УДК 575.852

Ф. Ф. Дедус, Л. И. Куликова, С. А. Махортых, Н. Н. Назипова, А. Н. Панкратов, Р. К. Тетуев

РАСПОЗНАВАНИЕ СТРУКТУРНО-ФУНКЦИОНАЛЬНОЙ ОРГАНИЗАЦИИ ГЕНЕТИЧЕСКИХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ[1]

(*кафедра математических методов прогнозирования факультета ВМиК, e-mail: radja@impb.ru*)

Данная работа посвящается преодолению серьезных проблем, которые встречаются на пути разработки новых подходов в задачах распознавания структурно-функциональной организации генетических последовательностей.

На рубеже начала XXI в. за относительно короткий срок в молекулярной биологии и генетике произошли изменения, повлиявшие на развитие связанных с биологией областей знаний. Исходным событием явилось осуществленное в 1953 г. Уотсоном и Криком [1] открытие двойной спирали ДНК. За этим последовало создание методов расшифровки аминокислотных и нуклеотидных последовательностей и определения пространственных структур этих биополимеров. Появились новые методы исследования, сделано множество открытий. Началась массовая расшифровка геномов, увенчавшаяся таким выдающимся достижением, как расшифровка генома человека. Были разработаны эффективные методические подходы, гарантирующие получение фундаментальных знаний о молекулярно-генетическом, клеточном, организменном уровнях организации жизни. Реализована трансформация этих знаний для нужд прикладных, научных отраслей.

Следует отметить, что вышеупомянутый технологический прорыв в молекулярной биологии и генетике был обусловлен участием в исследованиях ученых-физиков. В настоящее время состояние дел в этих науках таково, что для осмысления и обработки накопленных и лавинообразно образующихся новых данных необходимо участие математиков, особенно таких, которые имеют опыт распознавания сигналов в больших массивах данных.

Суммарные объемы первичных экспериментальных данных только по молекулярно-генетическому уровню организации жизни превышают сотни терабайт (T = 1 тера единиц = 10^{12} единиц). В результате расшифровки нуклеотидных последовательностей в молекулярной биологии и генетике за последние 20 лет произошел информационный скачок. Объемы получаемых данных поражают воображение.

Например, геном человека состоит более чем из 3 млрд пар оснований и содержит более 30 тыс. генов. При его расшифровке получены данные объемом в десятки терабайт о физических и цитогенетических картах хромосом, их нуклеотидных последовательностях, локализации генов, мутациях. На сегодняшний день выявлено не менее 1,5 млн мутаций, по которым геномы людей отличаются друг от друга.

Расшифрованы структуры геномных ДНК тысяч вирусов, десятков бактерий, геномы дрожжей, дрозофилы, ряда животных и растений. Расшифрованы аминокислотные последовательности миллионов белков и более 15 тыс. их пространственных структур. Технология ДНК-чипов позволяет количественно измерять экспрессию десятков тысяч генов одновременно в отдельной клетке. Разворачиваются исследования по протеомике, направленные на изучение белков и их взаимодействий в живых организмах. Они основаны на применении двумерных гель-электрофорезов, высокоэффективной жидкостной хроматографии и масс-спектроскопии. Экспериментальные данные накапливаются при изучении разнообразия геномов человека и животных. Не менее мощные массивы экспериментальных данных накапливаются в таких классических направлениях биологии, как зоология, ботаника, систематика, экология.

Современная молекулярная биология стала источником беспрецедентно больших объемов экспериментальных данных, осмысливание которых невозможно без привлечения современных информационных технологий и эффективных математических методов распознавания (анализа) данных и моделирования (прогнозирования поведения) биологических систем и процессов.

[1] Работа выполнена при финансовой поддержке грантов РФФИ № 04–07–90402, 06–01–08039, 06–07–89274, 06–07–89303.

Этим объясняется возникновение новой науки — **биоинформатики**. Основными объектами исследований биоинформатики являются: биологические макромолекулы — ДНК, РНК, белки; фундаментальные генетические процессы — репликация, транскрипция, трансляция, репарация и др. В последнее время появились публикации по моделированию работы генных сетей, которая обеспечивает выполнение всех функций организмов.

Биоинформатика относится к числу высоких технологий современной биологии. Ее основные задачи — разработка информационно-компьютерных и теоретических основ молекулярной биологии, молекулярной генетики, генетической и белковой инженерии, генетики и селекции, а также биотехнологии, медицинской генетики, генодиагностики, генотерапии — словом, тех наук, благодаря выдающимся достижениям которых биология превратилась в одну из лидирующих наук грядущего столетия.

Биоинформатика занимает в современной биологии ключевую и исключительно важную позицию. Ее предметом является исследование биологических систем на всех уровнях их организации — субклеточном, клеточном, надклеточном и организменном.

Обсуждаемая работа посвящена одной из задач биоинформатики — изучению структурно-функциональной организации геномов. Геном биологического организма — библиотеку всех его генов — можно представить в виде линейной символьной последовательности, заданной над алфавитом, состоящим из четырех букв. Каждая буква алфавита определяет тип мономеров, из которых состоят молекулы ДНК. Известно, что геномы сильно избыточны в том смысле, что доля участков, кодирующих белки и некоторые другие важные для организма продукты, ничтожно мала по сравнению с общей длиной генома. Видимо, в тех частях генома, которые не кодируют известные на сегодняшний день биологам продукты, располагаются участки, регулирующие различные процессы.

Однако до сегодняшнего дня достоверно неизвестно, какую функцию играют некодирующие части геномов. Есть предположения, что эти участки задействованы в эволюционном развитии организма и что именно там содержатся резервы для изменчивости.

С избыточностью текстов сильно связано понятие повторяемости. В литературе давно и широко обсуждаются [2] механизмы и характерные места появления простых и сложноорганизованных повторов в геномах, поддержания их гомогенности, характер отбора повторов в нетранскрибируемой ("молчащей") части генома, скорость дивергенции внутри повторов конкретного вида, внутри таксономических групп.

Одним из методов исследования статистических свойств полных геномов и их фрагментов является измерение количества информации (сложности) и связанной величины — избыточности. С практической точки зрения величины сложности и избыточности могут быть использованы для выделения периодичностей и статистически однородных зон при анализе протяженных символьных последовательностей.

Для оценки количества информации в геномных последовательностях использовались различные методы: классическая мера информации, алгоритмическая сложность Колмогорова [3], энтропия Шеннона [4], сложность в смысле алгоритмов сжатия (например, алгоритма Лемпеля–Зива [5]). Применялась также такая характеристика, как длина наибольшей строго повторяющейся подпоследовательности [6]. Была продемонстрирована внутренняя неоднородность полных геномов. Наблюдаемые значения количества информации фрагментов геномов связываются с их функциональной ролью.

Периодически повторяющиеся фрагменты символов самой разной длины, так называемые тандемные повторы, в последовательностях ДНК и белков встречаются достаточно часто. Иногда с ними связывают конкретные функции (связывание субстрата, межбелковые взаимодействия), пространственные структуры или характерные свойства, например такие, как гибкость белков или отдельных участков ДНК. Короткие тандемные повторы от 2 до 6 нуклеотидов — микросателлиты, видимо, используются в регуляции генов или формируют районы полиморфизма длины в геноме, на основании которых происходит идентификация личности или родства. Более длинные повторы, получившие название VNTRs (variable number of tandem repeats), представляют потенциальные сайты рекомбинации. Тандемные повторы можно отнести также к структурной периодичности символьных последовательностей. Структурная периодичность в ДНК и белках подвержена эволюционной дивергенции (изменчивости) и со временем теряет вид совершенных повторов. Таким образом, не всегда возможно однозначно выделить причины возникновения несовершенной периодичности: были ли это дупликации отдельных фрагментов или конвергенция последовательностей, направляемая отбором на достижение лучших функциональных свойств. Особенно в последнем случае периодичность может быть выделена не только благодаря преобладающим символам в позициях периодической единицы, но и на основа-

нии практически полного отсутствия символов конкретного типа в тех или иных позициях. Поэтому часто периодичность может быть выделена только в результате анализа статистики распределения символов по позициям периода.

Для поиска структурной и статистической периодичности в символьных последовательностях применяются методы, основанные на фурье-анализе, автокорреляционных функциях, теории информации. Эти методы требуют значительных затрат времени вычисления и не позволяют проводить поиск периодичности, искаженной вставками и делециями символов. Поскольку анализ последовательностей геномов и белков различных организмов и осмысление полученной информации происходят гораздо медленнее, чем просто накопление расшифрованных генетических последовательностей, необходимость в разработке простых и высокоэффективных методов и алгоритмов их распознавания остается по-прежнему весьма актуальной.

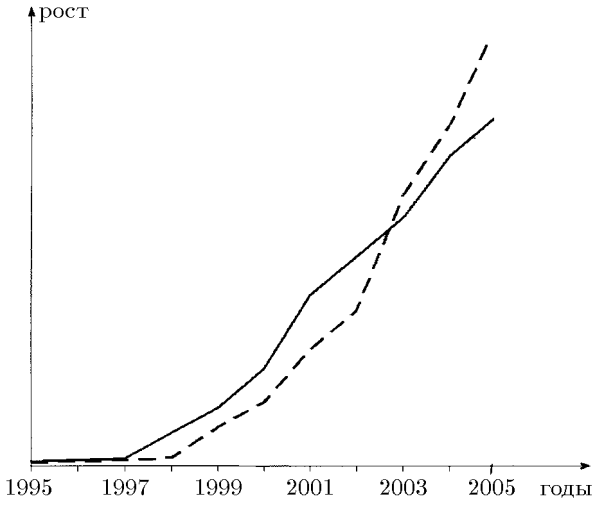
Упомянутые методы хорошо выявляют как микросателлитные (от 2 до 6 символов) повторы, так и более длинные тандемные повторы. Однако автокорреляционные функции обладают слабой чувствительностью к несовершенным, сильно размытым повторам, поэтому область их применения ограничена структурной или слабо размытой периодичностью. Применение фурье-анализа связано с разложением символьной последовательности в ряды, отражающие порядок следования символов одного типа в исходной последовательности, что приводит к снижению чувствительности к более длинным периодам, если на них есть периодичности в распределении символов одного конкретного типа. Лучшей чувствительностью к тандемным повторам обладает метод информационного разложения (Information Decomposition — ID) [7], основанный на вычислении взаимной информации между искусственной периодической последовательностью символов и анализируемой последовательностью. В основе этого метода лежит анализ распределения символов в матрице периодичности размерности $n \times m$ (n — число символов алфавита, m — длина тестируемого периода). Все перечисленные методы требуют значительных затрат времени вычисления и не позволяют проводить поиск периодичности, искаженной вставками и делециями символов. В настоящее время существуют и методы поиска тандемных повторов со вставками и делециями, но лежащие в их основе принципы либо не позволяют искать сильно размытые тандемные повторы [8], либо требуют предварительного задания маски совпадающих и несовпадающих позиций повторов [9].

Следует подчеркнуть, что основные этапы в истории становления и развития генетики и информатики хронологически в высокой степени совпадают. Также известно, что ряд важных открытий, сделанных во многих областях биологии, особенно в генетике, был бы невозможен без применения современных информационных технологий.

Тенденции в развитии современной вычислительной технической базы свидетельствуют о резком, экспоненциальном росте основных вычислительных характеристик, известном как "закон Мура" [10]: каждые два года возможности основных вычислительных ресурсов (количество микрочипов, быстродействие, объем памяти и т.д.) удваиваются. Это послужило основанием для ряда оптимистических заявлений о скорейшем разрешении большинства задач биоинформатики на основе уже имеющихся принципов обработки и методов анализа информационных массивов.

Однако, с другой стороны, рост получаемых сегодня информационных массивов, требующих обработки и детального анализа (генетических, белковых последовательностей и др.), оказался также экспоненциальным, но даже более стремительным, чем предполагалось исследователями ранее (см. рисунок). Добавим, что сами требования по обработке биологических данных продолжают претерпевать существенные качественные и количественные усложнения вслед за развитием самой теории и возникновением в биологии различных "рабочих" гипотез.

Такое положение не могло не вызвать существенное отставание современных информационных технологий от технических требований, предъявляемых рядом актуальных биологических задач, в особенности тех, что сопряжены с обработкой генетических последовательностей. Простой иллюстрацией может стать упоминание того факта, что, несмотря на существенный рост в сравнении с 1997 г. пропускной способности каналов связи и объемов постоянной памяти, только получение из глобальной сети и хранение целиком расшифрованного тогда генома кишечной палочки *Escherichia colii* (около 4 мегабайт) вызывало меньше проблем, чем для генома человека (около 3 терабайт) спустя пять лет. Ситуация в значительной степени усложняется тем, что на текущий момент алгоритмы линейной и сублинейной сложностей относительно объема входных данных удалось предложить и фактически реализовать лишь для весьма ограниченного числа некоторых элементарных задач обработки информационных массивов. Приведем в качестве типового примера сравнение оценок алгоритмической



Рост количества микрочипов в современных микропроцессорах (сплошная линия) и нуклеотидных баз GenBank'a (пунктирная линия) в общем масштабе

сложности в задаче поиска тандемных повторов в ДНК. Вычисления, основанные на простом переборе всех возможных вариантов точного совпадения повторов, дают квадратичную алгоритмическую сложность $O(n^2)$, где n — длина последовательности. Часто при малом числе содержащихся в генетической последовательности повторов z относительно n эта задача проще решаема методом Ландау и Шмидта [11] со сложностью $O(n \ln n + z)$. Однако если предположить количество возможных несовпадений в повторах равным k, то сложности подобных алгоритмов возрастают соответственно как $O(kn^2)$ и $O(kn \ln n + z)$.

Начало работы в рамках сформулированного проекта сопровождалось многократными обсуждениями, связанными с выбором математического аппарата, который мог бы стать основой для приемлемого решения задачи распознавания структуры генетических последовательностей.

Непрерывное усложнение большинства прикладных задач (особенно в микробиологии работа осуществляется с большими массивами данных генетических последовательностей), необходимость во многих случаях многократной цифровой обработки данных, а также преодоление случаев счетной неустойчивости сводят к минимуму достижения по увеличению быстродействия компьютеров. Поэтому введение в процесс обработки данных аналитических методов в соответствующих случаях могло бы существенно сократить и ускорить цифровые расчеты. Выполнение же на цифровой технике аналитических преобразований в процессе решения задач весьма затруднительно и неудобно в дальнейшем использовании.

Преодоление рассмотренных противоречий возможно только на пути применения комбинированных цифро-аналитических технологий, которые во многих прикладных задачах обеспечили бы получение приемлемых результатов по скорости и точности обработки данных.

В рамках обобщенного спектрально-аналитического метода (ОСАМ) [12] достигается естественное обобщение дискретности исходных массивов данных и непрерывности математического анализа, спектрального представления и лежащей в его основе изменчивости и повторяемости изучаемой системы. Сокращение объемов представления информации, необходимых вычислений при решении задач распознавания в широком смысле является следствием адаптивности подхода, основанной на его избирательности математического способа представления информации в широком классе базисных функций.

Исследования, проводимые в рамках данного проекта, направлены на существенное понижение сложности алгоритмов, применяемых при распознавании структурно-функциональной организации генетических последовательностей. Для осуществления этой цели потребовалось разработать новые приемы обработки, принципиально отличные от имеющихся ранее, по возможности отказываясь от применения метода динамического программирования и комбинаторных подходов. Преимущества предлагаемых методов особенно хорошо выражены при решении задач, предполагающих неточные совпадения и существенные искажения генетического кода.

Реализация предлагаемого нами решения данной задачи приводит к алгоритмической сложности, степень которой не зависит от предполагаемого числа несовпадений (это сказывается на оценках полученных повторов), причем результаты работы алгоритма устойчивы к возможным различным нелинейным искажениям входных данных, таким, как вставки и выпадения фрагментов текста. Более того, предлагаемый подход подразумевает возможность распараллеливания вычислений, что приведет к дополнительному ускорению поиска тандемных повторов при реализации на вычислительных кластерах алгоритмов сублинейной сложности. Используемая нами стратегия поиска периодичностей позволяет искать не только тандемные, но и разнесенные неточные повторы, более проста в сравнении с другими подходами. Создан алгоритм, осуществляющий поиск периодичности с использованием спектрально-аналитического метода при анализе профилей избыточности с варьированием границ искомого участка. Его программная реализация уже на начальном этапе исследований позволяет получать приемлемые результаты, не только не уступающие существующим методам распознавания структурно-функциональной организации генетических последовательностей, но, возможно, и опережающие их.

Первые опыты применения комбинированной вычислительной технологии вселяют уверенность в успешном решении поставленных задач распознавания. Однако объем требуемых исследований и разработка соответствующих алгоритмов и программного обеспечения потребуют большого времени и труда.

СПИСОК ЛИТЕРАТУРЫ

1. Watson J.D., Crick F.H.C. Molecular structure of deoxypentose nucleic acids // Nature. 1953. **171**. P. 737.
2. Александров А.А. и др. Компьютерный анализ генетических текстов. М.: Наука, 1990.
3. Колмогоров А.Н. Теория информации и теория алгоритмов. М.: Наука, 1987.
4. Шеннон К. Работы по теории информации и кибернетике. М.: Иностранная литература, 1963.
5. Lempel A., Ziv J. On the complexity of finite sequences // IEEE Transactions on Information Theory. 1976. V.IT-22. Issue 1. P. 75–81.
6. Горбань А.Н., Миркес Е.М., Попова Т.Г., Садовский М.Г. Новый подход к изучению статистических свойств генетических последовательностей // Биофизика. 1993. **38**. Вып. 5. С. 762–767.
7. Korotkov E.V., Korotkova M.A. Latent periodicity of DNA sequences from some human gene regions // DNA Seq. 1995. **5**. P. 353–358.
8. Benson G. Tandem repeats finder: a program to analyze DNA sequences // Nucl. Acids Res. 1999. **27**. Issue 2. P. 573–580.
9. Noé L., Kucherov G. Improved hit criteria for DNA local alignment // BMC Bioinformatics. 2004. **5**. P. 149 (http://www.biomedcentral.com/1471–2105/5/149).
10. Moore G.E. Cramming more components onto integrated circuits // Electronics Magazine. 1965. **38**. N 8. P. 114–117.
11. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. СПб.: Невский проспект, 2003.
12. Дедус Ф.Ф., Махортых С.А., Устинин М.Н., Дедус А.Ф. Обобщенный спектрально-аналитический метод обработки информационных массивов. М.: Машиностроение, 1999.

Поступила в редакцию
20.02.06