

Interdisciplinary Centre for Bioinformatics

Working Group 1: Databases and Data Integration – Selected Projects

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

Objectives

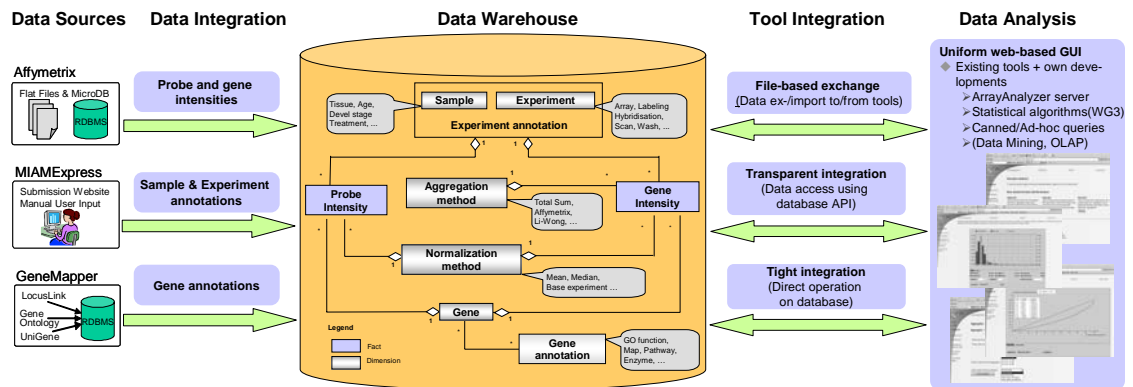
- Management and analysis of complex molecular-biological data of users for research networks with fast growing amount of data
- Design and implementation of flexible databases and analysis platforms for interdisciplinary projects and clinical studies
- Database research topics:
 - Integration of molecular-biological 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



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

GeneMapper Version 0.2

Step 1: Specify a file of or copy/paste accessions (File Open, Create, Copy, Tab and Enter as shortcuts)

Step 2: Select type of source accessions: (Affy)

Step 3: Check target sources for annotations:

Gene	Protein	Annotations
AF01	EMT2B	BIOLOGICAL_PROCESS
GENBANK	BL0232	CELLULAR_COMPONENT
ENSEMBL	BT029307	MOLECULAR_FUNCTION
LOCUSLINK	PFAM	CMB4
REFSEQ	BC09	BIOCHEMICAL_FUNCTION
GENBANK	U01549	
ALIAS		
EMBL		

Retrieved Associations

AFIX	GENBANK	ENSEMBL	BIOLOGICAL_PROCESS
1007_#	U01549	BN2500000121318	GO:0007155
1053_#	M27238	BN2500000045241	
111_#	X51757	BN2500000173110	GO:0006493
111_#	X06439	BN2500000172418	GO:0007735 GO:0007345 GO:0007387
1255_#	U36811		
1258_#	U13928	BN2500000124321	GO:0006484
1431_#	X02843	BN2500000126649	GO:0006118
1438_#	X02508	BN2500000171590	GO:0007145
1482_#	U36817		
1594_#	M23118	BN2500000171563	GO:0006118
1598_#	U13750	BN2500000171563	GO:0007130
16502_#	U64811	BN2500000173222	GO:0006508
1729_#	U41430	GO:0006917 GO:0007145	
1773_#	U06635	BN2500000142954	GO:0006503

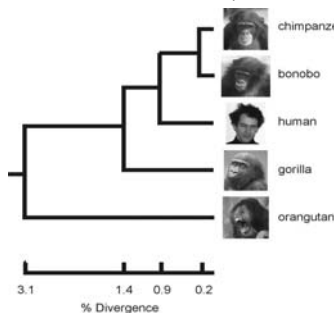
selected accession id's

mapped accession id's

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. Eitzold)
- Humboldt University of Berlin, Dept. of Computer Science (J.-C. Freytag, S. Heymann, U. Leser, F. Naumann)
- University of Leipzig, Dept. of Computer Science