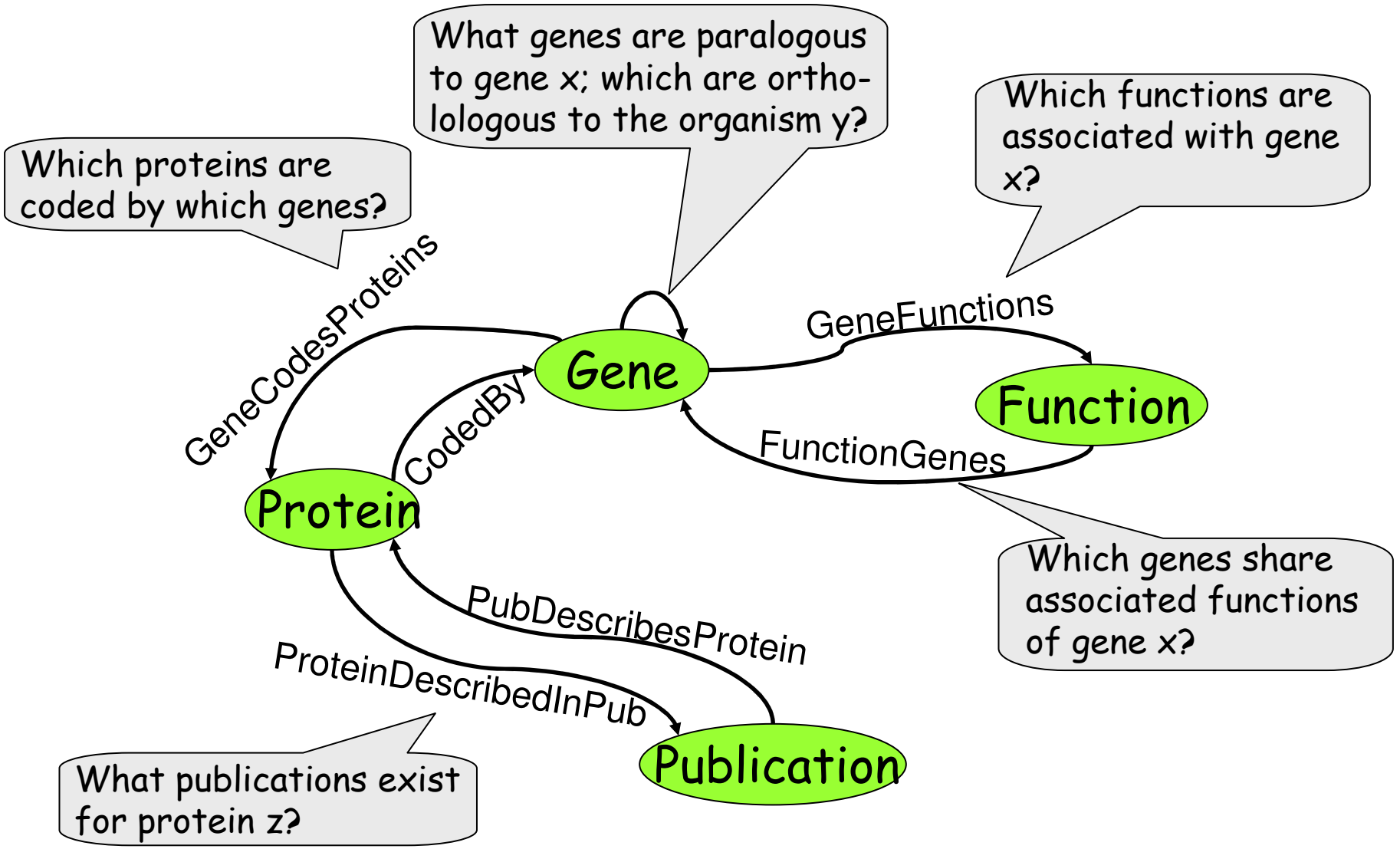
BioFuice: Current state and future developments

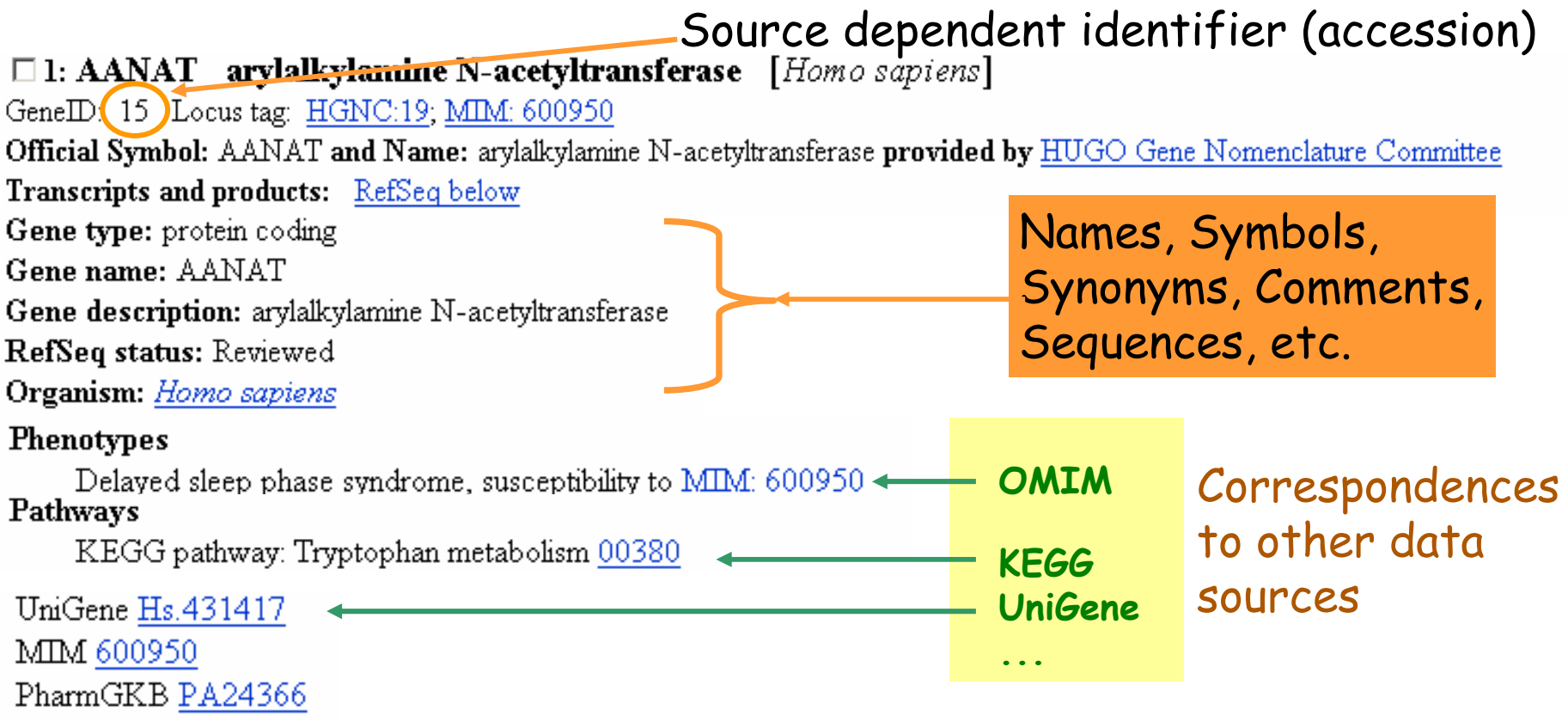
T. Kirsten, E. Rahm
University of Leipzig, Germany
www.izbi.de, dbs.uni-leipzig.de

BioDomain: Selected Object and Mapping Types



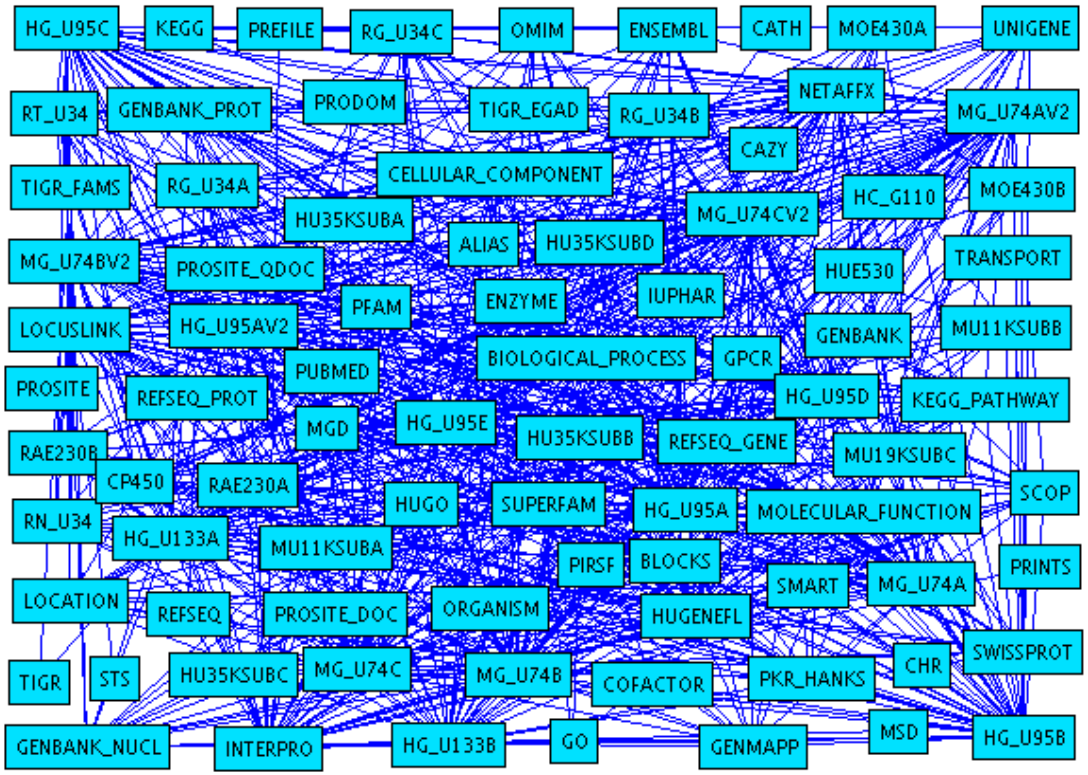
Characteristics of biological Sources

- Public available: Entrez, SwissProt, GeneOntology, …



Mapping Correspondences between two objects of two data sources

Data Integration Challenges



Annotation sources connected by primary attributes (accessions)

- **Many data sources**
- **Many mappings**

- **Heterogeneity**
  - Data formats
  - Schemas
  - Semantics

- **Data Quality**
  - Incompleteness
  - Data curation

- **Constant changes**
  - Data
  - Schemas

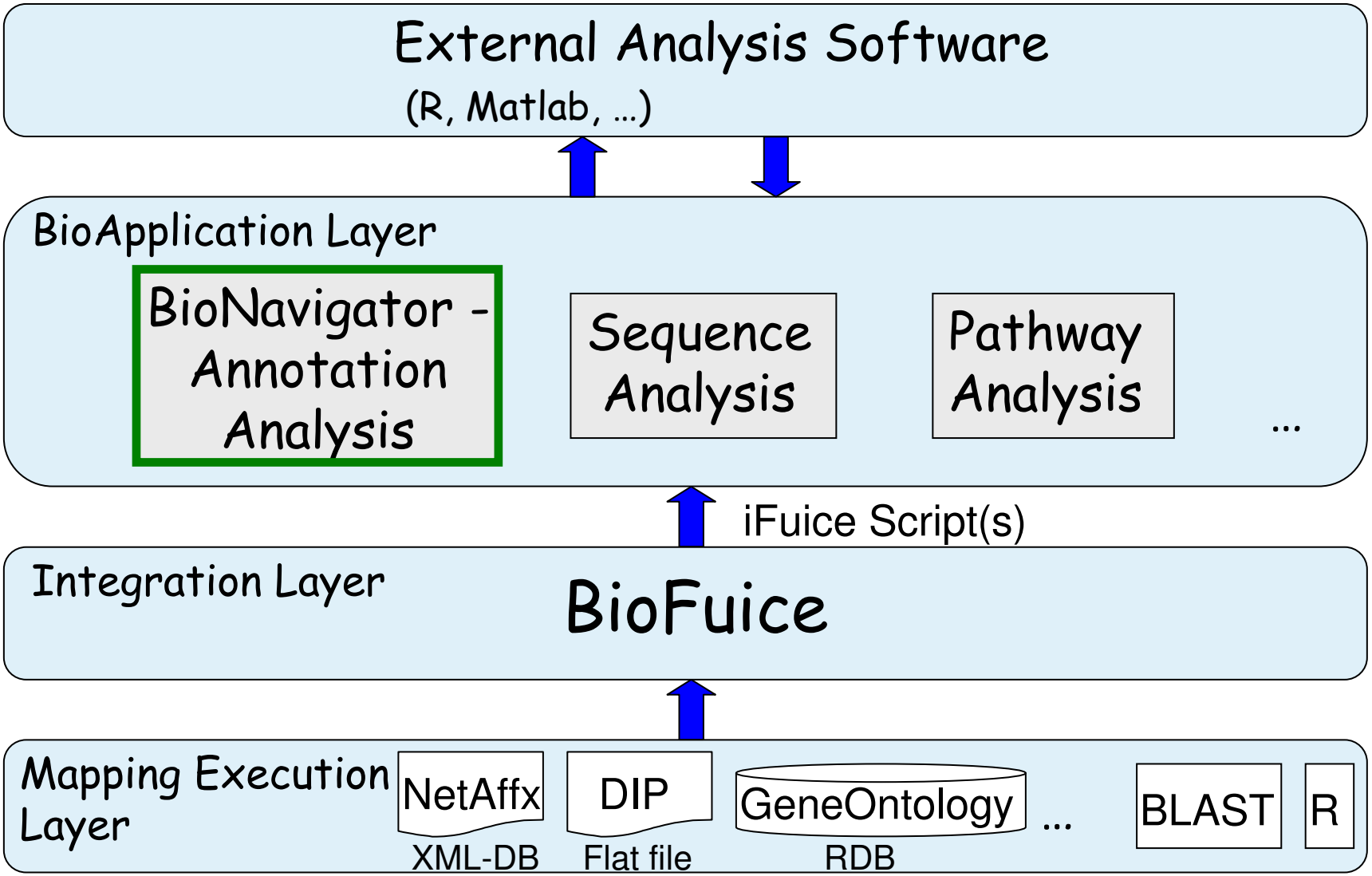
Outline

- Motivation
- BioFuice, BioApplications and integrated Sources
- Management of multiple SMMs
- Coupling with external Analysis Applications
- Conclusions

BioFuice

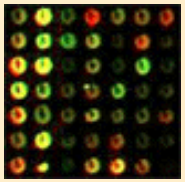
- BioFuice – Platform to fuse molecular biological data of different sources
  - Based upon the iFuice approach [Rahm et al, 2005]
  - P2P-like data management
  - Utilization of a domain model consisting of
    - Object types: Gene, Protein, Disease, Function, …
    - Mapping types: Semantic relationships between OT
  - Application of different generic mapping execution services, e.g. sql, java, xml-db, xml-file

BioApplications: Overview



BioApplication: Annotation Analysis with the BioNavigator

Experimental Data



Expression value,
P-Value,
...

GeWare

Data

Annotation Data

Objects: Gene, protein, ...
Attributes: Function,disease,
gene location, ...

Expression Analysis

Identification of relevant genes / proteins with expression data

Similar
expression patterns?

Analysis

gene groups

Annotation Analysis

Identification of relevant genes / proteins with annotation data

Functional groups?

BioApplication: BioNavigator - Example

- Example: Find 'Chemokine' related genes and return NetAffx identifiers
  - → Generate a gene group to focus on selected genes within the expression analysis

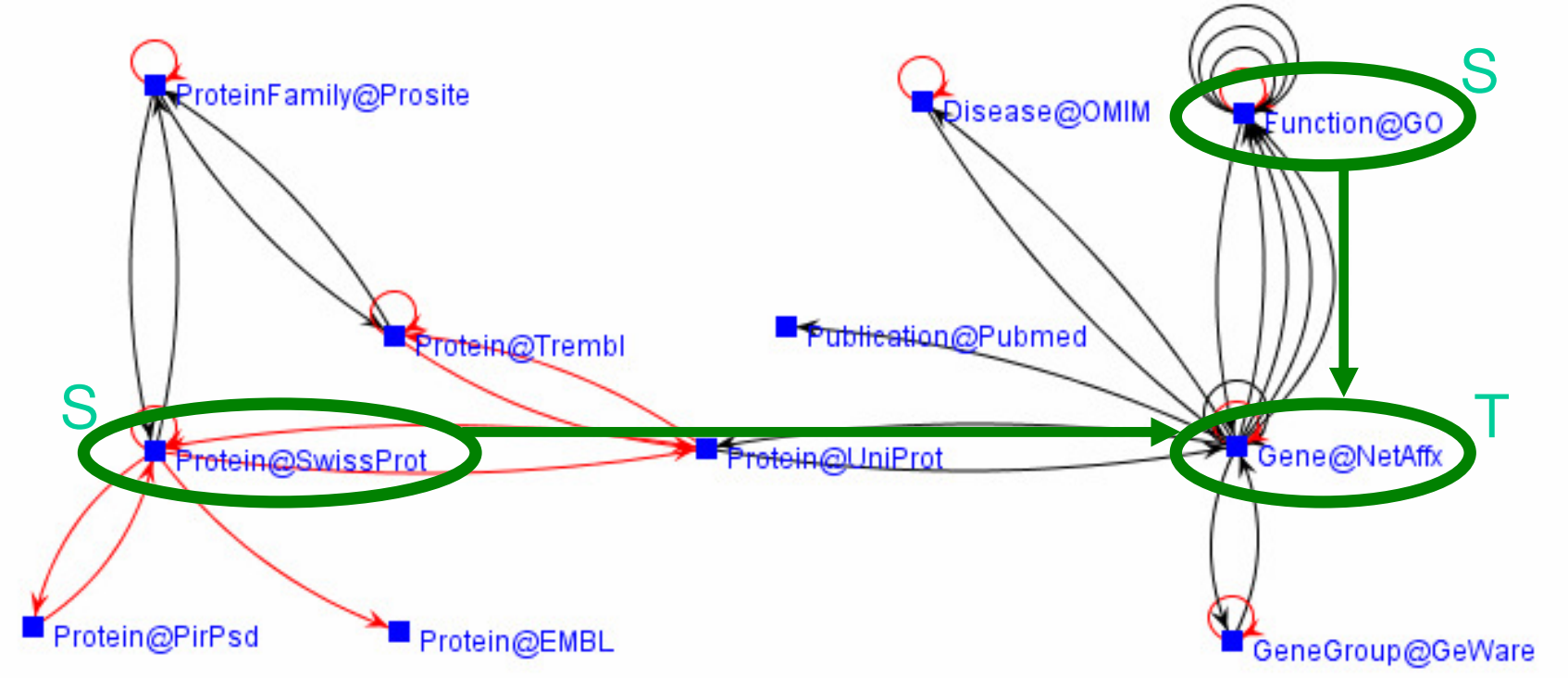
- Querying multiple sources due to incompleteness of single sources and mappings
  - Query source: SwissProt (Proteins)
  - Query source: GeneOntology (Functions, …)
  - Query target: NetAffx (Genes)
  - Intermed. source: UniProt (Proteins)

[Tanaka 2005]

Table 1. Chemokines and their receptors. Chemokines consisting of four major subfamilies (CXC, CC, XC, CX3C) are listed here together with their original names. Major receptors for each chemokine are also shown, although some chemokines may bind other receptors. Chemokines and their receptors identified in humans are listed here

New official name	Original name (other names may exist)	Receptor(s)
CXCL1	GROα – growth related oncogene α	CXCR2 > CXCR1
CXCL2	GROβ – growth related oncogene β	CXCR2
CXCL3	GROγ – growth related oncogene γ	CXCR2
CXCL4	PF-4 – platelet factor 4	Unknown
CXCL5	ENA-78 – epithelial cell derived neutrophil activating factor 78	CXCR2
CXCL6	GCP-2 – granulocyte chemoattractant protein 2	CXCR1, CXCR2
CXCL7	NAP-2 – neutrophil activating protein 2	CXCR1, CXCR2
CXCL8	IL-8 – interleukin 8	CXCR1, CXCR2
CXCL9	MIG – monokine induced by interferon-γ	CXCR3
CXCL10	IP-10 – γ interferon inducible ptrotein 10	CXCR3
CXCL11	I-TAC – interferon inducible T cell α-chemoattractant	CXCR3
CXCL12	SDF-1 – stromal cell derived factor 1	CXCR4
CXCL13	BCA-1–B cell activating chemokine 1	CXCR5
CXCL14	BRAK – breast and kidney chemokine	Unknown
CXCL15	Unknown	Unknown
CXCL16	SR-PSOX – scavenger receptor that binds phosphatidylserine and oxidized lipoprotein	CXCR6
CCL1	I-309	CCR8
CCL2	MCP-1 – monocyte chemoattractant protein 1	CCR2
CCL3	MIP-1α – macrophage inflammatory protein 1α	CCR1, CCR5
CCL4	MIP-1β – macrophage inflammatory protein 1β	CCR5
CCL5	RANTES – regulated on activation, normally T cell expressed and secreted	CCR1, CCR3, CCR5
CCL6	Unknown	CCR1, CCR2, CCR3
CCL7	MCP-3 – monocyte chemoattractant protein 3	CCR1, CCR2, CCR3
CCL8	MCP-2 – monocyte chemoattractant protein 2	CCR2, CCR3, CCR5
CCL9/10	Unknown	CCR1
CCL11	Eotaxin	CCR3
CCL12	Unknown	CCR2
CCL13	MCP-4 – monocyte chemoattractant protein 4	CCR1, CCR2, CCR3
CCL14	HCC-1 – hemofiltrate CC chemokine	CCR1

BioApplication: BioNavigator - Example cont.



- Different source formats and size
  - RDB: GeneOntology (≈350MB), GeWare
  - XML-DB: SwissProt (≈1,7GB), Trembl (>10GB), PirPsd (≈800MB), NetAffx (ca. 500MB per chip type)
  - Online access: Pubmed, OMIM
  - Special formats: Prosite → RDB (≈10MB)

BioApplication: BioNavigator - Example cont.

- iFuice Script

```
$ProteinGenes := queryTraverse(Protein@SwissProt,
                               [fn:contains($a/protein/name,'CXCR')],
                               {SwissProt2UniProt,UniProt2NetAffx});
$FunctionGenes := queryTraverse(Function@GeneOntology,
                               [term_name like 'Chemokine%'],
                               {Go2NetAffx});
$UnionGenes :=union($ProteinGenes, $FunctionGenes);
$Result := sort($UnionGenes, [accession] asc);
```

- Result set (accession, name, chromosomal location, …)

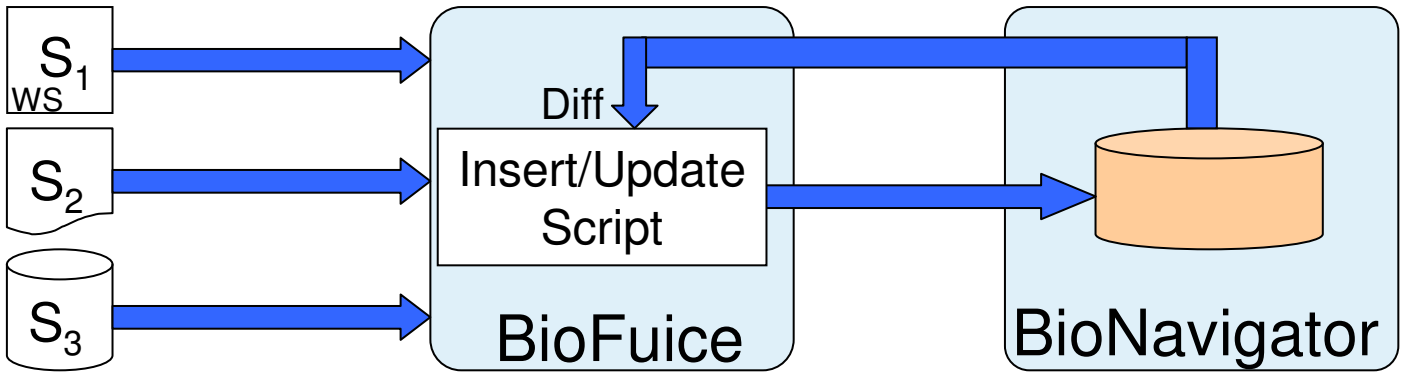
1405_i_at, *chemokine (C-C motif) ligand 5, 17q11.2-q12, …*
1569203_at, *chemokine (C-X-C motif) ligand 2, 4q21, …*
202859_x_at, *interleukin 8, 4q13-q21, …*
203666_at, *chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1), 10q11.1*
…

BioApplication: BioNavigator - Requirements

- Graphical interface for browsing
  - Tree based
  - Graphical script generation: Definition of sources, query conditions, (mappings) and target(s) based upon the SMM
- LDS specific keyword search
- Local management of selected objects (OI,MR,...)
  - Materialization of objects of interest, i.e. public data and "private" data (e.g. personal gene list)
  - Flexible schema management due to fast evolving SMMs
  - Incremental periodic updates of data
- Export capability / Coupling with existing analysis software

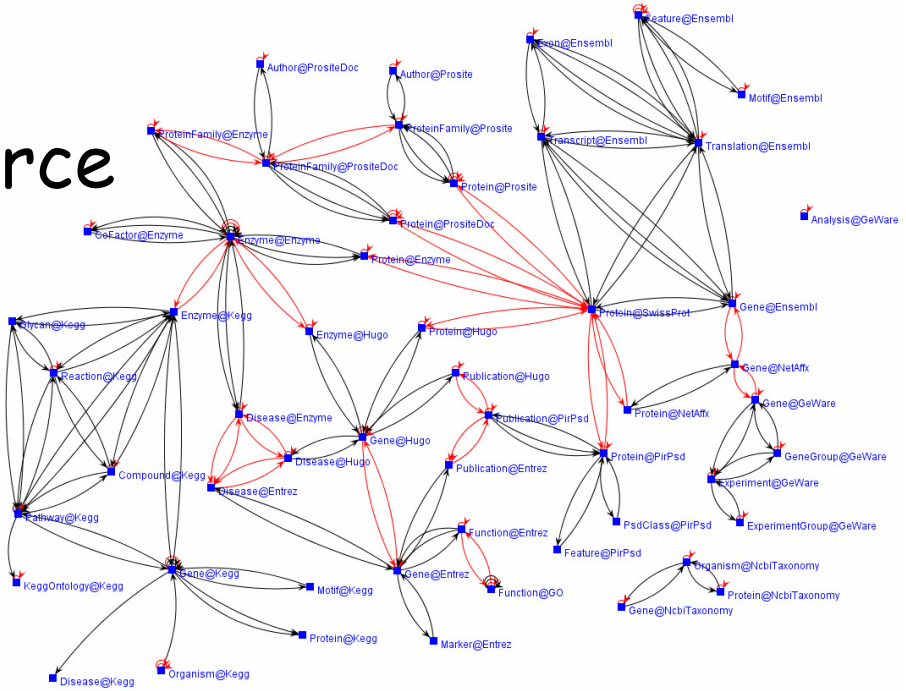
BioApplication: BioNavigator- Very early Ideas

- Schema Management
  - Utilization of SMM as BioNavigator schema
  - Schema changes:
    - Recognized by comparing BioFuice SMM and BioNavigator schema
    - Add new LDSs and attributes when BioFuice SMM provide new one
    - Complete local schema refresh by user action
- Data Management: Relation Database (XML-file)
  - Possibly incremental updates
  - Idea: Utilze the BioNavigator source and compute the difference within the iFuice "insert/update" script



Management of multiple SMMs: Motivation

- Complex domains (e.g. Bioinformatics) comprising numerous physical sources or object types
- User behavior: Extension of available SMM and include new "interesting" LDS and mappings
- Result: Ever growing source mapping model (SMM) with many LDS and mappings



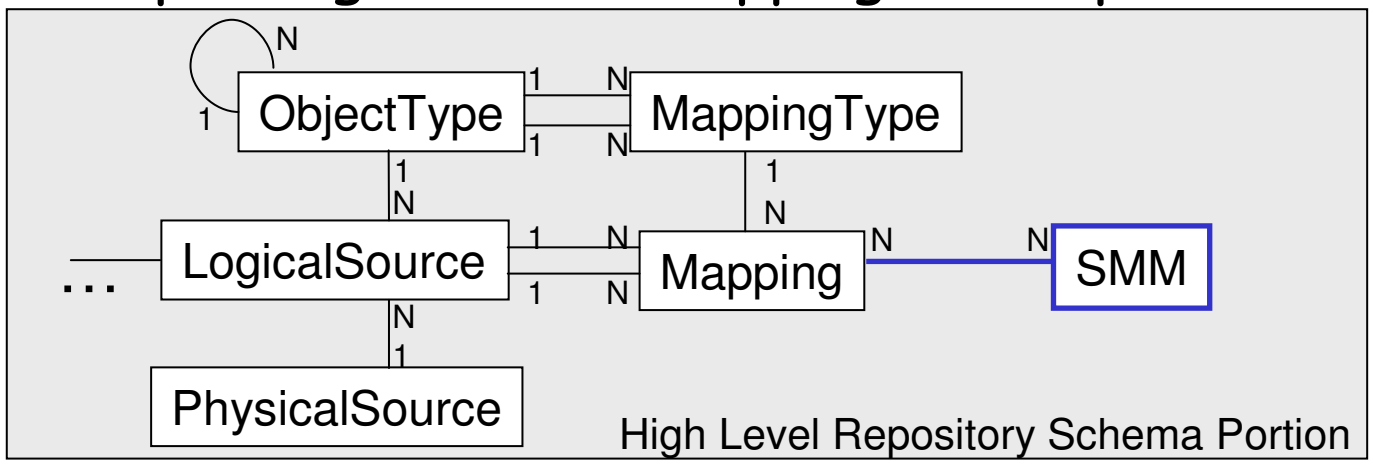
Management of multiple SMMs

- Solution: Application specific SMMs
  - SMM with as many as necessary mappings
  - Multiple applications → multiple SMMs

- Currently, file based definition at boot time:
  - Processing all files of a specific directory
  - Change by file move, cut & paste
  - Model change/exchange only by platform shutdown + reboot
  - → Laborious and error-prone

Management of multiple SMMs cont.

- Optionally XML-based SMM specification by sets of mapping (LDS, PDS) names
  - Single vs. multiple SMM specifications
  - iFuice initialization by using a default or specified SMM
  - Management within the iFuice repository
  - Online exchange and iFuice re-initialization without re-importing source + mapping descriptions



Coupling with external Analysis Applications

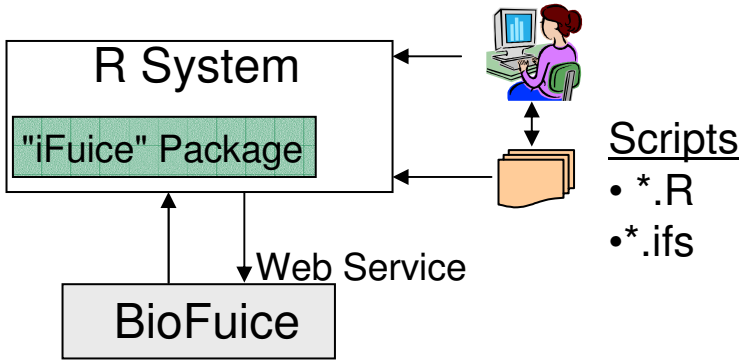
- Numerous analysis applications available
  - Fields of operation: Homology Search, …
  - Tools: R, Matlab, BLAST, RNA-Package, …
- Goal: Analysis of data from different sources with existing applications
  - Avoiding manual export/import
  - Automatic analysis workflows
- Two possible approaches:
  - iFuice as single data provider
    - Utilization of the iFuice interface, e.g. Web Service interface, to build application specific iFuice Wrappers
  - iFuice as Workflow Management System
    - Wrapping applications in mappings

Coupling with external Analysis Applications: "iFuice" Package for R

- R: Freely available, statistical software
  - Popular for analyzing expression (experimental) data in the Bioinformatics domain
  - Various packages providing different analysis routines and annotation data
  - BUT: Static and limited set of annotation data
    → periodic package updates necessary!
- Goal: Combined analysis of experimental and a wide range of annotation data with advanced routines, e.g. gene classification
- → Design of the "iFuice" package

Coupling with external Analysis Applications: "iFuice" Package for R cont.

- Retrieves annotation data from a connected iFuice web service instance



- Set of functions
  - Management and metadata functions:
    - connect(platform$_{iFuice}$), disconnect()
    - getLdsNames(), getMappingNames(), getOpertors()
    - clearCache(), clearVariables()
  - "get" functions:
    - executeCommand(cmd), executeScript(script name)
  - "set" functions (w.i.p.):
    - storeObjects(OI, OT, PDS, Var)
    - storeMappingResult(MR, OT_{input},PDS_{input},OT_{output},PDS_{output},Var)

Coupling with external Analysis Applications: "iFuice" Package for R cont.

R Script: MyFirstExpressionAnalysisUsingBioFuice.R

```
…
library(iFuice)
connect("http://ducati.izbi.uni-leipzig.de:8080/axis/services/IFuice")
queryResult ← executeScript("~/ChemokineNetAffxGenes.ifs")

# utilization of queryResult$Result for focussed expression analysis
…
```

```
$ProteinGenes := queryTraverse(Protein@SwissProt,
                               [fn:contains($a/protein/name,'CXCR')],
                               {SwissProt2UniProt,UniProt2NetAffx});
$FunctionGenes := queryTraverse(Function@GeneOntology,
                               [term_name like 'Chemokine%'],
                               {Go2NetAffx});
$UnionGenes :=union($ProteinGenes, $FunctionGenes);
$Result := sort($UnionGenes, [accession] asc);
```

iFuice Script: ChemokineNetAffxGenes.ifs

```
> queryResult$Result
   accession                                                    name
1   1405_i_at                                chemokine (C-C motif) ligand 5
2  1569203_at                               chemokine (C-X-C motif) ligand 2
3 202859_x_at                                                interleukin 8
4   203666_at chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor)
  chromosomal.location
1          17q11.2-q12
2                 4q21
3             4q13-q21
4              10q11.1
> 
```

Shell | Shell No. 2 | Shell No. 3

Conclusions

- Bioinformatics as complex domain, many sources & mappings
- BioFuice – a valuable integration platform
  - Based upon the iFuice approach to integrate data of different heterogeneous sources
- Different BioApplications
  - BioNavigator, Sequence Analysis, Pathway Analysis
- Management of multiple SMMs
- "iFuice" R package for coupling BioFuice with existing analysis applications