бактерий имеется некий механизм записи информации, зависящий от ориентации цепочки, несущей информацию. Этот механизм проявляется в виде симметричных по комплементарным основаниям статистических закономерностей. Некоторые из описанных закономерностей обнаружены и в геномах гораздо более сложных организмов – растений и животных, что свидетельствует об их важности для функционирования генетического аппарата и универсальности.

4.3 Комплементарность оснований и выбор модели описания аминокислотных последовательностей белков

Как отмечалось выше, комплементарность в записи оснований по одной нити ДНК хромосомы означает, что выполняются приближенные соотношения

$$n(\mathrm{A}) \approx n(\mathrm{T}), \quad n(\mathrm{C}) \approx n(\mathrm{G}), \tag{4.1}$$

где $n(j)$ – количество оснований j, $j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, вычисленных на одной нити [45, 46].

Для пар оснований выполняются следующие соотношения комплементарности

$$n(ij) \approx n(\bar{j}\,\bar{i}), \tag{4.2}$$

где $i, j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $\bar{\mathrm{A}} = \mathrm{T}$, $\bar{\mathrm{C}} = \mathrm{G}$, $\bar{\mathrm{T}} = \mathrm{A}$, $\bar{\mathrm{G}} = \mathrm{C}$. Заметим, что пары AT, TA, CG и GC не присутствуют в (4.2), поскольку они сами себе антикомплементарны.

Запись и считывание оснований у первой нити хромосомы ДНК выполняется слева направо в направлении $5' \to 3'$, а у комплементарной второй нити в направлении $5' \to 3'$ справа налево (рис. 4.10).

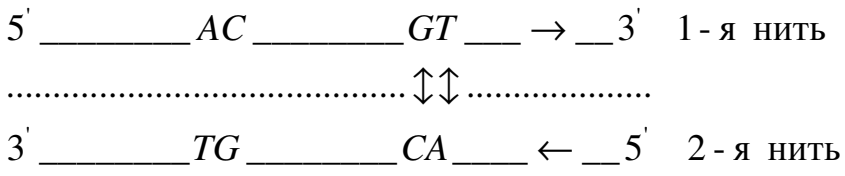

Рис. 4.10

Разделение на 1-ю и 2-ю нити условно (рис 4.10), оно здесь вводится для наглядности изложения.

Известно, что соотношения

$$\hat{p}(j \mid i) = \frac{n(i,j)}{n(i)}, \tag{4.3}$$

где $n(ij)$ – число пар (ij), $i, j \in \{\mathrm{A,C,G,T}\}$, $n(i)$ – число оснований i в цепи хромосомы, представляют собой оценки переходных вероятностей для стационарных цепей Маркова. В [4] показано, что для длинных цепей оценки (4.3) сходятся к значениям переходных вероятностей.

Из соотношений комплементарности (4.2) вытекает, что вторая комплементарная нить в направлении $5' \to 3'$ имеет такие же оценки переходных вероятностей $\hat{p}(j|i)$, что и исходная первая нить (на рис. 4.10 представлена пара AC и комплементарная ей пара GT). Отсюда следует, что вероятности двух противоположных нитей хромосомы, подсчитанные в модели однородной цепи Маркова, совпадают.

Пусть $x_1, x_2, ..., x_{n-1}, x_n$ – конечная последовательность оснований, записанных на первой нити, тогда $\bar{x}_n, \bar{x}_{n-1}, ..., \bar{x}_2, \bar{x}_1$ – комплементарная ей последовательность оснований, записанных на второй нити (рис. 4.11)

$5'$ _____ $x_1, x_2, ..., x_{n-1}, x_n$ _____ $\rightarrow$ __ $3'$ 1 - я нить

$3'$ _____ $\overline{x_1}, \overline{x_2}, ..., \overline{x_{n-1}}, \overline{x_n}$ _____ $\leftarrow$ __ $5'$ 2 - я нить

Рис. 4.11 Комплементарность нуклеотидных последовательностей

Для стационарной цепи Маркова порядка 1 выполняется следующее важное утверждение.

Теорема 4.1. Вероятность последовательности $x_1, x_2, ..., x_{n-1}, x_n$ совпадает с вероятностью последовательности $\overline{x}_n, \overline{x}_{n-1}, ..., \overline{x}_2, \overline{x}_1$, т.е.

$$P(x_1, x_2, ..., x_{n-1}, x_n) = P(\overline{x_n}, \overline{x_{n-1}}, ..., \overline{x_2}, \overline{x_1}). \quad (4.4)$$

Вероятность стационарной цепи Маркова определяется соотношением

$$P(x_1, x_2, ..., x_{n-1}, x_n) = P(x_1)P(x_2 \mid x_1)...P(x_n \mid x_{n-1}), \quad (4.5)$$

где $P(x_1)$ – вероятность начального состояния, $P(x_i \mid x_{i-1})$ – переходные вероятности, $i = 1,2,...,n$.

Заменив вероятность начального состояния частотой, а переходные вероятности $P(x_i \mid x_{i-1})$ в (4.5) их оценками (4.3), получим сразу же соотношение (4.4). Отсюда следует, что вероятности двух противоположных нитей, подсчитанные для модели стационарной цепи Маркова, совпадают.

Кодоны (тройки оснований) связаны следующими соотношениями комплементарности [46]:

$$n(i, j, k) \approx n(\overline{k}, \overline{i}, \overline{j}), \quad (4.6)$$

где $n(i,j,k)$ – число троек оснований (i,j,k), $(\bar{k},\bar{i},\bar{j})$ – антикодон кодона (i,j,k). Для 64 триплетов получаем 32 соотношения (4.6) типа кодон – антикодон.

Оценки переходных вероятностей для цепей Маркова порядка 2 определяются соотношениями

$$\hat{p}(k \mid i,j) = \frac{n(i,j,k)}{n(i,j)}, \tag{4.7}$$

где $n(i,j,k)$ – количество троек оснований (i,j,k), $n(i,j)$ – количество пар (i,j), $i,j,k \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$.

Из соотношений комплементарности (4.6) заключаем, что оценки переходных вероятностей (4.7) для обеих нитей, подсчитанные в направлении $5' \to 3'$, совпадают. Легко показать, что результат теоремы справедлив и для цепей Маркова высших порядков.

Белок-кодирующий участок записанный на первой нити, теряет свой смысл, если его считывать со второй нити. Наблюдения показывают, что белок-кодирующие участки, записанные на разных нитях, концентрируются у противоположных концов хромосомы и редко пересекаются (исключением в этом смысле являются вирусы, гены которых могут пересекаться в связи с маленькими размерами геномов).

Рассмотрим первую нить хромосомы (рис 4.10, 4.11), обозначим соответствующую ей матрицу переходных вероятностей – $\vec{P}$. Матрицу переходных вероятностей соответствующую второй (комплементарной) нити обозначим через $\overleftarrow{P}$. В силу комплементарности по одной нити матрицы $\vec{P}$ и $\overleftarrow{P}$ совпадают, т.е. $\vec{P} = \overleftarrow{P}$. Отметим, что комплементарные цепочки ДНК – это разные по составу последовательности, кодирующие разные белки. Такое совпадение наталкивает на мысль, что существует единая схема записи информации, которая описывается цепями Маркова.

Относительно простая схема записи белков в геномах бактерий позволяет достаточно просто выделить белок-кодирующие участки на каждой из нитей бактериальных хромосом. Обозначим через $\vec{P}^*$ матрицу переходных вероятностей, соответствующую всем белок-кодирующим участкам первой нити сложенным вместе. Заметим, что $\vec{P}^*$ отличается $\vec{P}$. Для белок-кодирующих участков второй нити аналогично получаем матрицу переходных вероятностей $\bar{\vec{P}}^*$ ($\bar{\vec{P}}^* \neq \vec{P}$). Оказывается, что матрицы переходных вероятностей белок-кодирующих участков комплементарных цепочек ДНК также совпадают, иными словами $\vec{P}^* = \bar{\vec{P}}^*$ (таблица 4.11). Таким образом, схема записи белок-кодирующих участков одинакова для комплементарных нитей.

Таким образом, можно сделать заключение, что цепь Маркова – удобная модель описания нуклеотидных последовательностей ДНК. Далее этот вывод переносится на аминокислотные последовательности белков.

Покажем, что если белок-кодирующие участки описываются цепями Маркова, то и аминокислотные последовательности соответствующих белков также описываются цепями Маркова. Напомним, что аминокислотная последовательность белка получается путем трансляции четырехбуквенного алфавита оснований (обозначим N) в 20 буквенный алфавит аминокислотных остатков (обозначим A) (таблица 4.1). Генетический код образует функцию, которая переводит непересекающиеся тройки оснований в одну из 20 аминокислот:

$$f\ (x_i, x_j, x_k) = y,\ \ x_i, x_j, x_k \in N,\ \ y \in A.$$

Покажем, что в результате действия генетического кода на марковскую цепь с матрицей переходных вероятностей $P^N = p^N_{i,j},\ i, j = 1,..,4$, получается марковская цепь с матрицей переходных вероятностей $P^A = p^A_{i,j},\ i, j = 1,..,20$.

Пускай произвольной переходной вероятности $P(y_{i+1} \mid y_i)$, $y \in A$ соответствует условная вероятность $P(x_{i+3}, x_{i+4}, x_{i+5} \mid x_i, x_{i+1}, x_{i+2})$, $x \in N$. Это означает, что аминокислота y_i кодируется триплетом нуклеотидов x_i, x_{i+1}, x_{i+2}, соответственно аминокислота y_{i+1} кодируется триплетом $x_{i+3}, x_{i+4}, x_{i+5}$. Следовательно шестерка нуклеотидов $x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}, x_{i+5}$ кодирует пару аминокислот y_i, y_{i+1}.

Тогда

$$P(y_{i+1} \mid y_i) = P(x_{i+3}, x_{i+4}, x_{i+5} \mid x_i, x_{i+1}, x_{i+2}) = \frac{P(x_i, x_{i+1}, x_{i+2}, x_{i+3}, x_{i+4}, x_{i+5})}{P(x_i, x_{i+1}, x_{i+2})} =$$

$$= \frac{P(x_i)P(x_{i+1} \mid x_i)P(x_{i+2} \mid x_{i+1})P(x_{i+3} \mid x_{i+2})P(x_{i+4} \mid x_{i+3})P(x_{i+5} \mid x_{i+4})}{P(x_i)P(x_{i+1} \mid x_i)P(x_{i+2} \mid x_{i+1})} =$$

$$= P(x_{i+3} \mid x_{i+2})P(x_{i+4} \mid x_{i+3})P(x_{i+5} \mid x_{i+4})$$

Условные вероятности $P(y_{i+1} \mid y_i)$ и $P(x_{i+1} \mid x_i)$ – это элементы матрицы переходных вероятностей P^A и P^N соответственно. Матрицы P^A и P^N могут зависеть от времени в случае нестационарных цепей.

Покажем, что сумма всех переходных вероятностей в строке матрицы P^A равна единице.

$$\sum_{y_i \in A} P(y_{i+1} \mid y_i) = \sum_{x_{i+3} \in N} \sum_{x_{i+4} \in N} \sum_{x_{i+5} \in N} P(x_{i+3} \mid x_{i+2})P(x_{i+4} \mid x_{i+3})P(x_{i+5} \mid x_{i+4}) =$$

$$= \sum_{x_{i+3} \in N} \sum_{x_{i+4} \in N} P(x_{i+3} \mid x_{i+2})P(x_{i+4} \mid x_{i+3}) = \sum_{x_{i+3} \in N} P(x_{i+3} \mid x_{i+2}) = 1 \qquad (4.8)$$

Суммирование в (4.8) проводится по всему пространству элементарных событий, т.е. по всему множеству N. В терминах цепей Маркова это эквивалентно суммированию по всем элементам строки матрицы переходных вероятностей P^N.

В том случае, когда одна аминокислота кодируется несколькими триплетами, соответствующие элементы матрицы переходных вероятностей суммируются. При этом сумма всех элементов в строке матрицы P^A остается равной единице, а размерность матрицы уменьшается.

Аналогично проводится доказательство для цепей высших порядков. Генетический код сохраняет свойство эргодичности цепи.

Выводы. В данном разделе обосновывается выбор цепей Маркова в качестве модели описания аминокислотных последовательностей белков. Для этого проводится статистический анализ геномов живых организмов, и на основе наблюдений, делается вывод, что нуклеотидные последовательности ДНК эффективно описываются моделями цепей Маркова. Кроме того приводятся методы определения параметров модели, таких как порядок цепи и стационарность. Эти методы сводятся к решению серии задач распознавания гипотез.

Принимая во внимание свойство комплементарности по одной нити ДНК показывается, что существует единый в рамках конкретного генома процесс записи генетической информации. Этот процесс описывается цепями Маркова.

Доказывается, что генетический код переводит цепь Маркова с четырехбуквенным пространством состояний в цепь Маркова с двадцатибуквенным пространством состояний, откуда следует, что цепи Маркова могут использоваться для описания аминокислотных последовательностей белков.

РАЗДЕЛ 5
РАСПОЗНАВАНИЕ ВТОРИЧНОЙ СТРУКТЫРЫ БЕЛКОВ

В данном разделе полученные ранее результаты применяются для решения задачи предсказания вторичной структуры белка. Предлагаемый нами метод относится к методам машинного обучения и, следовательно, работает на обучающих выборках. При решении задачи важно использовать эффективные методы, т.е. имеющие гарантированные полиномиальные оценки сложности от входа задачи (т.е. от размеров обучающей выборки и количества признаков).

Важным моментом при решении задачи является выбор модели описания аминокислотной последовательности белка. В данном случае используются модели цепей Маркова (обоснование использования цепей Маркова проводится в четвертом разделе диссертации).

5.1 Структура белка

На сегодняшний день общими усилиями ученых всего мира расшифровано геномы человека, шимпанзе, курицы, рыбы, некоторых растений и более трехсот бактерий. Основной вопрос современной молекулярной биологии: какую функцию выполняет определенный ген? Ген – это часть молекулы ДНК, которая кодирует соответствующий ему белок. Зная нуклеотидную последовательность гена, можно однозначно определить аминокислотную последовательность белка, так как каждая из 20 аминокислот кодируется определенным триплетом нуклеотидов (кодоном). После трансляции последовательности аминокислот из молекулы РНК белок сразу начинает сворачиваться в пространственную конфигурацию. Именно пространственная конфигурация белка определяет его функциональность, поскольку белки в живых организмах взаимодействуют как трехмерные объекты в пространстве. Поэтому в исследованиях белков и их функций

придерживаются доктрины «последовательность-структура-функциональность» [49]. Это означает, что функциональность белка однозначно определяется его пространственной структурой, а пространственная конфигурация однозначно задается его аминокислотной последовательностью. В 1993 году было показано, что математическая формулировка проблемы формирования структуры белка столь же трудна, что и известная NP –полная задача коммивояжера [50].

Существует четыре уровня организации структуры белка:

- первичная структура – линейная последовательность аминокислотных остатков в молекуле белка;
- вторичная структура – формирование на линейной последовательности локальных регулярных структур: α-спиралей и β-слоев;
- третичная структура – расположение элементов вторичной структуры (α-спиралей и β-слоев) в пространстве друг относительно друга;
- четвертичная структура – формирование белкового комплекса из отдельных белков.

Структура белка на каждом из уровней организации оказывает решающее влияние на формирование структуры на следующем уровне, т.е. первичная структура определяет вторичную, вторичная – третичную и т.д.

Первичная структура белка, т.е. его аминокислотная последовательность, определяется эксперементальным путем относительно просто. Определение вторичной структуры уже связано с большими трудностями, поскольку требует применения дорогих методов рентгено-структурного анализа и магнитно-ядерного резонанса.

Сложность экспериментального определения вторичной структуры белка способствует развитию математических методов ее предсказания. Задача ставится следующим образом: имеется первичная структура белка (т.е. линейная последовательность аминокислот), необходимо определить его

вторичную структуру, иными словами, поставить в соответствие каждой аминокислоте один из двух возможных типов регулярной структуры (α-спираль, β-слой), или ее отсутствие, т.е. нерегулярность (coil).

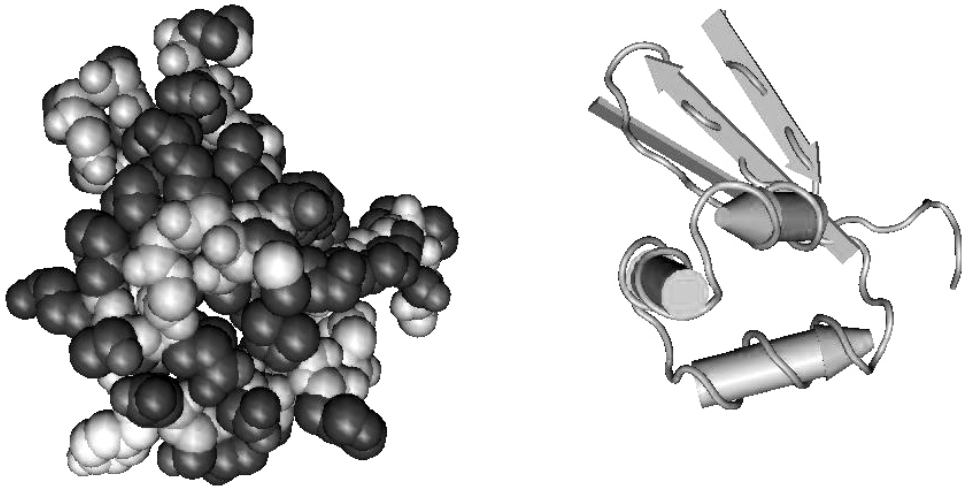

Рис. 5.1 Третичная и вторичная структура белка репрессора CRO.

На рис. 5.1 α-спирали обозначаются цилиндрами, а β-слои – стрелками. Вторичная структура белка во многом определяет его функции. Например, комбинация из трех α-спиралей часто используется в природе для узнавания определенной последовательности ДНК.

Далее будем использовать следующие обозначения: “h”(helix) для обозначения α-спиралей, “s”(sheet) – для β-слоев и “-” – для обозначения нерегулярности, т.е. отсутствие какой-либо структуры (coil). Таким образом, первичная и соответствующая ей вторичная структура белка CRO (рис. 5.1) представляется в виде

```
meqritlkdyamrfgqtktakdlgvyqsainkaihagrkifltinadgsvyaeevkpfpsn
-sssssshhhhhhh-hhhhhhhhh--hhhhhhhhhh--ssssssss-ssssssss------
```

5.2 Обзор методов предсказания вторичной структуры белка

Первые попытки предсказания вторичной структуры основывалисть на физике образования внутренних химических связей в процессе сворачивания белка. α-спирали и β-слои образуются в результате возникновения водородных связей между боковыми группами аминокислотных остатков, и таким образом влияют на свободную энергию белковой молекулы. Белок, как и любая физическая система, стремится занять наиболее выгодное энергетическое состояние, т.е. состояние с наименьшей энергией. Таким образом, белок формирует вторичную структуру, отвечающую найнизшему энергетическому состоянию.

Возникшая в результате таких рассуждений задача минимизации внутренней энергии белковой молекулы не решается классическими методами, поскольку имеет неполиномиальную сложность. Действительно, для среднего белка длиной 200 аминокислот существует 3^{200} возможных конформаций, из которых необходимо выбрать отвечающую состоянию с наименьшей энергией. Безусловно, некоторые из конформаций изначально не будут иметь физического смысла, но на порядок величины перебора это существенным образом не повлияет. Ситуация осложняется низкой точностью методов оценки свободной энергии белка.

До конца неясным остается и сам механизм сворачивания белка. Эта проблема еще называется «парадоксом Левинталя» и в упрощенном виде заключается в следующем. Средний белок (длиной 200 аминокислот) должен искать «свою» вторичную структуру из порядка 3^{200} возможных, каждой из которых соответствует определенный уровень свободной энергии. При этом белок может «почувствовать» стабильность структуры, попав прямо в нее. А так как переход от одной возможной конформации к другой происходит не менее чем за 10^{-13} секунды, перебор всех структур должен был бы занять порядка 10^{80} лет, на фоне которых время существования Вселенной – 10^{10} лет – величина бесконечно малая [51].

Другой подход к задаче предсказания вторичной структуры основывается на применении методов машинного обучения (machine learning approach) [1]. Постоянно растущие базы данных экспериментально установленных вторичных структур используются такими методами в качестве обучающих выборок. Именно в этом направлении получены значительные результаты в предсказания вторичной структуры белка. Тем не менее, начиная от момента своего рождения до настоящего времени, методы машинного обучения прошли значительную эволюцию [52]:

- Методы 1-го поколения основывались на статистиках единственной аминокислоты, оценивая вероятность вхождения аминокислоты в определенную вторичную структуру.
- Методы 2-го поколения стали использовать предположение, что на вхождение аминокислоты в определенную вторичную структуру влияет окружение из соседних аминокислот. При этом использовались самые разные модели и алгоритмы, среди которых следует выделить цепи маркова со скрытыми параметрами, генетические и энтропийные алгоритмы, нейронные сети. Точность таких методов достигала 60%.
- Методы 3-го поколения считаются самыми производительными на сегодняшний день, достигая точности 77%. Такие методы используют, наряду с методами машинного обучения, информацию о структуре эволюционно близких белков. Наблюдения показывают, что два природных белка имеющие 35 одинаковых аминокислот из 100 имеют и схожую структуру. Поиск гомологичных белков производится с помощью методов выравнивания последовательностей типа BLAST, которые позволяют установить степень схожести двух последовательностей. Ввиду ресурсоемкости алгоритмов выравнивания методы 3-го поколения реализуют на высокопроизводительных серверах поддерживающих параллельные вычисления.

Основной недостаток всех описанных методов – отсутствие обоснования. Методы используются скорее как ноу-хау: эффективность никак не исследуется. Методы 3-го поколения, дающие наивысшую точность предсказания, остаются слишком трудоемкими для использования на персональных компьютерах.

Предлагаемый нами метод относится скорее к 2-му поколению методов в том смысле, что он базируется только на предположении, что вхождение аминокислоты в конкретную вторичную структуру определяется ее окружением. В отличие от методов 3-го поколения он не использует трудоемких процедур выравнивания последовательностей, тем не менее, точность предсказания этого метода выше, чем у методов 3-го поколения. Более того, предлагаемый метод строго обоснован, имеет полиномиальные оценки погрешности в зависимости от размеров обучающей выборки а также является существенно неулучшаемым.

5.3 Предсказание вторичной структуры белков

На вход задачи поступает аминокислотная последовательность белка. Необходимо по поступившей на вход аминокислотной последовательности и имеющимся последовательностям с уже известной вторичной структурой (т.е. с помощью обучающей выборки) определить вторичную структуру соответствующую исходной последовательности. Для предсказания вторичной структуры белка будем последовательно определять вторичные структуры всех входящих в его состав аминокислот, используя предположение, что на вхождение аминокислоты x_s в определенную вторичную структуру влияет окружение из соседних аминокислот $x_1, x_2, ..., x_{s-1,} x_s, x_{s+1}, ..., x_n$.

Для решения этой задачи будем использовать байесовскую индуктивную процедуру обобщенную на дискретный случай. Пусть $x_1, x_2, ..., x_n$ — описание объекта, а вторичная структура аминокислоты

x_s (будем обозначать f_s) — искомый целевой признак. Следовательно по имеющейся обучающей выборке необходимо оценить вероятности вида

$$P(f_s \mid x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n \mid f_s) P(f_s)}{P(x_1, x_2, ..., x_n)}. \qquad (5.1)$$

Целевому признаку f_s назначаем такое значение вторичной структуры, при котором вероятность (5.1) достигает своего максимума.

Во втором разделе показано, что байесовская процедура распознавания имеет гарантированную полиномиальную оценку погрешности от размеров обучающих выборок и является субоптимальной для независимых признаков $x_1, x_2, ..., x_n$.

Для независимых случайных величин (цепь Маркова порядка 0) имеет место равенство

$$P(x_1, x_2, ..., x_n \mid f_s) = \prod_{i=1}^{n} P(x_i \mid f_s). \qquad (5.2)$$

Вероятности в (5.1), (5.2) заменяются частотами, подсчитанными на основе обучающих выборок.

Основная особенность анализа аминокислотных последовательностей белков состоит в том, что частота встречаемости соседних аминокислотных остатков в белках не является независимой. Поэтому в четвертом разделе рассматривался вопрос соответствия этих последовательностей модели Марковской цепи. Порядок цепи устанавливается методами правдоподобия (на основе подсчетов вероятностей аминокислотных цепочек) путем решения серии задач распознавания гипотез.

Для модели марковской цепи первого порядка вероятность цепочки $x_1, x_2, ..., x_n$ задается соотношением

$$P(x_1, x_2, ..., x_n | f_s) = P(x_1 | f_s) P(x_2 | x_1, f_s) ... P(x_n | x_{n-1}, f_s), \qquad (5.3)$$

где $P(x_k | x_{k-1}, f_s)$, $k = 2, ..., n$ – нестационарные переходные вероятности. В численных расчетах используются оценки переходных вероятностей, построенные в виде частот

$$\hat{p}(x_k = j | x_{k-1} = i, f_s) = \frac{m(x_{k-1} = i, x_k = j, f_s)}{m(x_{k-1} = i, f_s)}, \qquad (5.4)$$

где $m(x_{k-1} = i,\ x_k = j,\ f_s)$ – число последовательностей (окружений) $x_1, x_2, ..., x_n$, для которых на $(k-1)$- м месте находится аминокислота i, а на k-м месте – аминокислота j при заданном состоянии f_s; $m(x_{k-1} = i, f_s)$ – число последовательностей, для которых на $(k-1)$- м месте находится аминокислота i при условии f_s.

В третьем разделе исследовались свойства оценок переходных вероятностей (5.4). Величины $\hat{p}(x_k | x_{k-1}, f_s) - p(x_k | x_{k-1}, f_s)$ имеют асимптотическое нормальное распределение со средним 0 и дисперсией порядка $\frac{1}{m}$, где m – объем обучающей выборки, при этом оценки переходных вероятностей (5.4) асимптотически независимы.

В работе [10] исследовалась процедура предсказания вторичной структуры одиночной аминокислоты на основе известной формулы Байеса:

$$P(f_s | x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n | f_s) P(f_s)}{P(x_1, x_2, ..., x_n)}. \qquad (8)$$

Здесь f_s — состояние аминокислоты x_s, число классов f_s — 60, так как 20 – количество аминокислот, 3 – число вторичных структур. Тип

вторичной структуры определялся окружением $x_1, x_2, ..., x_n$ из соседних аминокислот, расположенных слева и справа от исследуемой аминокислоты x_s (рис. 5.2).

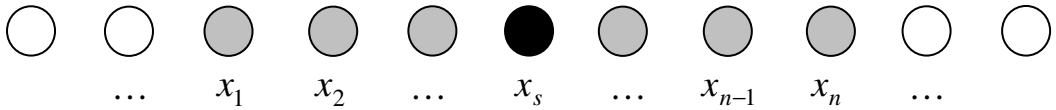

Рис. 5.2

В [10] представлены результаты численных расчетов предсказания вторичной структуры белка на основе байесовской процедуры распознавания на нестационарных цепях Маркова до 4-го порядка включительно.

Поскольку аминокислотные основания входящие в состав белка являются зависимыми, имеет смысл прогнозировать вторичную структуру сразу нескольких соседних (наиболее зависимых) оснований.

Распознавание пар состояний для двух текущих аминокислот.
В [11] распознавались состояния пары текущих аминокислот:

$$P(f_{s,s+1} \mid x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n \mid f_{s,s+1}) P(f_{s,s+1})}{P(x_1, x_2, ..., x_n)}. \qquad (9)$$

Здесь $f_{s,s+1}$ – состояние пары аминокислот x_s, x_{s+1}, число различных классов $f_{s,s+1}$ составляет 3600, так как 400 – количество различных пар аминокислот, а 9 – число вторичных структур пары аминокислот.

Тип вторичных структур определяется окружением (рис. 5.3)

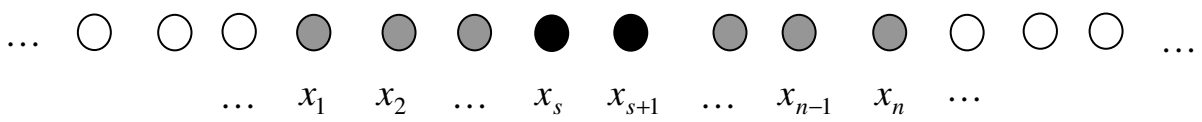

Рис. 5.3

Для определения вторичной структуры белка мы определяем вторичную структуру для всех пар соседних аминокислотных остатков, входящих в состав белка. При этом, в общем случае, каждая аминокислота входит в две пары. Действительно, обозначим α-спираль через α, β-слой – β и coil – c. Пусть имеется тройка аминокислот IKL, i,k,l – их состояния, $i,k,l \in \{\alpha,\beta,c\}$. Определим состояние текущей аминокислоты K на основе оценок вероятностей пар состояний для пар аминокислот IK и KL в последовательности соседних аминокислот IKL. Обозначим $\hat{P}_{\mathrm{IK}}(ik)$ оценки вероятностей пар состояний ik аминокислот IK, вычисленных по формулам (9).

Полагаем $\hat{P}_{\mathrm{IK}}(k) = \hat{P}_{\mathrm{IK}}(\alpha k) + \hat{P}_{\mathrm{IK}}(\beta k) + \hat{P}_{\mathrm{IK}}(ck)$. Аналогично для пары аминокислот KL имеем $\hat{P}_{\mathrm{KL}}(k) = \hat{P}_{\mathrm{KL}}(k\alpha) + \hat{P}_{\mathrm{KL}}(k\beta) + \hat{P}_{\mathrm{KL}}(kc)$.

Вычисляем

$$\hat{P}(\alpha) = \hat{P}_{\mathrm{IK}}(\alpha) + \hat{P}_{\mathrm{KL}}(\alpha) - \hat{P}_{\mathrm{IK}}(\overline{\alpha}) - \hat{P}_{\mathrm{KL}}(\overline{\alpha}),$$

$$\hat{P}(\beta) = \hat{P}_{\mathrm{IK}}(\beta) + \hat{P}_{\mathrm{KL}}(\beta) - \hat{P}_{\mathrm{IK}}(\overline{\beta}) - \hat{P}_{\mathrm{KL}}(\overline{\beta}),$$

$$\hat{P}(c) = \hat{P}_{\mathrm{IK}}(c) + \hat{P}_{\mathrm{KL}}(c) - \hat{P}_{\mathrm{IK}}(\overline{c}) - \hat{P}_{\mathrm{KL}}(\overline{c}),$$

где $\hat{P}_{\mathrm{IK}}(\overline{\alpha}) = \hat{P}_{\mathrm{IK}}(\alpha\beta) + \hat{P}_{\mathrm{IK}}(\beta\beta) + \hat{P}_{\mathrm{IK}}(c\beta) + \hat{P}_{\mathrm{IK}}(\alpha c) + \hat{P}_{\mathrm{IK}}(\beta c) + \hat{P}_{\mathrm{IK}}(cc)$

Аналогично определяются $\hat{P}_{\mathrm{KL}}(\overline{\alpha})$, $\hat{P}_{\mathrm{IK}}(\overline{\beta})$, $\hat{P}_{\mathrm{KL}}(\overline{\beta})$, $\hat{P}_{\mathrm{IK}}(\overline{c})$, $\hat{P}_{\mathrm{KL}}(\overline{c})$.

Состояние для текущей аминокислоты K определяем

$$k = \arg\max\left\{\hat{P}(\alpha), \hat{P}(\beta), \hat{P}(c)\right\}.$$

Результаты численных расчетов приведены в [11].

Распознавание троек состояний для трех текущих аминокислот. Рассматриваемый случай наиболее интересен, и, оказалось, дает наилучшие результаты. Процедуры распознавания строятся на основе формулы

$$P(f_{s-1,s,s+1} \mid x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n \mid f_{s-1,s,s+1}) P(f_{s-1,s,s+1})}{P(x_1, x_2, ..., x_n)} . \qquad (10)$$

Здесь $f_{s-1,s,s+1}$ – состояние тройки аминокислот x_{s-1}, x_s, x_{s+1}, число различных классов $f_{s-1,s,s+1}$ равно $216 \cdot 10^4$. Тип вторичной структуры определяется окружением $x_1, x_2, ..., x_n$ из соседних аминокислот, расположенных слева и справа от исследуемой тройки аминокислот x_{s-1}, x_s, x_{s+1} (рис. 5.4).

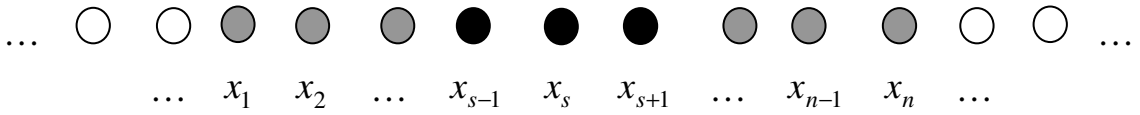

Рис. 5.4

Рассматриваемый случай сводится к предыдущему. Пусть имеется пятерка аминокислот NIKLM. Необходимо определить состояние аминокислоты K. Обозначим $\hat{P}_{NIK}(nik)$, $\hat{P}_{IKL}(ikl)$, $\hat{P}_{KLM}(klm)$ – оценки вероятностей троек состояний, вычисляемых по формулам (10) соответственно для троек аминокислот NIK, IKL и KLM. На основе этих оценок строим оценки вероятностей пар состояний для пар аминокислот IK и KL, что и будет соответствовать уже рассмотренному предыдущему случаю.

Оценки вероятностей пар состояний для пары аминокислот IK строим на основе оценок $\hat{P}_{NIK}(nik)$ и $\hat{P}_{IKL}(ikl)$. Обозначим

$$\hat{P}_{NIK}(ik) = \hat{P}_{NIK}(\alpha ik) + \hat{P}_{NIK}(\beta ik) + \hat{P}_{NIK}(cik),$$

$$\hat{P}_{IKL}(ik) = \hat{P}_{IKL}(ik\alpha) + \hat{P}_{IKL}(ik\beta) + \hat{P}_{IKL}(ikc).$$

Определяем оценки вероятностей пар состояний для пары аминокислот IK следующим образом

$$\hat{P}_{IK}(ik) = \hat{P}_{NIK}(ik) + \hat{P}_{IKL}(ik). \qquad (11)$$

Аналогичным образом оценки вероятностей пар состояний для пары аминокислот KL строятся на основе оценок $\hat{P}_{IKL}(ikl)$ и $\hat{P}_{KLM}(klm)$.

Обозначим

$$\hat{P}_{IKL}(kl) = \hat{P}_{IKL}(\alpha kl) + \hat{P}_{IKL}(\beta kl) + \hat{P}_{IKL}(ckl),$$

$$\hat{P}_{KLM}(kl) = \hat{P}_{KLM}(kl\alpha) + \hat{P}_{KLM}(kl\beta) + \hat{P}_{KLM}(klc).$$

Аналогично (11) определяем оценки вероятностей пар состояний для пары аминокислот KL

$$\hat{P}_{KL}(kl) = \hat{P}_{IKL}(kl) + \hat{P}_{KLM}(kl). \qquad (12)$$

На основе полученных оценок (11), (12) определяем состояние аминокислоты K, как это было описано в предыдущем случае для пар аминокислот.

5.4 Обучающая выборка и оценки точности

В качестве обучающей выборки используются постоянно растущие базы данных экспериментально установленных вторичных структур белков. Мы использовали базу данных NCBI [5], в которой по состоянию на апрель 2006 года насчитывалось около 80000 последовательностей.

Одним из основных недостатком в организации работы подобных баз данных является большое количество повторяющихся последовательностей. Это связано прежде всего с тем, что определение вторичной структуры белка – процесс достаточно трудоемкий и требующий времени. Часто несколько групп ученых одновременно работают над структурой одного и того же белка, и впоследствии заносят свои результаты в базу данных под разными именами.

После обработки обучающей выборки ее размер сократился до 23 тыс. последовательностей. В процессе обработки удалялись повторяющиеся последовательности и последовательности, полностью входящие в другие последовательности. Полученная в результате обучающая выборка и использовалась для обучения.

Точность распознавания вторичной структуры белка, в самом простом случае, определяется как отношение количества правильно предсказанных состояний аминокислотных оснований к длине белка (обозначается C_3). Для более точной оценки точности метода на определенном типе вторичной структуры (например α-спирали) используется коэффициент:

$$C_\alpha = \frac{p_\alpha n_\alpha - u_\alpha o_\alpha}{\sqrt{[n_\alpha + u_\alpha][n_\alpha + o_\alpha][p_\alpha + u_\alpha][p_\alpha + o_\alpha]}}$$

где:

p_α - количество верно определенный оснований, попадающих в α-спираль,

n_α - количество верно определенных оснований, не попадающих в α-спираль,

u_α - количество неверно определенных оснований ,попадающих в α-спираль,

o_α - количество неверно определенных оснований, не попадающих в α-спираль.

Обычно точность метода определяется на некотором множестве тестовых белков. В идеальном случае такое множество должно быль репрезентативным, и не должно пересекаться с обучающей выборкой. Удовлетворить этим условиям для белков практически невозможно, поскольку, белки имеют доменную структуру, т.е. сложные белки строятся из более простых белков. Мы использовали два подхода к определению точности метода.

Первый подход заключается в определении средней точности метода на всех белках из обучающей выборки. Для этого определяется вторичная структура каждого белка из обучающей выборки. При этом белок, чья вторичная структура определяется не принимает участия в обучении. Таким образом предсказывается вторичная структура каждого из 23 тыс. белков из обучающей выборки.

Второй подход заключается в определении средней точности метода на множестве белков, мало схожих с имеющимися в обучающей выборке. Такие множества получаются с помощью методов выравнивания типа BLAST. Одно из таких множеств, состоящее из 3 тысяч белков, доступно на сервере EVA [53].

Точность предсказания вторичной структуры описанного выше метода на 3 тыс. белков оказалась 79%, а на множестве из 23 тыс. белков – превышает 85%.

5.5 Примеры применения метода

Для примера приведем несколько результатов предсказания вторичной структуры белков. В качестве примеров будут использоваться классические белки.

Распознавание вторичной структуры белка Cro. Белок Cro (рис. 5.1) играет важную роль в жизненном цикле бактериофага λ – вируса паразитирующего на бактериях E. coli (кишечная палочка) [54]. После заражения бактериальной клетки у бактериофага существует два пути развития: литический и лизогенный. В первом случае гены вирусной ДНК продуцируют множество новых вирусных частиц, которое выходят наружу, разрушая бактериальную клетку. Во втором случае ДНК вируса встраивается в ДНК бактерии и жизненный цикл бактериофага переходит в спящую стадию. Спящая стадия – приём, выработанный в результате эволюции бактериофага. Когда колония E. coli находится в неблагоприятных условиях, вирусу выгоднее подождать в спящем режиме более оптимальных условий для бурного литического роста. Белок Cro выключает гены бактериофага ответственные за литическую стадию, инициируя спящий режим.

Здесь и далее в первой строке приводится аминокислотная последовательность белка (в нашем случае Cro). Вторая строчка соответствует экспериментально установленной структуре из базы данных [5], третья – получена в результате проведенных расчетов, указаны коэффициенты точности. Напомним C3 – общая точность распознавания. C(alpha/beta/coil) – точность распознавания конкретного типа вторичной структуры.

```
MEQRITLKDYAMRFGQTKTAKDLGVYQSAINKAIHAGRKIFLTINADGSVYAEEVKPFPSNKKTTA
-ssssshhhhhhhh-hhhhhhhh---hhhhhhhhhh--ssssssss-ssssssss----------
--ssssh--h---h-hhhhhhhh----hhhhhhhhh--ssssssss-ssssssss-s---------

C3:         0.878788
C(alpha):   0.815068
C(beta):    0.92674
C(coil):    0.74525
```

Распознавание вторичной структуры гемоглобина человека. Гемоглобин сложный железосодержащий белок эритроцитов животных и

человека, способный обратимо связываться с кислородом. Основная функция гемоглобина – перенос кислорода в ткани. Молекула гемоглобина состоит из двух пар одинаковых белков (рис. 5.5).

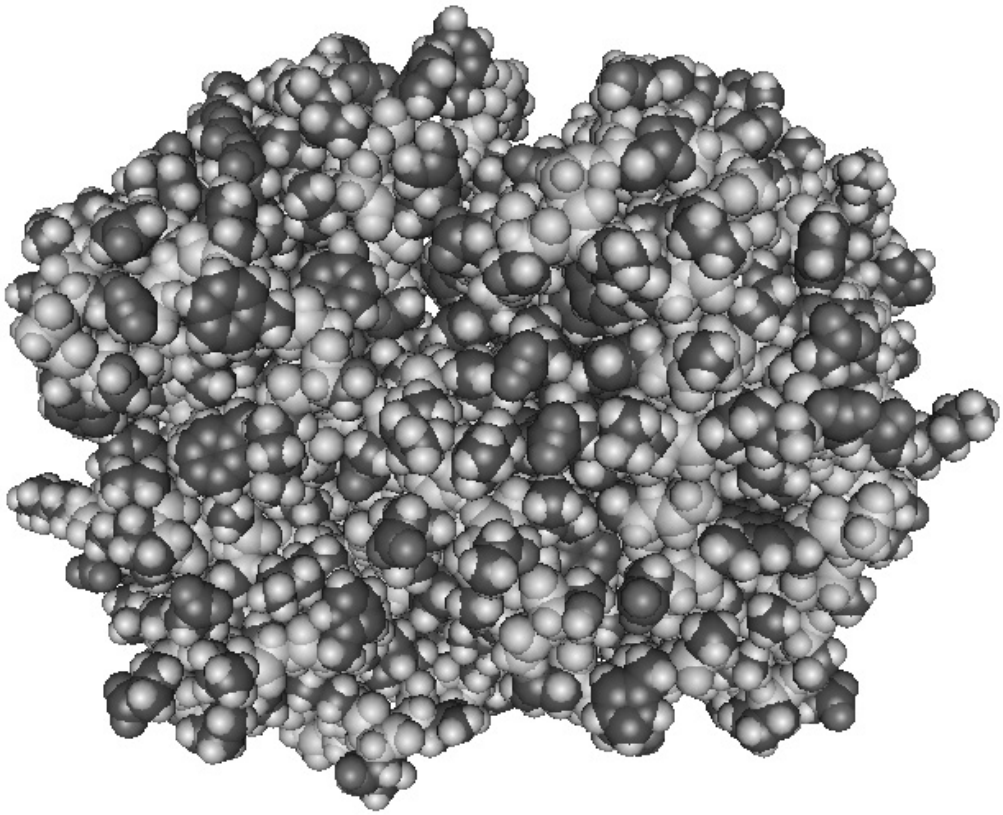

Рис. 5.5 Молекула гемоглобина человека

Это так называемые две α-цепи и две β-цепи (название цепей никак не связано с названием элементов вторичной структуры)[55]. Для определения вторичной структуры гемоглобина необходимо определить вторичную структуру каждого из входящих в его состав белков. Поскольку белки с одинаковой аминокислотной последовательностью формируют одинаковую вторичную структуру, то нам будет достаточно определить вторичную структуру одной α-цепи и одной β-цепи.

Распознавание вторичной структуры α-цепи гемоглобина (рис 5.6):

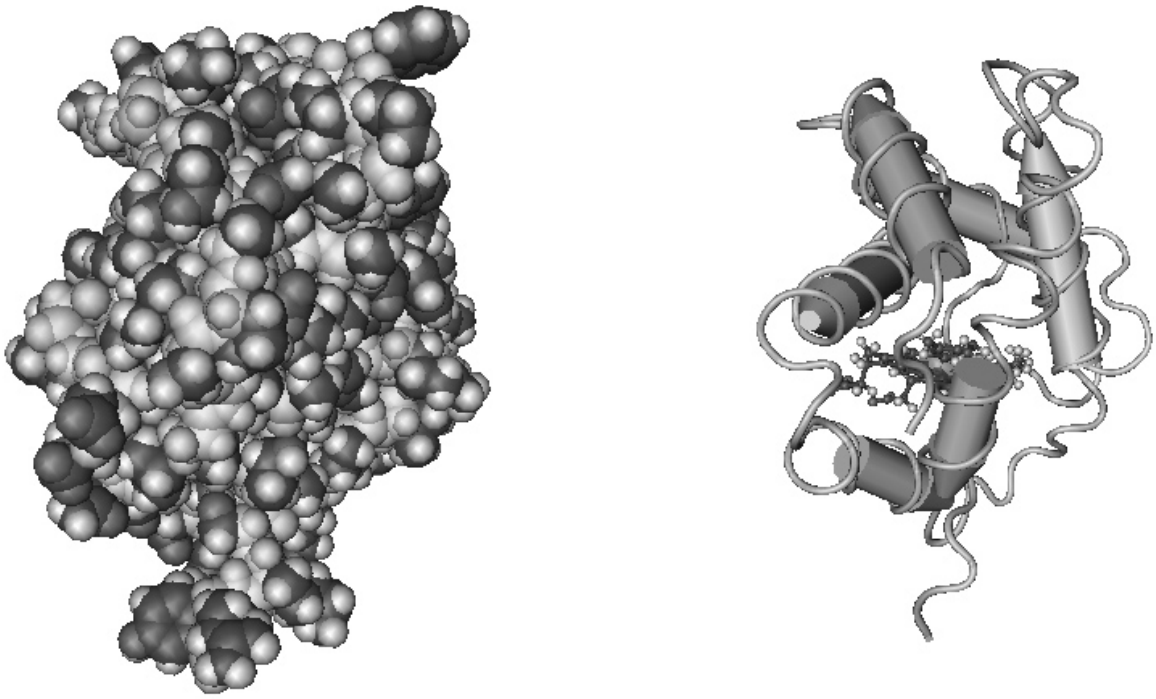

Рис 5.6 α-цепь молекулы гемоглобина человека

```
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAV
-----hhhhhhhhhhhh---hhhhhhhhhhhhh--------------------hhhhhhhhhhhhhhhhh
---hhhhhhhhhhhhh---hhhhhhhhhhhhhhh-----------------hhhhhhhhhhhhhhhhh

AHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY
h---------hhhhhhhh---------hhhhhhhhhhhh--------------------hhhhhhhh---
hh--------hhhhhhhh---------hhhhhhhhhhhhhhh------hhhhhhhhhhhhhhhhhh---

R
-
-

C3:         0.858156
C(alpha):   0.746525
C(beta):    -
C(coil):    0.746525
```

Распознавание вторичной структуры β-цепи гемоглобина (рис. 5.7):

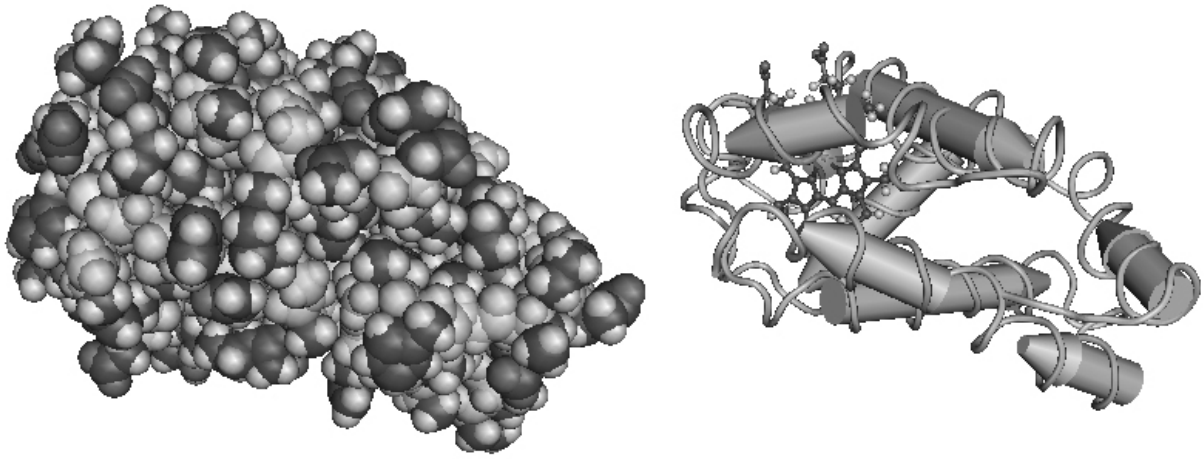

Рис. 5.7 β -цепь гемоглобина человека

```
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGA
----hhhhhhhhh---------hhhhhhhhhhhh------------------------hhhhhhhhhh
----hhhhhhhhhhh-----h-hhhhhhhhhhhh--------------------h--hhhhhhhhhhhh

FSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANA
hhhhhh---------hhhhhhhh---------hhhhhhhhhhhhhhh-----hhhhhhhh----hhhhhh
hh-h-h---------hhhhhhhh--------hhhhhhhhhhhhhhhhh-----hhhhhhhhhhhhhhhh

LAHKYH
hh----
hh----
```

```
C3:         0.890411
C(alpha):   0.786733
C(beta):    -
C(coil):    0.786733
```

Общая точность предсказания вторичной структуры всей молекулы гемоглобина приведена ниже.

```
C3:         0.874564
C(alpha):   0.764788
C(beta):    -
C(coil):    0.764788
```

Заметим, что коэффициент точности распознавания β-слоев не определен для гемоглобина. Это связано с тем, что гемоглобин не формирует β-слоев.

Распознавание вторичной структуры лизоцима человека. Лизоцим (мурамидаза) — антибактериальный агент, фермент класса гидролаз, разрушающий клеточные оболочки бактерий путём гидролиза мурамилглюкозамина клеточной стенки грам-положительных бактерий. Лизоцим (рис. 5.8) содержится, в первую очередь, в местах соприкосновения организма человека с окружающей средой — в слизистой оболочке желудочно-кишечного тракта, слёзной жидкости, слюне, слизи носоглотки и т. д.

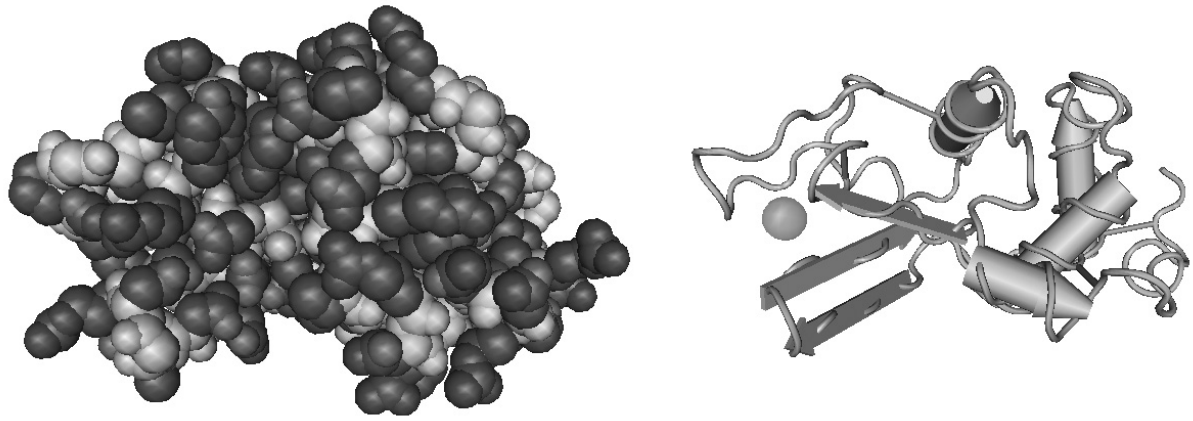

Рис. 5.8 Лизоцим человека

```
KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYGIFQINSRYWCNDGKN
----hhhhhhhhhh---------hhhhhhhhhh------sssssssssssss-ssssss--------
----hhhhhhhhhh---------hhhhhhhhhh----s-ssssss--ssssss---ss-s-s-s----

PGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV
------------------hhhhhhhhhhh-------hhhhhhhh--------------
------------------hhhhhhhhhhh-----------------------------

C3:        0.869231
C(alpha):  0.840393
C(beta):   0.751852
C(coil):   0.746826
```

Распознавание вторичной структуры белков p17 и p24 ВИЧ. ВИЧ – вирус иммунодефицита человека, вызывающий вирусное заболевание ВИЧ-инфекция, последняя стадия которой известна как синдром приобретённого иммунодефицита (СПИД). Иная точка зрения на развитие СПИД изложена в [56].

В состав оболочки вируса ВИЧ входит около 2000 копий белка p17 и приблизительно столько же копий белка p24. Именно белки p17 и p24 являются мощными антигенами, инициирующими иммунный ответ организма. Все современные методики тестирования на ВИЧ основаны на этом факте. Эффективное распознавание антигенов иммунной системой – залог успеха в борьбе организма с инфекцией.

Распознавание вторичной структуры белка p17 ВИЧ (рис. 5.9)

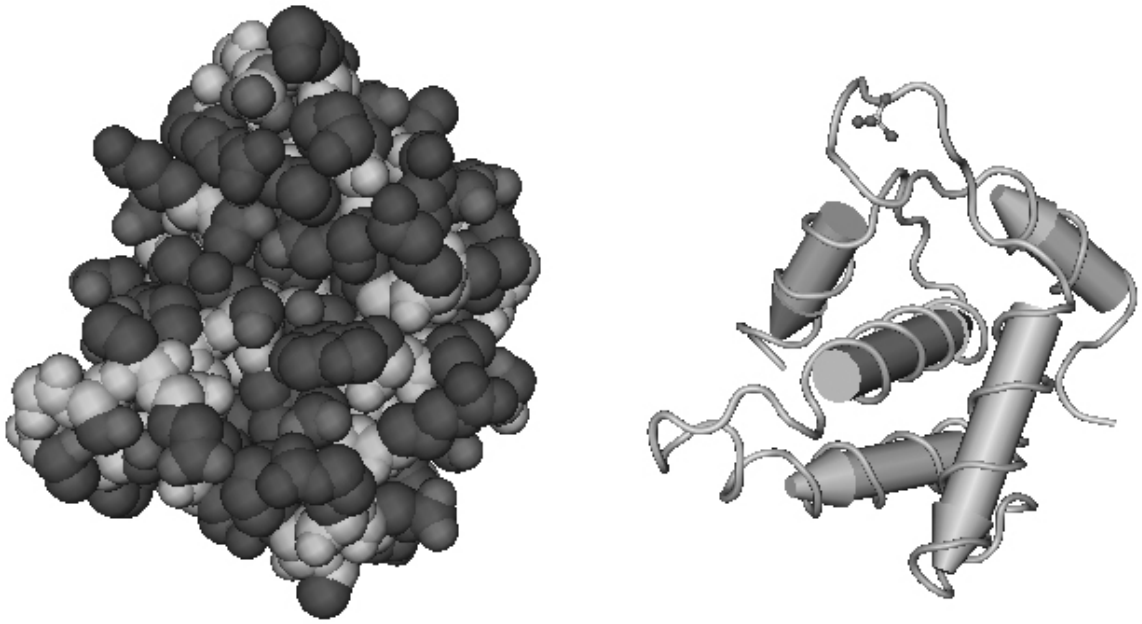

Рис. 5.9 Белок p17 ВИЧ

```
HMGARASVLSGGELDKWEKIRLRPGGKKQYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQ
----------hhhhhhhhh------------hhhhhhhhhhhhh---------hhhhhhhhhhh-----
-----------hhhhhhhh------------hhhhhhhhhhhhhh------------hhhhhhhh-----

TGSEELRSLYNTIAVLYCVHQRIDVKDTKEALDKIEEEQNKSKKKAQQAAADTGNNSQVSQNY
------hhhhhhhhhhhhhhh------hhhhhhhhh--------------------------
---hhhhhhhhhhhhhhhhhh------hhhhhhhhhhhhh--hh-h-hh-hh-----------
```

```
C3:         0.857143
C(alpha):   0.725128
C(beta):    -
C(coil):    0.725128
```

Распознавание вторичной структуры белка p24 ВИЧ (рис. 5.10)

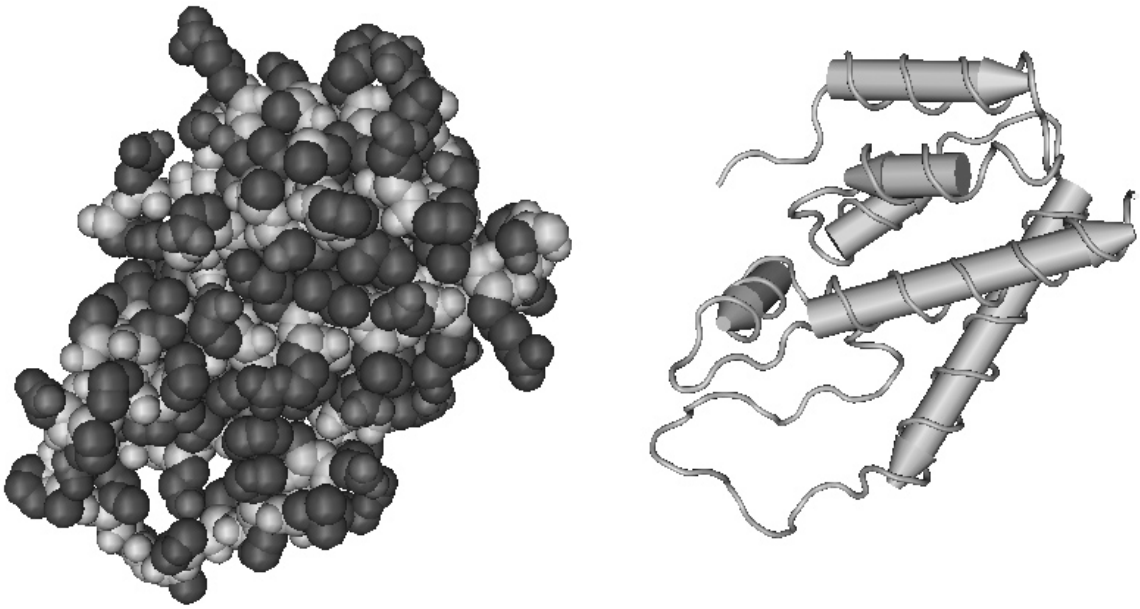

Рис. 5.10 Белок p24 ВИЧ

```
PIVQNLQGQMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLK
----------------hhhhhhhhhhhhh-------hhhhhhhh----hhhhhhhhh-----hhhhhhhh
ssssss-ssssss---hhhhhhhhhhhhh-------hhhhhhh-----hhhhhhhhh-----hhhhhhhh

ETINEEAAEWDRLHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTHNPPIPVGEIYKRWIILGLNK
hhhhhhhhhhhhhh--------------------------hhhhhhhhh------hhhhhhhhhhhhhhh
hhhhhhhhhhhhh---------------------------hhhhhhhhh------hhhhhhhhhhhhhhh

IVRMYS
hhhhh-
hh-h-h

C3:         0.883562
C(alpha):   0.923716
C(beta):    -
C(coil):    0.767189
```

Выводы. Численные расчеты показали, что байесовские процедуры распознавания на цепях Маркова довольно успешно предсказывают вторичную структуру белков. В отличие от известных методов предсказания вторичной структуры [1], используемые методы математически строго обоснованы, обладают полиномиальными оценками погрешности от размеров обучающих выборок и количества признаков, и являются оптимальными для цепей Маркова.

ВЫВОДЫ

В диссертации построены эффективные процедуры распознавания для цепей Маркова и независимых признаков. Получены точные верхние и нижние оценки погрешности в зависимости от размеров обучающей выборки, количества признаков и числа значений признаков.

Основные научные результаты диссертации.

1. Верхняя и нижняя оценки байесовской процедуры распознавания обобщены на дискретный случай.
2. Исследованы асимптотические свойства оценок переходных вероятностей для нестационарных цепей Маркова.
3. Обосновано использование цепей Маркова в качестве модели описания аминокислотных последовательностей белков.
4. Проведен статистический анализ геномов растений и бактерий. Проанализирована схема записи белок-кодирующих генов в геномах бактерий.
5. Разработаны полиномиальные методы распознавания вторичной структуры белков с учетом специфики исследуемых объектов.
6. Разработано необходимое программное обеспечение и проведены эксперименты на реальных данных с целью подтверждения теоретических результатов полученных в диссертации.
7. Разработана информационная технология распознавания вторичной структуры белков для кластерного компьютера.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Baldi P., Brunak S. Bioinformatics: machine learning approach. – Cambridge: MIT Press, 2001. – 452 p.
2. Гупал А.М., Пашко С.В., Сергиенко И.В. Эффективность байесовской процедуры распознавания // Кибернетика и системный анализ. – 1995. - № 4. – С.76-89.
3. Сергиенко И.В., Гупал А.М., Пашко С.В. О сложности задач распознавания образов // Кибернетика и системный анализ. – 1996. – № 4. – С.70-88.
4. Anderson T.W., Goodman L.A. Statistical inference about Markov Chains // The Annals of Mathematical Statistics. – 1957. – 28. – P. 89-110.
5. http://www.ncbi.nlm.nih.gov/
6. Б.А. Белецкий, А.М. Гупал. Статистический анализ геномов бактерий. Комплементарность оснований // Проблемы управления и информатики. – 2005. – №6 – С.135-140.
7. Б.А. Белецкий, А.М. Гупал. Статистический анализ геномов растений // Доповіді національної академії наук України. – 2006. – №7 – С.84-87.
8. Б.А. Белецкий, А.М. Гупал. Статистический анализ записи белков в геномах бактерий // Компьютерная математика. – 2006. – №3 – С.127-134.
9. Б.А. Белецкий, А.А. Вагис, С.В. Васильев, А.М. Гупал. Сложность байесовской процедуры индуктивного вывода // Проблемы управления и информатики. – 2006. – №6 – С.55-70.
10. Б.А. Белецкий, С.В. Васильев, А.М. Гупал. Предсказание вторичной структуры белков на основе байесовских процедур распознавания // Проблемы управления и информатики. – 2007. – №1 – С.61-69.
11. И.В. Сергиенко, Б.А. Белецкий, С.В. Васильев, А.М. Гупал. Предсказание вторичной структуры белков на основе байесовских процедур распознавания на цепях Маркова // Кибернетика и системный анализ. – 2007. - №2 – С.59-64.

12. Б.А. Белецкий, С.В. Васильев, А.А. Вагис, А.М. Гупал. Процедуры распознавания вторичной структуры белков // Проблемы управления и информатики. – 2007. – №4 – С.134-139.
13. Б.А. Белецкий, С.В. Васильєв. Предсказание вторичной структуры белков на основе байесовской процедуры распознавания на цепах Маркова // Міжнародний симпозіум «Питання оптимізації обчислень XXXIII». Праці симпозіуму. Смт. Кацивелі, Крим, Україна. 23-28 вересня 2007 г. С.31.
14. Воронцов К.В. Математические методы обучения по прецедентам.
15. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.
16. Langford J. Quantitatively tight sample complexity bounds. – 2002. – Carnegie Mellon Thesis.
17. Пытьев Ю. П. Возможность. Элементы теории и применения. – М.: Эдиториал УРСС, 2000.
18. Трауб Д., Васильковскии Г., Вожьняковский Х. Информация, неопределённость, сложность: Пер. с англ. – М.: Мир, 1988.
19. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. – М.: Наука, 1974.
20. Немировский А С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации. – М.: Наука, 1979, – 383 с.
21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // 14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. – 1995. – Pp. 1137–1145.
22. Mullin M., Sukthankar R. Complete cross-validation for nearest neighbor classifiers //Proceedings of International Conference on Machine Learning. – 2000.
23. Domongos P., Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss // Machine Learning. – 1997. – no. 29. – Pp. 103–130.

24. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов. –М.: Физматлит, 2004. – Т. 13. – С. 5–36.
25. Vapnik V. Statistical Learning Theory. – Wiley, New York, 1998.
26. Lugosi G. On concentration-of-measure inequalities. – Machine Learning Summer School, Australian National University, Canberra. – 2003.
27. Донской В. И. Колмогоровская сложность классов общерекурсивных функций с ограниченной емкостью // Таврический вестник информатики и математики. – 2005. – № 1. – С. 25–34.
28. Rissanen J. Modeling by shortest data description // Automatica. – 1978. –Vol. 14. – Pp. 465–471.
29. Vapnik V. The nature of statistical learning theory. – Springer-Verlag, New York, 1995.
30. Kearns M. J., Mansour Y., Ng A. Y., Ron D. An experimental and theoretical comparison of model selection methods // 8th Conf. on Computational Learning Theory, Santa Cruz, California, US. – 1995. – Pp. 21–30.
31. Herbrich R., Williamson R. Algorithmic luckiness // Journal of Machine Learning Research. – 2002. – no. 3. – Pp. 175–212.
32. Ruckert U., Kramer S. Towards tight bounds for rule learning // Proc. 21th International Conference on Machine Learning, Banff, Canada. – 2004. – P. ??
33. Boucheron S., Bousquet O., Lugosi G. Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. – 2005. – no. 9–_ Pp. 323–375.
34. Сергиенко И. В., Гупал А. М. Принципы построения процедур индуктивного вывода // Кибернетика и системный анализ. – 2006. - № 4. – С. 51-63.
35. Беликов. Б., Лбов Г. С. Байесовские оценки качества распознавания по конечному множеству событий // Доклады Академии Наук. – 2005, том 402, №1. – С. 10–13.

36. Чень Ч., Ли Р. Математическая логика и автоматическое доказательство теорем. – М.: Наука, 1983. – 360 с.
37. Хаусдорф Ф. Теория множеств. – М. Л. : Главная редакция технико-теоретической литературы, 1937, 304 с.
38. Марков А.А. Исчисление вероятностей. – М., 1924. – 592 с.
39. Марков А.А. Избранные труды. Теория чисел. Теория вероятностей. – Л.: Изд. АН СССР, 1951. – 720 с.
40. Ширяев А.Н. Вероятность. – М.: Наука, 1989. – 640 с.
41. Крамер Г. Математические методы статистики, 2-е изд. – М.: Мир, 1975. – 648 с.
42. Chernoff H. Large-sample theory: parametric case // The Annals of Mathematical Statistics. – 1956. – 27. – P. 1-22.
43. Сингер М., Берг П. Гены и геномы: В 2-х т. Т. 1. Пер. с англ. – М.: Мир, 1998. – 373 с.
44. Вейр Б. Анализ генетических данных. – М.: Мир, 1995. – 400 с.
45. Сергиенко И.В., Гупал А.М., Вагис А.А. Соотношения комплементарности в записи оснований по одной нити ДНК // Цитология и генетика. – 2005. – № 6. – С. 71-75.
46. Гупал А.М., Вагис А.А. Комплементарность оснований в хромосомах ДНК // Проблемы управления и информатики. – 2005. - № 5. – С. 153-157.
47. М.В. Гусев, Л А. Минеева. Микробиология. – М.: Изд-во МГУ, 1992. – 412 с.
48. Baisnée P.-F., Hampson S., Baldi P. Why are complementary DNA strands symmetric? // Bioinformatics – 2002. – 18, N. 8 – P. 1021 – 1031.
49. Ginalski, K., Grishin, N.V., Godzik,A. and Rychlewski,L. Practical lessons from protein structure prediction // Nucleic Acids Res. – 2005. - 33. – P. 1874-1891.
50. Casti J.L. Confronting science's logical limits // Scientific America. – October 1996. – P.78-81.

51. Финкельштейн А.В., Птицын О.Б. Физика белка: Курс лекций лекций с цветными и стереоскопическими иллюстрациями и задачами. – 3-е изд., испр. и доп. – М.: КДУ, 2005. – 456 с.
52. Rost B. Rising accuracy of protein secondary structure prediction // in ‘Protein structure determination, analysis, and modeling for drug discovery’ ed. D. Chasman, New York, 2003. –P. 207-249.
53. http://cubic.bioc.columbia.edu/eva/
54. Пташне М. Переключение генов. Регуляция генной активности и фаг λ: Пер. с англ. – М.: Мир, 1989. – 160 с.
55. Основы биохимиии. Ю.Б. Филлипович. М. Высшая Школа 1985. – 503 с.
56. Дмитревский А. А.,. Сазонова И.М. СПИД: приговор отменяется. – М.: ООО «Издательство «Олимп»: ООО «Издательство АСТ», 2003. – 365 с.
57. Яглом А.М., Яглом И.М. Вероятность и информация. – М.: Государственное издательство технико-теоретической литературы, 1957. – 160 с.