БАЗА ДАННЫХ ПО ПЕРЕНОСУ ЗАРЯДА В ДНК

Лахно В.Д., Соболев Е.В., Фиалко Е.В., Барышников Е.В., Киржеманов Е.В.

(Пущино, Московская область)

В ИМПБ РАН в течение нескольких лет ведутся работы по моделированию переноса заряда в ДНК, цель которых – выяснение функциональных параметров, определяющих скорость электронного переноса в нуклеотидных последовательностях (временные зависимости вероятностей распределения электронной плотности, смещений уотсон-криковских пар, сопутствующих электронному переносу и т.д.). В результате проведения вычислительных экспериментов накапливается большой массив данных, требующий систематизации, и все более остро встает вопрос о необходимости создания базы данных (БД). Такая база данных может быть использована для диагностики генетических заболеваний, расчета элементов схем наноэлектронных устройств на основе ДНК и т.д.

Database on charge transfer in DNA

V.D.Lakhno, E.V.Sobolev, N.S.Fialko, V.V.Baryshnikov, L.A.Kirzhemanov

(Pushchino, Moscow Region)

The Institute of Mathematical Problems of Biology, RAS has been engaged in the studies on modeling charge transfer in DNA for several years. The aim of the studies has been to identify functional parameters determining the electron transfer rate in nucleotide sequences, such as temporal dependencies of the probabilities of the electron density distribution, displacements of Watson-Crick pairs attendant the transfer process, etc. Numerical experiments provide a great body of information on electron transfer rates for different DNA fragments. These data require systematization which makes still more acute the problem of creating a relevant database. Such a

database could be used to diagnose genetic diseases, to calculate elements of circuits in DNA-based nanoelectronic devices, etc.

В настоящее время компьютеры и суперкомпьютеры в биологических исследованиях становятся все более важным инструментом познания и получения прикладных результатов. В современной компьютерной биологии используется множество методов прикладной математики, включая широкий спектр вычислительных методов. В основе математического моделирования биологических объектов лежит описание явлений с помощью систем дифференциальных уравнений и дискретных моделей. Исследование различных вариантов переноса заряда в биомолекулярных системах (в частности, в нуклеотидных цепочках) является одним из разделов компьютерной биологии. В ИМПБ РАН в течение нескольких лет ведутся работы по моделированию переноса заряда в ДНК. Основной целью моделирования переноса электрона в ДНК является выяснение функциональных параметров, определяющих скорость электронного переноса в нуклеотидных последовательностях (временные зависимости вероятностей распределения электронной плотности, одноэлектронного тока, смещений уотсон-криковских пар, сопутствующих электронному переносу, и т.д.). На основе полученных в результате моделирования значений параметров формируется база данных скоростей электронного переноса для фрагментов ДНК разной длины и содержащих различные последовательности нуклеотидов. Такая база данных может быть использована для диагностики генетических заболеваний, расчета элементов схем наноэлектронных устройств на основе ДНК и т.д.

В результате проведения многочисленных вычислительных экспериментов накапливается большой массив данных по скоростям электронного переноса для разнородных фрагментов ДНК, требующий систематизации, и все более остро встает вопрос о необходимости создания базы данных. Это объясняется, во-первых, потребностью в организации удобного доступа к полученным результатам (классификация результатов в зависимости от длины цепочки ДНК и расчетных параметров, наглядность представления результатов); во-вторых – большими объе-

мами результатов расчета (в среднем 150-200 Mb на один эксперимент – для цепочек длиной 100-200 сайтов, 5-10 Mb – для цепочек длиной 3-20 сайтов); в-третьих – уникальностью некоторых расчетных экспериментов (расчет даже на кластерах может длиться несколько десятков часов, а объем полученных результатов достигать 1-5 Gb. Учитывая достаточно высокую стоимость использования кластерного времени и трафика при передачи данных, можно сказать, что повторить такие эксперименты очень сложно).

Работа по созданию БД «Перенос заряда в ДНК» находится только в самом начале, поэтому хотим сразу оговориться – мы лишь излагаем наше видение данной системы и те направления, в которых собираемся двигаться.

1. Текущее положение дел

Разработанное в секторе квантово-механических систем ПО (БД «SOLITON») организует выходные данные в виде списка ссылок на файлы с результатами расчётов, позволяет производить просмотр и дополнительную обработку данных, однако оно может быть использовано только в пределах локальной сети ИМПБ РАН (Рис.1).

В настоящее время создан ряд программ для расчета динамики переноса заряда вдоль нуклеотидных цепочек различной длины. Программы написаны на языке программирования C++, исполняемый модуль компилируется в зависимости от операционной системы. Модельные системы дифференциальных уравнений решаются методом Рунге-Кутты, для суперкомпьютерных расчетов проводится «естественное распараллеливание» по параметру. Для хранения данных используется формат файлов СУБД «PARADOX», интерфейсная программа работы с БД написана на языке PASCAL.

Как видно из приведенной на Рис.1 схемы, можно выделить 4 основных этапа работы с информацией базы данных:

I. моделирование переноса заряда в ДНК;
II. занесение информации (полученные результаты расчетов и расчетные параметры) в основную базу данных в соответствии с выбранными классифицирующими признаками;
III. просмотр и обработка прикладным математиком информации, содержащейся в базе данных (стандартные пакеты

статистической обработки данных и т.д.);

IV. отбор, предварительная подготовка и периодическая передача расчетных данных и файлов параметров в базу.

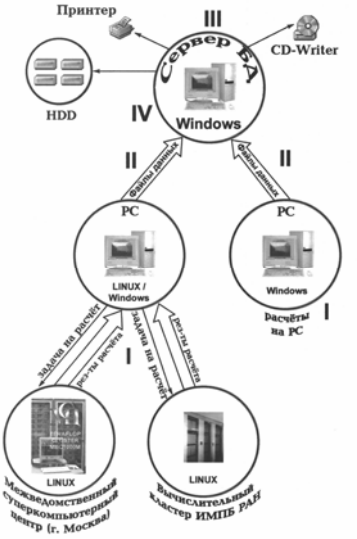

Рис. 1

Экспериментатору пока предоставлена только одна дополнительная возможность – создание видеоклипа по результатам вычислительного эксперимента (файл в формате AVI). Он может выбрать из основной БД нужный вычислительный эксперимент и проделать следующие операции (**Рис.2**):

- изменить параметры просмотра – масштаб (50-500%), шаг расчетной сетки, скорость вывода результатов на экран (1-5000 м/сек), направление просмотра данных (вперед, назад, просмотр отдельного фрагмента) и т.д.;
- сохранить в БД найденные параметры просмотра (тогда в дальнейшем эти параметры будут устанавливаться по умолчанию для данного вычислительного эксперимента);
- просмотреть расчетные данные (кнопки «Воспроизведение», «Пауза», «Стоп») и подобрать наиболее наглядный для записи в видеофайл режим просмотра;
- записать видеоклип (кнопка «Запись»).

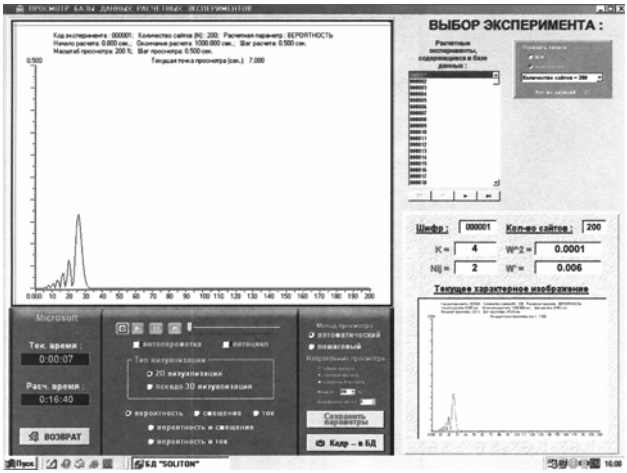

Рис. 2.

2. Построение системы интеграции данных, основанной на Web-технологиях

В дальнейшем в плане проведения расчетов переноса заряда в ДНК планируется провести следующую работу:

- теоретическая разработка и корректировка расчетной модели, написание соответствующих программ для новых модельных уравнений; объединение программ в пакеты;
- создание удобного пользовательского интерфейса (сейчас все файлы параметров расчетной цепочки ДНК создаются самим экспериментатором, в дальнейшем этот процесс будет до определенной степени автоматизирован);
- разработка программных средств для промежуточной визуализации и статистической обработки результатов численных экспериментов.

Поскольку данные представляют интерес для специалистов, работающих в различных областях науки (биомедицины, нанотехнологий и т.д.), возникает необходимость предоставления доступа всех заинтересованных сторон к этим данным с возможностью пополнения. Существует мировой опыт построения таких информационных систем на основе корпоративных сетей и клиент-серверных технологий.

Хранение данных планируется организовать в реляционной

базе данных под управлением SQL-ориентированной СУБД. При этом работа с данными организуется при помощи специального ПО, работающего на сервере приложений и доступного для пользователей через INTERNET. Схема системы представлена на Рис.3.

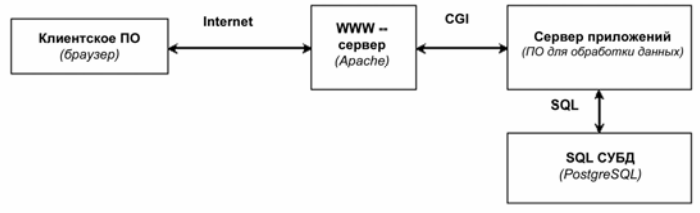

Рис. 3

Предлагаемая архитектура построения БД обладает следующими преимуществами:

- централизованное хранение данных и ПО для обработки данных;
- доступность клиентского ПО (необходимо наличие только http- и ftp-клиентов);
- унифицированный пользовательский интерфейс к данным;
- относительно малый объем информации, передаваемой по каналам связи.

Централизованное хранение данных и ПО для обработки данных позволит упростить администрирование системы, разработку и внедрение нового ПО, сохранение целостности данных. Доступность клиентского ПО, унифицированный интерфейс и удобное представление данных дадут возможность пользователям быстро начать работу с системой без дополнительных затрат и обучения. Кроме того, с системой сможет одновременно работать множество пользователей.

Предлагаемая интерактивная информационная система будет обладать следующими характеристиками:

1. **Независимость от аппаратной или программной платформы на стороне клиента**. Доступ к данным будет организован через Web-интерфейс и произойдет перенос всех вычислительных процедур на сторону сервера. Для работы по-

требуются только Web- и SSH-клиенты.

2. **Информационная безопасность**. Будет организовано разграничение прав доступа к данным для зарегистрированных и незарегистрированных пользователей. Для доступа к данным и к виртуальной среде (User Mode Linux) будут использоваться протоколы SSL (Secure Socket Layer) и SSH (Secure Shell). Зарегистрированный пользователь будет иметь возможность ввести свои собственные расчетных параметров, производить расчеты, а результаты экспериментов предоставить для публичного доступа. Незарегистрированный пользователь сможет только просматривать данные и результаты расчётов, но не водить свои собственные.
3. **Простота навигации**. Поскольку не все Web-клиенты поддерживают в полной мере существующие технологии представления данных в сети (Macromedia Flash, Javascript, ActiveX и т.д.), а также в силу аппаратной разнородности клиентов, все действия по формированию страниц производятся на стороне сервера, пользователю выдаётся лишь HTML-документ. Соответственно, пользователь может использовать для просмотра любой из существующих браузеров, включая и текстовые.
4. **Офисная среда**. Это позволит уменьшить и облегчить выполнение рутинных операций с данными (например, выполнять статистическую обработку данных, производить автоматическое сохранение результатов, настраивать отчетные формы и т.д.). Кроме того, зарегистрированным пользователям будет предоставляться собственный почтовый ящик
5. **Возможность индексирования содержания и поиска по ключевым словам или фразам**. Существующие разработки позволяют вести поиск внутри сайта по ключевым словам или выражениям на основе индексированных или неиндексированных документов
6. **Доступ к дополнительным источникам информации**. Дополнительными источниками информации являются ссылки на действующие специализированные сайты, электронная библиотека и возможность общения с коллегами в тематических форумах (приватных и общедоступных).
7. **Предоставление полнофункциональной операционной**

среды (User Mode Linux) для отладки собственных приложений. User Mode Linux – это ядро операционной системы Linux, запущенное в пользовательском режиме (user mode). Подобная организация работы позволит зарегистрированному пользователю отлаживать и запускать собственные приложения, запускать «ненадежные» сетевые службы (например, http и ftp) без опасения, что это может привести к краху или взлому основной системы.

Литература.

1. В.Д.Лахно. Динамика переноса дырки в нуклеотидных последовательностях // Компьютеры и суперкомпьютеры в биологии / Под ред. В.Д.Лахно, М.Н.Устинина. – Москва-Ижевск: Институт компьютерных технологий, 2002
2. В.Д.Лахно, Н.С.Фиалко. Перенос заряда в ДНК на большое расстояние // Компьютеры и суперкомпьютеры в биологии / Под ред. В.Д.Лахно, М.Н.Устинина. – Москва-Ижевск: Институт компьютерных технологий, 2002
3. В.И.Артемьев. Разработка Intranet-приложений // Центр Информационных Технологий, 1998
4. И.Д.Медведовский, П.В.Семьянов, В.В.Платонов. Атака через Internet // *НПО "*Мир и семья-95", 1997
5. Е.Игумнов. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных // Геокад Плюс (Новосибирск), 2001