Обобщенный спектрально-аналитический метод и его приложения

Дедус Ф. Ф.
ffdedus@impb.ru
Пущино, Институт математических проблем биологии РАН

Поиски новой вычислительной технологии для эффективности обработки данных привели к следующему заключению. Принятый математический метод для обработки данных должен быть комбинированным, численно-аналитическим. Он должен сочетать сильные стороны как числовых расчетов на ЭВМ, так и аналитических преобразований и выводов.

Создание такого комбинированного численно-аналитического метода началось свыше 20 лет тому назад. Уже тогда стало ясно, что, несмотря на универсальность цифровых расчетов, не всегда можно обеспечить требуемую точность обработки. Это объяснялось тем, что резкое возрастание объемов информационных массивов данных и увеличение сложности вычислений, в особенности, в ситуациях счетной неустойчивости, а также необходимость многократной числовой обработки исходных данных неизбежно приводило к росту не учитываемых ошибок округления. Поэтому гарантировать выполнение вычислений с требуемой точностью не представлялось возможным.

В результате решения многих и разнообразных прикладных задач окончательно сложилась следующая структура метода.

1. Поступающие в ЭВМ на обработку исходные массивы данных должны быть прежде всего описаны аналитически с заданной, равномерной погрешностью.
2. Результаты аналитического описания следует подвергать аналитическим преобразованиям и выводам в общем виде с целью получения формул, обеспечивающих вычисление нужных оценок, параметров и характеристик.

Затем выведенные соотношения вводятся программно в ЭВМ, которая обеспечивает получение конкретных количественных оценок и характеристик.

Обстоятельный анализ и поиск подходящего математического аппарата для аналитического описания исходных массивов данных, обеспечивающий равномерное приближение, привел к необходимости воспользоваться ортогональными разложениями на основе широкого применения полных ортонормированных базисов из числа классических полиномов и функций непрерывного и дискретного аргументов.

Теория классических ортогональных базисов (иногда их называют специальными функциями математической физики) есть обобщение тео-

рии рядов Фурье на алгебраические ортогональные полиномы непрерывного и дискретного аргументов. Эти базисы хорошо изучены.

Следует отметить, что введение процедур адаптации в процесс вычисления коэффициентов разложения, обеспечивает не только выполнение условий $N = N_{\min}$, но и в значительной степени способствует регуляризации условий описания данных по А. Н. Тихонову [1] в случае решения обратных некорректно поставленных задач.

В конце 19-го столетия было опубликовано условие замкнутости Ляпунова-Стеклова для алгебраических ортогональных базисов:

$$\sum_{n=0}^{\infty} A_n^2 = \int_a^b [f(x)]^2 \rho(x) dx.$$

Выполнение данного равенства для классических ортогональных базисов соответствует тому, что применяемые при аналитической аппроксимации ортогональные базисы являются полными (замкнутыми) системами. В этом случае члены отрезков ортогональных рядов являются линейно независимыми. А так как вся информация об аппроксимируемом сигнале сосредоточена в коэффициентах разложения, то появляется возможность полную обработку описываемых сигналов проводить в пространстве коэффициентов разложения.

Получены таблицы соотношений между исходными коэффициентами и коэффициентами разложений, соответствующие искомым оценкам и характеристикам.

Одним из направлений, требующих дополнительных исследований, является необходимость аналитического описания данных, получаемых с установок ядерно-магнитного резонанса. Массивы данных, соответствующие одному опыту здесь составляют от 2000 до 5000 отсчетов.

Потребовалось проведение специальных исследований классических ортонормированных базисов непрерывного аргумента для определения допустимой глубины разложения при описании больших массивов данных. Установлено, что использование ортонормированных базисов Чебышева, Лагерра и Эрмита позволяет вычислять до 3000–5000 коэффициентов разложения [2]. При этом точность аналитического описания подобных массивов данных находится в допустимых пределах. Было показано также, что вполне реально выполнять сверхглубокие разложения с вычислением до 10 000 членов ряда. Это весьма серьезное достижение.

К началу XXI века в молекулярной биологии и генетике произошли радикальные изменения. Они последовали после того, как в 1953 году Уотсон и Крик [3] открыли двойную спираль ДНК. Возникла актуальная необходимость создания методов расшифровки аминокислотных и нуклеотидных последовательностей и определения пространственных

структур указанных полимеров. В настоящее время состояние дел в этих научных областях таково, что для осмысления и обработки накопленных и лавинообразно поступающих данных необходимо активное участие математиков-прикладников, имеющих опыт распознавания сигналов в громадных массивах данных. Суммарные объемы первичных экспериментальных данных только по молекулярно-генетическому уровню организации живого превышают сотни терабайт (Тера = 10^{12}). Например, длина генома человека составляет более 3 миллиардов пар оснований. При его расшифровке получены данные объемом в десятки терабайт о физических, цитогенических картах, их нуклеатидных последовательностях, локализации генов, мутациях. В настоящее время выявлено более 1,5 миллиона мутаций, по которым геномы людей отличаются друг от друга. Первые результаты применения метода в таких грандиозных исследованиях внушают оптимизм. Наиболее важные проблемы, стоящие перед нами при решении, в частности, задач распознавания структурно-функциональной организации генетических последовательностей это разработка алгоритмов спектрального сжатия объемов данных и развитие методов решения сложных задач распознавания в пространстве коэффициентов разложения. В следующих докладах сотрудники нашей группы сформулируют актуальные задачи биоинформатики, которые могут быть решены на основе обобщенного спектрально-аналитического метода в пространстве коэффициентов разложения. Кроме того, будут доложены результаты теоретических разработок, обеспечивающих новые направления исследований в пространстве коэффициентов разложения для решения задач биоинформатики, и других важных прикладных исследований.

Работа поддержана РФФИ, проекты №,06-01-08039 и № 07-01-00564.

Литература

[1] *Тихонов А. Н.* О регуляризации некорректно поставленных задач // ДАН СССР. — 2006. — Т. 153, № 1.

[2] *Дедус Ф. Ф., Махортых С. А., Устинин М. Н., Дедус А. Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов. — Москва: Машиностроение, 1999. — 356 с.

[3] *Watson J. D., Crick F. H. C.* Molecular structure of nucleic acids // Nature. — 1953. — Т. 171, — С. 738–740.