IZBI

UNIVERSITÄT LEIPZIG

Interdisciplinary Centre for Bioinformatics

Working Group 1: Databases and Data Integration – Selected Projects

Erhard Rahm, Hong-Hai Do*, Toralf Kirsten*

Objectives

- Management and analysis of complex molecular-biological data of users for research networks with fast growing amount of data
- Design and implementation of flexible databases and analysis platforms for interdisciplinary projects and clinical studies
- Database research topics:
  - Integration of molecular-biological data and metadata (e.g. annotations)
  - Database coupling / integration of analysis algorithms and tools
  - Flexible, high performance data organization and querying

Current Results

- Comparative evaluation of microarray-based gene expression databases showed limitations of previous approaches (BTW2003)
- **GeWare**: Design and implementation of a data warehouse for gene expression analysis; first version of warehouse operational
- **GeneMapper**: Integration of gene annotations from different public sources; first version operational

Data Warehouse for Gene Expression Analysis / GeWare

- Flexible data management for gene expression analysis based on Affymetrix oligonucleotide arrays
- Large amounts of data (around 500 experiment series per year) generated by local user groups
- Innovative data warehouse approach:
  - Multidimensional data organization
  - Integration of sample/experiment and gene annotation data with expression data
  - Support for several normalization and aggregation algorithms
  - Integration of existing analysis tools

Data Sources · Data Integration · Data Warehouse · Tool Integration · Data Analysis

Annotation Integration / GeneMapper

- Public sources with annotations refer to different gene representations, i.e. identifiers
  - Public sources: LocusLink, Human Genome Browser, Ensembl, UniGene, GeneCards, GeneLynx, …
  - Vendor-based sources, e.g. NetAffx (Affymetrix): annotations of proprietary genes, i.e. probe sets
- Goals: Providing gene-oriented views on annotations by matching between different gene representations

Human-Chimpanzee Comparison Platform

E. Rahm, S. Pääbo, et al.

- Recent availability of draft versions of the human and chimpanzee genomes: first example of two closely related mammalian genomes
- Goals: Design and implementation of an integrated platform for comparative analysis between humans and chimpanzees

- Genome-wide comparison of sequence data, expression level, recombination rates, …
- High volume of data: currently approx. 1 TB microarray expression data in flat files from about 200 experiments available

Cooperation Partners

Experimental Partners

- **MPI for Evolutionary Anthropology** (S. Pääbo)
- Clinical trial groups (lymphoma, sepsis networks)
- **IZKF Leipzig**
- Groups at **Biotechnical-Biomedical Centre Leipzig**

IZBI — WG1 Databases and Data Integration, WG2 Tissue Organization, WG3 Signal Transduction Gene Expression, WG4 Genetic Evolution

Computer Science Partners

- **European Bioinformatics Institute** Cambridge, UK (A. Brazma, U. Sarkans)
- **Lion BioScience**, UK (T. Etzold)
- **Humboldt University of Berlin**, Dept. of Computer Science (J.-C. Freytag, S. Heymann, U. Leser, F. Naumann)
- **University of Leipzig**, Dept. of Computer Science