На правах рукописи

Тетуев Руслан Курманбиевич

АЛГЕБРА СПЕКТРАЛЬНЫХ ПРЕОБРАЗОВАНИЙ В ЗАДАЧАХ ОБРАБОТКИ ДАННЫХ

Специальность: 05.13.17 – Теоретические основы информатики

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Пущино – 2007

Работа выполнена в Институте математических проблем биологии РАН,
в филиале кафедры ММП факультета ВМиК МГУ им. М.В. Ломоносова.

Научный руководитель: доктор технических наук,
профессор
Флоренц Федорович Дедус

Официальные оппоненты: доктор технических наук,
профессор
Цимбал Владимир Анатольевич
кандидат физико-математических наук
Сенько Олег Валентинович

Ведущая организация: Межведомственный Суперкомпьютерный Центр РАН, г. Москва

Защита диссертации состоится _____ 2007г.
в _____ на заседании Диссертационного совета
Д002.017.02 Вычислительного центра им. А.А.Дородницына РАН
по адресу: 119991, Москва, ГСП-1, ул. Вавилова, дом 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан _____ 2007 г.

Ученый секретарь Диссертационного Совета Д002.017.02, доктор физико-математических наук
В.В. Рязанов

Актуальность темы

Во второй половине двадцатого столетия стремительное развитие вычислительной техники ознаменовало начало новой эры для всей математической теории и различных ее применений. Однако специфика ЭВМ диктовала свои условия, но целочисленные представления лишь отчасти, приближенно отвечали развитым к тому времени математическим, аналитическим понятиям и представлениям. Классическая техника математических вычислений привыкла оперировать понятиями вещественных чисел, аналитических функций, операций дифференцирования, интегрирования и т.д. Однако попытки реализовать эти представления на ЭВМ приводили к необходимости решения довольно сложных, нетривиальных задач.

Трудно переоценить влияние спектральных представлений функций на развитие современной прикладной математики, однако на практике спектральное представление довольно редко используется для проведения определенных аналитических преобразований функций. Однако, как показано ранее в работе [1], существует принципиальная возможность осуществления аналитических преобразований, при использовании лишь самих спектральных представлений функций. Более того, такой подход приводит к более приемлемым с практической точки зрения результатам в сравнении с теми, что были получены при помощи численных методов. Этот факт объяснен авторами свойствами высокой адаптивности метода при выборе базиса из семейства классических ортогональных полиномов, гладкостью аппроксимативного представления, заданного ограниченным ортогональным рядом и другими.

Однако основным недостатком спектральных преобразований функций является тот факт, что вычислительный процесс оказывается намного более сложным и громоздким. Помимо этого подобные алгоритмы требуют стадии предварительных вычислений и резервирования больших объемов памяти под вспомогательные коэффициенты.

Указанные недостатки являются существенным ограничением для внедрения такого похода на практике в задачах, требующих проведения быстрых, устойчивых и точных аналитических преобразований функций, сигналов и т.д. Такое отставание во времени способно вынудить разработчиков программ отдать предпочтение в пользу более грубых, но быстрых алгоритмов.

Вышеописанное положение демонстрирует потребность в создании альтернативного простого и точного математического аппарата для осуществления быстрых спектральных преобразований функций, соответствующих ряду основных аналитических преобразований функций и групп их суперпозиций. В качестве ортогональных базисов рассмотрены все системы из семейства классических ортогональных полиномов (полиномы Лагерра, Якоби, Эрмита, Лежандра, Сонина-Лагерра, Гегенбауэра, Чебышева первого и второго рода) и некоторые основные модификации этих базисов.

Цель работы

Решение проблемы создания математического аппарата, для осуществления на вычислительных машинах быстрых аналитических преобразований функций, заданных ограниченным ортогональным рядом. В качестве ортогональных базисов рассматриваются системы функций, построенные на классических ортогональных полиномах.

В соответствии с поставленной задачей определены **задачи диссертационной работы:**

1. Поиск общей вычислительной схемы спектральных преобразований рядов ортогональных полиномов.
2. Вывод требуемых аналитических соотношений для класса классических ортогональных полиномов и их некоторых модификаций.
3. Реализация программного комплекса для осуществления спектральных преобразований на практике.
4. Решение задачи контурного распознавания визуальных объектов средствами разработанной системы аналитических преобразований.
5. Решение задачи поиска тандемных повторов в геномах средствами разработанной системы аналитических преобразований.

Постановка научной задачи

Требуется найти группу алгебраических правил спектральных преобразований, соответствующих основным аналитическим преобразованиям функций, представленных ограниченными рядами, построенными на классических ортогональных полиномов. Найти решение при повышенных требованиях на вычислительную сложность реализуемых алгоритмов.

Научная новизна

В данной диссертационной работе впервые предложен вычислительный метод, позволяющий существенно понизить временную сложность вычислений над спектральным представлением функций, соответствующих ряду основных аналитических преобразований над функциями.

Сформулирована и доказана Теорема о достаточном условии существования алгоритма быстрого спектрального вычисления.

Показано соблюдение достаточного условия Теоремы для ряда линейных операторов и всех систем классических ортогональных полиномов и функций.

В ходе доказательств получены рекуррентные соотношения особого вида для полиномов Якоби, Гегенбауэра и функций Якоби, Гегенбауэра, Сонина-Лагерра, Лагерра, Чебышева первого и второго рода, Эрмита.

Сформулирована и доказана Теорема об обратимости линейных операторов, а также сформулирована и доказана Лемма о суперпозиции линейных операторов.

На основе применения спектральных преобразований предложено и реализовано вычисление инвариантных геометрических признаков, описанных

аналитически с помощью операторов дифференцирования высоких порядков.

Найдены и применены аналитически описанные признаки, инвариантные к выбору начальной точки и вычисляемые на основе отрезков ортогонального ряда.

Реализован программный комплекс “SpectralRevisor”, организующий локализацию участков тандемных повторов в ДНК последовательностях на основе вычисления и анализа некоторых первых производных от функций-профилей, построенных на данных генетических последовательностях.

Практическая значимость работы

1) Новый метод сопоставляет некоторым часто используемым на практике аналитическим преобразованиям функций, набор простых и быстрых вычислений для получения соответствующего преобразования спектрального представления. Это позволяет применять метод в различных прикладных областях, где требуются быстрые, точные и устойчивые аналитические преобразования сигналов различной природы.
2) Благодаря возможности рассмотрения суперпозиции простых преобразований несложно применять в прикладных задачах составные схемы вычислений, соответствующие более сложным преобразованиям функций. Многие вычисления в узлах выстраиваемой общей схемы вычислений, как показано в диссертационной работе, могут быть произведены одновременно, что позволяет дополнительно ускорить алгоритмы при реализации на ЭВМ с параллельной архитектурой исполнения команд. Это позволяет применять метод в задачах, сопряженных с необходимостью сверхбыстрой обработки данных:
    i) обнаружение и распознавание визуальных образов в реальном времени на основе контурного восприятия;
    ii) поиск тандемных повторов в сверхдлинных генетических последовательностях, таких, как геном человека ≈ 3 млрд. нуклеотидов и т.д.
3) В ходе проведения диссертационной работы были получены новые аналитические соотношения для ряда базисов, построенных на классических ортогональных полиномах, что может быть полезным для последующих исследований и разработок в смежных областях науки.

Апробация работы

Результаты работы докладывались

на международной конференции:

PRIA-7-2004, International Conference on Pattern Recognition and Image Analysis: New Information Technologies, St. Petersburg, Russia October 18-23;

на международной летней школе:

BGRS-2-2006, Bioinformatics of the Genome Regulation and Structure, International Summer School for young scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics", Novosibirsk, Russia July 12-15;

на всероссийских конференциях:

Математические Методы Распознавания Образов-11, Пущино, 20-26 ноября (2003), ММРО-12, Москва, 20-26 ноября (2005), ММРО-13, Пущино, 9-15 октября (2006);

на всероссийских школах-конференциях:

Пущинских школах-конференциях молодых ученных «Биология – наука XXI века», секция «Математическая биология» (2004), (2006).

Публикации

По теме диссертации опубликовано 14 печатных работ, в том числе 8 тезисов конференций, 4 трудов международных конференций, 6 статей в научных журналах (в т.ч. 4 в изданиях, рекомендованных ВАК), 1 учебное пособие, 1 препринт, 1 зарегистрированная программа для ЭВМ.

Объем и структура диссертации

Диссертация состоит из введения, 5 глав, заключения, списка литературы и 2 приложений, изложена на 88 страницах. Список литературы содержит 24 наименования. Работа содержит 29 рисунков и 7 таблиц.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

ВВЕДЕНИЕ

Во введении обосновывается актуальность проблемы, сформулированы цели и задачи, дан краткий обзор содержания диссертации, перечислены полученные новые результаты исследования.

ГЛАВА I. Аналитические преобразования над спектром

Первая глава представляет собой обзор литературы по теме работы. Она состоит из трех параграфов:

1. Основы спектрального представления функций.
2. Аналитические преобразования спектра. Постановка научной задачи.
3. Ранние попытки разрешения задачи. Матричный способ.

Первый параграф посвящен истории вопроса о возможности спектрального представления функций на практике. Второй параграф вводит в проблематику задач совершения аналитических преобразований функций в пространстве спектральных коэффициентов (коэффициентов разложения) функции. Третий параграф знакомит с ранними попытками решения научной задачи, обсуждается основные недостатки, приведшие к необходимости поиска нового способа практического расчета спектральных преобразований.

ГЛАВА II. Быстрые аналитические преобразования над спектром

Вторая глава посвящена описанию основных этапов разработки нового вычислительного подхода в задаче аналитического преобразования функ-

ций на практике. Здесь сформулированы и доказаны две теоремы и одна лемма, на основании которых предложен новый метод спектральных преобразований, основанный на двух введенных правилах вычислений и названный методом *спектрального каскада-диффузии.*

Ниже приведем указанные теоремы и лемму без доказательств, а также пример применения метода спектрального каскада-диффузии.

Теорема 1: *Пусть* $f(x)$ *- некоторая функция пространства* $L^2_\rho(a,b)$ *и* $\{\varphi_n(x)\}$ *- система ортогональных функций того же пространства, такая что:*

$$f(x) = \sum_{n=0}^{N+1} C_n \varphi_n(x).$$

Пусть далее $A(\ )$ *- некоторый линейный оператор, такой что* $A(f) \in L^2_\rho(a,b)$ *и*

$$A(f) = \sum_{n=0}^{N+q} C_n^* \varphi_n(x).$$

Тогда если для каждого $\varphi_{n+1}(x)$ *существует рекуррентное соотношение вида*

$$A(\varphi_{n+1}) = F_n\left(A(\varphi_n),\ldots,A(\varphi_{n-d}),\varphi_{n+q},\ldots,\varphi_{n-p}\right),$$

где $F_n(\ldots)$ *- линейная форма,*

$$d,q,p = const \in \mathbf{N},$$

то существует алгоритм линейной временной сложности для вычисления коэффициентов разложения $\{C_n^*\}$ *при известных* $\{C_n\}$.

Теорема 2 *(об обратимости линейных операторов)*: *Пусть* $\{\varphi_n(x)\}$ *– некоторая система ортогональных функций пространства* $L^2_\rho(a,b)$ *и* $A(\ )$ *– некоторый линейный оператор, определяющий преобразование:* $A: L^2_\rho(a,b) \to L^2_\rho(a,b)$. *Пусть далее* $A^{-1}(\ )$ *– обратный оператор. Тогда, если для каждого* $\varphi_{n+1}(x)$ *существует рекуррентное соотношение вида*

$$A(\varphi_{n+1}) = F_n\left(A(\varphi_n),\ldots,A(\varphi_{n-d}),\varphi_{n+q},\ldots,\varphi_{n-p}\right),$$

и для всякого $\varphi_n(x)$

$$A^{-1}\left(A(\varphi_n)\right) \in L^2_\rho(a,b),$$

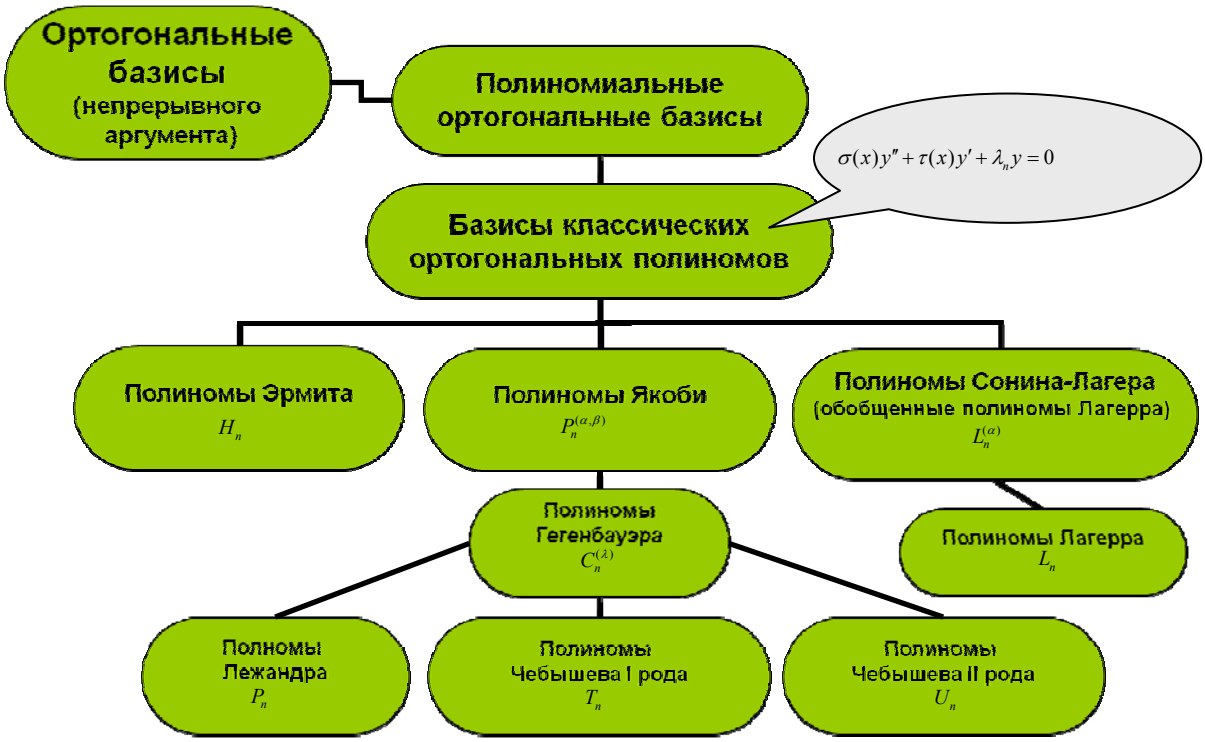
то для оператора $A^{-1}(\ )$ *также существуют рекуррентные соотношения вида*

$$A^{-1}(\varphi_{n+1}) = \overline{F_n}\left(A^{-1}(\varphi_n),\ldots,A^{-1}(\varphi_{n-p-q+1}),\varphi_{n-q+2},\ldots,\varphi_{n-d-q+1}\right),$$

где $F_n(\ldots), \overline{F_n}(\ldots)$ *– линейные формы,* $d,q,p = const \in \mathbf{N}$.

Лемма *(о суперпозиции линейных операторов)*: Если для осуществления внутриспектральных преобразований функций, соответствующих линейным операторам $F(f)$ и $G(f)$, существуют алгоритмы линейной временной сложности, то для осуществления преобразований спектра, соответствующих линейному оператору $A(f) = F\left(G(f)\right)$, также существует алгоритм линейной временной сложности.

Особое внимание уделено описанию общей схемы вывода требуемых для вычислений аналитических соотношений особого вида в случае использования определенных базисов. Соотношения были получены для множества всех базисов классических ортогональных полиномов (схема ниже) в отношении к основным, используемым на практике линейным операторам: умножение на рациональные функции, дифференцирование, интегрирование (см. ниже) и т.д.



Рекуррентные соотношения для оператора $A(\varphi_n) = \int \varphi_n dx$ *:*

$$\int P_n^{(\alpha,\beta)} dx = \frac{2(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)} P_{n+1}^{(\alpha,\beta)} + \frac{2(\alpha-\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)} P_n^{(\alpha,\beta)} -$$

$$- \frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)(n+\alpha+\beta)} P_{n-1}^{(\alpha,\beta)};$$

$$\int C_n^{\lambda} dx = \frac{C_{n+1}^{\lambda} - C_{n-1}^{\lambda}}{2(\lambda+n)}; \qquad \int L_n^{\alpha} dx = -L_{n+1}^{\alpha} + L_n^{\alpha}; \qquad \int L_n dx = -L_{n+1} + L_n; \qquad \int H_n dx = \frac{H_{n+1}}{2(n+1)};$$

$$\int T_n dx = \frac{T_{n+1} - T_{n-1}}{2n}; \qquad \int U_n dx = \frac{U_{n+1} - U_{n-1}}{2(n+1)}; \qquad \int P_n(x) dx = \frac{P_{n+1}(x) - P_{n-1}(x)}{2n+1}.$$

ГЛАВА III. Общая вычислительная схема для реализации быстрых спектральных преобразований

Здесь предложена система правил для осуществления на практике быстрых аналитических преобразований, называемая *алгеброй спектральных преобразований*. Отмечено, что разработанный подход осуществления аналитических преобразований, производимых над коэффициентами спектрального представления, лишен недостатков, присутствующих в ранних попытках решения данной теоретической задачи.

Пример: Пусть функция $f(x)$ представлена рядом ортогональных полиномов Лагерра:

$$f(x)=10L_0(x)-5L_1(x)-8L_2(x)+6L_3(x)$$

и требуется найти подобное же представление для результата применения оператора дифференцирования:

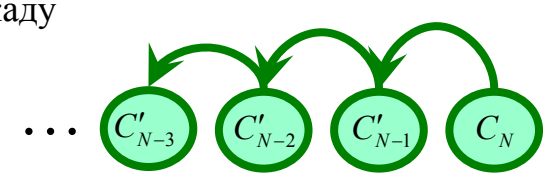
$$f'(x)=C_0^*L_0(x)+C_1^*L_1(x)+C_2^*L_2(x)+C_3^*L_3(x)\,,$$

если для полиномов Лагерра и оператора дифференцирования известно рекуррентное соотношение вида

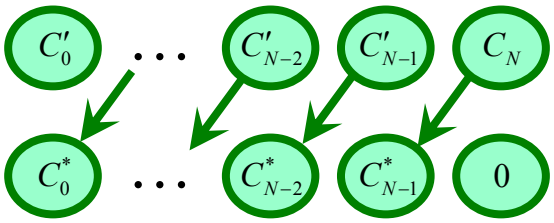
$$L'_{n+1}(x)=L'_n(x)-L_n(x)\,.$$

Последнее соотношение приводит к двум фазам вычисления

1) спектральному каскаду



2) спектральной диффузии



где начальные табличные данные $\{C_n\}$ содержат значения

$$\{C_n\}=[\,10,-5,-8,\quad 6]\,,$$

через $\{C'_n\}$ – обозначены промежуточные значения, полученные в таблице $\{C_n\}$ после проведения спектрального каскада и $\{C_n^*\}$ – новая таблица, в которой ожидаются конечные коэффициенты представления.

Согласно построенным схемам вычисления нам надо всего лишь на этапе каскада добавить значение старшего коэффициента в соседний младший, начиная с самого старшего, а затем, на этапе диффузии, просто сдвинуть все значения на одну ячейку к началу таблицу, при этом изменив знак каждого из значений на обратный. Так и поступим и после первой фазы вычислений получим промежуточные значения:

$$\{C'_n\}=[\,3,-7,-2,\quad 6]\,,$$

а после второй фазы получим искомые значения в виде

$$\{C_n^*\}=[\,7,\,2,-6,\quad 0]\,,$$

т.е.

$$f'(x)=7L_0(x)+2L_1(x)-6L_2(x)\,.$$

Если вспомнить что

$$L_0=1;\quad L_1=1-x;\quad L_2=1-2x+\frac{x^2}{2};\quad L_3=1-3x+\frac{3x^2}{2}-\frac{x^3}{6}\,,$$

то полученный результат несложно проверить. Действительно:

$$f(x) = 10L_0(x) - 5L_1(x) - 8L_2(x) + 6L_3(x) = 3 + 3x + 5x^2 - x^3;$$
$$f'(x) = 7L_0(x) + 2L_1(x) - 6L_2(x) = 3 + 10x - 3x^2.$$

ГЛАВА IV. Обобщения алгебры спектральных преобразований

В данной главе описаны основные направления дальнейшего развития теории алгебры спектральных преобразований в практических задачах. На сегодня наметилось три таких направления, каждому из которых посвящен отдельный параграф четвертой главы:

1. Быстрые нелинейные спектральные преобразования.
2. Быстрые спектральные преобразования для систем классических ортогональных полиномов дискретного аргумента.
3. Сверхбыстрые спектральные преобразование – вычисления на системах с параллельной архитектурой. Векторизация и конвейеризация вычислений.

В первом параграфе рассмотрена реализация аналитического преобразования над спектральными представлениями функций, соответствующего перемножению функций, т.е. перемножения ортогональных рядов, возведения функций в квадрат и т.д. Предложенная реализация, отличается от раннее известных, как показано там же, простотой реализации алгоритма и скоростью вычислений.

Быстрые спектральные преобразования для систем классических ортогональных полиномов дискретного аргумента являются естественным обобщением алгебры спектральных преобразований на случай использования полиномов Хана (Гана), Мейкснера, Шарлье, Кравчука и Чебышева дискретного аргумента (схема ниже). Перечисленные полиномы являются решениями разностного аналога гипергеометрического уравнения.

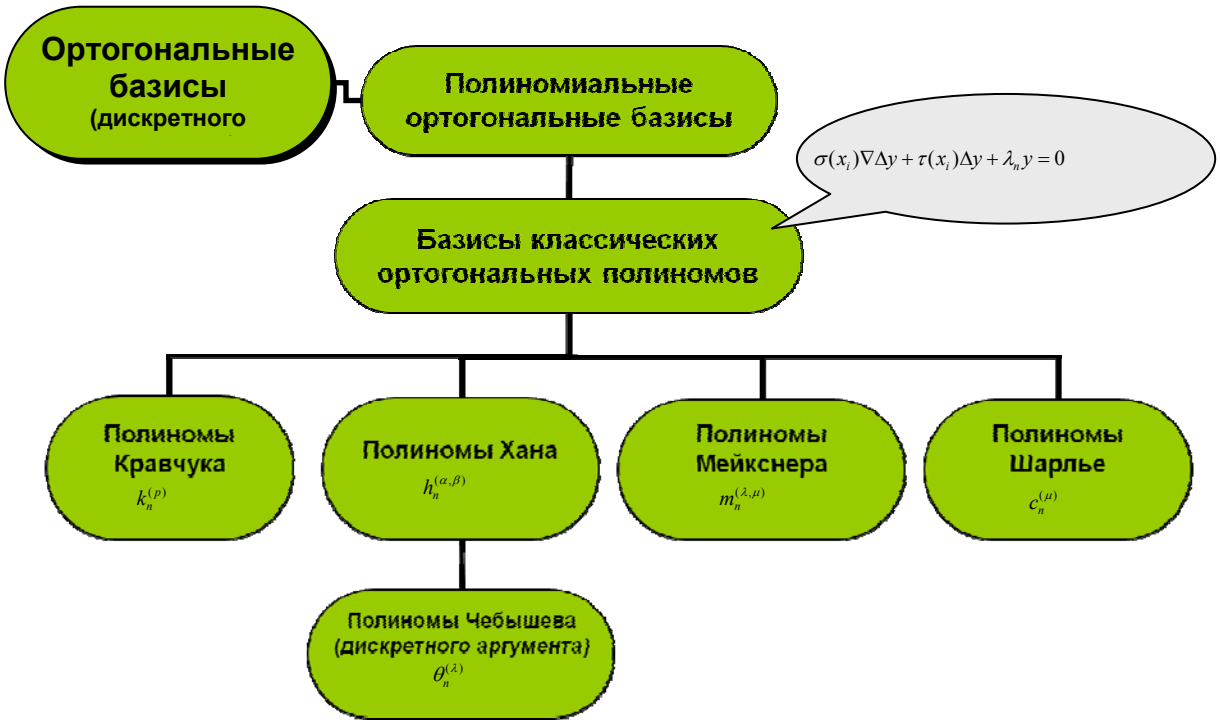


Таблица 1. Соответствие некоторых преобразований функций и операций над коэффициентами разложения по полиномам Кравчука.

N	Преобразования функций	Операции над коэффициентами разложения
1	$f = \alpha\, f_1$	$C_n = \alpha\ C_n^*$
2	$f = f_1 + f_2$	$C_n = C_n^* + C_n^{**}$
3	$f = f_1'$	$C_n \approx C_{n+1}^*$
4	$f = \int f_1 dx$	$C_n \approx C_{n-1}^*$
5	$f = x f_1$	$C_n = \gamma_{n+1} C_{n+1}^* + \beta_n C_n^* + \alpha_{n-1} C_{n-1}^*$

В третьем параграфе, посвященном сверхбыстрым спектральным преобразованиям, рассматривается возможность и дается оценка степени ускорения вычислений, при реализации спектральных преобразований на системах с параллельной архитектурой. Показано, что при этом применяются два принципа распараллеливания: векторизации и конвейеризации вычислений. Пример векторизации приведен ниже.

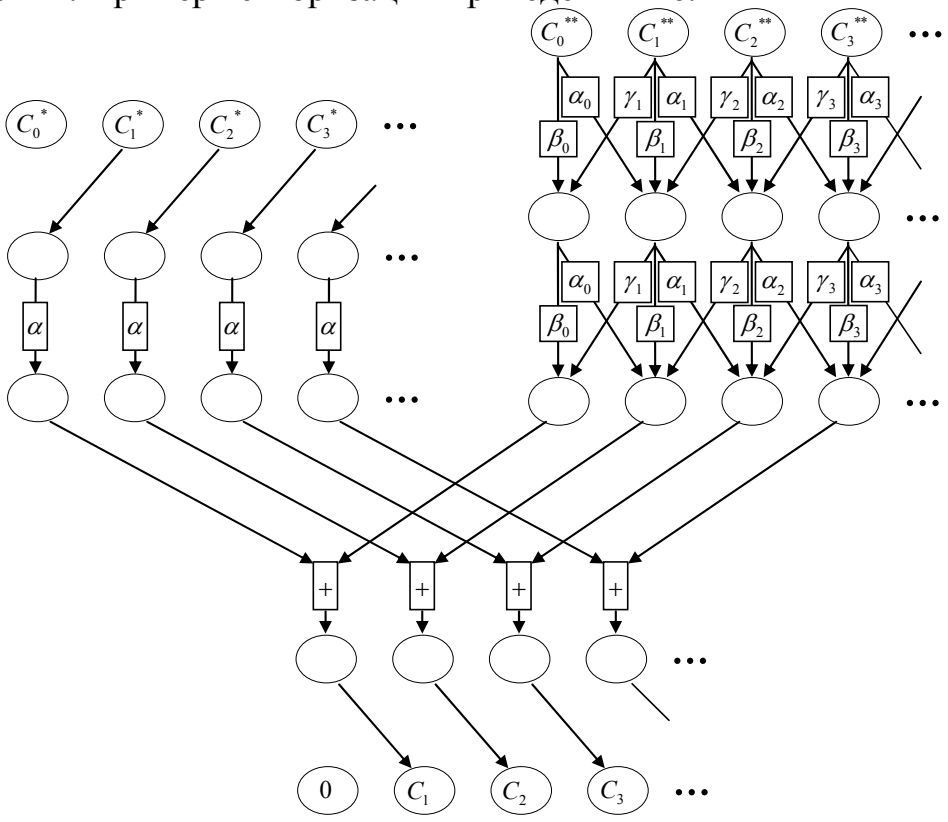


Рис. 1 Схема распараллеливания спектрального вычисления для оператора

$$f(x) = \int_a^b \left(\alpha f_1' + x^2 f_2\right) dx \approx \sum_{t=0}^{x} \left(\alpha\, \Delta f_1(t) + x^2 f_2(t)\right)$$

(функции представлены рядами по полиномам Кравчука).

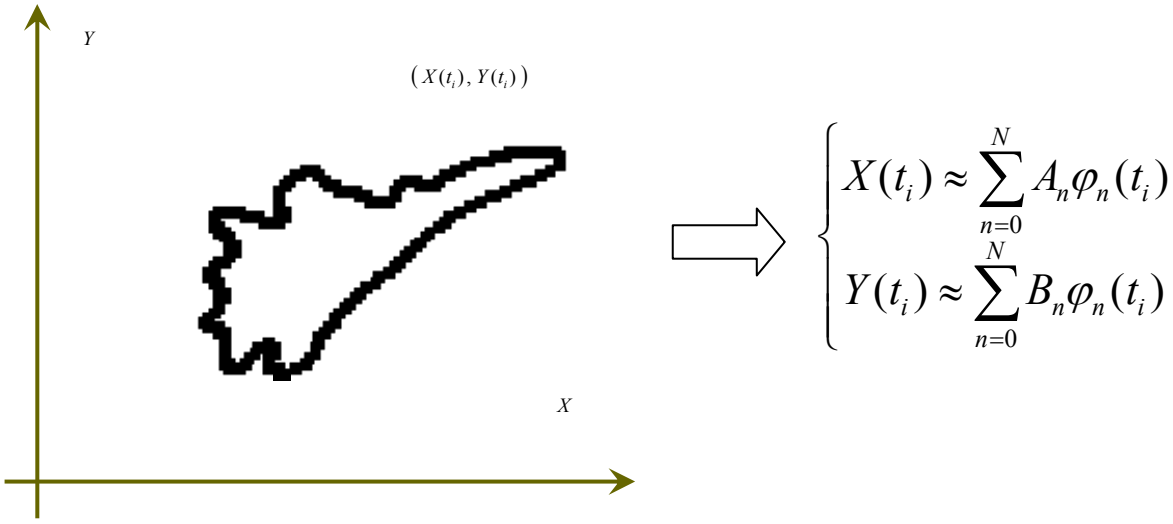
ГЛАВА V. Применение быстрых спектральных преобразований

Данное место диссертационной работы посвящено описанию решения двух практических задач, успешно разрешенных при помощи разработанного математического аппарата аналитических преобразований сигналов. Одна из задач связана с применением алгебры спектральных преобразований в распознавании визуальных образов, другая – с поиском структурных закономерностей в генетических последовательностях. Ниже дано краткое описание каждой их этих практических задач с последующим описанием предложенных способов решения, основанных на применении аналитических преобразований спектральных представлений функций.

Задача 1. Применение в задачах распознавания визуальных образов

Известно, что часто для распознавания визуального образа и идентификации объектов различной природы (см. рисунок ниже) достаточно располагать только контуром объекта, игнорируя цветовую составляющую визуального образа.

В таких случаях распознавание визуального образа можно рассматривать лишь как распознавание контура объекта. Далее считаем, что контур объекта фактически является замкнутой кривой на плоскости, которая может быть представлена в параметрическом виде как пара функций: $\left(X(t_i), Y(t_i)\right)$.



Каждую из упомянутых функций, в свою очередь, можно представить в спектральном виде как ограниченный ортогональный ряд по некоторому базису (см. рисунок выше). Причем, чем короче взят ортогональный ряд, тем более гладкое приближение контура мы получим при аппроксимативном восстановлении функций $\left(X(t_i), Y(t_i)\right)$.

Как уже упоминалось выше, аналитическая теория является хорошо разработанной областью математики и поэтому возникает устойчивое желание попытаться использовать в задаче распознавания контуров соотношений из аналитической геометрии. Сегодня для контуров существует множество аналитически описанных инвариантов, т.е. количественных характеристик, неизменных к различным преобразованиям фигуры: перемещению, вращению, сжатию и т.д. Однако многие из них, несмотря на ряд своих «аналитических» достоинств, не получили распространение, т.к. оказались практически бесполезными в реальных задачах. Один из примеров – аффинная длина дуги кривой:

$$l = \int_a^b \sqrt[3]{x''y' + x'y''}\, dt\,.$$

Подобные аналитические оценки до сих пор оставались по большей части теоретическими, ввиду точностной и временной сложности вычислений производных зашумленных функций высоких порядков. В представляемой диссертации это проблема была успешно решена.

Более простые оценки геометрических признаков также оказались легко вычисляемыми через спектральное представление контура. Например, следующие признаки:

Координаты центра масс контура:

$$x_c = \int_0^T x(t)\, dt\,,$$

$$y_c = \int_0^T y(t)\, dt\,.$$

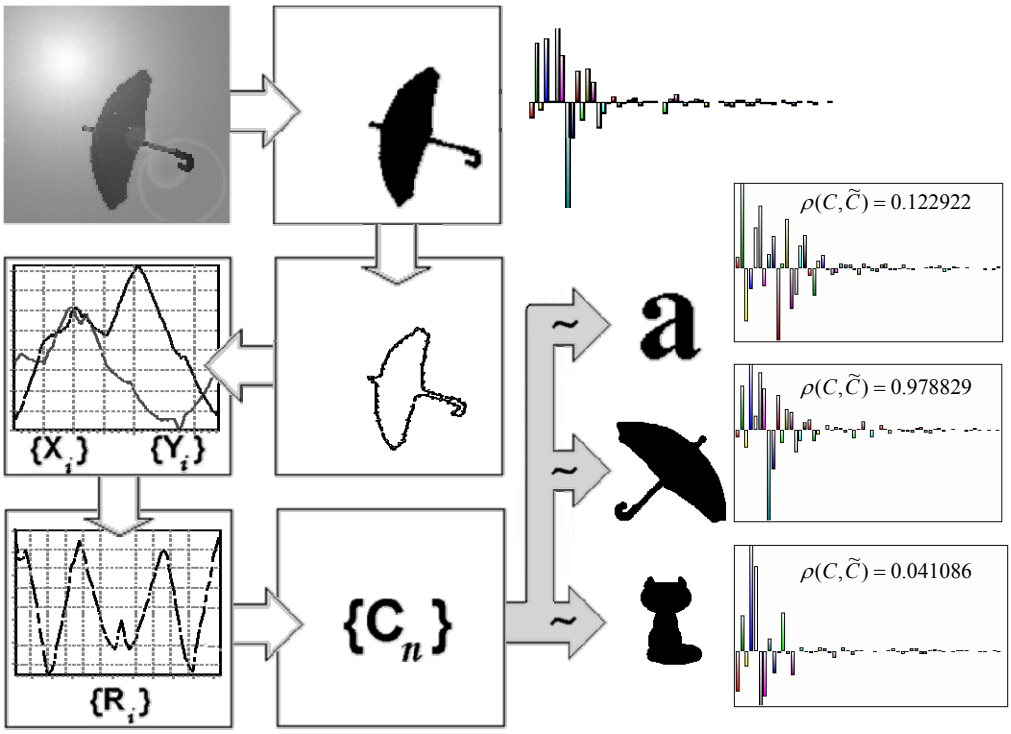
Площадь, ограниченная контуром:

$$S = \int_0^T y(t)x'(t)dt = \int_0^T x(t)y'(t)dt\,.$$

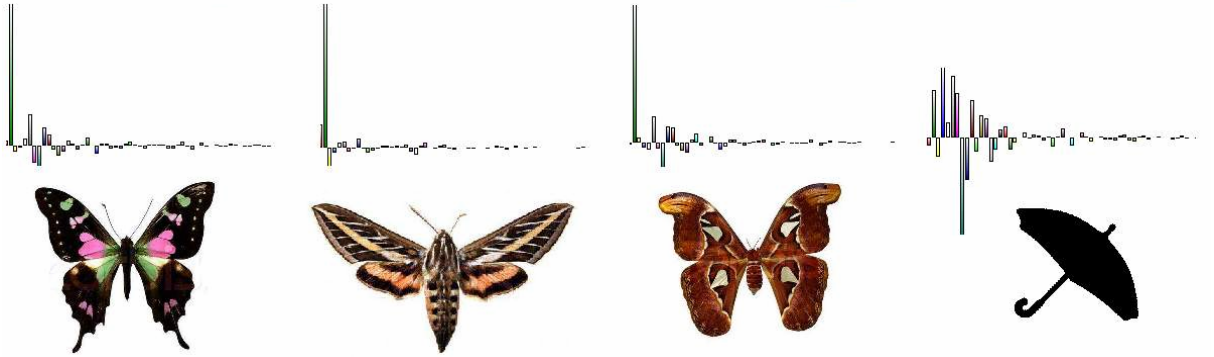
Периметр контура:

$$l = \int_0^T \sqrt{x'^2(t) + y'^2(t)}\, dt\,.$$

Автором диссертационной работы также был предложен ряд новых признаков и инвариантов контуров, рассчитываемых в пространстве коэффициентов, таких как инварианты к выбору начальной точки обхода контура и др.



Выше представлена общая применяемая схема распознавания контуров объектов. Здесь в качестве количественной оценки подобия используется коэффициент корреляции рядов признаков, причем значениям близким единице соответствует высокая вероятность подобия фигур.



Из расположенных выше примеров видно, что предложенный метод распознавания приводит к оценкам, адекватным действительности. В диссертационной работе также подробно рассмотрены другие случаи и примеры.

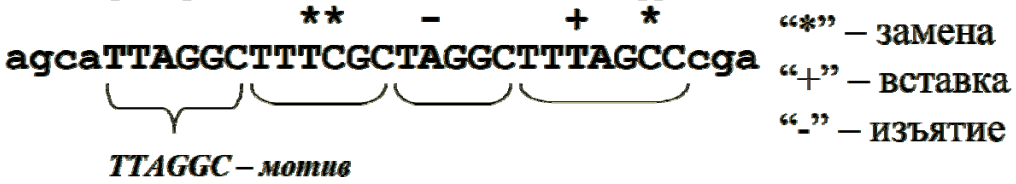
Задача 2. Применение в задачах биоинформатики

На основе применения разработанного математического аппарата в диссертационной работе была реализована алгоритмическая схема для решения одной из задач биоинформатики – задачи поиска тандемных повторов.

После того, как в молекулярной биологии стало возможно решение задачи секвенирования, целью которой является однозначное описание линейной структуры ДНК молекулы в виде текстовой последовательности из четырех букв (**A** – аденина, **T** – тимина, **G** – гуанина, **C** – цитозина), широко начали развиваться методы анализа генетических последовательностей, производимые посредством ЭВМ.

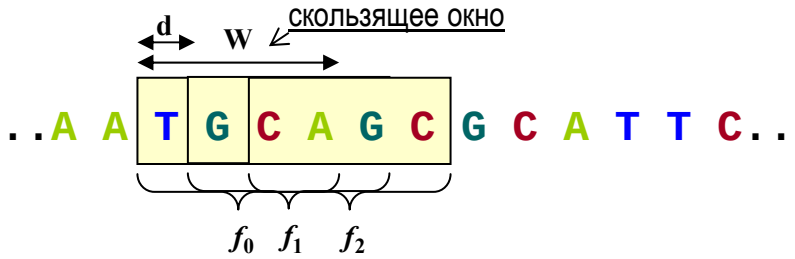
Известно, что ДНК молекулы подвержены многочисленным мутациям. Множественное последовательное дублирование части ДНК (*тандемные повторы*) – одни из мутаций, которые можно обнаружить в ходе численного анализа генетической последовательности:

ДНК подвержены также, так называемым, точечным мутациям, приводящим к трем видам нарушений в генетических последовательностях: замене/вставке/изъятию одной из букв. Присутствие подобных искажений в изначально точных повторах способно сильно усложнить их обнаружение. В случае применения средств динамического программирования каждое нарушение текстового фрагмента приводит к многократному увеличению количества проверок на стадии анализа ДНК фрагмента.



Ввиду этого в данной проблематике получили развитие методы предварительной оценки текста, быстро обнаруживающие вероятные тандемные повторы и лишь потом подтверждающие их присутствие (или отсутствие) на этапе длительного анализа.

Вообще, работа с текстовыми фрагментами в диссертационной работе не ведется вплоть до этапа анализа – все действия производятся с функциями особого вида, построенных как массивы некоторых количественных оценок последовательных фрагментов генетической последовательности, обычно называемых *профилями*.



Автором впервые предложено оценивать вероятность присутствия тандемных повторов по выраженности «периодичности» соответствующего фрагмента профиля, причем оценка осуществляется средствами разработанного математического аппарата.

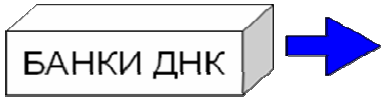
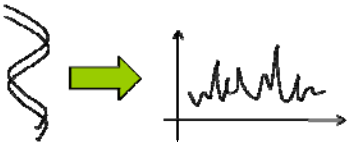
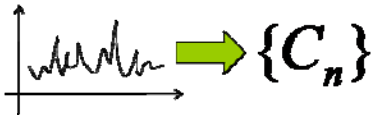
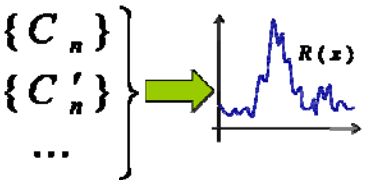
Результатом тестовых запусков программного комплекса стало обнаружение множества действительных тандемных повторов генома бактерии Staphylococcus haemolitycus. Среди множества выявленных подтвержденных фрагментов имеются ряд повторов, обнаруженных ранее. Однако, как и ожидалось на ранних стадиях основания проекта, были выявлены новые фрагменты, ранее не известные. Как правило, новые обнаруженные повторы искаженны мутациями и требуют на стадии проверки (стадия 7) приме-

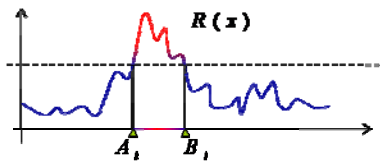
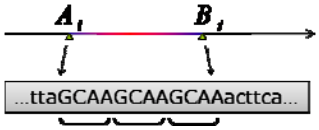
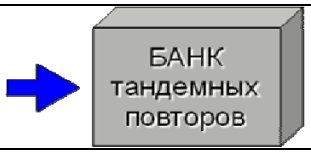
нения сторонних программ множественного выравнивания (Clustal и др.). Пример одного из таких, ранее не известных повторов представлен ниже.

```
                 ****      *     *   * *     *   *   *     ***
2586040:  -ACATAC--CCAG--GAGGA-CTTTTA-CATGA--CCCAAAA
2586073:  TACATTTTA-CACT-GAAAAGCGA--A-GA-GAT--TCAA-G
2586106:  --CATTATT-GAAT--ATT--CTG-TAA-AGGATGACCAAAG
2586139:  TTC-TAAAAATATTTTAACAACGGTATTTA--AT-GT-----
2586172:  --C-TAA-TGGAAA--ATCAACGAACA-GAATATATTCAA--
2586205:  --CAAAA---GAAT--ATGAACGAACA-GAAAACCGACAA-G
2586238:  T-CA-AC---GAAA---TG-GCTATTATGAGCGCAGCCAAAG
```

Обнаруженный фрагмент тандемных повторов
(после множественного выравнивания)

Ниже представлена общая применяемая схема поиска тандемных повторов в генетических повторах. В качестве количественной оценки «меры периодичности» профиля используется определенный функционал, зависящий от производных аппроксимации профиля высоких порядков.

№	Описание стадии процесса	Схематическое описание
1.	Загрузка фрагмента изучаемого генома из сети Internet (базы данных Genbank, Emboss, Entrez и др.)	
2.	Перевод текстового генетического фрагмента в функциональное пространство – вычисление «профиля» фрагмента (GC-состав)	
3.	Представление профиля в виде ортогонального ряда – перевод в пространство коэффициентов разложения $\{C_n\}$	
4.	Осуществление преобразований в пространстве коэффициентов разложения, соответствующих дифференцированию профиля	
5.	Построение функционала на основе сличения профиля с производными профиля высоких порядков. Вид функционала основан на применении соотношений корреляционного анализа	

6.	Считая значения построенного функционала оценкой присутствия периодичности в профиле, сохраним интервалы, на которых функционал принимает значения выше заранее определенной пороговой величины	
7.	Проверка на наличие тандемных повторов во фрагменте генетической последовательности, соответствующем максимуму функционала оценки периодичности профиля	
8.	После успешного множественного выравнивания осуществляется занесение подтвержденных тандемных повторов в банк данных	

В конце главы приведена сравнительная оценка полученных результатов с результатами исследований, полученных в данных областях науки ранее.

ЗАКЛЮЧЕНИЕ

В качестве достаточного условия реализуемости быстрых спектральных преобразований предложено условие существования рекуррентных соотношений особого вида. Данное предположение сформулировано автором диссертационной работы в виде теоремы и доказано.

Сформулирована и доказана теорема о существовании рекуррентных соотношений особого вида для линейного оператора, если для обратного к нему оператора существует хотя бы одно подобное соотношение.

Сформулирована и доказана лемма о существовании рекуррентных соотношений особого вида для суперпозиции двух линейных операторов, если подобные соотношения существуют для каждого из них.

Рекуррентные соотношения особого вида получены для множества всех базисов классических ортогональных полиномов в отношении к основным, используемым на практике линейным операторам (умножение на рациональные функции, дифференцирование, интегрирование и т.д.).

Предложена система правил для осуществления на практике быстрых аналитических преобразований над спектральным представлением функций, названная *алгеброй спектральных преобразований*. Вычисления, производимые над коэффициентами спектрального представления согласно двум основным из введенных правил, названы *спектральным каскадом* и *спектральной диффузией* коэффициентов.

Реализован ряд программных продуктов, осуществляющих быстрые аналитические преобразования сигналов на основе соотношений алгебры спектральных преобразований.

Применение разработанных программ в решении ряда практических задач показало действительную практическую ценность возможности осу-

ществления на практике быстрого аналитического преобразования функций (сигналов) в задачах обработки данных произвольной природы.

Зарегистрирована программа для ЭВМ *"SpectralRevisor"*, осуществляющая спектральный анализ данных на предмет обнаружения и локализации неточных периодов в сигналах.

ПРИЛОЖЕНИЯ

Приведенные в диссертационной работе приложения содержат рекуррентные соотношения особого вида, найденные для систем классических ортогональных полиномов непрерывного и дискретного аргументов. Данные соотношения приведены в табличной форме и требуются для реализации спектральных преобразований на практике. Отметим, что рекуррентные соотношения для полиномов Якоби и Гегенбауэра, соответствующие операторам дифференцирования и интегрирования, получены в рамках выполнения диссертационной работы.

ВЫВОДЫ

1. Создана система правил для аналитических преобразований сигналов, представленных рядами через классические ортогональные полиномы – *алгебра спектральных преобразований*.
2. Задача контурного распознавания визуальных объектов разрешена на основе аналитического вычисления инвариантных признаков контуров.
3. Задача поиска тандемных повторов в геномах разрешена на основе аналитической оценки степени «периодичности» профилей.
4. Применение в задачах обработки данных правил алгебры спектральных преобразований в пространстве коэффициентов разложения способно в значительной степени улучшить и ускорить результаты вычислений.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Ф.Ф.Дедус, Л.И.Куликова, А.Н.Панкратов, Р.К.Тетуев. Классические ортогональные базисы в задачах аналитического описания и обработки информационных, сигналов, Москва, издательский отдел факультета ВМК им.Ломоносова, 2004.
2. Ф.Ф. Дедус, Л.И. Куликова, С.А. Махортых, Н.Н. Назипова, А.Н.Панкратов, Р.К. Тетуев Аналитические методы распознавания повторяющихся структур в геномах. Доклады Академии Наук 2006, том 411, №5, с. 599-602.
3. Ф.Ф. Дедус, Л.И. Куликова, С.А. Махортых, Н.Н. Назипова, А.Н. Панкратов, Р.К. Тетуев, Распознавание структурно-функциональной организации генетических последовательностей, ТРУДЫ ВМК МГУ 2007, №2.
4. R. K. Tetuev, Recognition of Lines Detected in the Image Plane on the Basis of the Generalized Spectral-Analytical Method, Pattern Recognition and Image Analysis, Vol. 15, No. 2, 2005, pp. 334–337.
5. R. K. Tetouev, Contour Recognition Based on Spectral Methods. Solution of the Problem of Choice of the Start-Point, Pattern Recognition and Image Analysis, Vol. 17, No. 2, 2007, pp. 227–235.
6. Р.К. Тетуев, Ф.Ф. Дедус, Классические ортогональные полиномы. Применение в задачах обработки данных, препринт, М.: 11-й ФОРМАТ, 2007, 60 с.
7. Р.К. Тетуев “Аналитическое описание зашумленных исходных сигналов по функциям Сонина-Лагерра и получение их первых производных” Математические Методы Распознавания Образов-11, Пущино, 20-26 ноября, 2003.
8. Ruslan Tetuev “About Curves Recognition Based on Generalized Spectral-Analytical Method” 7-th International Conference on Pattern Recognition And Image Analysis: New Information Technologies, St. Petersburg, Russia October 18-23, 2004.

9. Р. К. Тетуев “Вычисление некоторых геометрических характеристик плоских кривых на основе спектральных методов” Математические Методы Распознавания Образов-12, Москва, 20-26 ноября, 2005.
10. Ruslan Tetuev, Florencz Dedus, Lyudmila Kulikova, Sergey Makhortikh, Anton Pankratov, Nafisa Nazipova “Analytical methods in problems of recognition the structural and functional organization of genetic sequences”, The 2006 BGRS (Bioinformatics of the Genome Regulation and Structure) International Summer School for young scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics", Novosibirsk, Russia July 12-15, 2006.
11. Р. К. Тетуев, Ф. Ф. Дедус, Л. И. Куликова, С. А. Махортых, Н. Н. Назипова, А. Н. Панкратов “Аналитические методы в проблемах распознавания структурно-функциональной организации генетических последовательностей” Математическая Биология и Биоинформатика, Пущино, 9-15 октября 2006.
12. Р.К. Тетуев, Ф.Ф. Дедус, Н.Н. Назипова, С.А. Махортых, Л.И. Куликова, А.Н. Панкратов, М.М. Ольшевец Свидетельство Роспатента об официальной регистрации программы для ЭВМ №2007611639 «Спектральный анализ данных, поиск неточных периодов в сигналах «SpectralRevisor».