Межинститутский центр коллективного пользования «Биоинформатика»

Цели и задачи

Современная биология стала источником беспрецедентно огромных объемов экспериментальных данных. Их осмысление и практическое применение невозможны без привлечения современных информационных технологий, эффективных методов анализа данных и моделирования биологических систем и процессов на различных иерархических уровнях организации живой материи: от молекулярно-генетического, клеточного, организменного до экосистемного и биосферного. В ответ на этот вызов возникла компьютерная системная биология как союз экспериментальных подходов с биоинформатикой.

Основные направления работы

1. Создание и поддержка информационной и программно-аппаратной инфраструктуры, объединяющей через специализированные сети передачи данных суперкомпьютерные вычислительные комплексы и объемные хранилища данных с уникальными высокопроизводительными экспериментальными установками для решения задач геномики, протеомики, масс-спектрометрии, микроскопии и томографии.

2. Компьютерная и информационная поддержка исследований и разработок, выполняемых в рамках программы «Геномика, протеомика, биоинформатика» включая решение следующих задач:

- обработка первичных данных, получаемых с использованием экспериментальных технологий геномики, протеомики, масс-спектрометрии, микроскопии, томографии;
- обеспечение доступа к распределенным мировым информационным ресурсам в области геномики, транскриптомики, протеомики, метаболомики, генетики, молекулярной и клеточной биологии, физиологии, фармакологии, биомедицины, биотехнологии и др.;
- разработка алгоритмов и методов обработки и хранения биологических данных;
- поддержка высокопроизводительных, параллельных, распределенных вычислений и Web-сервисов в биоинформатике;
- компьютерное моделирование живых систем на различных уровнях их иерархической организации;
- планирование экспериментов с использованием методов биоинформатики и системной биологии;
- разработка новых экспериментально-компьютерных технологий для системной биологии.

3. Повышение квалификации сотрудников в области компьютерной биологии, биоинформатики и высокопроизводительных вычислений в науках о жизни; подготовка специалистов, владеющих методами биоинформатики, теоретического и компьютерного анализа и моделирования, необходимыми для решения широкого круга фундаментальных и прикладных проблем биологии, физиологии и биомедицины, фармакологии и биотехнологии.

Инфраструктура

Центральным инфраструктурным объектом ЦКП является высокопроизводительный вычислительный комплекс, приобретенный в рамках программы СО РАН «Геномика, протеомика, биоинформатика» (рис. 1–3), в следующей комплектации:

- вычислительный кластер с 64 двойными блейд-серверами HP BL2×220G6 пиковой производительностью 10,362 ТФлопс (всего 128 вычислительных модулей, каждый из которых включает два четырехъядерных процессора Intel Xeon E5540, 2,53 Ггц и 16Гб оперативной памяти). Технологически вычислительный кластер объединен с вычислительными ресурсами Сибирского суперкомпьютерного центра (32 двойных блейд-сервера HP BL2×220 G5: 64 вычислительных модуля по 2 процессора Intel Xeon E5450, 3,00 Ггц). Общая производительность объединенного кластера – 16,5 ТФлопс;
- параллельная кластерная система хранения данных емкостью 48 Тбайт, используемая при вычислениях;
- файловый сервер (сервер HP DL360G6 в комплектации: 2 четырехъядерных процессора Intel® Xeon® X5550 2,66 ГГц; 12 Гбайт оперативной памяти) с долговременным хранилищем данных (емкость – 36 Тбайт);
- система резервного копирования в виде ленточной библиотеки HP MSL4048 (стандарт LTO-4), поддерживающей одновременную установку 48 картриджей;
- специализированный сервер баз данных на основе HP DL380G6 (2 четырехъядерных процессора Intel® Xeon® X5560 2,80 ГГц, оперативная память 36 Гбайт) с дисковым массивом HP StorageWorks D2700 емкостью 3,6 Тбайт. Установлены СУБД Oracle 11g, MySQL, PostgreSQL;
- технологический сервер баз данных на основе HP DL380 G6 в следующей комплектации: 2 четырехъядерных процессора Intel® Xeon® X5550 2,66 ГГц; оперативная память 12 Гбайт;
- специализированный Web-сервер для доступа к информационным и вычислительным ресурсам ЦКП «Биоинформатика»;
- высокоскоростная сеть для объединения экспериментальных установок и локальных компьютеров с высокопроизводительным вычислительным комплексом;
- терминальный класс в ИЦиГ СО РАН (рабочие станции);
- графическая рабочая станция HP Z800 Xeon, оперативная память 8 Гбайт, 2 монитора: рабочий монитор Dell U2410, для работы монитор с 3D изображениями ACER G245HQBID 23,6", и 3D устройство (рис. 3).

Программное обеспечение включает лицензионное системное программное обеспечение (операционные системы, ПО для управления файловыми системами, компиляторы и отладчики, СУБД «Oracle 11g», MySQL, PostgreSQL), а также специализированные пакеты программ для решения задач биоинформатики, геномики, протеомики, метаболомики, транскриптомики, в том числе:

- Accelrys discovery studio для решения задач протеомики, молекулярного дизайна и конструирования лекарств, предсказания свойств больших и малых молекул и т. д.;
- пакет прикладных программ MatLAB, ориентированный на параллельные вычисления на кластере со специализированными модулями для решения задач биоинформатики;
- пакет прикладных программ Mathematica 7 Professional, в рамках которого работает множество систем обработки биологических данных и моделирования биологических

Кластер – 16,5 ТФлопс. Параллельная система хранения данных – 48ТБ. Файловое хранилище данных – 36ТБ. Система резервного копирования. Сервер баз данных, Web-сервер.

Рис. 1. Инфраструктура технологического кластера, объединяющего ресурсы организаций-учредителей ЦКП «Биоинформатика».

систем и процессов (Cellerator и xCellerator);
- пакеты программ для решения задач молекулярной динамики (amber, GROMACS с LAM/MPI и др.).

Программное обеспечение для решения задач биоинформатики и системной биологии, разработанное в ИЦиГ СО РАН и других организациях-учредителях ЦКП «Биоинформатика», включает:
- Web-портал по биоинформатике и компьютерной системной биологии, содержащий пакеты программ по решению задач в области геномики, регуломики, транскриптомики, моделирования клетки и дизайна искусственных генетических конструкций, трансгенеза, анализа полиморфизмов;
- программную систему «Protein Structure Discovery» для решения задач протеомики;
- программные системы ANDCell и ANDVisio, позволяющие извлекать знания о взаимодействии биомолекул из текстов научных публикаций и интегрировать эти знания в виде семантических сетей;
- программные системы проведения вычислительных экспериментов в области эволюционной биоинформатики, система «Эволюционный конструктор» для имитационного моделирования эволюционных процессов в популяциях;
- наиболее распространенные пакеты программ для решения задач биоинформатики, в том числе EMBOSS, boost, Blast, Blast, BioC++ libraries, Bio++ Program Suite, BioJava, BioPerl, BioPython, R – среда программирования с большим числом ППП для статистического анализа и графического представления экспрессионных данных.

Для решения задач биоинформатики также используются ресурсы организаций-учредителей ЦКП «Биоинформатика» (рис. 2, 3) включая:
- вычислительные ресурсы ССКЦ (ИВМИМГ СО РАН), объединенные с вычислительным кластером ИЦиГ СО РАН;

Рис. 2. Стойка управления кластером НКС-30Т.

Рис. 3. Графическая станция HP Z800.

- высокопроизводительный вычислительный кластер НГУ (производительность – 13,2 ТФлопс);
- хранилище данных ИВТ СО РАН – 127 Тбайт.

Принципы организации и структура

ЦКП «Биоинформатика» является межинститутской организационной структурой. Базовыми подразделениями Центра в ИЦиГ являются лаборатория теоретической генетики и сектор высокопроизводительных вычислений. Кадровый состав ЦКП «Биоинформатика» комплектуется из сотрудников и аспирантов организаций-учредителей.

Учредителями ЦКП являются:
- Институт цитологии и генетики СО РАН (ИЦиГ СО РАН);
- Институт вычислительной математики и математической геофизики СО РАН (ИВМиМГ СО РАН);
- Институт «Международный томографический центр» СО РАН (ИМТЦ СО РАН);
- Институт химической биологии и фундаментальной медицины СО РАН (ИХБФМ СО РАН);
- Институт математики СО РАН (ИМ СО РАН);
- Институт вычислительных технологий СО РАН (ИВТ СО РАН);
- ГОУ ВПО «Новосибирский государственный университет» (НГУ).

При возникновении новых научных направлений и задач в число учредителей могут быть включены другие организации на основании дополнений к Договору о научно-техническом сотрудничестве в рамках Межинститутского распределенного ЦКП «Биоинформатика».

В рамках деятельности ЦКП предполагаются проведение совместных конференций, семинаров и рабочих совещаний, научных школ, организация стажировок, учебной и научной практики для студентов и аспирантов. О своей деятельности ЦКП отчитывается перед Научно-техническим советом ЦКП, а также представляет отчеты в ученые советы организаций-учредителей, Объединенный ученый совет по биологическим наукам, Объединенный ученый совет СО РАН по нанотехнологиям и информационным технологиям, Объединенный ученый совет СО РАН по математике и информатике, совет СО РАН по супервычислениям.

В рамках ЦКП функционируют следующие научно-технологические секции:

- компьютерная геномика и транскриптомика (ИЦиГ СО РАН);
- компьютерная протеомика (ИЦиГ СО РАН, МНТЦ);
- статистический анализ биологических данных (ИЦиГ СО РАН);
- математическое моделирование биологических систем и процессов (ИЦиГ СО РАН, ИМ СО РАН);
- эволюционная биоинформатика (ИЦиГ СО РАН);
- популяционно-генетический и сегрегационный анализ (ИЦиГ СО РАН);
- моделирование макромолекул методами молекулярной динамики и механики (ИЦиГ СО РАН, ИХБФМ СО РАН);
- анализ текстов научных публикаций (ИМ СО РАН, ИЦиГ СО РАН);
- астробиология (ИК СО РАН);
- высокопроизводительные вычисления в биоинформатике (ИЦиГ СО РАН и ИВМиМГ СО РАН);
- математические проблемы биоинформатики (ИЦиГ СО РАН, ИМ СО РАН);
- информационно-телекоммуникационные технологии в биоинформатике (ИВТ СО РАН);
- программно-аппаратная поддержка ЦКП «Биоинформатика» (ИВМиМГ СО РАН).

Организуются другие секции, необходимость которых определяется в ходе работы ЦКП «Биоинформатика», в частности, для задач анализа данных томографии, анализа текстов научных публикаций в прикладных областях. В задачи научных секций входит поддержание сферы компетенции по кругу научных задач и методов суперкомпьютерных (параллельных) вычислений включая проведение специализированных и рабочих семинаров; техническая подготовка задач биоинформатики для счета.

Материально-техническое обеспечение работы ЦКП «Биоинформатика» осуществляется организациями-учредителями из собственных бюджетов, средств программы СО РАН «Геномика, протеомика и биоинформатика», средств приборной комиссии СО РАН, а также из других источников, образуемых в результате деятельности Центра, в частности, при осуществлении НИР в рамках программ научного сотрудничества, поддерживаемых российскими, зарубежными и международными фондами и прочими источниками, обеспечивающими финансовую поддержку научных исследований, на основании дополнительных соглашений сторон.

Контактная информация

Председатель Научно-технического Совета ЦКП «Биоинформатика» и научный руководитель – академик РАН Николай Александрович Колчанов
тел.+7(383) 363-49-80
e-mail: kol@bionet.nsc.ru
Руководитель ЦКП «Биоинформатика» к.б.н. Юрий Львович Орлов
г. Новосибирск, просп. Академика Лаврентьева, 10, ИЦиГ СО РАН
тел. +7 (383) 363-49-22
e-mail: orlov@bionet.nsc.ru