UNIVERSITÄT LEIPZIG

BioFuice: A decentralized approach to integrate molecular-biological data

Toralf Kirsten[1], Erhard Rahm[1,2]

[1] Interdisciplinary Centre for Bioinformatics, University of Leipzig
http://www.izbi.de

[2] Department of Computer Science, University of Leipzig
http://dbs.uni-leipzig.de

Motivation

- **Objective**: Integration of complex molecular-biological data from different sources allowing a
  - Source overspanning data analysis, e.g. sequence analysis and pathway analysis,
  - Combination with experimental data for joint analysis and interpretation

Integration Challenges

- Large and growing amount of molecular-biological data sources, e.g. Entrez, OMIM, Pubmed, GeneOntology, …
- High inter-connectivity of sources by means of cross-references (mappings)
- Dealing with different schemas, formats (e.g. XML, Fasta, GenBank), and semantics (e.g. gene definitions)
- Incompleteness of sources and their connecting mappings
- Constant schema and data changes → adaption and updates necessary

Example: Gene entry of Entrez

1: **AANAT** **arylalkylamine N-acetyltransferase** [*Homo sapiens*]
GeneID: 15 Locus tag: HGNC:19; MIM: 600950
Official Symbol: AANAT **and Name:** arylalkylamine N-acetyltransferase **provided by** HUGO Gene Nomenclature Committee
Transcripts and products: RefSeq below
Gene type: protein coding
Gene name: AANAT
Gene description: arylalkylamine N-acetyltransferase
RefSeq status: Reviewed
Organism: *Homo sapiens*
Phenotypes
Delayed sleep phase syndrome, susceptibility to MIM: 600950
Pathways
KEGG pathway: Tryptophan metabolism 00380
UniGene Hs.431417
MIM 600950
PharmGKB PA24366

Kegg
UniGene
OMIM
…

Cross-references between Entrez and other data sources

Related Work

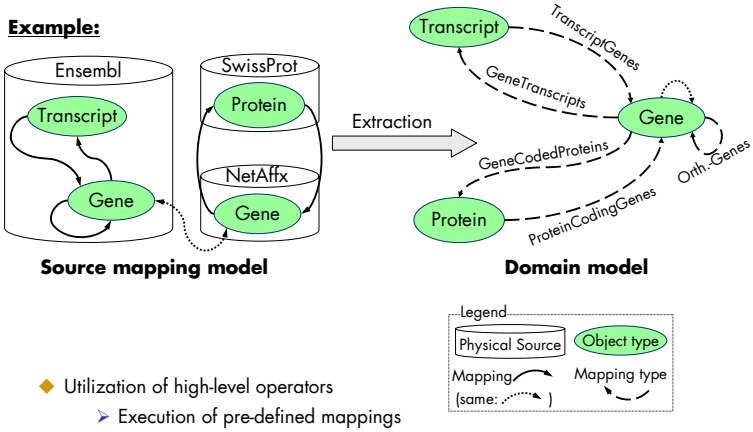
- Application-specific schema for consistent representation and access of integrated data [Stein 2003, Lacroix 2003]
  - Physical integration: IGD, GMS, GeneExpress
  - Virtual federated integration: TAMBIS, DiscoveryLink, K2/Kleisli
  - Difficult construction / maintenance of global schema, low scalability
- Sacrifice global schema for more flexibility [Lacroix 2003]
  - SRS, DBGET/LinkDB: Uniform query interface for many sources
  - Web-links: useful for interactive navigation, but not for large-scale analysis

Integration Approach

BioFuice = Biological Data Fusion utilizing Instance Correspondences and Peer Mappings

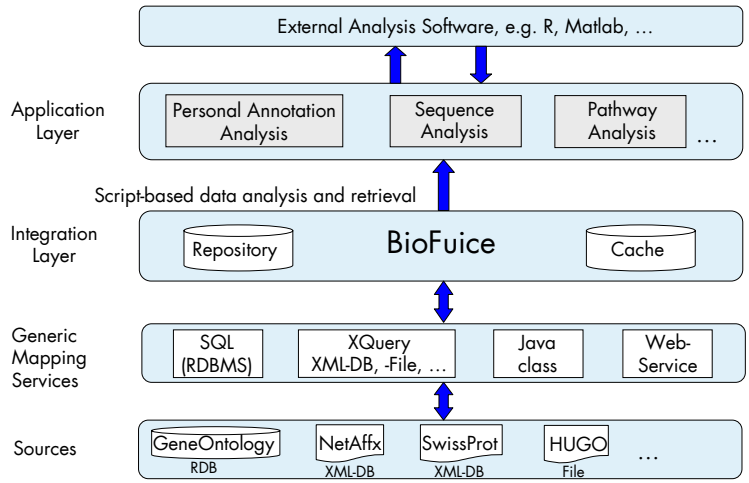
- Based on iFuice integration approach [Rahm 2005]
- **Bottom-up integration**: Prevention of designing a global target schema
- **P2P-like** infrastructure
  - Mappings between autonomous data sources
  - Easy link-up of a new source "where it fits best"
- Integration by using a **domain model** comprising
  - **Object types**: Refer to a real world entity, such as gene, protein etc.
  - **Mapping types**: Semantic correspondences between object types

Example:



- Utilization of high-level operators
  - Execution of pre-defined mappings
  - Combination within scripts to perform complex intregration tasks

BioFuice Integration Architecture



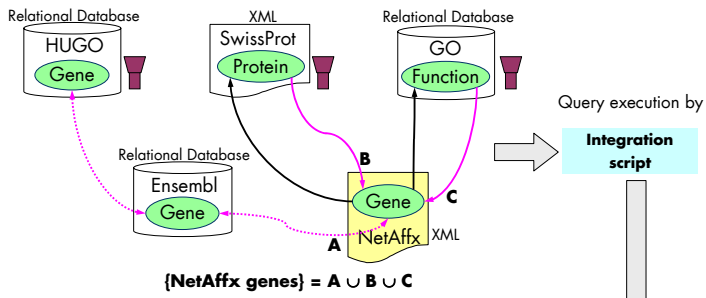
Annotation Analysis Example

Analysis Goal

Find human 'Chemokine' related NetAffx genes
→ Focused microarray-based gene expression analysis

- Querying multiple sources to overcome the incompleteness of single sources and mappings:
  - HUGO: Genes for which the chemokine relationship is documented
  - SwissProt: Already known chemokine proteins, e.g. published by [Tanaka 2005]
  - GeneOntology: Genes sharing the molecular function 'Chemokine'

Query Processing and Result



Result: List of NetAffx genes

1405_i_at, *chemokine (C-C motif) ligand 5, 17q11.2-q12, …*
1569203_at, *chemokine (C-X-C motif) ligand 2, 4q21, …*
202859_x_at, *interleukin 8, 4q13-q21, …*
203666_at, *chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1), 10q11.1*
…

References

[Rahm 2005] Rahm, Erhard et al.: iFuice – Information Fusion utilizing Instance Correspondences and Peer Mappings. 8th International Workshop on the Web & Databases (WebDB) in conjunction with SIGMOD 2005, Baltimore, 2005

[Lacroix 2003] Lacroix, Zoe; Critschlow, Terence: Bioinformatics – Managing scientific data. Morgan Kaufmann Publishers, 2003

[Stein 2003] Stein, Lincoln: Integrating Biological Databases. Nature Review Genetics, 4(5): 337-45, 2003

[Tanaka 2005] Tanaka, Toshiyuki et al.: Chemokines in tumor progression and metastasis. Cancer Science, Volume 96, Number 6, June 2005, pp. 317-322(6)

Printed by Universitätsrechenzentrum Leipzig