УДК 004.75, 004.032.24, 577.323, 528.9.

Массовые вычисления карт молекулярных поверхностей спиральных белков и нуклеиновых кислот

О. А. Афанасьев[*], П. В. Зрелов[*], В. В. Иванов[*], В. А. Степаненко[*], Р. В. Полозов[†], Ю. Н. Чиргадзе[‡]

[*] *Лаборатория информационных технологий*
Объединённый институт ядерных исследований
ул. Жолио-Кюри, д.6, Дубна, Московская область, 141980, Россия
[†] *Институт теоретической и экспериментальной биофизики РАН*
142290, г. Пущино Московской обл.
[‡] *Институт белка РАН*
ул. Институтская, д.4, г. Пущино, Московская область, 142290, Россия

Разработан подход для организации массового расчёта карт молекулярной поверхности спирализованных белков и нуклеиновых кислот в распределённых вычислительных средах. Три программы SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact, представляющие собой модифицированные Linux версии программных кодов SURFACE-2008, PROT-Z и DNA-RNA-Z, были разработаны для расчёта карт поверхностей спиральных белковых молекул и спиральных ДНК/РНК-молекул. Для того, чтобы организовать массовый счёт большого набора карт, из программных кодов SURFACE-2008, PROT-Z и DNA-RNA-Z были исключены графический интерфейс и ввод управляющих параметров в диалоговом режиме. Ввод управляющих параметров и запуск программ SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact был реализован с помощью специальной скрипт-программы. Для графического представления и дальнейшего анализа полученных таким образом карт используются соответствующие полные версии программ SURFACE-2008, PROT-Z и DNA-RNA.

Ключевые слова: распределённые вычислительные системы, спиральные белки, нуклеиновые кислоты, картографирование, скрипт-программа.

1. Введение

В работе [1] были разработаны программы картографирования молекулярных поверхностей белков и нуклеиновых кислот. Такие карты нужны для изучения процессов взаимодействия белковых комплексов, ДНК и РНК. Данные программы предоставляют возможность детального анализа как отдельных участков, так и полных поверхностей спиральных молекул белков и нуклеиновых кислот в атомном приближении. Молекула рассматривается в виде атомной модели, состоящей из набора всех неводородных атомов, координаты которых взяты из банка белковых данных PDB (Protein Data Bank) [2][1]. Выбор типов карт, а также системный подход к анализу функциональной раскраски поверхности белка были описаны в работах [4, 5]. Картографирование глобулярных белков реализуется программой SURFACE-2008 в проекции Аитова-Хаммера [6], спирализованных — программой PROT-Z, а построение карт молекул ДНК и РНК в цилиндрической проекции осуществляется в программе DNA-RNA-Z.

Целью настоящей работы является реализация массового счета карт поверхностей белков, ДНК, РНК с помощью консольных программ SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact. Процесс счета должен управляться специальной скрипт-программой. Данный подход позволит существенно уменьшить время вычислений большого объёма входных данных.

Для осуществления этого были решены следующие задачи:

Статья поступила в редакцию 28 ноября 2009 г.
Работа поддержана грантом РФФИ 07-07-234.

[1] Описание банка белковых данных PDB см. в [3].

1. Разработка нового программного кода SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact в ОС Linux путём выделения расчётной части и исключения графического интерфейса из программ SURFACE-2008, PROT-Z и DNA-RNA-Z.
2. Разработка скрипт-программы, управляющей массовым счётом карт с помощью программ SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact.

2. Консольные версии программ картографирования

Разработанные в [1] программы SURFACE-2008, PROT-Z и DNA-RNA-Z написаны в объектно-ориентированной среде программирования Delphi6 под ОС Windows с использованием языка OBJECT PASCAL. Чтобы осуществить разработку консольных версий программ под ОС Linux, использовался пакет Lazarus [7][1].

Реализация консольных версий SURFACE-2008-compact, PROT-Zcompact и DNA-RNA-Zcompact подразумевает под собой процесс исключения графического интерфейса из исходного кода соответствующих программ SURFACE-2008, PROT-Z и DNA-RNA-Z. Результатом вычислений являются три файла:

1) файл с расширением CHT — сохранённая карта белкового комплекса или ДНК/РНК;
2) файл с расширением SAV, содержащий в себе точную копию PDB-файла (необходим при визуализации карты);
3) файл INFO.txt содержит краткую информацию о результате счета (тип карты, время счета, дата счета, имя директории с результатами).

В реализованном варианте составляющими модулями каждой консольной версии являются:

1) SURFACE-2008-compact.lpr, PROT-Zcompact.lpr (для белковых версий), DNA-RNA-Zcompact.lpr (для ДНК/РНК версии) — файлы-проекты;
2) Pr_U1.pas — модуль, содержащий основные процедуры и функции расчёта входных PDB-данных;
3) Pr_LoadPar.pas — модуль обработки текстового файла, эмулирующего настройки интерфейса (размеры окна, масштаб, выбор типа карты, высота координатной оси OZ; раскраска атомов, остатков, рельефа; выбор цвета фона; координаты меток);
4) Pr_SaveChart.pas — модуль, сохраняющий результаты вычислений в CHT-файл и копию PDB-файла в SAV-файл, необходимого при загрузке карты.

Время компиляции консольных версий занимает около 20-30 секунд. При запуске программы на счёт должны быть указаны следующие параметры: имя входного PDB-файла, имя входного текстового файла эмуляции настроек интерфейса, имя карты.

Пример запуска программы в командной строке:

```
./ SURFACE-2008-compact pdb4gcr.pdb loadpar.txt pdb4gcr
```

3. Управляющая скрипт-программа

Запуск задач на счёт осуществляется из временных scratch-каталогов, доступных на всех машинах Linux кластера с помощью скрипт-программы, в которой в качестве инструмента мы использовали команды системы пакетной обработки заданий Portable Batch System (PBS) [8]. Эта система осуществляет определение

[1]Lazarus — система с открытым исходным кодом, которая построена на компиляторе Free Pascal Compiler с добавлением Интегрированной среды разработки (IDE), которая совместима с Библиотекой визуальных компонентов Delphi (VCL). Кросплатформенность пакета позволяет установить его на большинство известных на сегодняшний день операционных систем, в частности, на ОС Scientific Linux 5.

конкретной фермы Linux кластера и машины в ней, управление выполнением задачи и отправку результатов пользователю. С целью длительного хранения результатов счета карт молекулярных поверхностей молекул белков и ДНК/РНК и организации базы данных использовался домашний каталог пользователя, определённый в AFS. Такое место хранения является наиболее безопасным с точки зрения защищённости от несанкционированного доступа и различных сбоев. База данных результатов счета представляет собой следующую структуру:

- [каталог_результаты]
- [каталог_тип_данных (глобулярные белки, спирализованные белки, ДНК/РНК)]
- [каталог_дата_создания]
- [каталог_название_pdb-файла (содержит непосредственно результаты)]
- CHT-файл, SAV-файл, INFO.txt.

Структура базы данных задаётся с помощью управляющей скрипт-программы и в любой момент может быть изменена, если в этом возникнет необходимость.

В ходе разработки скрипт-программы её структура была разделена на два модуля (скрипт-модули).

Функции первого скрипт-модуля:

1. Инициализирует все переменные.
2. Осуществляет поиск входных PDB-файлов, поиск файла эмуляции интерфейса loadpar.txt и саму консольную программу.
3. Создаёт соответствующую названиям PDB-файлов структуру каталогов во временной директории:
   /scr/u/afanoleg/protein2008 (для глобулярных белков),
   /scr/u/afanoleg/protein_z (для спирализованных белков),
   /scr/u/afanoleg/dnarna_z (для ДНК/РНК); копирует в эти разделы соответственно по одному входному файлу белкового образца, loadpar.txt, исполняемую программу и второй скрипт-модуль.
4. Реализует запуск второго скрипт-модуля с передачей ему параметров по каждому из PDB-файлов из каждой временной директории с помощью команды qsub системы управления заданиями PBS.
5. Создаёт иерархию каталогов для результатов в домашнем каталоге пользователя HOME.

Функции второго скрипт-модуля:

1. Копирует на ферму кластера исполняемую программу, файл белковых данных и модуль эмуляции интерфейса из временной директории.
2. Запускает на счёт консольную программу, вычисляет время счёта и сохраняет, добавляя в INFO.txt информацию о дате создания и времени вычислений входного PDB-фрагмента.
3. Копирует результаты в указанный в нём каталог домашней директории.

На рис. 1 представлена схема всего вычислительного процесса, реализуемого с помощью обоих скрипт-модулей.

Для графического представления результатов (карт) и их дальнейшего анализа используются соответствующие полные версии программ SURFACE-2008, PROT-Z и DNA-RNA. На этапе тестирования развитого нами подхода были рассчитаны карты для 27 файлов входных данных:

- 9 комплексов с глобулярными белками, сильно отличающихся по типу структуры;
- 9 комплексов с узнающими альфа-спиралями спирализованных белковых факторов транскрипции, вырезанных из PDB-файлов координат комплексов белок-ДНК;
- 9 комплексов, содержащих фрагменты узнающих участков ДНК, вырезанных из PDB-файлов координат комплексов белок-ДНК.

Результаты вычислений сохраняются в домашнюю директорию в такой же иерархии как при копировании во временный каталог (рис. 2). В названия файлов включены также обозначения интервалов вырезанных участков белков/ДНК.

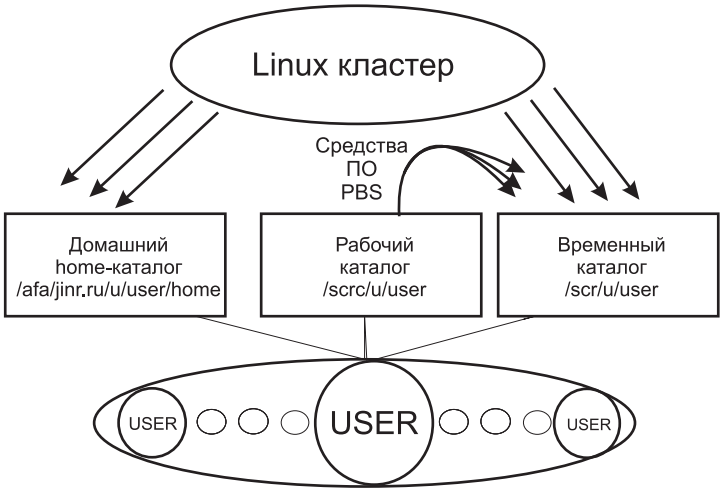

Рис. 1. **Схема работы управляющей скрипт-программы на кластере ЦИВК ОИЯИ**

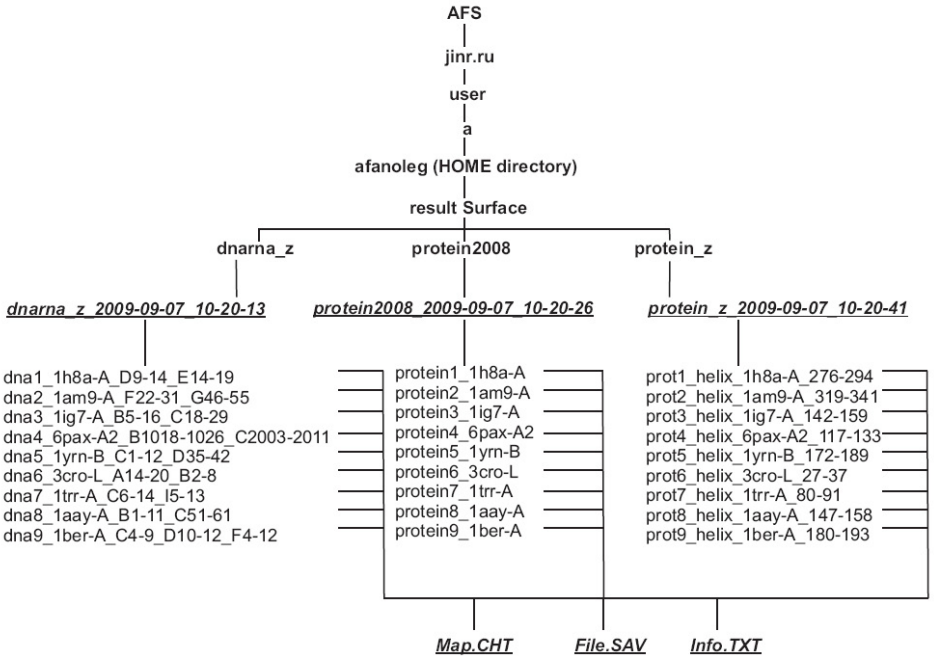

Рис. 2. **Структура HOME-директории с результатами**

4. Заключение

Разработаны консольные программы SURFACE-2008-compact, PROT-Zcompact, DNA-RNA-Zcompact и скрипт-программа, управляющая процессом счёта и созданием соответствующих структур выходных данных в распределённой вычислительной среде. С помощью средств ПО системы пакетной обработки заданий (PBS) на Центральном информационно-вычислительном комплексе (ЦИВК) ОИЯИ реализован массовый счёт карт молекулярных поверхностей спиральных белков, ДНК и РНК. Данный подход позволяет существенно сократить время на исследования структур белков и нуклеиновых кислот.

Литература

1. Software Complex for Computing Surface Maps of Helical Biopolymer Molecule Proteins and Nucleic Acids / O. A. Afanasiev, V. V. Ivanov, V. A. Stepanenko et al. // Book of abstracts, of Inter. Conf. "Mathematical Modeling and Computational Physics, 2009" / JINR, Laboratory of Informational Technologies. — Dubna, 2009. — P. 171.
2. An Information Portal to Biological Macromolecular Structures. — `http://www.rcsb.org/pdb/home/home.do`.
3. Structural Bionformatics, second edition / Ed. by P. E. Bourne, J. Gu. — John Wiley & Sons, Inc., Hoboken, NJ, 2009. — Pp. 271–291.
4. *Chirgadze Y. N., Kurochkina N., Nikonov S.* Molecular Cartography of Proteins: Surface Relief Analysis of the Calf Eye Protein Gamma-Crystallin // Protein Engineering. — 1989. — Vol. 3. — Pp. 105–110.
5. *Чиргадзе Ю. Н., Ларионова Е. А.* Определяющая роль кластеров полярных остатков в структурах белковых факторов при узнавании большой бороздки двухспиральной В-ДНК // Мол. биология. — 2003. — Т. 37, № 2. — С. 266–276.
6. *McDonnell P. W.* Introduction to Map Projections. — Marcel Dekker, Inc. (New York), 1979.
7. Lazarus. — `http://www.lazarus.freepascal.org/`.
8. Система управления заданиями. — `http://rsusu1.rnd.runnet.ru/opbs/ipbs.html`.

UDC 004.75, 004.032.24, 577.323, 528.9.

On Massive Calculations of Maps of Molecular Surface of Helical Proteins and Nucleic Acids

O. A. Afanasiev[*], P. V. Zrelov[*], V. V. Ivanov[*], V. A. Stepanenko[*], R. V. Polozov[†], Yu. N. Chirgadze[‡]

[*] *Laboratory of Information Technologies*
Joint Institute for Nuclear Research
Joliot-Curie 6, 141980 Dubna, Moscow region, Russia

[†] *Institute of Theoretical and Experimental Biophysics*
Russian Academy of Sciences
142290 Puschino, Moscow region, Russia

[‡] *Institute of Protein Research*
Russian Academy of Sciences
Institutskaja str., 4, 142290 Puschino, Moscow region, Russia

An approach has been developed for organizing massive calculations of the maps of a molecular surface of helical proteins and nucleic acids in the distributed computing media. Three new program codes SURFACE-2008-compact, PROT-Zcompact and DNA-RNA-Zcompact that represent modified Linux versions of codes SURFACE-2008, PROT-Z and DNA-RNA-Z were elaborated to calculate the surface maps of the helical protein molecules and the helical DNA-RNA molecules. In order to organize massive computing of a large set of molecules, the graphical interface and the input of control parameters in a dialog mode are eliminated from the SURFACE-2008, PROT-Z and DNA-RNA-Z codes. To input the control parameters and to run codes SURFACE-2008-compact, PROT-Zcompact and DNA-RNA-Zcompact, a special script-program has been implemented. For graphical presentation and further analysis of the maps obtained in such a way corresponding full versions of codes SURFACE-2008, PROT-Z and DNA-RNA-Z are used. The investigation has been supported by a grant of the RFBR 07-07-234.

Key words and phrases: distributed computing systems, helical proteins, nucleic acids, mapping, script-program.