

# **Высокопроизводительные вычисления в биоинформатике**

# Особенности предметной области

1. Большой темп накопления знаний. Появление новых высокопроизводительных экспериментальных установок.
2. Большой темп роста числа гетерогенных источников данных - баз данных.
3. Тенденция к усложнению моделей предметной области.
4. Расширение области применения молекулярно-генетических знаний: биомедицина, фармакология, нанобиоинженерия и т. д.
5. Необходимость решать задачи, требующие больших вычислительных ресурсов.
6. Необходимость решать задачи, требующие интеграции больших объемов гетерогенных источников данных.

# Системная биология

*Цель - изучение организации и механизмов развития и функционирования живых систем на основе информации, закодированной в их геномах, в ходе их взаимодействия с окружающей средой.*

Описание в базах данных и интеграция огромных объемов гетерогенной экспериментальной информации, характеризующей живые системы на различных уровнях их структурно-функциональной организации

Крупномасштабный анализ экспериментальных данных

Построение математических моделей организации и функционирования живых систем

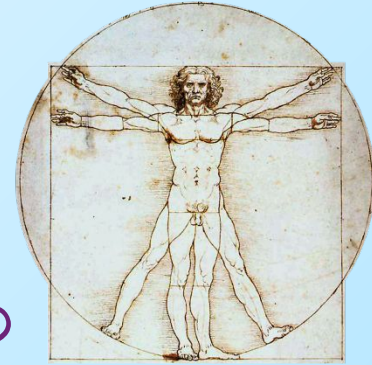
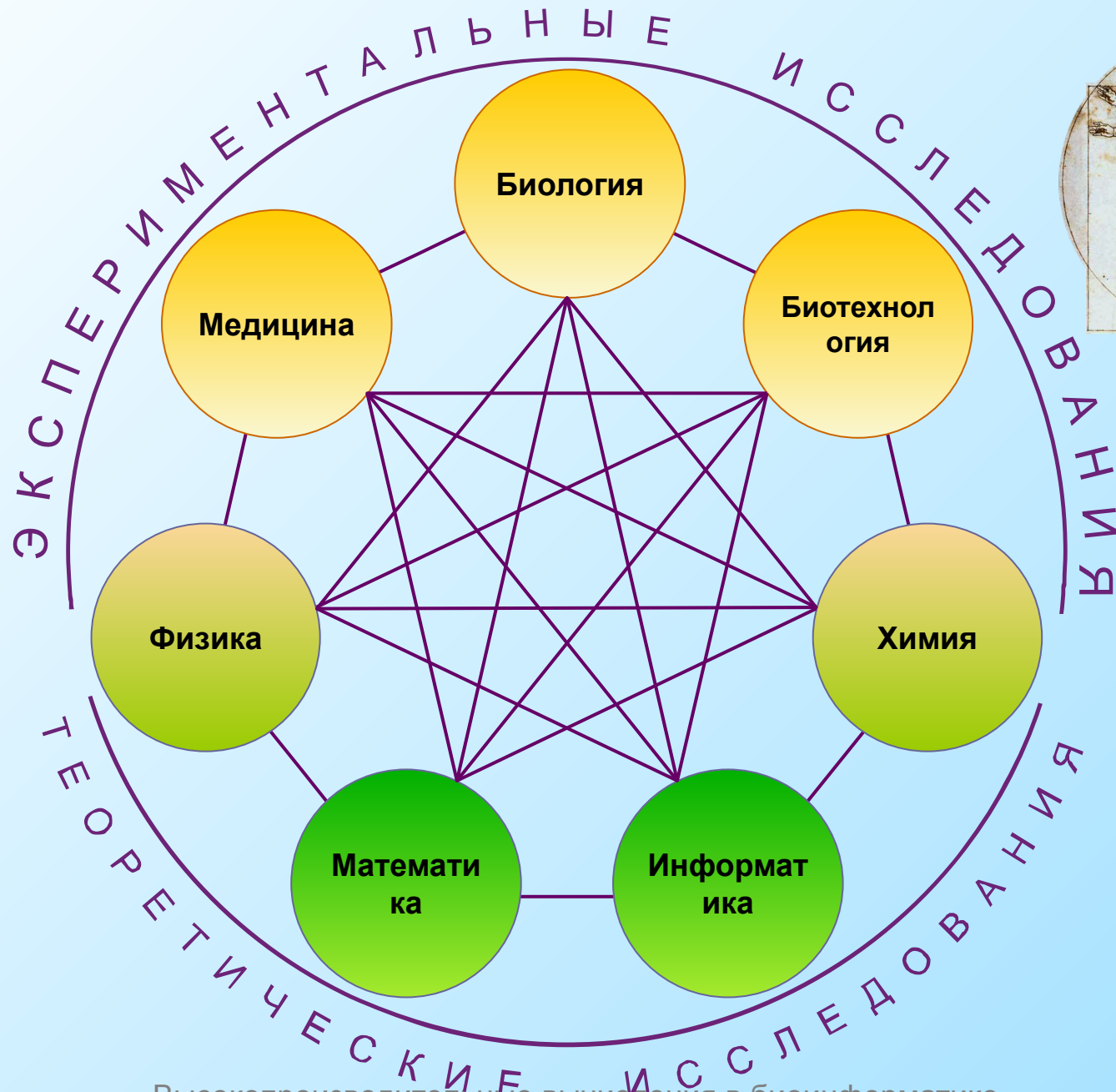
Предсказание новых особенностей организации и функционирования живых систем

Планирование экспериментов по проверке результатов предсказания

Проведение экспериментов и получение новых данных и знаний

**СИСТЕМНАЯ БИОЛОГИЯ ВОЗНИКЛА, КОГДА ОНА СТАЛА ПРЕДСКАЗАТЕЛЬНОЙ НАУКОЙ**  
Высокопроизводительные вычисления в биоинформатике

# Системная биология – интегративная наука



Высокопроизводительные вычисления в биоинформатике

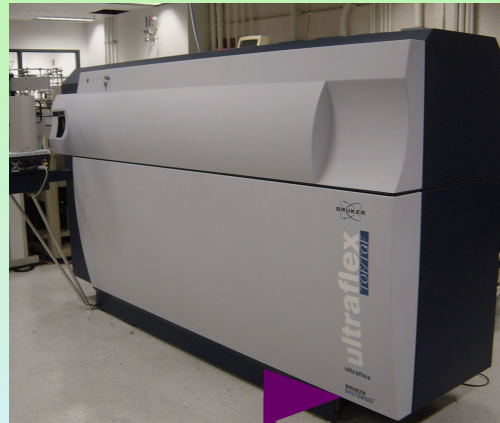
# Экспериментально-вычислительная база системной биологии

## Кластер «Системная биология» Новосибирского научного центра СО РАН

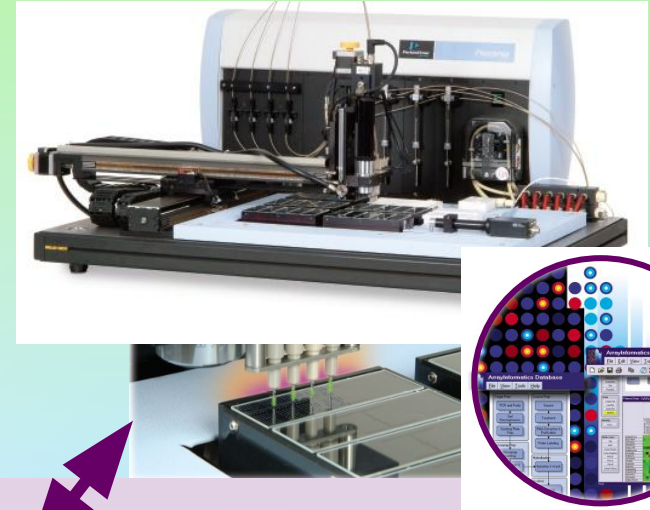
Геномика: автоматический секвенатор



Протеомика: массовый анализ белков и метаболитов



ТРАНСКРИПТОМИКА: производство биочипов высокой плотности и анализ профилей экспрессии генов



Клеточная биология: лазерный сканирующий микроскоп

LSM510 META

Высокопроизводительные вычислительные

Высокопроизводительные вычисления в биоинформатике

Прижизненная томография экспериментальных животных

# Объемы молекулярно-биологических данных и комбинаторная сложность задач биоинформатики

Источник данных	Объем данных	Задачи
Секвенированные последовательности ДНК	~40 млн. последовательностей, $10^{12}$ пар оснований	Функциональная аннотация
Белковые последовательности	~5.5 $10^6$ последовательностей (~300 аминокислот каждая)	Сравнительный анализ. Выявление консервативных мотивов
Структуры макромолекул	50000 структур (~3000 атомных координат каждая)	Предсказание, выравнивание, измерение геометрии, докинг
Геномы	Около 1200 геномов прокариот, более 160 геномов эукариот	Сборка полных геномов; Функциональная аннотация; Сравнительный анализ
Экспрессия генов в различных тканях, стадиях развития, состояний организма и т.д.	Сотни тысяч образцов с тысячами вариантов измерений для десятков тысяч генов. ~ $10^{13}$ измерений.	Анализ механизмов регуляции коэкспрессирующихся генов. Связь с последовательностями, структурными и биохимическими данными.
SNP (однонуклеотидные мутации в ДНК)	Только одна база данных dbSNP содержит информацию о $10^8$ мутациях в 23 геномах.	Анализ связи с заболеваниями
Молекулярные взаимодействия, метаболические пути и генные сети	Более $10^6$ молекулярных взаимодействий описано в публикациях. Более ста тысяч метаболических путей и генных сетей представлено в базах данных.	Моделирование молекулярно-генетических процессов и систем
Публикации	Десятки миллионов публикаций	Поиск и извлечение знаний

# Список некоторых наиболее затратных задач биоинформатики и потребности в вычислительных и информационных ресурсах

<b>Ассемблирование полных геномов</b>	Реконструкция последовательности полного генома человека, животных или растений.	10 TFlops	30 TB of trace files per genome
<b>Анализ полных геномов</b>	Сравнительный анализ полных геномов	10 TFlops	5 TB
<b>Предсказание структуры белка</b>	Анализ всех белков бактериального генома за одни сутки	100 TFlops	10 TB
<b>Молекулярная динамика</b>	Моделирование ДНК-белковых взаимодействий (20000 атомов, до 1 мс)	100 TFlops	30 TB of trace files
<b>Молекулярная динамика (с учетом квантовомеханических взаимодействий)</b>	Моделирование реакции для фермент активного сайта (200 атомов, 1 нс) за одни сутки.	1000 TFlops	100s TB of trace files
<b>Докинг белковых молекул</b>	Моделирование взаимодействия белок-легант. Предсказание функции белка. Поиск новых лекарственных средств.	>10 TFlops	5 TB





# Технологии ускорения решения задач

1. Использование высокопроизводительных вычислительных кластеров или суперкомпьютеров:

- Распараллеливание по данным
- Распараллеливание по процессам

2. Использование специальных процессоров:

- **FPGA** (Field Programmable Gate Array)
- **MPPA** (Massively Parallel Processor Array)
- **GPU** (Graphics Processing Unit)

3. Использование гибридных вычислительных систем, объединяющих в вычислительных узлах CPU вместе со спецпроцессорами, GPU или FPGA.

Пример: IBM Roadrunner. Процессор PowerXCell 8i.

# Когда эффективно GPU?

**GPU демонстрируют хорошие результаты при:**

- 1. Параллельной обработке данных**
  - Когда одна и та же последовательность действий, применяется к большому объёму данных
- 2. Расчетах с высокой плотностью арифметики**
  - Когда велико отношение числа арифметических инструкций к числу обращений к памяти

**Одни и те же вычисления означают меньшие требования к управлению исполнением (flow control)**

**Высокая плотность арифметики и большой объём данных означают возможность покрытия латентности памяти вычислениями (вместо больших кэшей на CPU)**



# CUDA ZONE

USA - United States

Search NVIDIA

- DOWNLOADS
- WHAT IS CUDA
- CUDA U
- DEVELOPING WITH CUDA
- FORUMS
- NEWS AND EVENTS

## LATEST CUDA NEWS NVIDIA And NEC Collaborate To Deliver GPU Computing Solutions To HPC Market



Accelerating Leukocyte Tracking using CUDA

29 x



Ultrasound goes GPU: real-time simulation using CUDA

270 x



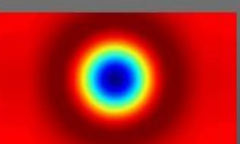
Particle Swarm Optimization on GPU

270 x




Accelerated Image Registration With CUDA

100 x



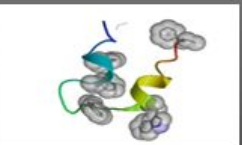
Jacket: GPU Engine for MATLAB

50 x



Smith Waterman algorithm

3.5 x



Accelerating Molecular Dynamic Simulations on GPUs Using OpenMM

100 x



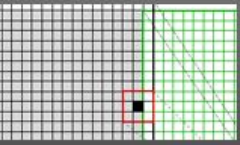
Manufacturing Computations Lab

100 x



GPUGRID.NET

100 x




Biomedical Image Analysis

13 x



3D Particle Boltzmann Solver

120 x




Creation parallel dotplots for suite of protein sequences

120 x



LISSOM


9 x



SI	3 min (13.0x)	Rig
Base	39 min	
SI	9 min (9.0x)	Fl
Base	1hr, 21 min	

Silicon Informatics Protein Docking

20 x



Folding@home

100 x

Search

Sort by Release Date

Submit Your Work to CUDA Zone



### Filter by Application Type

Reset

- Computatio...
- Digital Cont...
- Electronic D...
- Finance
- Game Physics
- Imaging
- Numerics
- Life Sciences
- Libraries
- Oil & Gas
- Science
- Signal Proce...
- Video & Audio
- Other

### Filter by Content Type

- Application
- Code
- Multimedia
- Paper
- Presentation

### Filter by Organization Type

# Примеры приложений GPU CUDA и их эффективность

- Генетический алгоритм оптимизации. Монте-Карло (300-1000)
- Анализ текстов, поиск регулярных выражений. (10-35).
- Сравнительная геномика. Филогения (15)
- Smith Waterman, BLAST, ClustalW (30-70)
- Скрытые марковские процессы. HMMer (25-30)
- Множественное выравнивание (30).
- распознавание образов(100), K-ближайших соседей (470), SVM(150),
- Нейросети (15);
- Алгоритмы на графах (20)
- Дискретное моделирование биологических систем (200)
- Молекулярная динамика (100-150),
- Молекулярный докинг (16)
- Молекулярный фолдинг (100)
- Медицинская томография (300)
- Анализ изображений (100)
- Решение систем линейных уравнений (50)
- Сингулярная декомпозиция (60)

**Благодарю за внимание!**

# Классы задач, решаемых в СО РАН

1. Компьютерный анализ результатов секвенирования и ассемблирование полноразмерных геномов.
2. Структурно-функциональная аннотация полногеномных последовательностей прокариот и эукариот.
3. Сравнительный анализ полногеномных последовательностей.
4. Молекулярная эволюция. Филогения.
5. Широкомасштабный компьютерный анализ протеомов.
6. Компьютерный анализ и моделирование структурно-функциональной организации ДНК, РНК, белков и их комплексов.
7. Функциональная аннотация белковых макромолекул. Молекулярный скрининг. Молекулярный докинг и молекулярный дизайн медицинских препаратов.
8. Дизайн самоорганизующихся ДНК/РНК наноструктур.
9. Молекулярная эпидемиология. Анализ полиморфизмов.
10. Компьютерное моделирование сложных молекулярно-генетических систем и процессов в норме и патологии.
11. Компьютерно-информационная поддержка экспериментального дизайна искусственных бактериальных молекулярно-генетических конструкций.
12. Компьютерный анализ изображений.

# CUDA™ Toolkit – среда разработки для GPU, основанная на языке C

- **CUDA** (Compute Unified Device Architecture) -- это технология от компании NVidia, предназначенная для разработки приложений для массивно-параллельных вычислительных устройств (в первую очередь для GPU начиная с GeForce 8800, а также Quadro и Tesla).
- Основными плюсами CUDA являются ее бесплатность (SDK для всех основных платформ свободно скачивается с [developer.nvidia.com](http://developer.nvidia.com)), простота (программирование ведется на "расширенном C") и гибкость.
- **GPU – сопроцессор для CPU** (хоста)
- **У GPU есть собственная память**
- GPU с CUDA работает либо как **гибкий потоковый процессор**, где тысячи вычислительных программ, называемых потоками, или *threads*, вместе решают сложные задачи, либо как потоковый процессор в специфических приложениях, например, для вывода изображения, где потоки не связаны между собой.
- GPU способен одновременно обрабатывать **множество потоков данных** одним и тем же алгоритмом
- Для осуществления расчётов при помощи GPU хост должен осуществить запуск вычислительного ядра, которое определяет конфигурацию GPU в вычислениях и способ алгоритм получения.
- Процессы GPU (в отличие от CPU) очень просты и многочисленны (~ 1000 для полной загрузки GPU)