На правах рукописи

Шеленков Андрей Александрович

РАЗРАБОТКА АЛГОРИТМОВ И ПРОГРАММ ДЛЯ ИЗУЧЕНИЯ РЕГУЛЯРНОГО СТРОЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

05.13.18 – математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2008

Работа выполнена в Центре «Биоинженерия» Российской академии наук

Научный руководитель: доктор биологических наук, профессор
Коротков Евгений Вадимович

Официальные оппоненты: кандидат физико-математических наук
Андреев Сергей Григорьевич,
Институт биохимической физики РАН

доктор биологических наук
Лисица Андрей Валерьевич,
ГУ НИИ биомедицинской химии им. В.Н.
Ореховича РАМН

Ведущая организация: Институт биоорганической химии им. академиков
М.М. Шемякина и Ю.А. Овчинникова РАН

Защита диссертации состоится «*26*» *ноября* 2008 г. в *15* часов на заседании Диссертационного совета Д 212.130.09 в Московском инженерно-физическом институте (государственном университете) по адресу: 115409, Москва, Каширское шоссе, 31, тел (495) 324-84-98, (495) 323-92-56

С диссертацией можно ознакомиться в библиотеке Московского инженерно-физического института (государственного университета)

Просим принять участие в работе совета или прислать отзыв в одном экземпляре, заверенный печатью организации.

Автореферат разослан «*23*» *октября* 2008 г.

Ученый секретарь Диссертационного совета

доктор физико-математических наук, профессор Леонов А.С

Общая характеристика работы

Предметом исследований диссертационной работы являются структурные свойства последовательностей оснований нуклеиновых кислот (ДНК) и их связь с функциональной значимостью этих последовательностей. Основное внимание уделялось обнаружению последовательностей, имеющих регулярное строение, в частности, последовательностей, обладающих скрытой периодичностью и регулярностью других типов.

Актуальность работы

В конце XX века с появлением новых технических средств такие области науки, как молекулярная биология и генетика, вышли на совершенно новый уровень. Рост объемов получаемых биологических данных, в частности, последовательностей геномов различных организмов, приобрел экспоненциальный характер. С наступлением нового века эта тенденция сохранилась. Основным носителем наследственной информации являются молекулы дезоксирибонуклеиновой кислоты (ДНК), представляющих собой двойную спираль, состоящую из двух цепочек азотистых оснований – нуклеотидов. В молекулах ДНК присутствует четыре типа нуклеотидов, обозначаемых буквами A, T, C и G. Объем наиболее известного банка данных последовательностей ДНК - Genbank - превышает 85 млрд. нуклеотидов.

Однако, определение последовательности генома является лишь первым шагом на пути к пониманию принципов функционирования генетического аппарата. В настоящее время достоверно известна биологическая роль лишь небольшого числа участков геномов различных организмов. Экспериментальные методы проведения аннотации (выявления функциональной значимости) требуют значительных затрат времени и ресурсов, кроме того, в ряде случаев число рассматриваемых вариантов взаимодействия функциональных элементов настолько велико, что экспериментальное исследование становится невозможным. В связи с этим, на первый план выступают математические методы анализа генетических последовательностей, которые позволяют эффективно использовать значительные вычислительные мощности, применяемые в настоящее время для подобных исследований. Таким образом, современная биология превращается из описательной науки в вычислительную, что ознаменовалось появлением биоинформатики как совокупности математических методов, алгоритмов и программного обеспечения, предназначенных для анализа биологических данных. В настоящее время биоинформатика является главным научным направлением во многих мировых научных центрах, а появление новых методов в этой области неизменно вызывает широкий резонанс в среде ученых-экспериментаторов. Несомненно, компьютерные методы не могут полностью заменить эксперименты, однако, полученные теоретически результаты способны значительно сократить объемы необходимых лабораторных опытов, а в ряде случаев могут способствовать выявлению общих закономерностей, ускользающих от внимания экспериментаторов

Одной из важнейших задач аннотации является предсказание генов – участков ДНК, кодирующих белок, а также предсказание функций, выполняемых этим белком. Однако, в геномах высокоразвитых организмов, таких как растения, насекомые и млекопитающие, доля кодирующих последовательностей в геноме составляет не более 10%. Экспериментальные исследования показали, что в некодирующих областях располагаются участки, принципиальным образом влияющие на активность генов и саму возможность их правильного функционирования К числу таких участков относятся промоторы – важнейшие регуляторные элементы. Кроме того, некодирующие области генома также содержат большое число повторяющихся последовательностей с различной длиной периода [1] Несмотря на то, что такие последовательности на первый взгляд представляются бесполезными, они также играют определенную роль в функционировании генетического аппарата, в том числе, в обеспечении эволюционной гибкости вида, то есть, его способности реагировать на изменяющиеся внешние условия [1] Кроме того, мутационное изменение общей длины микросателлитной последовательности в некоторых случаях может быть связано с серьезными заболеваниями. В работе [2] было показано, что при наличии большого числа микросателлитов вида $(CAG)_n$ (более 22) в гене андрогенового рецептора возрастает риск возникновения рака простаты, тогда как при числе повторов менее 20 риск значительно снижается. Таким образом, изменение числа повторяющихся элементов всего на две единицы может говорить о наличии заболевания, что делает обнаружение и анализ микросателлитов важным диагностическим инструментом.

Предсказание функциональной значимости участка ДНК естественным образом предполагает выявление общих структурных свойств последовательностей, характерных для определенных элементов генома (гены, промоторы, повторы и т.д.). В качестве характеристического свойства может выступать периодичность. Для обнаружения периодичности в последовательностях ДНК было разработано большое число методов, использующих различные математические алгоритмы, такие как преобразование Фурье [3, 4], динамическое программирование [5], исследование статистических свойств распределений символов [6], информационные подходы [7] и другие алгоритмы (например, [8]). Однако у всех ранее разработанных алгоритмов есть достаточно серьезные ограничения по выявлению периодичности в нуклеотидных последовательностях.

Основным недостатком использования преобразования Фурье при поиске периодичности в символьных последовательностях является необходимость перекодировки символьной последовательности в числовую. Эту перекодировку можно рассматривать как введение разных весов для равноправных символов, что в конечном итоге может приводить к невозможности обнаружения некоторых типов периодичности при использовании преобразования Фурье [7]. Кроме того, такие методы не способны обнаруживать периодичность при наличии вставок и делеций и они не дают возможность получить матрицу или некоторую другую характеристику типа периодичности, которая могла бы использоваться в дальнейших вычислениях.

При использовании динамического программирования и некоторых других подходов серьезным ограничением для выявления периодичности является поиск идентичных совпадений символов между последовательностями при выявлении повторов. Под идентичными совпадениями понимаются совпадения вида $s(t)s(t)$, $t=1,\ldots,h$, где $s(i)$ – символ алфавита последовательности, h – размер алфавита символьной последовательности. В случае динамического программирования поиск преимущественно идентичных повторов задается при помощи весовой матрицы совпадений символов, для нуклеотидных последовательностей - это матрица идентичности (Identity matrix) или подобные ей матрицы. В этих матрицах веса идентичных совпадений (aa, tt,cc,gg для нуклеотидной последовательности) значительно выше, чем веса всех других видов парных совпадений. Это приводит к тому, что сильно размытые повторяющиеся последовательности, которые можно обнаружить на статистически значимом уровне только при наличии в последовательности многих периодов (>2), не могут быть выявлены этими методами [7]. Образование таких последовательностей в реальной ДНК может происходить посредством множественных замен оснований, а также путем образования делеций и вставок символов.

Программы поиска тандемных повторов в геномных последовательностях, представленные в пакете EMBOSS [9], находят только те повторы, которые принадлежат к ограниченному множеству возможных типов периодичности (некоторые микросателлиты). Некоторые алгоритмы демонстрируют сильную чувствительность к наличию вставок и делеций, таким образом, они могут обнаруживать только повторы, подчиняющиеся очень строгим правилам [8, 10].

Таким образом, ни один из существующих на данный момент методов поиска периодичности в последовательностях ДНК не может претендовать на универсальность.

Поиск периодичности в промоторах является намного более сложной задачей, чем поиск микросателлитов. На данный момент не выявлено теоретических или экспериментальных предпосылок к тому, что промоторы обладают периодической структурой. Таким образом, исследование последовательностей промоторов в основном сводится к выявлению некоторого «сигнала» (консервативного участка, определенного нуклеотидного состава и т.п.), который позволил бы разделить участки ДНК с неизвестной функцией на гипотетические промоторные последовательности и участки, наличие промоторов в которых маловероятно.

Однако, анализ экспериментальных данных показал [11], что вопрос выбора правильных биологических сигналов, используемых в программах предсказания промоторов, все еще остается открытым. Ни один из использованных сигналов не описывает все разнообразие промоторов, и каждый признак, полученный на основе изучения промоторных последовательностей, имеет свои ограничения в использовании [12]. Таким образом, существует необходимость поиска некоторой новой характеристики последовательностей промоторов, которая являлась бы специфичной по отношению к этим элементам, но при

этом обладала бы достаточной гибкостью для того, чтобы соответствовать многообразию видов таких последовательностей.

Цель и задачи исследования

Целью представленной работы является разработка и программная реализация алгоритмов выявления регулярных структур в последовательностях ДНК, способных обнаруживать периодичность и регулярность, сильно размытые в ходе эволюционного процесса и по этой причине не обнаруживаемые существующими методами поиска периодичности.

Методы исследования

Предлагается использовать комбинированные алгоритмы для выявления регулярности строения генетических последовательностей, включающие применение как статистических методов (информационное разложение, критерий серий), так и методы динамического программирования (профильный анализ). Под регулярностью понимается статистически значимое отклонение распределения символов на участке последовательности от ожидаемого для случайной последовательности с тем же символьным составом. Периодичность является частным случаем регулярности, наиболее ярко ее иллюстрирующим.

Основными задачами диссертационной работы являются:

1. Разработка и программная реализация алгоритма классификации скрытой периодичности, обнаруженной в последовательностях ДНК с помощью информационного разложения (ИР); классификация скрытой периодичности из банка данных Genbank на основании частотных матриц периодичности.

2. Выявление сильно размытой периодичности с использованием полученных классов методом модифицированного профильного анализа (МПА) в различных геномах; выявление функциональной значимости обнаруженной периодичности

3. Создание базы данных по потенциальным мини- и микросателлитным последовательностям ДНК на основе результатов поиска скрытой периодичности в геномах различных организмов.

4. Разработка Интернет-сервера для поиска скрытой периодичности, реализующего метод МПА.

5. Разработка и программная реализация алгоритма выявления регулярности последовательностей ДНК, основанного на использовании критерия серий.

6. Применение разработанного алгоритма для выявления регулярности в последовательностях промоторов из различных геномов.

Научная новизна работы:

1. Разработан новый алгоритм классификации скрытой периодичности в

последовательностях ДНК

2. Предложен новый алгоритм выявления скрытой периодичности, сочетающий в себе преимущества трех математических методов: расширенного подобия, весовых функций и динамического программирования

3 Разработана база данных, содержащая около 3 млн. последовательностей, обладающих скрытой периодичностью с периодом 2-100, для всех групп организмов.

4. Разработан веб-сервер для поиска скрытой периодичности, реализующий новый алгоритм выявления скрытой периодичности.

5. Введено понятие регулярности последовательности, расширяющее и дополняющее понятие скрытой периодичности и разработан алгоритм выявления регулярных последовательностей ДНК.

Достоверность результатов работы подтверждена проведенными исследованиями нуклеотидных последовательностей из банков данных Genbank и EPD и сравнением с экспериментальными данными.

Апробация работы

Основные результаты и положения диссертации докладывались и обсуждались на международных конференциях «Биология – наука XXI века», Пущино, в 2004 (17-21 мая) и 2005 (18-22 апреля) гг., Bioinformatics of genome regulation and structure (Новосибирск, 25-30 июля 2004 г.), I и II международной конференции «Математическая биология и биоинформатика» (Пущино, 9-15 октября 2006 г. и 7-13 сентября 2008 г.), российско-французском научном симпозиуме по аннотации бактериальных геномов (Тулуза, Франция, 5-6 октября 2006), на международной школе-конференции молодых ученых «Системная биология и биоинженерия» (Звенигород, 28 ноября – 2 декабря 2005 г.), а также на ежегодных конкурсах-конференциях аспирантов и сотрудников Центра «Биоинженерия» РАН в 2004-2008 годах.

Основные результаты диссертации опубликованы в 12 работах: 4 статьях в рецензируемых отечественных и зарубежных научных журналах, 7 сборниках материалов научных конференций и 1 монографии (в период с 2004 по 2008 гг.)

Практическая значимость работы

Практическая значимость работы заключается в разработке и программной реализации алгоритмов выявления и классификации скрытой периодичности и сильнодивергированных повторов в нуклеотидных последовательностях. Результаты, полученные в ходе изучения реальных последовательностей с помощью разработанных алгоритмов, имеют глубокий биологический смысл и несомненное значение для последующего развития методик анализа последовательностей ДНК, позволяющих существенным образом сократить объемы необходимых экспериментальных исследований.

На защиту выносятся:

1. Алгоритм эффективной классификации скрытой периодичности в последовательностях оснований нуклеиновых кислот.
2. Алгоритм поиска сильно размытой периодичности в условиях наличия вставок и делеций символов с использованием классов периодичности.
3. База данных по потенциальным мини- и микросателлитным последовательностям ДНК.
4. Веб-сервер для выявления скрытой периодичности.
5. Алгоритм выявления регулярности последовательностей ДНК.
6. Результаты поиска регулярных последовательностей в геномах различных организмов.

Структура и объем диссертации

Диссертация состоит из введения, трех глав, заключения и списка литературы из 114 наименований. Общий объем диссертации составляет 108 страниц; диссертация содержит 21 рисунок и 13 таблиц.

Краткое содержание работы

В **первой главе** приведены общие сведения по структурной организации молекул генетического аппарата, рассматриваются основные математические методы поиска периодичности в последовательностях ДНК, а также их программная реализация в виде веб-серверов и полученные с их использованием результаты, представленные в общедоступных базах данных. Также рассмотрены существующие методы классификации периодических последовательностей ДНК.

В **разделе 1.1** рассмотрены основные молекулы, участвующие в передаче наследственной информации. Приведен краткий обзор их структуры и функций, а также описано функционирование генома в целом. Особое внимание уделено биологической роли повторяющихся последовательностей ДНК (в частности, микросателлитов) и регуляторных элементов – промоторов.

В **разделе 1.2** рассмотрены и критически проанализированы основные существующие методы поиска периодических последовательностей в целом и микросателлитов в частности. Рассмотрены методы, основанные на использовании статистических закономерностей, преобразования Фурье, динамического программирования и профильного анализа, а также комбинированных подходов Приведены достоинства и недостатки каждой группы методов, а также рассмотрены условия, в которых их применение является целесообразным. Сделан вывод об отсутствии в настоящее время универсального метода поиска периодичности

В **разделе 1.3** приведено описание существующих методов классификации периодических последовательностей ДНК. Следует отметить, что подавляющее большинство методов классификации периодических последовательностей предназначено для разделения повторов по длине или числу повторяющихся элементов. При этом не

делается попыток определить классы последовательностей на основании более сложных особенностей повторяющихся элементов, возможно, являющихся общими для геномов различных групп организмов.

В **разделе 1.4** приведены сведения по общедоступным базам данных тандемных повторов и микросателлитных последовательностей. Указаны методы, использованные при создании баз данных, а также адреса в сети Интернет, по которым эти базы доступны в настоящее время.

В **разделе 1.5** рассмотрены веб-сайты поиска периодичности в последовательностях ДНК. Указаны используемые при их создании методы и кратко рассмотрены их достоинства и недостатки.

В **разделе 1.6** приведены имеющиеся на момент написания работы экспериментальные данные по периодичности в геномах бактерий и растений, а также в промоторных участках ДНК. Эти сведения необходимы для проверки адекватности применения разрабатываемых математических методов.

Во **второй главе** рассматриваются алгоритмы, разработанные и использованные в ходе выполнения диссертационной работы. Приводится обоснование оценки статистической значимости получаемых результатов, а также представлены результаты численного моделирования, проведенного для получения пороговых значений статистик критерия.

Разработка алгоритмов поиска регулярности в последовательностях ДНК состоит из двух частей. Первая часть работы связана с выявлением потенциальных микросателлитных последовательностей в геномах различных организмов. В целях выявления общих свойств повторяющихся последовательностей проведена классификация периодичности с использованием разработанных алгоритмов, в результате чего были получены матрицы классов для различных групп организмов и длин периода. Далее полученные классы использовались для поиска сильно размытой периодичности в условиях возможного присутствия вставок и делеций символов. Поиск проводился с помощью нового разработанного алгоритма – модифицированного профильного анализа (МПА), основанного на динамическом программировании и применении весовых функций. Использование комбинации информационного разложения и модифицированного профильного анализа позволяет обнаружить периодичность, которая не выявлялась ни одним из существующих алгоритмов поиска повторяющихся последовательностей.

Вторая часть работы заключалась в поиске последовательностей, обладающих менее четкой структурной организацией, но при этом также имеющих регулярное строение. Предварительно полученные данные позволяли сделать предположение о наличии регулярных последовательностей в важных функциональных элементах ДНК – промоторах. Для обнаружения таких последовательностей был разработан и программно реализован новый алгоритм, основанный на использовании критерия серий.

В **разделе 2.1** приведено описание алгоритма поиска регулярных последовательностей, применявшегося в работе для анализа промоторов.

Назовем «регулярными последовательностями» такие последовательности ДНК, которые имеют близкое или одинаковое количество нуклеотидов в каждом периоде при одинаковом распределении нуклеотидов по длине каждого периода. Под периодами длины *n* понимаются расположенные друг за другом фрагменты ДНК одинаковой длины Пусть нам дана последовательность $S=s_1s_2 \ldots s_L$, которая содержит символы из алфавита $Q=\{a,t,c,g\}$. Периоды в ней можно получить, если последовательность ДНК разбить на чередующиеся друг за другом фрагменты длины *n*. Вначале символ *F* ставится на нулевую позицию и на *L+1* позицию последовательности *S*. Затем отсчитывается *n* нуклеотидов, начиная с первой позиции последовательности *S*, после чего ставится еще один символ *F*, и так до конца последовательности *S*. В результате получаем последовательность $S'=Fs_1s_2 \ldots s_nFs_{n+1}s_{n+2}\ldots s_{2n}F\ldots\ldots s_LF$. Затем формируется последовательность категорий посредством замены символа *F* на ноль, а других символов - на единицу, и подсчитывается число серий. Под серией понимается последовательность символов одного типа, ограниченная символами другого типа, либо началом или концом последовательности. Например, последовательность 000111 содержит две серии, как и последовательность 01.

Если число нуклеотидов в каждом периоде (последовательности между символами *F*) будет одинаковым или близким для каждого типа нуклеотидов, и одновременно с этим нуклеотиды будут наблюдаться примерно в одних и тех же позициях каждого периода, то будем говорить о том, что последовательность *S* обладает регулярностью. Фактически, наличие регулярности означает, что распределение некоторого нуклеотида по последовательности такое же, как и распределение символа *F*. Для определения количественной меры регулярности последовательности ДНК используется критерий серий.

Пример расчета числа серий приведен на рис. 1, где последовательность ДНК разбивается на периоды длиной 3 при помощи символа *F*.

А. Последовательность aatccttttctacccttcaaaaac

```
  1   2   3   4   5   6   7   8    Номера периодов
FaatFcctFtttFctaFcccFttcFaaaFaacF  ДНК последовательность
011 0   0   0  1    0   0111011 0  последовательность категорий
0110001001110110  - 9  серий
```

Б. Периодическая последовательность atcatcatcatcatcatcatcatc

```
 1   2   3   4   5   6   7   8      Номера периодов
FatcFatcFatcFatcFatcFatcFatcFatcF   ДНК последовательность
01  01  01  01  01  01  01  01  0   последовательность категорий
01010101010101010  - 17 серий
```

В. Периодическая последовательность с делецией нуклеотида с в третьем периоде atcatcatatcatcatcatcatca

```
 1   2   3   4   5   6   7   8      Номера периодов
FatcFatcFataFtcaFtcaFtcaFtcaFtcaF   ДНК последовательность
01  01  01 10  10  10  10  10  10   последовательность категорий
010101101010101010  - 17 серий
```

Рис.1. Применение критерия серий для определения неслучайности распределения аденина (а) по периодам длиной три основания Изучаемая последовательность разбивается на отдельные периоды посредством вставки буквы F через каждые три основания, начиная с нулевой позиции последовательности. На периодической последовательности (Б) число серий будет всегда больше или равно, чем число серий на непериодической последовательности (А). Это позволяет использовать число серий как меру периодичности, слабо чувствительную к делециям и вставкам нуклеотидов. Это продемонстрировано на рисунке (В), где произведена делеция c в 9-й позиции периодической последовательности Видно, что число серий не изменилось при этом и осталось равным 17.

Для введения количественной меры создадим четыре последовательности категорий $K(i)$ из последовательности S' Последовательность категорий $K(i)$, i=1,2,3,4, создается из последовательности S' путем замены F на 0 и символа $q(i)$ на 1, остальные символы последовательности S' при этом игнорируются. Здесь $q(i)$ – символ алфавита Q. Для каждой из введенных последовательностей категорий рассчитывалось число серий $m(i)$. Для определения статистической значимости полученного числа проводилось имитационное моделирование методом Монте-Карло. Для этого выполнялось случайное перемешивание символов исходной последовательности S, после чего для вновь полученной последовательности строилась S' и определялось число серий $m(i)$, i=1,2,3,4. В результате применения метода Монте-Карло было получено N значений для числа серий $m(i)$ для случайных последовательностей. После этого для полученного множества значений для каждого символа рассчитывалось среднее значение $\mu_r(i)$ и стандартное отклонение $\sigma_r(i)$ Тогда формула расчета статистической значимости $Z(i)$ числа серий $m(i)$ будет иметь вид:

$$Z(i)=\frac{m(i)-\mu_r(i)}{\sigma_r(i)} \qquad (1)$$

Выше было рассмотрено применение критерия серий для поиска регулярности одного нуклеотида в изучаемой последовательности. Однако, для проведения более полного сравнительного анализа последовательностей необходимо иметь возможность оценить регулярность по всем четырем символам одновременно. Для этого сначала производится вычисление статистики Z (формула (1)) для каждого из нуклеотидов в отдельности. Обозначим результаты для a, t, c и g как Z_a, Z_t, Z_c и Z_g, соответственно. Все указанные величины имеют распределение, близкое к нормальному (это было показано в результате проведения численного моделирования для случайных последовательностей) Воспользуемся свойством стандартного нормального распределения, чтобы получить новую суммарную величину, также распределенную нормально. Получаем:

$$Z_{sum}=\frac{Z_a+Z_t+Z_c+Z_g}{\sqrt{4}}=\frac{Z_a+Z_t+Z_c+Z_g}{2}\sim N(0;1) \qquad (2)$$

В результате сначала проводился расчет величины Z для каждого из нуклеотидов по формуле (1), а затем рассчитывалось объединенное значение Z_{sum} по формуле (2).

Поиск регулярности в нуклеотидной последовательности проводился следующим образом. Последовательность ДНК сканировалось с помощью окна длиной 500 нуклеотидов, что соответствовало длине изучаемых промоторов. Выбиралась длина периода n и для последовательности, попавшей в окно, создавались четыре последовательности S' (для каждого нуклеотида) путем введения символов F. Для каждой этих 4-х последовательностей определялось максимальное значение Z, как было описано выше. Затем для полученных значений Z_a, Z_t, Z_c и Z_g по формуле (2) определялось значение Z_{sum}.

В пределах окна определялся максимум для значения Z_{sum} посредством варьирования границ изучаемой последовательности, а именно, правая и левая ее границы независимо изменялись с шагом 2, и каждый раз рассчитывалась статистика Z_{sum}

Модифицируем приведенный выше пример – пусть последовательность S имеет вид *tcgcgaaaattaaacaaaag*, *S'=FtcgcgFaaaatFtaaacFaaaagF*. В этом случае вариационный ряд для нуклеотида *a* будет иметь вид {0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0}. Данный пример показывает, что при увеличении числа символов в каждом из периодов число серий не увеличивается. Для больших длин периода (>4) это может привести к тому, что некоторые последовательности, обладающие регулярностью, не будут выявлены.

В целях разрешения данной проблемы вносились дополнительные символы F в последовательность S' В результате все периоды последовательности S делятся на части (или «ящики», по аналогии с моделями, используемыми в комбинаторике) одинаковым образом. Для примера, показанного выше, последовательность S' после внесения дополнительного символа F в каждый период может выглядеть как *FtcgFcgFaaaFatFtaaFacFaaaFagF* (жирным шрифтом выделены новые символы). Рассматривались все возможные позиции размещения новых символов, и выбиралось такое положение нового символа в периоде (во всех периодах идентично), которое обеспечивало наибольшее значение числа серий. Затем заново определялась статистическая значимость посредством имитационного моделирования методом Монте-Карло, как это было описано выше

В результате проведения имитационного моделирования было получено пороговое значение статистики критерия Z_{sum}= 4 0. Данное значение гарантирует, что на исследуемом множестве промоторов будет найдено не более одной случайной последовательности, обладающей регулярностью.

«Схемой регулярности» называется схематичное изображение расположения символов F в последовательности, для которого распределение нуклеотидов оказалось наиболее неслучайным, то есть, значение Z_{sum} для данной последовательности достигает максимальной величины. Например, приведенная ниже схема означает, что максимальное значение статистики критерия было получено для последовательности S', имеющей нуклеотид *a* в любой позиции периода с 1-го по 4-й нуклеотид, а также в любой из 5-й и 6-й позиций

периода. Нуклеотид *c* может наблюдаться в 1-3 позициях и в 4-6 позициях периода, символ 'g' – в 1, 2, 3-5 и в 6-й позициях периода. Нуклеотид *t* в исследуемой последовательности отсутствовал. Регулярность в данном случае наблюдается по трем символам – "acg" при длине периода в шесть нуклеотидов.

```
F----F--F        a
F---F---F        c
F-F-F---F-F      g
```

В **разделе 2.2** приведено описание алгоритма классификации периодичности последовательностей ДНК.

Каждому периодическому участку, найденному методом информационного разложения, соответствует позиционно-частотная матрица *M*. Элементы данной матрицы представляют собой частоту встречаемости каждого нуклеотида в каждой позиции периода. Поскольку участки последовательностей, обладающие скрытой периодичностью, имели различную длину, то все рассматриваемые матрицы были нормализованы на единицу (то есть, значение каждого элемента матрицы делилось на сумму всех ее элементов) для того, чтобы классификация этих матриц не зависела от длины последовательности со скрытой периодичностью. В качестве меры подобия матриц (или, другими словами, расстояния между матрицами) была выбрана статистику Пирсона. Блок-схема алгоритма классификации скрытой периодичности приведена на Рис. 2.

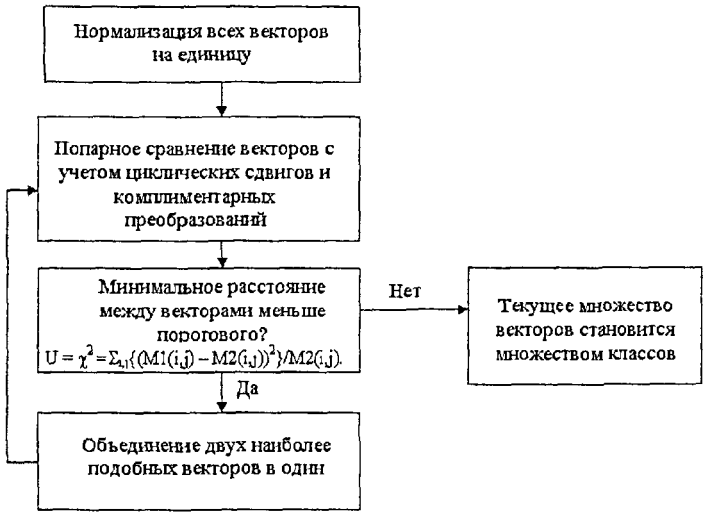

Рис. 2. Блок-схема классификации матриц скрытой периодичности.

Процесс сравнения и объединения векторов продолжался до тех пор, пока минимальное значение статистики Пирсона для всего множества не становилось больше порогового, соответствующего 5% уровню случайности для χ^2. Вектора, оставшиеся на этот момент, считались классами периодичности. Для увеличения адекватности отражения классами закономерностей периодичности исключались из рассмотрения классы, содержащие менее трех векторов.

Статистическая значимость объединения матриц обеспечивалась заданием порогового уровня для объединения в 95%. Это означает, что процесс объединения матриц автоматически заканчивается при подобии матриц в классах, составляющем менее чем 95%.

В **разделе 2.3** приведено описание профильного *анализа* - *основного* алгоритма, использованного для поиска периодичности в последовательностях ДНК в условиях наличия вставок и делеций символов.

Использовался метод динамического выравнивания профилей, в котором рассматривается дивергенция последовательностей, вызванная точечными мутациями, а также наличием вставок и делеций Данный подход объединяет в себе алгоритмы динамического программирования для нахождения оптимального выравнивания с *анализом* позиционной специфичности нуклеотидов в профиле.

При проведении поиска периодических последовательностей в базе данных Genbank проводилось построение позиционно-специфичных матриц (ПСМ) частот нуклеотидов для каждого из классов периодичности, полученных ИР. В целях предотвращения искажений частот в изотипичных последовательностях (эффект неравномерного представления) было проведено взвешивание каждой последовательности обратно пропорционально числу последовательностей, имеющих с ней высокий уровень сходства Полученные ПСМ были преобразованы в позиционно-специфичные весовые матрицы (ПСВМ) по формулам:

$$w'(S,j) = f(S,j)\ln\{f(S,j)/p(S)\}$$
$$w(S,j) = w'(S,j) - \tilde{w}'(j) \qquad (3)$$

где S = {A,T,C,G}, f(S,j) – элемент матрицы ПСМ, $p(S) = \sum_j f(S,j)$,

$\tilde{w}'(j) = 0.25\sum_S w'(S,j)$ - средний вес нуклеотидов в j –м столбце ПСМ матрицы, w(S,j) – вес нуклеотида S в ПСВМ.

Для выравнивания анализируемой последовательности из Genbank относительно ПСВМ использовался динамический алгоритм нахождения локального подобия, также известный как алгоритм Смита-Уотермана (Smith-Waterman). Элементы матрицы выравнивания определялись по формулам:

$$F(i,j) = \max\{\max_{1\le k\le dmax}\{F(i-k,j) - v_d(1+\log(k))\}; \max_{1\le l\le dmax}\{F(i,j-l) - v_d(1+\log(l))\};$$
$$F(i-1,j-1) + w(S(i),j), 0.0\};$$
$$F(0,0) = 0.0; F(i,0) = F(0,0) - v_d(1+\log(i)); F(0,j) = F(0,0) - v_d(1+\log(j)) \qquad (4)$$

где i – позиция нуклеотида в анализируемой последовательности, j – позиция в консенсусной последовательности, $dmax$ = 40 – максимальное анализируемое количество вставок и делеций, v_d = 1.0 – штраф за открытие делеции и $w(S(i),j)$ – элемент ПСВМ, рассчитываемый по формулам (3), $S(i)$ – нуклеотид в i-й позиции анализируемой последовательности.

Сканирование базы Genbank с помощью ПСВМ проводилось с шагом в 20 оснований. Консенсус-последовательность, полученная на основе класса периодичности, воспроизводилась необходимое для соответствия длине максимальной подпоследовательности количество раз. На каждом шаге проводилось выравнивание анализируемой последовательности относительно ПСВМ Матрица $F(k,j)$ заполнялась полностью, затем определялся ее максимальный элемент $f_{max}(k_m , j_m)$. На основании положения f_{max} определялось оптимальное выравнивание как путь от максимального элемента до первого нулевого элемента, координаты которого обозначались как (k_0 , j_0).. Полученное выравнивание задает «максимальную подпоследовательность» S_m, при этом положение соответствующей ПСВМ отмечено нуклеотидами, имеющими больший вес.

Для определения статистической значимости найденного выравнивания последовательности из Genbank необходимо оценить вероятность существования выравнивания случайных последовательностей с тем же нуклеотидным составом относительной заданной ПСВМ Таким образом, заполнялась матрица F' по формулам

$$F'(i,j) = \max\{\max_{1\le k\le dmax}\{F'(i-k,j) - v_d(1+\log(k))\}; \max_{1\le l\le dmax}\{F'(i,j-l) - v_d(1+\log(l))\};$$
$$F'(i-1,j-1) + w(S(i),j)\}; \qquad (5)$$
$$F'(0,0) = 0.0; F'(i,0) = F'(0,0) - v_d(1+\log(i)); F(0,j) = F'(0,0) - v_d(1+\log(j)),$$

то есть, аналогично формулам (4), но без рассмотрения нулевых значений при выборе максимума. Далее рассчитывалась величина $d = f(k_m , j_m) - f(k_0 , j_0)$, распределение которой при использовании данных формул близко к нормальному. Данная гипотеза была подтверждена путем проведения имитационного моделирования (метод Монте-Карло). При этом было сгенерировано 100 последовательностей с тем же символьным составом, что и в рассматриваемой последовательности (то есть, с теми же частотами появления символов и триплетной корреляцией), а затем матрица F' была заполнена для каждой последовательности.

В качестве меры статистической значимости было определено Z-значение, то есть, нормализованное отклонение веса найденного выравнивания анализируемой последовательности от среднего веса выравниваний случайных последовательностей относительно ПСВМ:

$$Z = (W - M(W_{rnd}))/\sigma(W_{rnd}) \qquad (6)$$

где W – вес найденного выравнивания, W_{rnd} – вес выравнивания случайной последовательности, M и σ – среднее значение и стандартное отклонение W_{rnd}, соответственно.

Поскольку МПА обеспечивает статистическую значимость подобия последовательностей, а не их периодичность, были проведены дополнительные статистические испытания последовательностей, обнаруженных с помощью данного алгоритма. Вначале рассчитывался спектр ИР для всех последовательностей. Затем был использован метод Монте-Карло для оценки статистической значимости, а именно, для каждой последовательности было сгенерировано 200 последовательностей путем случайного перемешивания ее символов, рассчитаны значения средней величины, дисперсии и, наконец, Z-значение, как это было описано выше. Считалось, что рассматриваемая последовательность является периодической с заданной длиной периода, если соответствующее этой длине значение Z было максимальным в спектре ИР при дополнительном условии Z>=7.0. Использование данного критерия позволяет утверждать, что последовательность, ему удовлетворяющая, обладает периодичностью с заданной длиной периода на статистически значимом уровне.

В **разделе 2.4** приведены методы создания базы данных потенциальных микросателлитных последовательностей, обнаруженных с помощью применения разработанных алгоритмов для анализа банка данных Genbank.

Разработанное программное обеспечение представляет собой комплекс, состоящий из собственно базы данных (информации, хранимой в виде таблиц), системы управления базой данных (СУБД), программных компонент обработки запросов к базе данных, программных средств ведения базы данных и веб-интерфейса пользователя.

База данных по потенциальным мини- и микросателлитным последовательностям содержит следующую информацию: участки ДНК, обладающие скрытой периодичностью; координаты последовательностей в соответствующем локусе (участке последовательности с уникальным идентификатором) Genbank, тип и параметры найденной периодичности. Обеспечивается возможность поиска последовательностей по различным параметрам, получения информации о наличии подобных последовательностей на некотором заранее заданном участке ДНК, просмотра результатов запросов посредством веб-обозревателя и сохранения результатов запроса в локальный файл на компьютере пользователя. Веб-интерфейс пользователя реализован в виде Интернет-страницы. Данная страница содержит элементы, обеспечивающие возможность создания запросов к базе данных и отображения результатов запросов. Также поддерживается возможность сохранения результатов запроса в файл на компьютере пользователя. Взаимодействие программных компонентов в виде диаграммы потоков данных представлено на рис. 3.

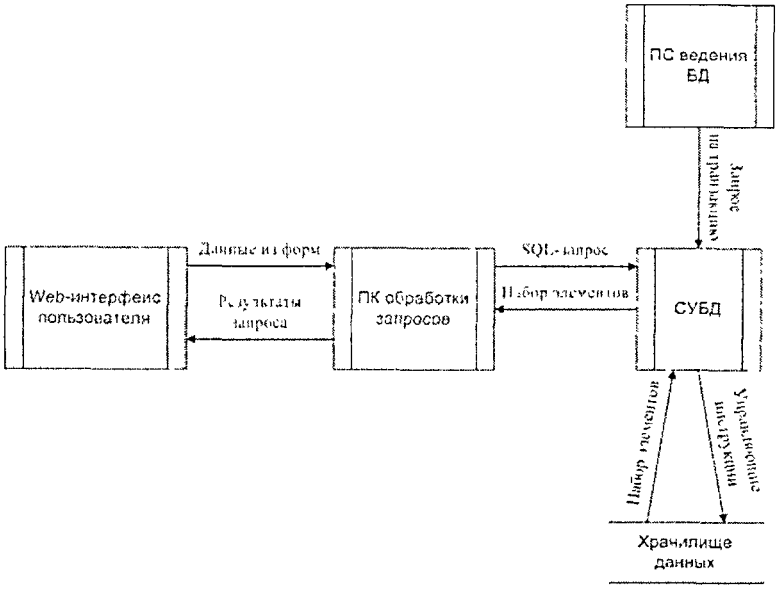

Рис.3. Взаимодействие программных компонентов (ПК) и подсистем (ПС) базы данных.

В **разделе 2.5** описаны методы создания веб-сервера поиска скрытой периодичности. Веб-сервер LEPSCAN (LAtent Periodicity SCANner) позволяет осуществлять поиск скрытой периодичности, тип которой относится к заранее определенному множеству классов. Классы были получены для различных групп организмов и различных длин периода с помощью алгоритмов, описанных в разделе 2.2. Основным алгоритмом, используемым при сканировании последовательности в целях выявления скрытой периодичности, является МПА. Сервер реализован на платформе openSuse Linux 10.0; для проведения вычислений используется суперкомпьютер с кластерной архитектурой, состоящий из 114 блоков, укомплектованных процессорами AMD Opteron и Pentium. Проведение параллельных вычислений реализовано с помощью технологии MPI (Message Passing Interface) с использованием системы очередей Cleo [13] Веб-интерфейс пользователя реализован на основе технологии CGI (Common Gateway Interface) Также с использованием этой технологии реализован доступ к базе данных, содержащей классы периодичности. Язык реализации – Perl. Основной вычислительный модуль реализован на языке C++ с использованием MPI, вспомогательные программы представляют собой Perl-сценарии (скрипты). Базы данных реализованы под управлением СУБД PostgreSQL версии 8.04. Диаграмма потоков данных представлена на рис.4.

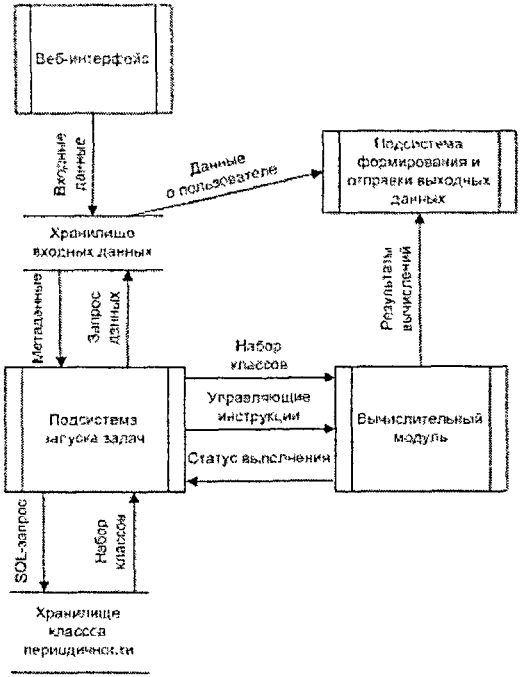

Рис. 4. Диаграмма потоков данных программного комплекса LEPSCAN.

Следует отметить, что использование параллельных вычислений с помощью технологии MPI позволяет проводить поиск скрытой периодичности одновременно для нескольких поступивших запросов, что существенным образом ускоряет проведение расчетов и, соответственно, позволяет пользователю быстрее получить результаты.

В **третьей главе** приведены результаты использования разработанных алгоритмов поиска регулярных последовательностей, а также детально описана реализация пользовательского интерфейса базы данных и веб-сервера поиска скрытой периодичности. Обсуждается биологическая значимость и адекватность полученных результатов, приведено сравнение с экспериментально полученными данными.

В **разделе 3.1** приведены результата поиска регулярных последовательностей в эукариотических промоторах из базы данных EPD [14] версии 93 (общее количество последовательностей, содержащихся в базе – 4809). Было выбрано 2236 последовательностей, представляющих все группы организмов, при этом уровень подобия любых двух последовательностей из этого множества составлял не более 50% (опция базы данных).

При поиске регулярности были использованы следующие параметры сканирования. размер окна=500, шаг внутри окна=2 Длина обнаруживаемых регулярных последовательностей находилась в диапазоне 50-500 нуклеотидов. В 1342 промоторах была выявлена регулярность на статистически значимом уровне. Таким образом, более 60% промоторов содержат последовательности с регулярной структурой в диапазоне длин от 2 до 16 нуклеотидов. Регулярность в остальных промоторных последовательностях также была обнаружена, но уровень статистической значимости для этих последовательностей был меньше порогового.

Важным для понимания биологической значимости регулярности является изучение распределения местоположения регулярных последовательностей в промоторах. Длина промотора была разбита на интервалы длиной 10 нуклеотидов, после чего определялось число последовательностей, попавших в каждый из этих интервалов. Было использовано имитационное моделирование (метод Монте-Карло) для определения, является ли распределение регулярных последовательностей случайным Для этого все найденные регулярные последовательности 200 раз размещались случайным образом по длине промотора.

На основе полученных данных для каждого интервала были рассчитаны среднее значение и дисперсия, после чего для каждого интервала рассчитывалось значение Z с помощью формулы (1). Число регулярных последовательностей, найденных в интервале с 400 по 500 нуклеотид (от -99 до +1 относительно начала гена), заметно превышает уровень, ожидаемый для случайных последовательностей. При этом наибольшее превышение наблюдается для мест связывания факторов транскрипции и РНК-полимеразы (-45, +5). Начиная с -100-ого нуклеотида, значение Z превышает уровень 5.0. Кроме того, в интервалах с -389 по -169 нуклеотид наблюдалось отклонение значений Z в отрицательную сторону, что связано с преимущественной локализацией районов с регулярностью в интервале с -99 по +1 нуклеотид.

Тестирование разработанного алгоритма показало, что он способен выявлять как явную, так и скрытую периодичность со сравнительно небольшим числом делеций или вставок. Однако, используемый подход может пропустить длинные вставки нуклеотидов в район ДНК с имеющейся регулярностью В этом случае последовательности категорий $K(i)$ могут содержать достаточно большое количество пустых «ящиков», что будет приводить к заметному снижению статистической значимости района регулярности с протяженной вставкой Такое явление может быть причиной того, что в ряде промоторных последовательностей регулярность в районе (-99, +1) была обнаружена при значениях $Z<4.0$.

Предложенный алгоритм поиска регулярности в значительной степени нечувствителен к перестановкам нуклеотидов, при которых соседние нуклеотиды меняются местами, в то время как для алгоритмов, основанных на динамическом программировании, такие перестановки являются критичными.

В целом проведенные расчеты показывают, что регулярное строение является неотъемлемым свойством промоторных районов ДНК.

В **разделе 3.2** приведены результаты классификации скрытой периодичности Проведение классификации скрытой периодичности, обнаруженной в различных геномах с помощью ИР, преследует две основные цели. Во-первых, выявление общих закономерностей повторов, поскольку в классы входят последовательности, находящиеся в различных местах генома, но при этом обладающие идентичными или очень похожими матрицами периодичности. Во-вторых, классификация позволяет получить исходные данные, повышающие эффективность последующего применения МПА.

Ниже представлены основные результаты классификации для длин периодов 2, 4 и 5.

Таблица 1. Результаты классификации для длин периода 2, 4, 5. Верхнее значение соответствует число обнаруженных периодических последовательностей, а нижнее значение – число полученных классов периодичности для соответствующей группы организмов и длины периода.

Группа организмов / Длина периода	2	4	5
Бактерии	454 45	883 34	867 25
Беспозвоночные	29430 199	17966 73	6113 62
Вирусы и фаги	306 29	171 12	136 8
Грызуны	131265 211	107013 189	12821 135
Позвоночные	17170 136	17536 120	3601 88
Приматы	168164 188	107297 174	10736 329
Растения	14766 141	8170 160	2107 131

Как видно из таблицы, число классов значительно меньше числа исходно обнаруженных периодических последовательностей – от 10 раз для небольшого числа исходных последовательностей до 100 и более раз для большого количества периодичностей, обнаруженных посредством использования информационного разложения. Таким образом, вторая цель проведения классификации (сокращение числа исходных данных для профильного анализа без потери их представительности), очевидно, достигнута. Рассмотрим теперь, каким образом полученные классы отражают общие закономерности периодичности соответствующих групп организмов. В качестве примера рассмотрим классификацию динуклеотидной (длина периода = 2) периодичности геномов растений. Более половины последовательностей попало в три самых больших класса (содержащих 3391, 2285 и 1951

последовательность, соответственно) Следовательно, свойства периодичности, определяемые этими классами, являются общими для большого числа последовательностей.

Рассмотрим свойства матриц, образующих три самых больших класса скрытой динуклеотидной периодичности растений Тип скрытого периода для самого большого класса показан на рис. 16.

A1	T1	C1	G1	A2	T2	C2	G2
1	31	0	0	31	0	0	0

	1	2
A	0,016	0,492
T	0,492	0,000
C	0,000	0,000
G	0,000	0,000

Рис. 5. Матрица класса (в виде вектора и в нормализованной форме), включившего наибольшее число периодических последовательностей из геномов растений.

Очевидно, что в первой позиции преобладает тимин, а во второй – аденин. Таким образом, условный консенсус для данного периода можно записать в виде {t}{a}. Рассмотрим теперь второй класс по числу элементов. Он имеет консенсус {a, g}{t}, причем частота гуанина меньше частот аденина и тимина. Третий по численности класс имеет консенсус {c}{t}, то есть, он отличается от предыдущих двух классов. Мотив AT является наиболее часто встречаемым в микросателлитах растений, что подтверждается полученными нами результатами. В геномах *Oryza sativa и Arabidopsis thaliana* преобладает мотив GA. Последовательности данных геномов составляют существенную часть Genbank (около 50% от длины всех растительных геномов), и число обнаруженных в них периодических последовательностей также велико. Следовательно, наличие большого класса с тем же консенсусом (CT является комплементарным по отношению к GA) согласуется с данными, полученными экспериментально.

В **разделе 3.3** приведены результата поиска микросателлитных последовательностей ДНК посредством модифицированного профильного анализа.

Данные по количеству обнаруженных с помощью разработанного алгоритма последовательностей и сравнение с результатами, полученными посредством информационного разложения, приведены в табл. 2 (на примере геномов растений и бактерий, приведено число обнаруженных неперсекающихся последовательностей).

Таблица 2. Результаты поиска скрытой периодичности алгоритмами ИР (верхнее значение в каждой ячейке) и МПА (нижнее значение) в геномах растений и бактерий.

Группа организмов / Длина периода	2	4	5
Бактерии	454 6593	883 7242	867 3732
Растения	14766 80396	8170 32165	2107 8132

Рассмотрим подробнее результаты поиска динуклеотидной периодичности в геномах растений. Был проведен поиск скрытой периодичности посредством МПА для 141 класса, полученного ранее. Каждая из матриц классов была использована для поиска во всех последовательностях ДНК из геномов растений, представленных в Genbank. При этом использовалось пороговое значение уровня значимости Z, равное 7.0, обеспечивающее неслучайность данного поиска. Число найденных неперекрывающихся последовательностей составило 103556. После фильтрации, проведенной согласно процедуре, описанной в разделе 2.2., число последовательностей составило 80396. Как и в случае бактерий, число полученных последовательностей намного превосходит исходное значение (14766).

Рассмотрим распределение числа периодических последовательностей, принадлежащих некоторым функциональным элементам. Геномы растений имеют намного больший размер, чем геномы бактерий или дрожжей, что сильно затрудняет проведение их аннотации (определение функциональной значимости участков генома) с помощью экспериментальных методов. Поэтому вполне ожидаемым является тот факт, что большинство найденных периодических последовательностей (более 57000) были обнаружены в ранее неаннотированных участках генома. Таким образом, подтверждается возможность проведения аннотации с помощью используемых алгоритмов Также следует отметить, что более 7000 последовательностей перекрываются с повторами, ранее обнаруженными экспериментально. Небольшое количество периодов с длиной два нуклеотида в генах (на эти области приходится около 12% найденных последовательностей) позволяет сделать вывод о предпочтительности формирования участков с размытой периодичностью в некодирующих областях

Кроме того, несомненный интерес представляет исследование распределения найденных периодических участков по организмам, последовательности которых представлены в Genbank. Общее число организмов, в геномах которых была обнаружена динуклеотидная периодичность, составило 2202. Полученные данные свидетельствуют о том, что распределение найденных периодических последовательностей не является случайным, поскольку в противном случае их общая длина была бы пропорциональна длине локусов исследуемых организмов.

В **разделе 3.4** рассмотрена реализация базы данных потенциальных микро- и минисателлитных последовательностей. MMsat представляет собой аналитическую базу данных по скрытой периодичности в последовательностях Genbank. Последовательности, обладающие скрытой периодичностью с длиной периода 2-100, могут рассматриваться как потенциальные микро- и минисателлиты. База данных размещена в свободном доступе в сети Интернет по адресу http://victoria.biengi.ac.ru/mmsat. Языки интерфейса - русский и английский.

В качестве результатов запроса пользователь может увидеть локализацию потенциальных микро- и минисателлитных последовательностей, а также спектр статистической значимости полученных последовательностей.

База данных содержит 2851428 последовательностей, обладающих скрытой периодичностью. Распределение последовательностей по длине периода представлено в табл. 3.

Таблица 3 Распределение количества последовательностей, содержащихся в базе данных, по длине периода.

Длина периода	2	3	4	5	6	7	8	9	10
Число последовательностей	366739	1323348	261544	36553	87649	22973	54555	16954	26397
Длина периода	11-20	21-50	51-80	81-100					
Число последовательностей	133337	269863	196295	60085					

Поиск потенциальных микро- и минисателлитных последовательностей можно производить с использованием следующих задаваемых пользователем параметров: идентификатор локуса, частотная матрица участка последовательности, ключевые слова - функциональные свойства участка последовательности, интервал значений статистической значимость периодичности, интервал длин периода, интервал для положения периодических участков в последовательности, группа организмов (соответствует группам Genbank).

Для того, чтобы выполнить поиск в базе, необходимо заполнить значения всех желаемых параметров (любой из них можно оставить незаполненным) и нажать кнопку 'Послать запрос'.

В **разделе 3.5** рассмотрена программная реализация веб-сервера поиска скрытой периодичности LEPSCAN (http://victoria.biengi ac.ru/lepscan).

Для работы с сервером необходимо задать последовательность, в которой будет осуществляться поиск периодичности (обязательно), адрес электронной почты (обязательно), а также заполнить значения желаемых параметров (любой из них можно оставить незаполненным), после чего нажать кнопку 'Послать запрос'. Результаты поиска высылаются на указанный адрес электронной почты в течение 24 часов в виде zip-архива, содержащего данные по найденным периодическим участкам исходной последовательности, заданной пользователем. Каждой длине периода соответствует один файл результата, содержащийся в архиве. Формат имени файла в архиве - PER[длина_периода]_final. Исходная последовательность также находится в архиве в файле P000.

Поиск потенциальных микро- и минисателлитных последовательностей можно производить с использованием задаваемых пользователем параметров: группы организмов (соответствуют группам, используемым в базе данных Genbank) и длины периода (возможные значения 2-20). Поиск осуществляется с использованием классов периодичности, полученных для соответствующих длин периода и групп организмов

Информация по каждой из найденных последовательностей включает в себя внутренний идентификатор последовательности, координаты в исходной последовательности, координаты в консенсус-последовательности, длину периода, частотную матрицу класса найденной периодичности, полученное выравнивание, а также статистическую значимость обнаруженной периодичности.

Заключение

Применение разработанных алгоритмов к анализу последовательностей ДНК позволило получить уникальные результаты, проливающие свет на эволюцию и строение геномов и позволяющие предсказывать функциональную роль вновь получаемых нуклеотидных последовательностей Это делает данную работу актуальной, а созданное программное обеспечение востребованным в экспериментальных исследованиях в области молекулярной биологии и генетики.

В частности, в рамках настоящей работы:

1 Разработан и программно реализован алгоритм классификации скрытой периодичности в последовательностях оснований нуклеиновых кислот; проведена классификация периодических последовательностей ДНК из банка данных Genbank для различных длин периода и групп организмов

2. Проведен поиск сильно размытой периодичности с использованием полученных классов посредством разработанного алгоритма модифицированного профильного анализа в различных геномах; выявлена функциональная значимость обнаруженной периодичности

3. Создана база данных по потенциальным мини- и микросателлитным последовательностям ДНК на основе результатов выявления скрытой периодичности в геномах различных организмов с помощью разработанных алгоритмов.

4. Разработан Интернет-сервера для поиска скрытой периодичности, реализующий алгоритм модифицированного профильного анализа.

5. Разработан и программно реализован алгоритм поиска регулярности последовательностей нуклеиновых кислот, основанный на использовании критерия серий

6. Разработанный алгоритм применен для поиска регулярности в последовательностях промоторов из различных геномов.

Список литературы

1. *Toth G., Gaspari Z., Jurka J.* Microsatellites in different eukaryotic genomes: survey and analysis // *Genome Res.* - 2000. - Vol. 10. - P. 967-981
2. *Mishra D., Thangaraj K., Mandhani A., Kumar A., Mittal R.* Is reduced CAG repeat length in androgen receptor gene associated with risk of prostate cancer in Indian population? // *Clin. Genet.* - 2005. – Vol. 68, No 1 - P. 55-60.
3 *Makeev V.Y., Frank G.K., Tumanyan V.G.* Statistics of periodic patterns in the sequences of human introns // *Biophysics.* – 1996. – Vol. 41, No 1. – P. 263-268.
4 *Chechetkin V.R., Lobzin V.V.* Levels of ordering in coding and noncoding regions of DNA sequences // *Phys. Lett. A* -1996 – Vol. 222. – P 354-360
5. *Benson G.* Tandem repeats finder: a program to analyse DNA sequences // *Nucleic Acids Res.* – 1999. – Vol. 27. - P. 573-580.
6 *Herzel H., Trifonov E.N., Weiss O., Grobe I* Interpreting correlations in biosequences // *Physica A.* – 1998 – Vol. 249. - P.449-459.
7. *Korotkov E.V., Korotkova M.A., Kudryashov N.A* Information decomposition method to analyze symbolical sequences // *Phys. Let. A.* - 2003. - Vol. 312. - P.198-210.
8. *Landau G., Schmidt J., Sokol D.* An algorithm for approximate tandem repeats // *J. Comp. Biol.* - 2001 - Vol. 8. - P. 1–18.
9. *Rice P., Longden I., Bleasby A* EMBOSS: The european molecular biology open software suite // *Trends Genet.* - 2000. - Vol. 16. - P. 276–277.
10. *Subramanian S., Mishra R.K., Singh L* Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions // *Genome Biol* - 2003. - Vol. 4, No. 2. - P R13
11. *Bajic V.B. et al.* Promoter prediction analysis on the whole human genome // *Nat. Biotechnol.* - 2004 – Vol 22. – P 1467–1473.
12. *Xie X., Wu S., Lam K.-M., Yan H.* PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm // *Bioinformatics.* – 2006. - Vol. 22. – P. 2722-2728.
13. *Воеводин Вл. В, Жуматий С.А.* Вычислительное дело и кластерные системы. - М: Изд-во МГУ, 2007. - 150 с.
14 *Schmid C.D., Perier R, Praz V., Bucher P.* EPD in its twentieth year: towards complete promoter coverage of selected model organisms // *Nucleic Acids Res.* - 2006. – Vol 34. - D82-5.

Основные положения диссертации изложены в публикациях:

1. Шеленков А.А, Коротков Е В Классификация скрытой периодичности нуклеотидных последовательностей из банка данных Genbank // Материалы 8-й Пущинской школы-конференции молодых ученых «Биология – наука XXI века» – Пущино. - 2004. - С. 245.
2. Shelenkov A.A., Chaley M B , Korotkov E.V. Revelation and classification of dinucleotide periodicity of bacterial genomes using the methods of information decomposition and modified profile analysis // Proceedings of the 4th International Conference on Bioinformatics of Genome Regulation and Structure – Novosibirsk. - 2004. - Vol 2. – P. 293-296.
3 Shelenkov A A , Chaley M.B., Korotkov E.V. Revelation and classification of dinucleotide periodicity of bacterial genomes using the method of information decomposition // Материалы 9-й Пущинской школы-конференции молодых ученых «Биология – наука XXI века». – Пущино. - 2005. - С. 323.
4 Шеленков А А. Классификационный анализ скрытой периодичности последовательностей оснований нуклеиновых кислот // Материалы международной школы-конференции молодых ученых «Системная биология и биоинженерия». – Звенигород. - 2005. - С. 92-93.
5. Shelenkov A A., Korotkov E.V. Search and Classification of Potential Minisatellite Sequences from Bacterial Genomes // Доклады I Международной конференции «Математическая биология и биоинформатика». – Пущино. - 2006. - С. 187-188.
6. Шеленков А.А , Коротков Е.В. Поиск регулярных последовательностей в эукариотических промоторах // Доклады II Международной конференции «Математическая биология и биоинформатика». – Пущино. – 2008. - С. 94-95.
7 Шеленков А.А LEPSCAN – веб-сервер поиска скрытой периодичности // Доклады II Международной конференции «Математическая биология и биоинформатика». – Пущино – 2008. - С. 100-101.
8 Shelenkov A A , Chaley M.B., Skryabin K.G., Korotkov E.V. Revelation and classification of dinucleotide periodicity of bacterial genomes using the method of information decomposition / Bioinformatics of Genome Regulation and Structure II. Eds . Kolchanov N , Hofestaedt R., Milanesi L. - New-York: Springer Science+Business Media Inc. – 2006. - P. 179-188.
9. Shelenkov A.A., Skryabin K.G., Korotkov E.V. Search and Classification of Potential Minisatellite Sequences from Bacterial Genomes // *DNA Res.* – 2006. – Vol. 13, No. 3. – P. 89-102.
10. Шеленков А А , Скрябин К.Г., Коротков Е.В. Классификационный анализ скрытой динуклеотидной периодичности геномов растений // *Генетика.* - Т. 44, №1. - С.120-136.
10a. Shelenkov A. A., Skryabin K. G., Korotkov E. V. Classification Analysis of a Latent Dinucleotide Periodicity of Plant Genomes // *Rus. J. Genet.* – 2008. - Vol. 44, No. 1. - P.101-114.
11. Shelenkov A A., Korotkov A.E , Korotkov E.V. MMsat - a database of potential micro- and minisatellites // *Gene.* – 2008 – Vol. 409, No. 1-2 – P.53-60.
12. Шеленков А А., Коротков Е.В Поиск регулярных последовательностей в промоторах из геномов различных групп организмов с использованием критерия серий // *Математическая биология и биоинформатика.* – 2008. - Т .3, №1. - С. 1-15.

Подписано в печать 20.10 2008 г.

Печать трафаретная

Заказ № 999
Тираж: 100 экз.

Типография «11-й ФОРМАТ»
ИНН 7726330900
115230, Москва, Варшавское ш., 36
(499) 788-78-56
www.autoreferat.ru