ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № ?, 2014, с. ??–??

УДК 577:004.4:004.94

Б. М. Глинский, Н. В. Кучин, И. Г. Черных, Ю. Л. Орлов,
Н. Л. Подколодный, В. А. Лихошвай, Н. А. Колчанов

Суперкомпьютерные технологии в решении задач биоинформатики

Аннотация. С 2001 года в ИВМиМГ СО РАН функционирует Центр коллективного пользования «Сибирский суперкомпьютерный центр» (ССКЦ) с пиковой производительностью кластеров 115 TFlops. Основные задачи центра: разработка и использование суперкомпьютерных технологий для математического моделирования различных задач, решаемых в институтах СО РАН; обеспечение работ институтов СО РАН и университетов Сибири по математическому моделированию в фундаментальных и прикладных исследованиях; обучение специалистов СО РАН и студентов университетов методам параллельных вычислений на суперкомпьютерах, методам моделирования больших задач. Одним из основных потребителей ресурсов является Центр коллективного пользования "Биоинформатика", созданный на базе Института Цитологии и Генетики СО РАН. В рамках совместных работ центров коллективного пользования, были разработаны программные пакеты по наиболее актуальным научным направлениям биоинформатики. Работа посвящена обзору ресурсов ССКЦ и прикладным программным пакетам по биоинформатике.

Ключевые слова и фразы: Суперкомпьютеры с гибридной архитектурой, биоинформатика, компьютерная геномика, эволюция, прикладные программные пакеты.

1. Введение

В настоящее время Центр Коллективного Пользования Сибирский Суперкомпьютерный Центр (ЦКП ССКЦ) имеет два кластера,

Работа была выполнена при поддержке гранта РФФИ 13-07-00589.

которые используются в режиме коллективного пользования институтами СО РАН. Один из кластеров построен на основе вычислительных узлов с Intel Xeon (архитектура MPP), пиковая производительность 30 TFlop/s, программирование с применением MPI и OpenMP, другой с гибридным расширением на GPU NVIDIA Tesla M2090 (архитектура GPGPU), пиковая производительность 85 TFlop/s, параллельное программирование при помощи C/C++ CUDA и OpenCL. Особенностью программирования задач на кластере с MPP-архитектурой, ориентированной на решение больших задач, прежде всего 3-D, является применения параллельных языков MPI и OpenMP, поскольку это обусловлено архитектурой кластера, построенного с использованием многопроцессорных серверов с общей памятью (SMP). При таком подходе внутри каждого вычислительного модуля формируются несколько потоков с помощью OpenMP. Поддерживаются две современных парадигмы параллельных вычислений – MPI для систем с распределенной памятью (кластеров) и OpenMP для систем с общей памятью. Схема вычислений предусматривает запуск на каждый вычислительный узел кластера по одному MPI-процессу, который запускает внутри каждого вычислительного модуля несколько потоков с помощью OpenMP. Другая технология высокопроизводительных вычислений связана с реализацией алгоритма на гибридной архитектуре: суперкомпьютер состоит из набора соединенных между собой узлов, для обмена данными используется MPI; каждый узел состоит из 2-х многоядерных CPU и 3 GPU; на каждом узле запускается 1 процесс MPI, управляющий вычислениями (процесс выполняется на CPU); из MPI процесса запускаются нити (threads) OpenMP, каждая из которых управляет работой одного GPU. Другой вариант: запускаются три MPI процесса на узел, каждый управляет закрепленным за ним GPU. ЦКП ССКЦ СО РАН предоставляет вычислительные и консалтинговые услуги 21 академическим институтам Сибирского отделения и 5 университетам, более 160 пользователей используют ресурсы центра для решения своих задач. Решается большое количество задач из различных областей знаний, в том числе, определенных приоритетными направлениями развития науки и техники.

2. Архитектурные особенности ЦКП ССКЦ

В настоящее время в ССКЦ имеются два кластера, которые используются в режиме коллективного пользования институтами СО

РАН. Один из кластеров построен на основе вычислительных узлов с Intel Xeon (архитектура MPP), пиковая производительность 30 TFlop/s, программирование с применением MPI и OpenMP, другой с гибридным расширением на GPU NVIDIA Tesla M2090 (архитектура GPGPU), пиковая производительность 85 TFlop/s, параллельное программирование при помощи C/C++ CUDA и OpenCL. Имеется кластерная файловая система Ibrix, содержащая 4 сервера и 32 Тбайта памяти. Кроме того, в состав ССКЦ входит сервер с общей памятью HP ProLiant DL980 G7 с восемью 10-ядерными процессорами Intel E7-4870 с тактовой частотой 2,4 ГГц, оперативной памятью 1024 Гбайт и 8 SAS дисками по 300 Гбайт. Пиковая производительность сервера в текущей конфигурации составляет 768 Гфлопс. В апреле 2012 года сервер включён в кластер НКС-30Т как нестандартный вычислительный узел. В состав кластера входят: 576 процессоров (2688 ядер) Intel Xeon E5450/E5540/X5670; 120 процессоров GPU - Tesla M 2090 (61440 ядер); SMP сервер с общей памятью hp DL980 G7 (8 процессоров, 80 ядер) Intel E7-4870, оперативная память 1024 Гбайт); кластерная файловая система IBRIX (4 сервера, 32 Тбайта). Таким образом, в состав гетерогенного кластера входят вычислительные блоки с MPP-архитектурой, гибридной архитектурой с использованием карт NVIDIA Tesla M2090 (40 узлов, на каждый узел 3 карты) и SMP-архитектурой. Все узлы кластера связаны между собой через Infiniband QDR. Такая структура кластера отвечает требованиям центров коллективного пользования, поскольку приходится решать самые разнообразные задачи из различных областей знаний и наличие нескольких архитектур в центре даёт возможность выбрать оптимальную исходя из специфики решения задачи. Например, для задач биоинформатики часто используют SMP-архитектуру, так как объем входных данных может достигать 5 Терабайт и наличие 1 Терабайт оперативной памяти дает возможность более эффективно их обрабатывать, благодаря отсутствию пересылок данных через сеть. Принципиально имеется возможность, при такой схеме построения центра, задействовать все ресурсы гетерогенного кластера при решении одной задачи. Подробнее о составе технических и программных средств, пакетах прикладных программ можно посмотреть на сайте ССКЦ http://www2.sscc.ru/.

3. Решение задач биоинформатики с помощью ресурсов ЦКП ССКЦ

С помощью оборудования ЦКП ССКЦ решается ряд важных научных задач биоинформатики:

- компьютерная геномика и транскриптомика;
- компьютерная протеомика;
- моделирование биологических процессов на молекулярном уровне;
- эволюционная биоинформатика;
- молекулярная динамика;
- математические проблемы биоинформатики;
- обработка текстовых данных для биологии.

На базе ССКЦ установлен ряд программных пакетов по молекулярной динамики и квантовой химии, такие как: Gaussian [1], Gromacs [2] и др. Однако, наибольший интерес представляют специализированное программное обеспечение, разработанное пользователями ССКЦ. Авторами статьи предложены два программных пакета, которые внедрены в ЦКП, для решения задач моделирования молекулярно-генетических систем и анализа символьных последовательностей геномики.

3.1. Моделирование молекулярно-генетических систем (МГС). MGSmodeller.

Для реконструкции математических моделей МГС используется система MGSmodeller (http://modelsgroup.bionet.nsc.ru/?page_id=491). Математические модели реконструируются в формате и по правилам стандарта SibML [3] в рамках обобщенного химико-кинетического подхода [4,5]. Анализ результатов моделирования производится средствами системы MGSmodeller и программами Matlab, Gnuplot. На рисунке 1 представлена схема организации MGSmodeller.

Модули компиляции и численного исследования реализованы на языке Fortran. Модули аннотации и редактирования языка SibML, а также постобработки результатов реализованы на языке Java. Математические модели в компьютерной среде MGSmodeller представлены в рамках стандарта SibML как совокупность элементарных подсистем молекулярно-генетических систем. Их реконструкция в рамках среды моделирования производится на основе блочного принципа. Сначала производится декомпозиция исследуемого объекта до

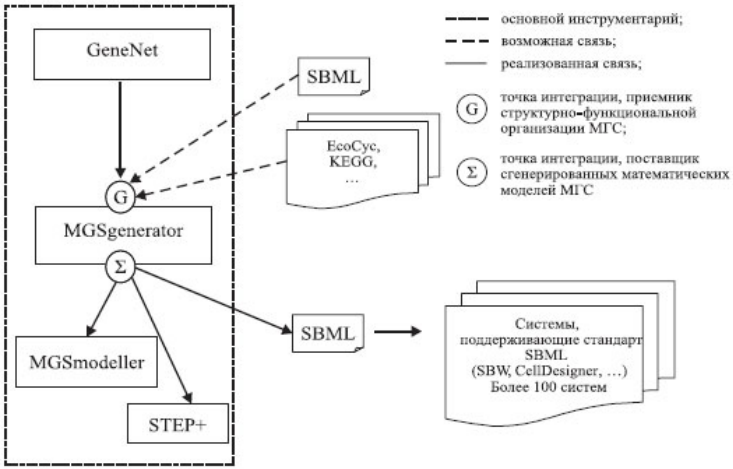

Рис. 1. Схема организации компьютерной модульной системы моделирования на основе программы MGSgenerator системы MGSmodeller.

уровня элементарных подсистем, которыми могут быть реакции ферментативного синтеза, подсистемы регуляции экспрессии генов, системы сплайсинга, транспорта, трансляции, процессы созревания и модификации белков, деградации макромолекул и др. Далее описываются математические модели каждой подсистемы, из которых формируется база элементарных моделей. На этой основе исследователь конструирует из элементарных моделей, как из строительных блоков, модель исследуемого объекта. Для этого описывается сценарий сборки модели – файл, содержащий заданную структурно-функциональную организацию модели целевого объекта, в котором указывается система отношений компартментов (структурный уровень организации целевого объекта) и для каждого компартмента указываются подсистемы, которые должны быть включены в него (функциональный уровень организации объекта). В результате численного эксперимента для моделей больших размерностей исследователь, как правило, получает большие объемы информации, и возникает проблема их интерпретации, анализа и визуализации. В случае если не хватает возможностей базовых средств визуализации, в рамках системы MGSmodeller результаты моделирования представ-

лены в структурированном виде. Организация атрибутов переменных модели, задающих ассоциацию с контекстом моделирования, позволяет проводить постобработку данных сторонними программами, в том числе, используя специализированные инструменты визуального анализа (http://www.gnuplot.info). Более детально возможности пакета и пример его использования изложено в [6].

3.2. Программный комплекс анализа символьных последовательностей геномики. ICGenomics

Программный комплекс ICGenomics предназначен для компьютерной поддержки исследований в геномике, молекулярной биологии, биотехнологии и биомедицине. Основное назначение – функциональная аннотация геномных последовательностей, получаемых в результате массового высокопроизводительного секвенирования на уровне нуклеотидных и аминокислотных последовательностей. Рабочее название – экспериментальный образец программного комплекса анализа символьных последовательностей геномики (ЭОПК АСПГ). Важная технологическая проблема обработки и анализа данных высокопроизводительного геномного секвенирования требует разработки специализированных компьютерных средств. Развитие новых экспериментальных методов геномики, прежде всего, секвенирования, привело к стремительному росту объемов экспериментальных данных, «информационному взрыву». Основная задача компьютерного анализа геномных данных состоит в их функциональной аннотации, интеграции результатов с молекулярно-биологическими информационными ресурсами. В связи с этим большую актуальность приобретает разработка информационно-компьютерных технологий автоматического анализа и функциональной аннотации геномных последовательностей. Для решения задачи был разработан ряд программ для извлечения и интеграции данных, а также визуального представления накопленной информации в форме геномных профилей, представленных на серверах крупнейших международных научных центров NCBI (http://www.ncbi.nlm.nih.gov/), UCSC Genome Browser (http://genome.ucsc.edu/), EBI (http://www.ebi.ac.uk/). Важнейшим объектом теоретической и прикладной геномики являются молекулярно-генетические системы , координирующие функцию геномов, генов, РНК, белков, генных и метаболических путей на различных иерархических уровнях жизни: клеточном, тканевом, органном, организменном, популяционном. Основным источником дан-

ных являются нуклеотидные последовательности, получаемые в результате массовых экспериментов высокопроизводительного секвенирования [7]. Программный комплекс ICGenomics позволяет выполнять следующие логически различные функции:

- процессинг (обработку) протяженных последовательностей нуклеотидов из данных секвенирования, полученных с помощью установок секвенирования нового поколения, в том числе: процессинг данных секвенирования платформ 454 и Illumina, процессинг данных секвенирования платформы SOLiD и обработку полногеномных профилей ChIP-seq, включая выделение пиков и предсказание ССТФ;
- аннотацию геномных нуклеотидных последовательностей, включая: разметку положения нуклеосом на основе вейвлет-преобразования полногеномных профилей предсказания, сайтов формирования нуклеосом и распознавания сайтов формирования нуклеосом с помощью данных полногеномного секвенирования линкерной ДНК; поиск экзонов во вновь секвенированных последовательностях; поиск промоторов генов миРНК в нуклеотидных последовательностях на основе специфичных структурных мотивов;
- предсказание аллергенности белков по их структурным и функциональным свойствам на основе метода функциональной аннотации пространственных структур белков, в том числе предсказания функциональных сайтов в пространственных структурах белков;
- исследование режимов эволюции белок-кодирующих генов, включая реконструкцию эволюционной истории белков на основе предсказания ортологов в секвенированных геномах, филогенетический анализ и исследование режимов эволюционного отбора.

Программный комплекс состоит из модуля управления (программной компоненты ICGenomics-web и управляющей программы ICGenomics-start) и 4 программных компонент ICGenomics-Processing, ICGenomics-GenomeAnnotation, ICGenomics-Allergen и ICGenomics-Evolution (рис. 2), которые отвечают за функционал пакета. Общий интерфейс представлен на рисунке 3.

Более детально возможности пакета и пример его использования изложено в [8].

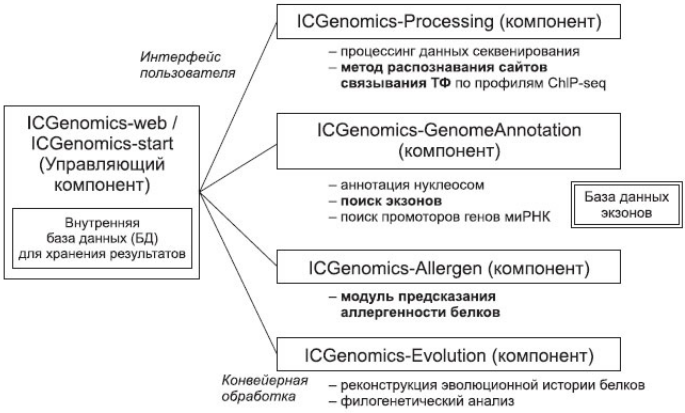

Рис. 2. Структура программного комплекса ICGenomics.

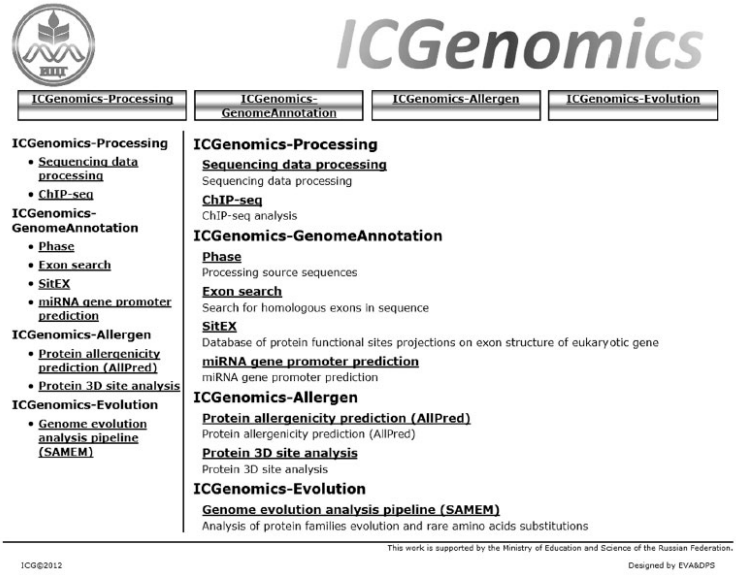

Рис. 3. Пример интерфейса управляющего модуля, содержащего функциональные компоненты.

4. Заключение

С 2001 года запущен и успешно функционирует Центр Коллективного Пользования Сибирский Суперкомпьютерный Центр. Центр обеспечен двумя кластерами с классической и гибридной архитектурами суммарной мощностью порядка 115 Терафлопс. ЦКП ССКЦ СО РАН предоставляет вычислительные и консалтинговые услуги 21 академическим институтам Сибирского отделения и 5 университетам, более 160 пользователей используют ресурсы центра для решения своих задач. Одним из наибольших потребителей ресурсов ЦКП является Институт Цитологии и Генетики СО РАН. Для решения задач биоинформатики сотрудниками института были разработаны программные пакеты для решения задач моделирования молекулярно-генетических систем и анализа символьных последовательностей геномики.

Список литературы

[1] *http://www.gaussian.com/*. ↑4

[2] *http://www.gromacs.org//*. ↑4

[3] М. В. В. Казанцев Ф.В. Новоселова Е.С. и др. *Язык моделирования молекулярно-генетических систем SiBML* // Параллельные вычислительные технологии (ПаВТ) 2012, 2012, с. 722. ↑4

[4] М. Ю. Г. Лихошвай В.А. Ратушный А.В. и др. *Обобщенный химико-кинетический метод моделирования генных сетей* // Молекуляр. биология, 2001. Т. **3 (6)**, с. 1072-1079. ↑4

[5] К. Ф. В. Акбердин И.Р. Омельянчук Н.А.. *Математическое моделирование метаболизма ауксина в клетке меристемы побега растения* // Информ. вестник ВОГиС., 2009. Т. **13 (1)**, с. 170-175. ↑4

[6] А. И. Р. Казанцев Ф.В. Н.Л. Подколодный. *НОВЫЕ ВОЗМОЖНОСТИ СИСТЕМЫ MGSmodeller* // Вавиловский журнал генетики и селекции, 2012. Т. **16 (4/1)**, с. 799-804. ↑6

[7] D. P. S. Ivanisenko V.A. Pintus S.S. et al. *Computer analysis of metagenomic data-prediction of quantitative value of specific activity of proteins* // Dokl. Biochem. Biophys., 2012. Vol. **443**, p. 76-80. ↑7

[8] Орлов Ю.Л. и др. *ICGenomics: программный комплекс анализа символьных последовательностей геномики* // Вавиловский журнал генетики и селекции, 2012. Т. **16 (4/1)**, с. 732-741. ↑7

Об авторах:

Борис Михайлович Глинский

Окончил Новосибирский Государственный Университет в 1967 г., профессор, доктор технических наук. Область научных интересов: вычислительные системы, моделирование сейсмических полей, имитационное моделирование.

e-mail: gbm@sscc.ru

Николай Владимирович Кучин

Окончил Новосибирский Государственный Университет в 1971 г., главный специалист по системному программному обеспечению ИВМиМГ СО РАН. Область интересов: высокопроизводительные вычислительные системы, системное программное обеспечение кластеров

e-mail: kuchin@sscc.ru

Игорь Геннадьевич Черных

Окончил Новосибирский Государственный Университет в 2002г., кандидат физико-математических наук. Область научных интересов: суперкомпьютерные вычисления, химическая кинетика.

e-mail: chernykh@ssd.sscc.ru

Юрий Львович Орлов

Окончил Новосибирский Государственный Университет в 1991г., кандидат биологических наук. Область научных интересов: биоинформатика, компьютерная геномика, эволюция.

e-mail: orlov@bionet.nsc.ru

Николай Леонтьевич Подколодный

Окончил Новосибирский Государственный Университет в 1974г. Область научных интересов: разработка программно-информационных систем для научных исследований.

e-mail: pnl@bionet.nsc.ru

Виталий Александрович Лихошвай

Окончил Новосибирский Государственный Университет в 1976г. Область научных интересов: математическое моделирование биологических систем, теория генных сетей, теория моделирования.

e-mail: likho@bionet.nsc.ru

Николай Александрович Колчанов

Директор Института цитологии и генетики СО РАН, заведующий Отделом системной биологии ИЦиГ СО РАН. Область научных интересов:информационная биология, молекулярная биология, молекулярная генетика, компьютерный анализ структурно-функциональной организации и эволюции геномов, генетических макромолекул - ДНК, РНК и белков и молекулярно-генетических систем геномов.

e-mail: kol@bionet.nsc.ru

Образец ссылки на эту публикацию:

Б. М. Глинский, Н. В. Кучин, И. Г. Черных, Ю. Л. Орлов, Н. Л. Подколодный, В. А. Лихошвай, Н. А. Колчанов. *Суперкомпьютерные технологии в решении задач биоинформатики* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. ??, № ?, с. ??–??.

URL: http://psta.psiras.ru/read/

Boris Glinskiy, Nikolay Kuchin, Igor Chernykh, Yuriy Orlov, Nikolay Podkolodnyi, Vitaly Likhoshvai, Nikolay Kolchanov. *Bioinformatics and High Performance Computing.*

Abstract. This article presents Siberian Supercomputer Center (SSCC) as a computational center for bioinformatics. Siberian Supercomputer consists from two cluster supercomputers, especially designed for bioinformatics workstation with symmetric multiprocessing architecture and data center. There are detailed hardware and software architectures described in article. Some success stories of SSCC usage for bioinformatics problems also presented. (in Russian).

Key Words and Phrases: high performance computers; bioinformatics; big data.