МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА

Механико–математический факультет

На правах рукописи
УДК 621.391.15

Воронина Анна Никитична

ТЕОРЕТИКО-ВЕРОЯТНОСТНЫЕ И КОМБИНАТОРНЫЕ ЗАДАЧИ ТЕОРИИ КОДИРОВАНИЯ ДНК-ПОСЛЕДОВАТЕЛЬНОСТЕЙ

01.01.05 — теория вероятностей и математическая статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата физико–математических наук

МОСКВА
2010

Работа выполнена на кафедре теории вероятностей механико-математического факультета Московского государственного университета имени М. В. Ломоносова.

Научный руководитель	доктор физико–математических наук, профессор Дьячков Аркадий Георгиевич
Официальные оппоненты	доктор физико–математических наук, профессор Гуревич Борис Маркович
	кандидат физико–математических наук Виленкин Павел Александрович
Ведущая организация	Институт Проблем Передачи Информации РАН имени А. А. Харкевича

Защита диссертации состоится « 9 » апреля 2010 г. в 16 часов 45 минут на заседании диссертационного совета Д 501.001.85 при Московском государственном университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, механико-математический факультет, аудитория 16–24.

С диссертацией можно ознакомиться в библиотеке механико-математического факультета (Главное здание, 14 этаж).

Автореферат разослан « 9 » марта 2010 г.

Ученый секретарь диссертационного
совета Д 501.001.85 при МГУ,
доктор физико-математических наук,
профессор И.Н. Сергеев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы В диссертации исследуются теоретико-вероятностные и комбинаторные задачи теории кодирования, возникающие в приложениях молекулярной биологии при рассмотрении экспериментов, существенно использующих специальные свойства молекул ДНК (а также РНК и некоторых других).

В упрощенном виде молекулу (одинарную цепочку) ДНК можно считать последовательностью с элементами из алфавита $\{A, C, G, T\}$. ДНК-цепочки являются направленными, а их важнейшим свойством является способность разнонаправленных цепочек сращиваться в *двойные спирали ДНК*, или *ДНК-дуплексы* в процессе *ДНК-гибридизации*. Основой этого процесса является образование водородных связей между так называемыми комплементарными элементами ДНК-цепочек, а именно элементами A и T, а также C и G. При этом прочность образовавшегося дуплекса определяется величиной, называемой *энергией гибридизации* и зависящей от числа образовавшихся водородных связей. *Сопряженной* к данной ДНК-цепочке называется ДНК-цепочка, полученная путем изменения направления исходной ДНК-цепочки и последующей замены всех элементов в ней на комплементарные. Известно, что максимальная энергия ДНК-гибридизации достигается при образовании *дуплексов Ватсона-Крика*, то есть дуплексов, состоящих из взаимно-сопряженных ДНК-цепочек.

В биологических экспериментах, использующих свойство ДНК-гибридизации, возникновения *кроссгибридизации* – то есть сращивание ДНК-цепочек, не являющихся взаимно-сопряженными, – следует избегать, так как она приводит к ошибкам в результатах опытов. Следовательно, желательно формирование таких наборов одинарных ДНК-цепочек (*ДНК-ансамблей*), что при заданных температурных (энергетических) условиях эксперимента только энергия гибридизации между сопряженными ДНК-цепочками будет достаточна для формирования устойчивых дуплексов, а кроссгибридизация будет невозможна. Это обстоятельство, наряду с другими предпосылками, привело к пониманию, что возникающие в таких биологических экспериментах ДНК-ансамбли, с математической точки зрения, могут быть естественным образом интерпретированы как специальный вид кодов.

В самом общем виде одну из основных задач теории кодирования можно описать следующим образом. Фиксируется некоторый q-ичный алфавит $\mathcal{A}_q \triangleq \{0, \ldots, q-1\}$ и рассматривается пространство q-ичных последовательностей длины n: $\mathcal{A}_q^n \triangleq \{\boldsymbol{x} = (x_1, \ldots, x_n),\ x_i \in \mathcal{A}_q\}$. *Ко-*

дом $\mathcal{X}$ называется некоторый набор таких последовательностей (*кодовых слов*) – подмножество $\mathcal{A}_q^n$: $\mathcal{X} \subset \mathcal{A}_q^n$. На множестве $\mathcal{A}_q^n$ задается функция расстояния $\rho(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}_q^n$. Как правило, на $\mathcal{A}_q^n$ также можно определить комплементарную $\rho(\boldsymbol{x}, \boldsymbol{y})$ функцию сходства. Говорят, что код $\mathcal{X}$ является *кодом с расстоянием* $D > 0$, если $\rho(\boldsymbol{x}, \boldsymbol{y}) \geq D$ для любых кодовых слов $\boldsymbol{x} \neq \boldsymbol{y}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

Пусть $N(n, D)$ есть максимальный *объем* кода (то есть максимальное число кодовых слов в коде) с фиксированными расстоянием D и длиной кодовых слов n. Ставится задача исследования *скорости* кода $R(d)$, то есть логарифмической асимптотики числа $N(n, dn)$ при фиксированной *доле расстояния* $d > 0$ и растущей длине кодовых слов:

$$R(d) \triangleq \varlimsup_{n\to\infty} \frac{\log_q N(n, dn)}{n}, \qquad d > 0.$$

На сегодняшний день эта задача все еще не имеет окончательного решения, а основные достижения состоят в построении *верхних и нижних границ* скорости кодов. Следует, кроме того, отметить, что основным направлением исследования было изучение данных вопросов для так называемого расстояния Хэмминга[1], которое определяется как число несовпадающих позиций между двумя q-ичными последовательностями.

Вторым важным направлением исследований является поиск кодовых конструкций, обладающих параметрами, близкими к теоретически достижимым, либо какими-то специальными свойствами, в зависимости от предполагаемых приложений. Здесь не существует общей теории построения кодов, а каждый пример конструкции уникален, причем обычно существует лишь для весьма узкого диапазона параметров кода.

Задача исследования указанной проблематики для "нехэммингов-ских" функций расстояния и сходства изучена слабо, главным образом в связи с отсутствием до последнего времени практической потребности в результатах такого рода.

Рассматриваемые нами постановки возникли из некоторых прикладных вопросов молекулярной биологии и ДНК-программирования. Принято считать, что первый пример практического алгоритма, позволяющего эффективно решать математическую задачу (так называемую "задачу почтальона") методами ДНК-программирования был по-

[1] Ф. Дж. Мак-Вильямс, Н. Дж. А. Слоэн, *Теория кодов, исправляющих ошибки*, Москва: Связь, 1979.

строен в работе Адлемана[2] в 1994г. Вследствие относительной новизны данной области исследований устоявшихся канонов математического моделирования ДНК-цепочек и их взаимодействия между собой еще не сложилось. Основные принципы, однако, ясны: ДНК-цепочки могут быть представлены как последовательности из четверичного алфавита $\mathcal{A}_4 \triangleq \{0, 1, 2, 3\}$. При этом основные свойства ДНК-цепочек, а также ограничения, вытекающие из потребностей биологических приложений, должны быть отражены в специальных требованиях к конструкции (определению) кода, а энергия ДНК-гибридизации сопоставляется соответствующей функции сходства на пространстве четверичных последовательностей $\mathcal{A}_4^n$. Первые попытки создания модели ДНК-кодов были предприняты в рамках уже хорошо изученной метрики Хэмминга[3,4]. Специфика задачи учитывалась путем определения понятия *пар взаимно сопряженных* ДНК-последовательностей и наложения на код некоторых дополнительных условий.

Этот подход, однако, никак не учитывал особенностей *взаимодействия* цепочек ДНК и в дальнейшем исследования пошли по пути разработки функции сходства, наиболее близко к действительной ситуации моделирующей реальные принципы гибридизации цепочек ДНК, с одной стороны, и поддающейся математическому изучению, с другой.

Первая функция сходства, учитывающая особенности взаимодействия цепочек ДНК, была предложена в работах Дьячкова[5] и Виленкина[6] (здесь и далее, ссылаясь на работы, имеющие больше одного авторов, мы для краткости будем указывать в тексте только первого автора). Она была близка к сходству Хэмминга с тем отличием, что каждому *типу* совпадающих символов присваивался свой вес $w(a)$, $a \in \mathcal{A}_4$, в общей сумме, определяющей такое сходство.

Следующим шагом в эволюции моделей ДНК-кодов стало примене-

---

[2]L. Adleman, "Molecular Computation of Solutions to Combinatorial Problems," *Science*, vol. 266. pp. 1021–1024, 1994.

[3]A. Marathe, A. E. Condon, R. M. Corn, "On combinatorial DNA design," *J. Comp. Biol.*, vol. 8, pp. 201–219, 2001.

[4]V. V. Rykov, A. J. Macula, C. M. Korzelius, D. C. Englehart, D. C. Torney, P. S. White, "DNA sequences constructed on the basis of quaternary cyclic codes," в *Proc. of the 4th World Multiconference on Systematics, Cybernetics, and Informatics, SCI 2000/ISAS 2000*, Орландо, Флорида, 2000.

[5]A. G. D'yachkov, D. C. Torney, "On similarity codes," *IEEE Trans. Inform. Th.*, vol. 46, n. 4, pp. 1558–1664, 2000.

[6]П. А. Виленкин, "Асимптотические задачи комбинаторной теории кодирования и теории информации," *Диссертация на соискание ученой степени к.ф.-м.н.* Москва, МГУ им. М.В.Ломоносова. 2000.

ние *неаддитивной* функции, так называемого *сходства выпадений*[7]. Здесь сходство между двумя ДНК-цепочками определяется как длина наибольшей общей подпоследовательности (элементы которой стоят на любых, не обязательно одинаковых, позициях в этих ДНК-цепочках). Такой подход позволяет учесть возможные на практике сдвиги цепочек друг относительно друга и образование петель на одинарных цепочках (так называемую *вторичную структуру* ДНК-дуплекса).

Еще один вариант определения сходства можно найти в работе Дьячкова 2005 г.[8]. Сходство между двумя ДНК-цепочками, называемое *сходством блоков*, в этом случае определялось как длина общей *блочной* подпоследовательности, то есть такой, что любые два ее соседние элемента либо одновременно являются соседними, либо одновременно разделены в исходных ДНК-цепочках.

Наиболее продвинутой *биологической* моделью, позволяющей с максимальной точностью вычислить энергию гибридизации ДНК-цепочек, на сегодняшний день считается так называемая модель "ближайшего соседа"[9]. В общих чертах, согласно этой теории, энергия гибридизации может быть рассчитана как сумма *термодинамических весов* всех образовавшихся при гибридизации *стеблей*. Стебель формируется, когда два последовательных основания (элемента) одной ДНК-цепочки связываются с двумя последовательными основаниями второй цепочки; термодинамические веса разных видов (в зависимости от вида входящих в него видов оснований) стеблей известны заранее из постановочных экспериментов.

Таким образом, как мы видим, "хорошая" функция сходства для ДНК-кодов должна зависеть не от числа совпадающих символов в ДНК-последовательностях, а от количества совпадающих стеблей, или блоков длины 2 в этих последовательностях. В настоящей работе мы исследуем три функции сходства, построенные с учетом этого ключевого принципа модели "ближайшего соседа" (подробнее о них можно прочитать во второй части настоящего автореферата).

---

[7] A. G. D'yachkov, P. L. Erdos, A. J. Macula, V. V. Rykov, D. C. Torney, C. S. Tung, P. A. Vilenkin, P. S. White, "Exordium for DNA Codes," *J. Comb. Optimization*, vol. 7, n. 4, pp. 369–379, 2003.

[8] А. Г. Дьячов, П. А. Виленкин, И. К. Исмагилов, Р. С. Сарбаев, А. Макула, Д. Торни, С. Уайт, "О ДНК кодах," *Проблемы Передачи Информации*, т. 41, н. 4, с. 57–77, 2005.

[9] K. J. Breslauer, R. Frank, H. Blocker, L. A. Markey, "Predicting Duplex DNA Stability from the Base Sequence," *Proc. National Academy of Sciences USA*, vol. 83, pp. 3746–3750, 1986.

Требуемые на практике ДНК-ансамбли, таким образом, могут рассматриваться как коды с расстоянием для специальной функции расстояния, комплементарной соответствующей "стебельной" функции сходства. При этом, исходя из практических соображений, эти коды должны состоять из пар взаимно-сопряженных ДНК-последовательностей и не должны содержать самосопряженных кодовых слов. Такие коды мы называем *ДНК-кодами*. Отметим, что указанное ограничение является достаточно специфичным и накладывает дополнительные условия как при исследовании асимптотических границ скорости кодов, так и при поиске конкретных кодовых конструкций.

Для всех изучаемых нами функций сходства мы будем интересоваться поведением скорости ДНК-кодов, а для аддитивного стебельного **1**-сходства также построим два примера кодовых конструкций.

Отметим, что существуют и другие направления исследований ДНК-последовательностей методами теории кодирования. Так, например, в работе Миленкович[10] изучается вопрос о построении ансамблей ДНК-цепочек, с вероятностью, близкой к единице (эмпирически), исключающих самогибридизацию, то есть образование склеек внутри одной цепочки.

В целом, существенное внимание уделяется практическим алгоритмам решения конкретных задач, как например, алгоритмы определения длины наибольшей общей подпоследовательности между двумя цепочками ДНК[11]. Разрабатываются комбинаторные, эвристические, биологические методы нахождения ДНК кодов – их можно найти, например, в работах Кадерали[12] и многих других авторов, программные решения для генерирования ДНК-кодов предложили Андронеску[13], Бишоп[14] и другие исследователи.

---

[10]O. Milenkovic, N. Kashyap, "DNA Codes that Avoid Secondary Structures," в *Proc.* 2005 *IEEE Int. Symp. Information Theory*, Аделаида, Южная Австралия, Австралия, 2005, pp. 288–292.

[11]S. B. Needleman, C. D. Wunsch, "A general method applicable to the search for similarities in the amino-acid sequences of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, 1970.

[12]L. Kaderali, A. Deshpande, J. Nolan, P. White, "Primer-design for multiplexed genotyping," *Nucleic Acids Res.*, vol. 31, pp. 1796–1802, 2003.

[13]M. Andronescu, R. Aguirre-Hernandez, A. Condon, et al., "RNAsoft: a suite of RNA secondary structure prediction and design software tools *Nucleic Acids Res.*, vol. 31, pp. 3416–3422, 2003.
Также доступно на: http:// www.rnasoft.com.

[14]M. Bishop, A. Macula, T. Renz, SynDCode Suite, 2006.

Как уже отмечалось, рассматриваемая задача возникла при построении математических моделей некоторых экспериментов молекулярной биологии. ДНК-коды находят много новых применений в активно развивающихся направлениях науки, таких как определение функции генов[15], самосборка наноструктур[16], ДНК-программирование[2,17], хранение информации[18], ДНК-метки (DNA taggants)[19] и другие.

Данная область исследований является сравнительно новой и бурно развивающейся, и единой устоявшейся модели ДНК-кодов как математического объекта на сегодняшний день нет. Отметим, однако, что для указанных приложений интерпретация требуемых ДНК-ансамблей как специального вида кодов с заданным минимальным расстоянием для некоторой нетипичной функции сходства возникает совершенно естественным образом. Предлагаемые нами функции сходства, в отличие от предыдущих примеров, учитывают ключевой принцип вычисления энергии гибридизации в модели "ближайшего соседа". Особо укажем, что адекватность рассматриваемой нами модели лишний раз подтверждается тем обстоятельством, что идеологически она совпадает с *биологическими* определениями ДНК-кодов, данными, например, в работе Кадерали[12], и отвечает принципам, на которых построены прототипы ДНК-алгоритмов вычисления комбинаторных задач высокой сложности[2,17]. Таким образом, задача изучения ДНК-кодов является актуальной и востребованной, а предлагаемые в диссертации математические модели позволяют с наибольшей на сегодняшний день точностью описывать возникающие в приложениях биологические объекты.

Доступно на: http://syndcode.geneseo.edu/.

[15]R. Eason, N. Pourmand, W. Tongprasit, et al., "Characterization of synthetic DNA bar codes in Saccharomyces cerevisiae gene-deletion strains *PNAS*, vol. 101, pp. 11046–11051, 2004.

[16]M. Valignat, O. Theodoly, J. Crocker, et al., "Reversible self-assembly and directed assemblyof DNA-linked micrometer-sized colloids," *PNAS*, vol. 102, pp. 4225–4229, 2005.

[17]Q. Ouyang, P. D. Kaplan, S. Liu, A. Libchaber, "DNA solution of the maximal clique problem," *Science*, vol. 278, pp. 446–449, 1997.

[18]M. Mansuripur, P. K. Khulbe, S. M. Kuebler, J. W. Perry, M. S. Giridhar, N. Peyghambarian, "Information Storage and Retrieval using Macromolecules as Storage Media," *Optical Data Storage Conference*, Ванкувер, Канада, 2003.

[19]A. Macula, S. Gal, C. Andam, T. E. Renz, M. A. Bishop, "PCR nonadaptive group testing of DNA libraries for biomolecular computing and taggant applications," *Discrete Mathematics, Algorithms and Applications*, vol. 1, n. 1, pp. 59-69, 2009.

Цель работы К основным целям настоящей диссертации относятся: изучение асимптотического поведения максимального объема ДНК-кодов для трех стебельных функций сходства; вычисление объемов сфер для метрики, задаваемой аддитивным стебельным **1**-сходством, в пространстве q-ичных последовательностей; исследование вопроса о построении кодовых конструкций для ДНК-кодов, основанных на аддитивном стебельном **1**-сходстве.

Научная новизна Основные результаты диссертации являются новыми и состоят в решении следующих задач для специального класса кодов – ДНК-кодов – для трех различных функций сходства на пространстве q-ичных последовательностей:

1. Получены две конструкции ДНК-кодов, основанных на аддитивном стебельном **1**-сходстве. Найден максимальный объем ДНК-кодов, основанных на аддитивном стебельном **1**-сходстве, для случая фиксированного сходства.

2. Найдены явные формулы для выражения объема сфер для метрики, задаваемой аддитивным стебельным **1**-сходством, в пространстве q-ичных последовательностей $\mathcal{A}_q^n$, получены неасимптотические оценки для таких объемов сфер, а также асимптотические формулы для случая постоянного радиуса.

3. Сформулированы верхние и нижние оценки максимального объема ДНК-кодов, основанных на аддитивном стебельном **1**-расстоянии, для случая постоянного расстояния (нулевой доли расстояния). Построены и исследованы верхние и нижние границы скорости ДНК-кодов, основанных на аддитивном стебельном **1**-расстоянии.

4. Разработан метод случайного кодирования с применением цепей Маркова. Опираясь на него, построены и исследованы верхние и нижние границы скорости ДНК-кодов, основанных на аддитивном стебельном w-расстоянии.

5. С применением специального класса ДНК-ансамблей Фибоначчи построена нижняя граница скорости ДНК-кодов, основанных на неаддитивном стебельном w-расстоянии.

Методы исследования Результаты диссертации получены с использованием различных методов построения границ скорости кодов

с расстоянием (Варшамова–Гилберта, Хэмминга, Плоткина, Элайеса, Синглтона), стандартных методов исследования логарифмической асимптотики, теорем о больших уклонениях для сумм независимых и зависимых случайных величин, методов выпуклого анализа. При доказательстве границы Варшамова-Гилберта для аддитивного стебельного w-расстояния был разработан метод случайного кодирования на основе марковских цепей. Это потребовало развития новой техники доказательства теорем подобного рода. При вычислении явных формул для объемов сфер для метрики, задаваемой аддитивным стебельным **1**-сходством, были применены комбинаторные методы подсчета числа q-ичных последовательностей специального вида, в частности, найдены рекуррентные уравнения и явные выражения для числа таких последовательностей. Для построения графиков границ скорости ДНК-кодов также были использованы некоторые численные методы.

Теоретическая и практическая ценность Диссертация носит теоретический характер. Полученные результаты и разработанные техники исследования кодов могут представлять интерес для специалистов, занимающихся теоретико-вероятностной и комбинаторной теорией кодирования, и, кроме того, могут быть использованы для некоторых практических задач молекулярной биологии.

Апробация работы Результаты диссертации неоднократно докладывались на Большом семинаре кафедры теории вероятностей мехмата МГУ (2008г. и 2009г., руководитель – член-корреспондент РАН А. Н. Ширяев), на семинаре по теории кодирования в Институте Проблем Передачи Информации РАН (2008г. и 2009г., руководитель – профессор Л. А. Бассалыго) и на семинаре по теории вероятностей и статистической физике в МГУ им. Ломоносова (2007г. и 2010г., руководитель – профессор Б. М. Гуревич). Полученные результаты были представлены на международном симпозиуме по алгебраической и комбинаторной теории кодирования ACCT-11 (Пампорово, Болгария, 2008), вошли в сборник трудов конференции 31-ой конференции молодых ученых и специалистов ИППИ РАН ИТиС'08.

Публикации По теме диссертации опубликовано 4 работы, список которых приведен в конце настоящего автореферата.

Структура и объем диссертации Диссертация состоит из списка обозначений, шести глав, разбитых на параграфы, списка литературы, содержащего 63 наименования, списка работ автора по теме диссертации и оглавления. Общий объем диссертации составляет 181 страницу.

КРАТКОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

В **главе 1** рассказывается о конструкциях, возникших из потребностей молекулярной биологии, которые приводят к изучению кодов специального вида, являющихся главным объектом рассмотрения диссертации. Мы приводим необходимые сведения из молекулярной биологии и описываем реальный эксперимент, из которого и возникает исследуемая нами задача. Далее делаются некоторые общие предположения, используемые в математической модели рассматриваемого явления. Там же обсуждается специфика сформулированной математической модели по сравнению с классическим объектом изучения теории кодирования и описывается применяемый во всей работе метод случайного кодирования для ДНК-кодов. В конце главы приводится краткая сводка основных результатов диссертации.

Символом $\triangleq$ будем обозначать равенство по определению. Отметим, что нумерация формул в настоящем автореферате не соответствует нумерации в диссертации.

Введем общие определения. Пусть $[n] \triangleq \{1, 2, \ldots, n\}$ обозначает множество целых чисел от 1 до n. Пусть $q = 2, 4, \ldots$ – произвольное фиксированное четное число, а $\mathcal{A}_q \triangleq \{0, 1, \ldots, q-1\}$ – стандартный алфавит объема $|\mathcal{A}_q| = q$. Стандартный символ $\lfloor u \rfloor$ ($\lceil u \rceil$) обозначает наибольшее (наименьшее) целое число $\leq u$ ($\geq u$).

Определим, что для любого $x \in \mathcal{A}_q$ *комплементарным* будет элемент $\bar{x} \triangleq (q-1) - x \in \mathcal{A}_q$ (при построении одного из примеров кодовых конструкций мы будем считать, что для любого $i = 0, 1, \ldots, q/2$ комплементарными являются элементы $2i$ и $2i+1$; понятно, что такое изменение не сказывается существенным образом на вводимой нами математической модели, а только изменяет порядок сопоставления алфавитов $\{A, C, G, T\}$ и $\mathcal{A}_4 = \{0, 1, 2, 3\}$). Для последовательности $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n-1}, x_n) \in \mathcal{A}_q^n$ определим *сопряженную* к ней последовательность $\widetilde{\boldsymbol{x}} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in \mathcal{A}_q^n$.

Пусть $w = w(a, b) \geq 0$, $a, b \in \mathcal{A}_q$, является *весовой функцией* такой, что

$$w(a, b) = w(\bar{b}, \bar{a}), \quad a, b \in \mathcal{A}. \tag{1}$$

Условие (1) означает, что функция $w(a, b)$ инвариантна относительно преобразования Ватсона-Крика.

Введем формальные определения рассматриваемых нами функций сходства.

Определение 3.1: Для двух произвольных q-ичных последовательностей $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{A}_q^n$ и $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{A}_q^n$ длины

n, $n \geq 2$, число

$$\mathcal{S}_{\mathbf{1}}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_{i=1}^{n-1} s_i^{\mathbf{1}}(\boldsymbol{x}, \boldsymbol{y}), \quad \text{где}$$

$$s_i^{\mathbf{1}}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \begin{cases} 1, & \text{если } x_i = y_i,\ x_{i+1} = y_{i+1}, \\ 0, & \text{в остальных случаях,} \end{cases} \quad i \in [n-1], \qquad (2)$$

называется *аддитивным стебельным* $\mathbf{1}$*-сходством* между $\boldsymbol{x}$ и $\boldsymbol{y}$.

Определение 5.1: Для двух произвольных q-ичных последовательностей $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{A}_q^n$ и $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{A}_q^n$ длины n, $n \geq 2$, число

$$\mathcal{S}_w(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_{i=1}^{n-1} s_i^w(\boldsymbol{x}, \boldsymbol{y}), \quad \text{где}$$

$$s_i^w(\boldsymbol{x}, \boldsymbol{y}) \triangleq \begin{cases} w(a,b), & \text{если } x_i = y_i = a,\ x_{i+1} = y_{i+1} = b, \\ 0, & \text{в остальных случаях,} \end{cases}$$

называется *аддитивным стебельным* w*-сходством* между $\boldsymbol{x}$ и $\boldsymbol{y}$.

Для двух произвольных q-ичных последовательностей $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{A}_q^n$ и $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{A}_q^n$ символом

$$\boldsymbol{z} = (z_1, z_2, \ldots, z_\ell) \in \mathcal{A}_q^\ell, \quad \ell \in [n],$$

обозначим *общую подпоследовательность* длины $|\boldsymbol{z}| \triangleq \ell$ между $\boldsymbol{x}$ и $\boldsymbol{y}$, что означает, что существуют две ℓ-последовательности целых чисел

$$1 \leq k_1 < k_2 < \cdots < k_\ell \leq n, \quad 1 \leq j_1 < j_2 < \cdots < j_\ell \leq n,$$

такие что $z_u = x_{k_u} = y_{j_u}$, $u \in [\ell]$.

Определение 6.1: Пусть $2 \leq r \leq n$ – произвольные целые числа. ДНК r-последовательность $\mathbf{a} = (a_1, a_2, \ldots, a_r) \in \mathcal{A}_q^r$, называется *общим блоком для последовательностей* $\boldsymbol{x}$ *и* $\boldsymbol{y}$ (кратко, *общим* $(\boldsymbol{x}, \boldsymbol{y})$*-блоком*) длины r, если последовательности $\boldsymbol{x}$ и $\boldsymbol{y}$ (одновременно) содержат $\mathbf{a}$ как подпоследовательность, состоящую из r последовательных элементов $\boldsymbol{x}$ и $\boldsymbol{y}$.

Определение 6.2: Пусть $2 \leq \ell \leq n$ – некоторое целое число. Последовательность $\boldsymbol{z} = (z_1, z_2, \ldots, z_\ell) \in \mathcal{A}_q^\ell$ называется *общей блоковой подпоследовательностью* длины $|\boldsymbol{z}| \triangleq \ell$ между $\boldsymbol{x}$ и $\boldsymbol{y}$, если $\boldsymbol{z}$ является *упорядоченным набором* непересекающихся общих $(\boldsymbol{x}, \boldsymbol{y})$-блоков и длина каждого общего $(\boldsymbol{x}, \boldsymbol{y})$-блока в этом наборе ≥ 2. Пусть $\mathcal{Z}(\boldsymbol{x}, \boldsymbol{y})$ обозначает множество всех общих блоковых подпоследовательностей

между $\boldsymbol{x}$ и $\boldsymbol{y}$. Для любого $\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{x}, \boldsymbol{y})$ обозначим через $k(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y})$, $1 \leq k(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}) \leq |\boldsymbol{z}|/2$, *минимальное число* общих $(\boldsymbol{x}, \boldsymbol{y})$-блоков, *составляющих* данную подпоследовательность $\boldsymbol{z}$.

Определение 6.4: Пусть $\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{x}, \boldsymbol{y})$ имеет вид

$$\boldsymbol{z} \triangleq \left(\boldsymbol{z}^1, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^{k(\boldsymbol{z},\boldsymbol{x},\boldsymbol{y})}\right), \qquad |\boldsymbol{z}| = \sum_{m=1}^{k(\boldsymbol{z},\boldsymbol{x},\boldsymbol{y})} |\boldsymbol{z}^m| = \sum_{m=1}^{k(\boldsymbol{z},\boldsymbol{x},\boldsymbol{y})} r_m,$$

где

$$\boldsymbol{z}^m \triangleq \left(z_1^m, z_2^m, \ldots, z_{r_m}^m\right) \in \mathcal{A}^{r_m}, \quad m = 1, 2, \ldots, k(\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}),$$

является упорядоченным набором общих $(\boldsymbol{x}, \boldsymbol{y})$-блоков, составляющих $\boldsymbol{z}$. Для ДНК-последовательностей $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}_4^n$ число

$$\mathcal{S}^{(w)}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \max_{\boldsymbol{z} \in \mathcal{Z}(\boldsymbol{x},\boldsymbol{y})} \left\{ \sum_{m=1}^{k(\boldsymbol{z},\boldsymbol{x},\boldsymbol{y})} \sum_{i=1}^{r_m - 1} w\left(z_i^m,\, z_{i+1}^m\right) \right\}$$

называется *неаддитивным стебельным w-сходством* между $\boldsymbol{x}$ и $\boldsymbol{y}$. Для постоянной весовой функции: $w(a, b) = 1$ для любых $a, b \in \mathcal{A}_q$ – будем использовать обозначение $\mathcal{S}^{(\mathbf{1})}(\boldsymbol{x}, \boldsymbol{y})$.

Пусть $\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}_q^n$, обозначает произвольную стебельную функцию сходства для ДНК-последовательностей. Тогда соответствующее ей стебельное расстояние определяется по следующей формуле:

$$\mathcal{D}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \mathcal{S}(\boldsymbol{x}, \boldsymbol{x}) - \mathcal{S}(\boldsymbol{x}, \boldsymbol{y}).$$

Пусть $\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)$, где $\boldsymbol{x}(j) \triangleq (x_1(j), x_2(j), \ldots, x_n(j)) \in \mathcal{A}_q^n$, $j \in [N]$, являются *кодовыми словами* q-ичного *кода* $\mathcal{X} = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)\}$ *длины* n и *объема* N, где $N = 2, 4, \ldots$ *четно*. Пусть D, $0 < D \leq \max\limits_{\boldsymbol{x} \in \mathcal{A}_q^n} \mathcal{S}(\boldsymbol{x}, \boldsymbol{x})$, – произвольное положительное число.

Определение 1.1: Код $\mathcal{X}$ называется ДНК-кодом с расстоянием D для стебельного сходства $\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})$, если выполнены следующие два условия:
(i). Для любого индекса $j \in [N]$, найдется $j' \in [N]$, $j' \neq j$, такой что $\boldsymbol{x}(j') = \widetilde{\overline{\boldsymbol{x}(j)}} \neq \boldsymbol{x}(j)$.
(ii). Для любых $j \neq j'$, расстояние

$$\mathcal{D}\left(\boldsymbol{x}(j), \boldsymbol{x}(j')\right) \geq D. \tag{3}$$

ДНК-коды длины n с расстоянием D, основанные на аддитивном стебельном $\mathbf{1}$-сходстве, будем кратко называть $(n, D)_{\mathbf{1}}$-кодами, соответствующие ДНК-коды, основанные на аддитивном стебельном w-

сходстве, – $(n, D)_w$-кодами, а ДНК-коды, основанные на неаддитивном стебельном w-сходстве, – $(n, D)^{(w)}$-кодами. Такие же индексы будут использоваться и для всех прочих вводимых величин для различных функций сходства.

Обозначим через $N(n, D)$ *максимальный* объем ДНК-кода, основанного на стебельном сходстве $\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})$. Для фиксированного параметра $d > 0$ число

$$R(d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_q N(n, dn)}{n}, \qquad d > 0, \tag{4}$$

называется *скоростью ДНК-кодов* для *доли расстояния* $d > 0$.

Глава 2 В этой главе рассматривается вопрос построения кодовых конструкций для ДНК-кодов. Нами получены следующие 2 примера конкретных кодовых конструкций для аддитивного стебельного **1**-сходства:

Теорема 2.1: *1. Если $n = kq$, где $k = 1, 3, 5, \ldots$ – нечетно, то код*

$$\mathcal{X} = \mathcal{M}_q(n) = \left\{ \boldsymbol{x} \in \mathcal{A}_q^n : x_1 + \cdots + x_n \equiv 0 \pmod q \right\}$$

является $(n, 2)_{\mathbf{1}}$-кодом объема q^{n-1}.
2. Если $n = kq$, где $k = 2, 4, 6, \ldots$ – четно, то код $\mathcal{M}_q(n)$ содержит подкод

$$\mathcal{M}'_q(n) \triangleq \left\{ \boldsymbol{x} \in \mathcal{A}_q^n : x_i \in \mathcal{A}_q,\ x_{n-i+1} = \overline{x_i},\ i \in [n/2] \right\},$$

состоящий из самосопряженных слов, а код $\mathcal{X} = \mathcal{M}_q(n) / \mathcal{M}'_q(n)$ является $(n, 2)_{\mathbf{1}}$-кодом объема $q^{n-1} - q^{n/2}$.

Пусть α обозначает примитивный элемент поля $\mathrm{G}F(4)$, а v_i, $i \in [m]$ – переменные, на которых заданы многочлены, определяющие код Рида-Маллера первого порядка $R_4(1, m)$.

Теорема 2.2: *Код Рида-Маллера первого порядка $R_4(1, m)$ содержит 4^m самосопряженных слов вида*

$$a + b_1 v_1 + b_2 v_2 + \cdots + b_m v_m, \quad \text{где } a \in \mathrm{G}F(4) \ \text{ и } \ \sum_{i=1}^{m} b_i = \alpha.$$

*Остальные $4^{m+1} - 4^m = 3 \cdot 4^m$ кодовых слов составляют ДНК-код объема $N = 3 \cdot 4^m$, длины $n = 4^m$ с аддитивным стебельным **1**-расстоянием $3 \cdot 4^{m-1}$. При этом если $\boldsymbol{x} = a + b_1 v_1 + \cdots + b_m v_m$ и*

$\widetilde{\widetilde{\boldsymbol{x}}} = a^s + b_1^s v_1 + \cdots + b_m^s v_m$, *то выполняются следующие равенства:*

$$\begin{cases} b_i^s = b_i, \quad i \in [m], \\ a^s = a + \alpha^2 \sum\limits_{i=1}^{m} b_i + 1. \end{cases}$$

Кроме того, для кодов с фиксированным сходством оказывается верно следующее утверждение:

Теорема 2.3: *Пусть $s \in [n-2]$ – некоторое фиксированное положительное число. Если длина n ДНК-кода с аддитивным стебельным $\mathbf{1}$-расстоянием $D = n - s - 1$ удовлетворяет неравенству:*

$$n > \frac{s \cdot (q^2+2)(q^2+1)}{4} + 1 \quad \textit{для четных } s,$$

либо

$$n > \frac{(q^2+2)\left(s\left(q^2+1\right)-1\right)+2q}{4} + 1 \quad \textit{для нечетных } s,$$

то максимальный объем

$$N_{\mathbf{1}}(n, n-s-1) = q^2.$$

В конце главы мы приведем также пример субоптимальной конструкции для ДНК-кодов, основанных на неаддитивном стебельном $\mathbf{1}$-сходстве (то есть неаддитивном стебельном w-сходстве для постоянной весовой функции), из которого будет следовать следующий результат:

Теорема 2.4: *Если $n = 4m$, $m = 1, 3, 5, \ldots$, то*

$$\frac{4^{n-1}+4}{2} \leq N^{(\mathbf{1})}(n, 2) \leq 4^{n-1}.$$

Глава 3 В этой главе рассматриваются неасимптотические (по расстоянию D) задачи для ДНК-кодов, основанных на аддитивном стебельном $\mathbf{1}$-сходстве. Нами будут установлены основные свойства этой функции сходства, и в частности мы покажем, что аддитивное стебельное $\mathbf{1}$-сходство является метрикой в пространстве q-ичных последовательностей длины n, а объем сферы в такой метрике не зависит от ее центра и определяется только ее радиусом.

Определим с помощью рекуррентных соотношений следующие три типа последовательностей:

$$F_q^i(t) \triangleq (q-1)F_q^i(t-1) + (q-1)F_q^i(t-2), \quad t \geq 3, \ i = 1, 2, 3,$$

с начальными условиями:

$$F_q^1(1) \triangleq q, \qquad F_q^2(1) \triangleq q-1, \qquad F_q^3(1) \triangleq q-1,$$
$$F_q^1(2) \triangleq q^2-1; \qquad F_q^2(2) \triangleq (q-1)^2; \qquad F_q^3(2) \triangleq q(q-1).$$

Кроме того, пусть $\mathbf{t}^{(k)} \triangleq (t_1, t_2, \ldots, t_k)$, $k = 1, 2, \ldots$, обозначает упорядоченное множество k целых чисел. Для фиксированных целых s, $1 \le s \le n-1$, и k, $1 \le k \le \min\left\{s\,;\,\left\lceil \frac{n-s}{2} \right\rceil\right\}$, введем множество:

$$T_2(s,k) \triangleq \left\{ \begin{array}{c} \mathbf{t}^{(k+1)} \,:\, t_1 \ge 0,\ t_{k+1} \ge 0, \quad t_i \ge 1,\ i = 2,3,\ldots k, \\ \sum\limits_{i=1}^{k+1} t_i = n-(s+k) \end{array} \right\}.$$

Мы получим следующие явные формулы для объемов сфер

$$\mathbf{S_1}(n,r) \triangleq \left|\left\{ \boldsymbol{y} \in \mathcal{A}_q^n \,:\, \mathcal{D}_\mathbf{1}(\mathbf{0}, \boldsymbol{y}) = r \right\}\right|, \quad 1 \le r \le n-1,$$

где $\mathbf{0} = (0, \ldots, 0) \in \mathcal{A}_q^n$:

Теорема 3.1: *Пусть n – некоторое целое число, $n \ge 2$. Если $r = n-1$ и $n \ge 3$, то $\mathbf{S}_q(n, n-1) = F_q^1(n)$. Если $1 \le s \le n-3$, то*

$$\mathbf{S}_q(n, n-1-s) = \sum_{k=1}^{\min\left\{s\,;\,\left\lceil \frac{n-s}{2} \right\rceil\right\}} \binom{s-1}{k-1} \sum_{T_2(s,k)} \left\{ F_q^3(t_1) \prod_{i=2}^{k} F_q^2(t_i)\, F_q^3(t_{k+1}) \right\}.$$

Отсюда будут выведены неасимптотические оценки для объемов сфер:

Теорема 3.2: *Для любых целых n, $n \ge 2$, и D, $1 \le D \le n-2$,*

$$\sum_{k=1}^{M} \binom{n-D-2}{k-1} \binom{D-k+2}{k} (q-1)^{D-k+1}\, \gamma^{-(k+1)} \times$$
$$\times \left[\left(\frac{\gamma+1}{2}\right)^{D-2k+3} - \left(\frac{\gamma-1}{2}\right)^{D-2k+3} \right] \le$$
$$\le \mathbf{S}_q(n, D) \le$$
$$\le \sum_{k=1}^{M} \binom{n-D-2}{k-1} \binom{D-k+2}{k} (q-1)^{D-k+1}\, \gamma^{-1} \times$$
$$\times \left[\left(\frac{\gamma+1}{2}\right)^{D-2k+3} + \left(\frac{\gamma-1}{2}\right)^{D-2k+3} \right], \quad (5)$$

где

$$M \triangleq \min\left\{n-D-1\,;\,\left\lceil\frac{D+1}{2}\right\rceil\right\},\qquad \gamma \triangleq \sqrt{\frac{q+3}{q-1}},\quad \gamma>1.$$

Асимптотические выражения для объема сферы при растущей длине кодовых слов n и постоянном радиусе D имеют вид:

Теорема 3.3: *Для любого четного положительного D, $D = 2,4,\ldots,$*

$$\mathbf{S}_q(n,D) = \frac{n^{d_1}(q-1)^{d_1}}{d_1!}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty;$$

для любого нечетного положительного D, $D = 1,3,\ldots,$

$$\mathbf{S}_q(n,D) = \frac{(d_1+2)n^{d_1}(q-1)^{d_1+1}}{d_1!}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty,$$

где $d_1 \triangleq \left\lfloor\frac{D}{2}\right\rfloor$.

Теорема 3.3 позволяет получить следующие оценки максимального объема $N_{\mathbf{1}}(n,D)$ ДНК-кодов, основанных на аддитивном стебельном **1**-сходстве, для случая постоянного расстояния D и растущей длины кодовых слов n:

Теорема 3.4: (Граница случайного кодирования) *Для любого нечетного D, $D = 3,5,\ldots,$ максимальный объем $(n,D)_{\mathbf{1}}$-кода удовлетворяет неравенству*

$$N_{\mathbf{1}}(n,D) \geq \frac{q^n d_2!}{4n^{d_2}(q-1)^{d_2}}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty,$$

а для любого четного D, $D = 2,4,\ldots,$ – неравенству

$$N_{\mathbf{1}}(n,D) \geq \frac{q^n d_2!}{4n^{d_2}(q-1)^{d_2}\left[1+(d_2+2)(q-1)\right]}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty,$$

где $d_2 \triangleq \left\lfloor\frac{D-1}{2}\right\rfloor$.

Теорема 3.5: (Граница Хэмминга) *Для любого фиксированного $D = 5,6,9,10,13,14,\ldots,$ максимальный объем $N(n,D)_{\mathbf{1}}$ $(n,D)_{\mathbf{1}}$-кода удовлетворяет неравенству*

$$N_{\mathbf{1}}(n,D) \leq \frac{q^n d_3!}{n^{d_3}(q-1)^{d_3}}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty,$$

а для любого фиксированного $D = 3,4,7,8,11,12,\ldots$ – неравенству

$$N_{\mathbf{1}}(n,D) \leq \frac{q^n d_3!}{n^{d_3}(q-1)^{d_3}\left[1+(d_3+2)(q-1)\right]}\left(1+\overline{\overline{o}}(1)\right),\quad n\to\infty,$$

где $d_3 \triangleq \lfloor \frac{D-1}{4} \rfloor = \lfloor \frac{d_2}{2} \rfloor$.

Теорема 3.6: (Граница Синглтона) *Для любого целого* n, $n > 0$, *и любого* $D \in [1, n-1]$ *максимальный объем* $N_{\mathbf{1}}(n, D)$ $(n, D)_{\mathbf{1}}$*-кода удовлетворяет неравенству*

$$N_{\mathbf{1}}(n, D) \leq q^{n-(D-1)}.$$

Глава 4 В данной главе изучаются границы скорости ДНК-кодов, основанных на аддитивном стебельном **1**-сходстве. Применяя описанный в первой главе диссертации метод случайного кодирования для ДНК-кодов мы находим следующий аналог классической границы Варшамова-Гилберта:

Теорема 4.2: (Граница случайного кодирования). 1). *Если* $0 < d \leq (q^2-1)/q^2$, *то скорость*

$$R_{\mathbf{1}}(d) \geq L_{\mathbf{1}}(d) \triangleq \max_{u \leq 0}\{ud - \mu_{\mathbf{1}}(u)\}, \qquad \mu_{\mathbf{1}}(u) \triangleq \log_q \lambda_{\mathbf{1}}(u), \tag{6}$$

$$\lambda_{\mathbf{1}}(u) \triangleq \frac{1 + (q-1)q^u + \sqrt{[1+(q-1)q^u]^2 - 4(q-1)q^u(1-q^u)}}{2q}. \tag{7}$$

2). *Нижняя граница* $L_{\mathbf{1}}(d) > 0$ *при* $0 \leq d < (q^2-1)/q^2$. *Кроме того,* $L_{\mathbf{1}}(0) = 1$ *и* $L_{\mathbf{1}}\left(\frac{q^2-1}{q^2}\right) = 0$. *При этом* $L_{\mathbf{1}}(d)$ *является убывающей* $\bigcup$*-выпуклой функцией и задается параметрическими уравнениями*

$$L_{\mathbf{1}}(d) = u\mu'_{\mathbf{1}}(u) - \mu_{\mathbf{1}}(u), \qquad d = \mu'_{\mathbf{1}}(u),$$

где $\mu'_{\mathbf{1}}(u)$, $u \leq 0$, *обозначает производную функции* $\mu_{\mathbf{1}}(u)$, *определяемой* (6)-(7).

Верхние границы представлены следующими аналогами классических теорем Плоткина, Хэмминга и Элайеса:

Теорема 4.1: (Граница Плоткина). *Скорость*

$$R_{\mathbf{1}}(d) \leq \mathcal{P}_{\mathbf{1}}(d) \triangleq 1 - \frac{d\,q^2}{q^2-1}, \quad \textit{если} \quad 0 < d \leq 1 - 1/q^2.$$

Теорема 4.3: (Граница Хэмминга). *Если* $0 < d < \frac{2(q^2-1)}{q^2}$, *то скорость* $R_{\mathbf{1}}(d)$ *ДНК* $(n, dn)_{\mathbf{1}}$*-кодов удовлетворяет неравенству*

$$R_{\mathbf{1}}(d) \leq L_{\mathbf{1}}(d/2),$$

где $L_{\mathbf{1}}(d)$ определено в (6).

Теорема 4.4: (Граница Элайеса). *Если* $0 < d \leq \frac{q^2-1}{q^2}$, *то скорость*

$$R_{\mathbf{1}}(d) \leq E_{\mathbf{1}}(d),$$

где верхняя граница $E_{\mathbf{1}}(d)$ *задается при* $u \leq 0$ *и* $0 < d \leq (q^2-1)/q^2$ *параметрическими уравнениями*

$$E_{\mathbf{1}}(d) = u\mu_{\mathbf{1}}'(u) - \mu_{\mathbf{1}}(u), \quad d = \mu_{\mathbf{1}}'(u)\left[2 - \mu_{\mathbf{1}}'(u)\frac{q^2}{q^2-1}\right],$$

в которых функция $\mu_{\mathbf{1}}(u)$, *определена в (6)-(7).*

Глава 5 В этой части диссертации изучается асимптотическое поведение максимального объема ДНК-кодов, основанных на аддитивном стебельном w-сходстве, для случая произвольной весовой функции $w(a,b)$, $a,b \in \mathcal{A}_q$.

Верхняя граница скорости $R_w(d)$ дается аналогом классической границы Плоткина. Пусть $\mathsf{p} \triangleq \{\, p(a,b),\ a,b \in \mathcal{A}_q\}$ – произвольное *совместное* распределение вероятностей на множестве упорядоченных пар $(a,b) \in \mathcal{A}_q^2$, т.е.

$$\sum_{a,b \in \mathcal{A}_q} p(a,b) = 1, \qquad p(a,b) \geq 0 \quad \text{для любых} \quad a,b \in \mathcal{A}_q,$$

а символы

$$p_1(a) \triangleq \sum_{b \in \mathcal{A}_q} p(a,b) > 0, \qquad p_1(b|a) \triangleq \frac{p(a,b)}{p_1(a)},$$

$$p_2(a) \triangleq \sum_{b \in \mathcal{A}_q} p(b,a) > 0, \qquad p_2(b|a) \triangleq \frac{p(b,a)}{p_2(a)}$$

обозначают соответствующие *маргинальные* и *условные* вероятности. При описании границ скорости $R_w(d)$ будем рассматривать распределения p, для которых совпадают маргинальные вероятности, т.е. для любого $a \in \mathcal{A}_q$

$$p_1(a) = \sum_{b \in \mathcal{A}_q} p(a,b) \;=\; \sum_{b \in \mathcal{A}_q} p(b,a) \;=\; p_2(a) > 0 \tag{8}$$

и, кроме того, функция $p(a,b)$, как и весовая функция $w(a,b)$, является инвариантной относительно преобразования Ватсона-Крика, т.е.

$$p(a,b) = p(\overline{b},\overline{a}) \quad \text{для любых} \quad a,b \in \mathcal{A}_q. \tag{9}$$

Для фиксированной весовой функции $w(a,b)$ введем величины

$$T_w \triangleq \max_{(8)\text{-}(9)} T_w(\mathsf{p}), \quad T_w(\mathsf{p}) \triangleq \sum_{a,b\in\mathcal{A}_q} \big(p(a,b) - p^2(a,b)\big)\, w(a,b). \qquad (10)$$

В конце главы 5 мы описываем итерационный метод решения выпуклой задачи максимизации (8)-(10), который можно применять при вычислении T_w для функций аддитивного стебельного w-сходства для приведенных в главе образцов весовой функции. Там же будет показано, что определение числа T_w равносильно определению

$$T_w \triangleq \max_{(8)} T_w(\mathsf{p}),$$

поскольку для любой весовой функции $w(a,b)$, инвариантной относительно преобразования Ватсона-Крика, экстремальное распределение p, получаемое в качестве решения последней задачи максимизации, удовлетворяет также и условию (9).

Теорема 5.1: (Граница Плоткина). 1). *Если* $d > T_w$, *то* $R_w(d) = 0$. 2). *Скорость*

$$R_w(d) \le \mathcal{P}_w(d) \triangleq 1 - \frac{d}{T_w}, \qquad 0 < d \le T_w.$$

На множестве упорядоченных пар $(\mathbf{a},\mathbf{b})$, $\mathbf{a} \triangleq (a_1,a_2,a_3,a_4) \in \mathcal{A}_q^4$, $\mathbf{b} \triangleq (b_1,b_2,b_3,b_4) \in \mathcal{A}_q^4$, определим $(q^4\times q^4)$-матрицу $\mathbf{P}_\xi = ||P(\mathbf{b}|\mathbf{a})||$ с элементами

$$P(\mathbf{b}|\mathbf{a}) = P\left((b_1,b_2,b_3,b_4)\,|\,(a_1,a_2,a_3,a_4)\right) \triangleq$$

$$\triangleq \begin{cases} p_1(b_2|b_1)p_1(b_4|b_3), & \text{если } b_1 = a_2 \text{ и } b_3 = a_4, \\ 0, & \text{если } b_1 \ne a_2 \text{ или } b_3 \ne a_4. \end{cases} \qquad (11)$$

По заданным значениям весовой функции $w(a,b)$, $a,b \in \mathcal{A}_q$, на множестве упорядоченных четверок $\mathbf{b} \in \mathcal{A}_q^4$ определим функцию

$$f_w(\mathbf{b}) = f_w(b_1,b_2,b_3,b_4) \triangleq \begin{cases} 0, & \text{если } b_1 = b_3,\ b_2 = b_4, \\ w(b_1,b_2), & \text{иначе.} \end{cases} \qquad (12)$$

Для матрицы (11) и функции (12) введем $(q^4\times q^4)$-матрицу Маркова

$$\mathbf{M}_w(u,\mathsf{p}) \triangleq \left|\left| P(\mathbf{b}|\mathbf{a})\, q^{u f_w(\mathbf{b})} \right|\right|, \quad \mathbf{a}\in\mathcal{A}_q^4, \quad \mathbf{b}\in\mathcal{A}_q^4, \qquad u\in\mathbb{R}. \qquad (13)$$

Пусть для распределения p, определяемого условиями (8)-(9), дополнительно известно, что матрица переходных вероятностей (11) задает

цепь Маркова, удовлетворяющую *условию Маркова* $\mathcal{M}$, а именно, что для любой пары состояний $\mathbf{a}, \mathbf{b} \in \mathcal{A}_q^4$ найдется такое целое $m > 0$, что $\mathbf{P}_\xi^m(\mathbf{a}, \mathbf{b}) > 0$, то есть за m шагов можно перейти из состояния $\mathbf{a}$ в состояние $\mathbf{b}$.

Символом $\lambda_w(u, \mathsf{p})$, $u \in \mathbb{R}$, обозначим максимальное собственное значение матрицы (13), т.е. такое положительное собственное значение, что для любого другого ее собственного значения $\hat{\lambda}$ абсолютная величина $|\hat{\lambda}| < \lambda_w(u, \mathsf{p})$. Число $\lambda_w(u, \mathsf{p})$ существует согласно теореме Перрона-Фробениуса.

Пусть величина $T_w^{\mathcal{M}}$ определяется равенством:

$$T_w^{\mathcal{M}} \triangleq \max_{(8)\text{-}(9),\,\mathcal{M}} T_w(\mathsf{p}), \tag{14}$$

где $T_w(\mathsf{p})$ определены в (10). Нижнюю границу скорости $R_w(d)$ дает

Теорема 5.2: *Для любого распределения вероятностей* p*, удовлетворяющего условиям* (8)-(9) *и условию Маркова* $\mathcal{M}$*, и любой доли расстояния* d*,* $0 < d \le T_w(\mathsf{p})$*, скорость*

$$R_w(d) \ge L_w(d, \mathsf{p}) \triangleq \inf_{0 \le z \le d} \sup_{u \in \mathbb{R}} \{uz - \mu_w(u, \mathsf{p})\},$$

$$\mu_w(u, \mathsf{p}) \triangleq \log_q \lambda_w(u, \mathsf{p}),$$

где функция $L_w(d, \mathsf{p}) > 0$ *для любого* d*,* $0 < d < T_w(\mathsf{p})$*, и значение* $L_w\left(T_w(\mathsf{p}), \mathsf{p}\right) = 0$.

Очевидно, теорема 5.2 означает, что справедливо

Следствие 5.1: (Граница Варшамова-Гилберта). *Для любого* $d > 0$ *скорость*

$$R_w(d) \ge \underline{R}_w(d) \triangleq \max_{(8)\text{-}(9),\mathcal{M}} L_w(d, \mathsf{p}).$$

Если $0 < d < T_w^{\mathcal{M}}$*, то нижняя граница* $\underline{R}_w(d) > 0$*, т.е. для любого* d*,* $0 < d < T_w^{\mathcal{M}}$*, скорость* $R_w(d) > 0$.

Если величины T_w и $T_w^{\mathcal{M}}$ совпадают, то соответствующая весовая функция называется *регулярной*, и *нерегулярной* в противном случае. Из теоремы 5.1 и следствия 5.1 вытекает основной результат главы 5 о *критической* доле расстояния $(n, dn)_w$-кодов для регулярных весовых функций:

Следствие 5.2: *Пусть* $w(a, b)$*,* $a, b \in \mathcal{A}_q$*, – регулярная весовая функция. При* $0 < d < T_w$ *скорость* $R_w(d) > 0$*. Если* $d \ge T_w$*, то* $R_w(d) \equiv 0$*. Другими словами, максимальный объем* $(n, dn)_w$*-кодов*

возрастает экспоненциально с ростом n тогда и только тогда, когда $0 < d < T_w$.

Далее также мы дадим дополнительные образцы весовых функций и проведем их сравнительный анализ на основе результатов случайного кодирования для каждого весового образца.

Глава 6 В последней главе работы мы получим нижнюю границу скорости четверичных ДНК-кодов, основанных на неаддитивном стебельном w-сходстве для случая произвольной весовой функции $w(a, b)$, $a, b \in \mathcal{A}_4$.

Нам потребуются специальные классы $(n, D)^{(w)}$-кодов, называемые *L-ансамблями Фибоначчи.* Символ L будет обозначать любое подмножество $L \subset \mathcal{A}_4^2$ 2-блоков букв ДНК-алфавита $\mathcal{A}_4$, *замкнутое относительно преобразования сопряжения.* Так, например, $L = L_k$, $k = 0, 1, 2, 3, 4, 6, 8$, где

$$L_0 \triangleq \emptyset, \quad L_1 \triangleq \{TA\}, \quad L_2 \triangleq \{TA, AT\},$$
$$L_4 \triangleq \{TA, AT, AA, TT\},$$

и

$$L_3 \triangleq \{TA, AA, TT\}, \quad L_6 \triangleq \{TA, AT, AA, TT, AG, CT\},$$
$$L_8 \triangleq \{TA, AT, AA, TT, AG, CT, GA, TC\}.$$

Пусть $DNA(n, L)$ (коротко, $[n, L]$) обозначает множество (ансамбль) всех ДНК-последовательностей длины n, которые *не содержат стеблей из L*. Такое множество $[n, L]$, очевидно, замкнутое относительно преобразования сопряжения, называется *L-ансамблем Фибоначчи.* Обозначим через $\lambda_L(n) \triangleq |[n, L]|$ *объем* множества $[n, L]$.

Определение 6.8: Пусть $N_L(n, D)$ обозначает *максимальный объем* ДНК $(n, D)^{(1)}$-кодов $\mathcal{X} \subseteq DNA(n, L)$ для постоянной весовой функции $w(a, b) = 1$ для любых $a, b \in \mathcal{A}_4$. Если доля расстояния $d > 0$ есть некоторое фиксированное число, то

$$R_L(d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_4 N_L(n, dn)}{n}$$

называется *скоростью* ДНК кодов для L-ансамбля Фибоначчи.

В главе 6 для множеств L_1, L_2 и L_4 будут найдены верхние оценки следующего типа (для множеств L_3, L_6 и L_8 верны аналогичные оценки более сложного вида, как показано в конце главы):

$$\lambda_{L_i}(n) \leq C\, r^n \left[1 + \omega\, \alpha^n\right], \quad i = 1, 2, 4.$$

Пусть

$$\rho_L \triangleq \log_4 r, \quad \rho'_L \triangleq \log_4 \frac{r}{C^3(1+\omega\,\alpha^2)(1+\omega\,\alpha)^2}.$$

Для L-ансамблей Фибоначчи мы получим следующую нижнюю границу скорости $R_L(d)$:

Теорема 6.1: *Для любой доли расстояния $0 < d < d_L$, скорость $R_L(d)$ удовлетворяет неравенству*

$$R_L(d) \geq \underline{R}_L(d) \triangleq (1-d)\rho_L - E_L(d) > 0,$$

где

$$\begin{aligned} E_L(u) &\triangleq \max_{0\leq v\leq \min\{u,\,1-u\}} E^L(v,u), \\ E^L(v,u) &\triangleq -\rho'_L \cdot v + (1-u)\,h_4\left(\frac{v}{1-u}\right) + 2\,u\,h_4\left(\frac{v}{u}\right), \\ h_4(u) &\triangleq -u\log_4 u - (1-u)\log_4(1-u) \end{aligned}$$

и d_L, $0 < d_L \leq 1$, является единственным на интервале $(0\,;1)$ корнем уравнения $\underline{R}_L(d) = 0$, или $(1-d)\rho_L = E_L(d)$.

Пусть

$$d(w) \triangleq \max_L \{\underline{w}_L \cdot d_L\}, \quad \underline{w}_L \triangleq \min_{(a,b)\notin L} w(a,b).$$

Применение ансамблей Фибоначчи для случайного кодирования вкупе с результатом теоремы 6.1 позволяет существенно улучшить нижнюю границу скорости $R^{(w)}(d)$ ДНК-кодов, основанных на неаддитивном стебельном w-сходстве, что отражено в

Теорема 6.2: *Если $0 < d < d(w)$, то скорость ДНК $(n, dn)^{(w)}$-кодов $R^{(w)}(d) > 0$ и верна нижняя граница*

$$R^{(w)}(d) \geq \underline{R}^{(w)}(d) \triangleq \max_L \left\{\underline{R}_L\left(\frac{d}{\underline{w}_L}\right)\right\}, \qquad 0 < d < d(w).$$

Введенные в главе 5 весовые образцы будут классифицированы в зависимости от того, какой ансамбль Фибоначчи позволяет получить экспоненциально растущие коды для наибольшего интервала значений доли расстояния d для данного образца.

Автор выражает глубокую благодарность своему научному руководителю, доктору физико–математических наук, профессору Аркадию Георгиевичу Дьячкову за постановку задач, постоянное внимание и помощь в работе.

Работы автора по теме диссертации

[1] А. Г. Дьячков, А. Н. Воронина, "ДНК-коды для аддитивного стебельного сходства," *Проблемы Передачи Информации*, т. 45, н. 2, стр. 56–77, 2009.

Дьячкову А. Г. принадлежит метод случайного кодирования для ДНК-кодов (неравенство (48)) и доказательство теоремы 3. Остальные результаты принадлежат Ворониной А. Н.

[2] А. Н. Воронина, "Об объёмах сфер для стебельного расстояния," *Проблемы Передачи Информации*, т. 46, н. 1, стр. 9–19, 2010.

[3] A. G. D'yachkov, A. N. Voronina, "DNA Codes Based on Stem Hamming Similarity," *Proc. 11th Int. Workshop Algebraic and Combinatorial Coding Theory*, Пампорово, Болгария, 2008, стр. 85–91.

Дьячкову А. Г. принадлежит метод случайного кодирования для ДНК-кодов (предложение 1). Остальные результаты принадлежат Ворониной А. Н.

[4] А. Г. Дьячков, А. Н. Воронина, "ДНК-коды, основанные на весовом стебельном сходстве Хэмминга," *Сборник трудов ИТиС'08*, Геленджик, Россия, 2008, стр. 316–320.

Дьячкову А. Г. принадлежит метод случайного кодирования для ДНК-кодов (предложение 1). Остальные результаты, в том числе численные методы и алгоритм нахождения максимизирующих распределений и границ случайного кодирования для различных распределений, принадлежат Ворониной А. Н.