An Overview of the Bioinformatics Platform Based on HPC in BGI

Jiandong Sun
Genomics & Bioinformatics Institute CAS
Beijing Genomics Institute
Outline

- A brief introduction of BGI
- Bioinformatics on Genomics
  - Sequencing Data Management
  - Assembly and Annotation
  - Comparative Genomics
  - EST/cDNA pipeline
- Bioinformatics on Proteomics
  - Data Collection and Management Module
  - LC/MS/MS data analysis Module
  - Protein 3D structure Prediction Module
  - Proteomics Database Module
- Acknowledge
A brief introduction of BGI

- Genomics & Bioinformatics Institute CAS
  Beijing Genomics Institute

Founded: 1999, 9, 9, 9:9 am
A brief introduction of BGI

- 4 directors
- 8 board members
- More than 400 people

Beijing Genomics Institute
A brief introduction of BGI

4 Major Departments ---- Genome Sequencing

◆ 80 MegaBACE sequencers
◆ Every Day :
  ➢ 60000 reactions
  ➢ 300,000,000 bps sequenced
  ➢ 15G Byte Raw data file
A brief introduction of BGI

4 Major Departments ---- Proteomics

◆ 2 Sets of 2D-PAGE
◆ 4 Mass Spectrometers
◆ Every Day :
  ➢ 35 2D-PAGE images
  ➢ 6500 protein identified by LC/MS/MS
  ➢ 3500 proteins identified by MALDI-TOF
◆ Projects:
  ➢ Rice, ThermoBacteria
  ➢ Cancer, Snake Venom
4 Major Departments ---- Drug Screening

- Just the beginning
- Setting up Cell line environment for screening
- Focus on Chinese Traditional Medicine
- 600 compounds have been extracted from Chinese Traditional Medicines
A brief introduction of BGI

4 Major Departments ---- Bioinformatics ( I )

- 63 people
- Average age : 24
4 Major Departments ---- Bioinformatics (II)

High Performance computers:

- 3000 (176 CPU, 96 G, 2 T) --- Dawning
- E10k (64 CPU, 64 G, 10 T) --- SUN Microsystem
- P690 (32 CPU, 256G, 5T) --- IBM
4 Major Departments ---- Bioinformatics ( III )

◆ Genomics Area :

assembly, Gene annotation, comparative genomics, SNP, software development, algorithm research, repeat analysis……

◆ Proteomics Area :

2D-PAGE image analysis, MS data analysis, proteomics and genomics relative, protein 3D structure, metabolic pathway……

◆ Biology Databases :

Genomics database, proteomics database, system biology……
Major Project:

- Human Genome (1%)
- Rice Genome
- Pig Genome
- *Thermoanaerobacter tengcongensis* Genome
- Spirulina Genome
- ......
Three Cover Papers

5 papers with 72.5 SCI factor in 2.5 years
A brief introduction of BGI

- National High Performance computing Center
- SUN Center of Excellence for Bioinformatics
  Genomics & Bioinformatics Institute (Beijing, Hangzhou)
- BGI-DAWNING Joint Bioinformatics Institute
- Sister Center of WhiteHead Genome Center / MIT
LIMS --- Sequencing Data Collection and Management
  collect, quality report, backup, basecalling, vector mask……

Assembly ---- Whole Genome Shot-gun sequencing data assembly
  ThermoBacteria, HGP, Rice Genome, Pig Genome
  A new strategy and a new software ( Genome Research )

Annotation ---- Gene finding, regulation factor, alternative splicing, repeats…
  GC3 Codon bias in Plant ( Genome Research )

Comparative Genomics ---- rice

EST / cDNA pipeline ---- clustering, assembly if needed, alignment ……
Structure ---- Data Collection and Management Module

- Detect data generation and gain data automatically
- Bar Code control to avoid miss-named
- Different data will drop in different raw database
- Backup automatically at setting time
- Logs and software status are real-time put on intranet
- Daily report and summary are email to manager automatically
The Integrated Bioinformatics Platform on Proteomics

- Sample Preparation
  - MALDI-TOF
  - 2D-PAGE
  - LC MS/MS

- Automatic Data Collection System
  - Super Computer (Sun10000, Dawn3000)
  - Database
    - MALDI-TOF Database
    - 2D-PAGE Database
    - LC MS/MS Database

Beijing Genomics Institute
The Integrated Bioinformatics Platform on Proteomics

Structure ---- LC/MS/MS data analysis Module

- **Database search engine**
  - Comprehensive data source
    - public database, EST assembly, genome self-sequenced, gene-self-development.....
  - New algorithm of Clustering spectra by peaks pattern to pick up sample spectra
  - Accurate pattern matching

- **De Novo Sequencing**
  - Dynamic Programming
  - Ions recognition algorithm

- **Web Base Interface**

- **Task profile and schedule design**
Structure ---- Protein 3D structure Prediction Module

- Based on PROSPECT
  - PROtein Structure Prediction and Evaluation Computer Toolkit
  - one of the top 6 performers in CASP4 contest
  - Parallelized on Dawning 3000
  - 80 sequences less than 500 amino acids can be calculated at the same time.

- Task queue management
- Auto-notice when finished prediction
- Result displays on Browser
Structure ---- Protein 3D structure Prediction Module
Structure ---- Proteomics Database Module

◆ Fast Internet Connection to Public Database

◆ Self-generated Data
  ➢ 2D-PAGE images database
  ➢ LC/MS/MS database including analysis result
  ➢ Peptides and Theoretic CID spectra Pattern Database
  ➢ Protein 3D-structure database
  ➢ Genomics integrated database
  ➢ Metabolic Pathway database
A “Private” Grid System

- Computing Service ---- HPCs
- Data Management ---- Integrated Database
- Applications ---- self-developed on HPCs
- Network ---- fast connection between from Beijing to Hangzhou

Basic Foundation of E-Science
Acknowledgment