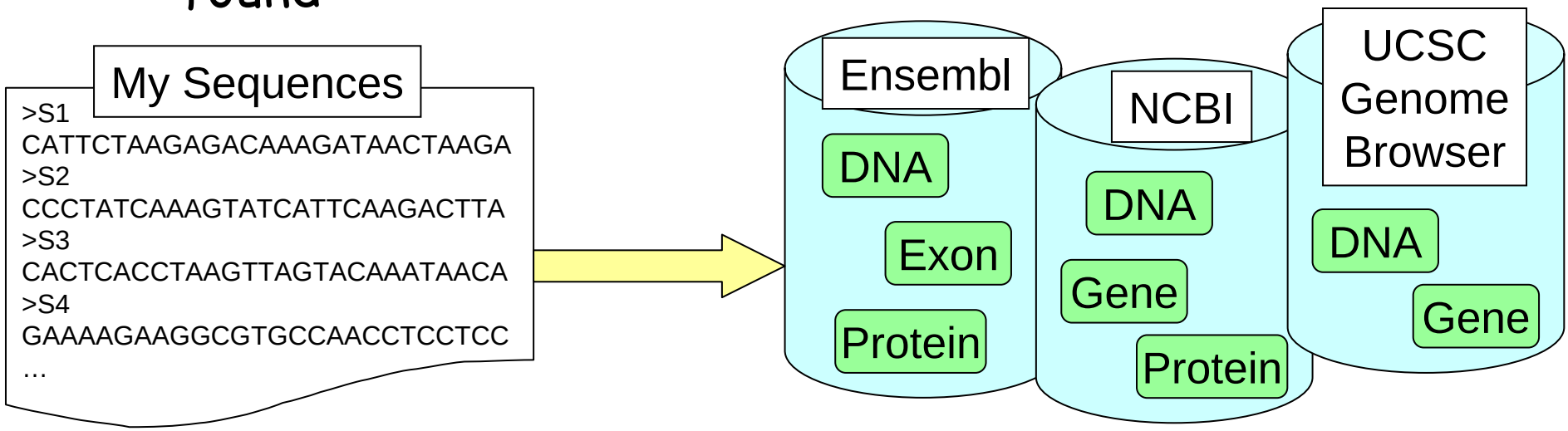
BioFuice: Mapping-based Data Integration in Bioinformatics

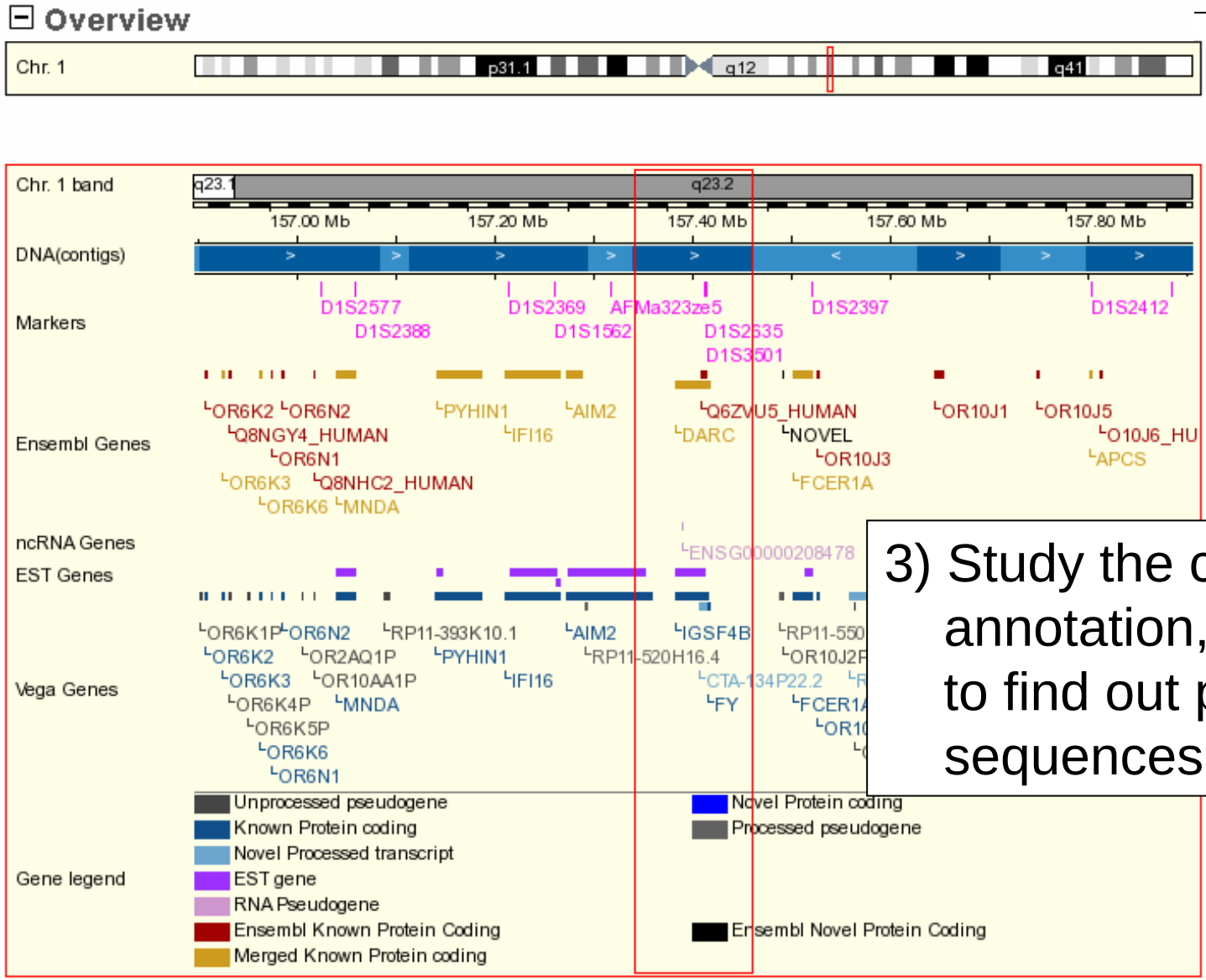
T. Kirsten, E. Rahm
University of Leipzig, Germany
www.izbi.de, dbs.uni-leipzig.de

Motivating Scenario

- Goal: Given a set of DNA sequences, classify them into
  - Sequences associated with protein-coding DNA (show more info)
  - Sequences associated with non-coding DNA
  - Sequences for which no corresponding DNA can be found



Manual Solution



sequence file
embl system

3) Study the corresponding annotation, e.g. genes, to find out protein coding sequences

ned sequences

Data Integration Challenges

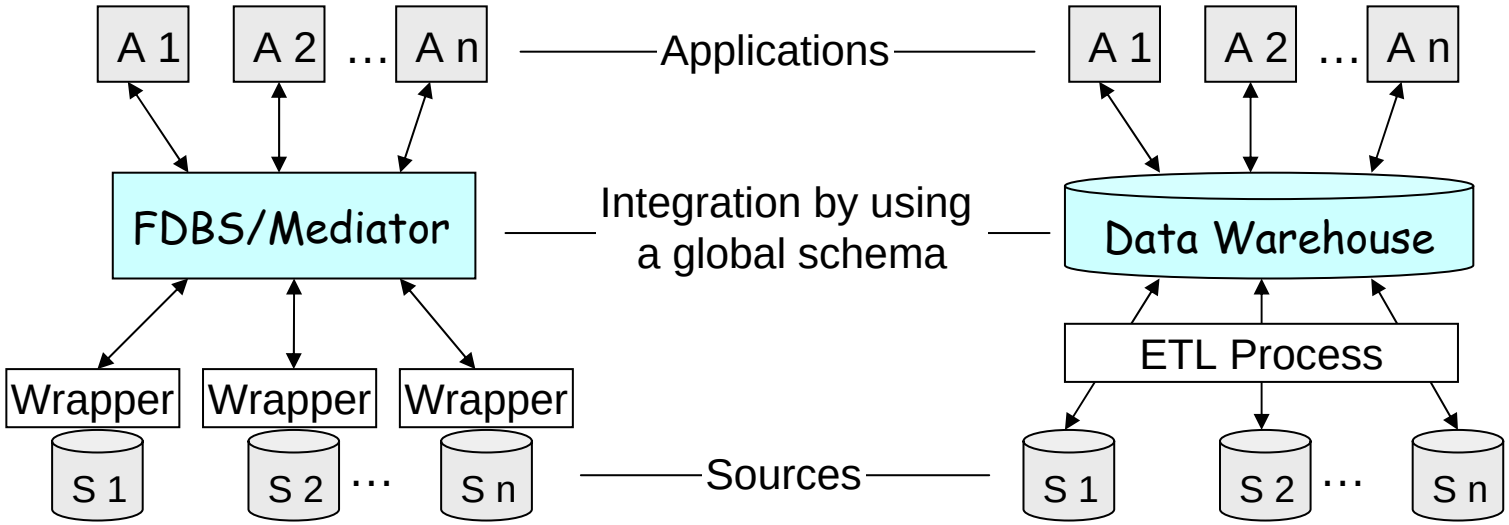
- Many sources, high connectivity
- High heterogeneity
  - Data formats (syntax)
  - Schemas
  - Semantics
- Constant changes of schemas and data
- Data quality (incompleteness of sources and their interconnections)
- Integration of local data sources, e.g. private gene list
- Support of ad-hoc workflows

Outline

- Motivating Scenario and Integration Challenges
- Typical Integration Approaches
- Mapping-based Data Integration
- BioFuice Architecture
- Query Processing
- Conclusions

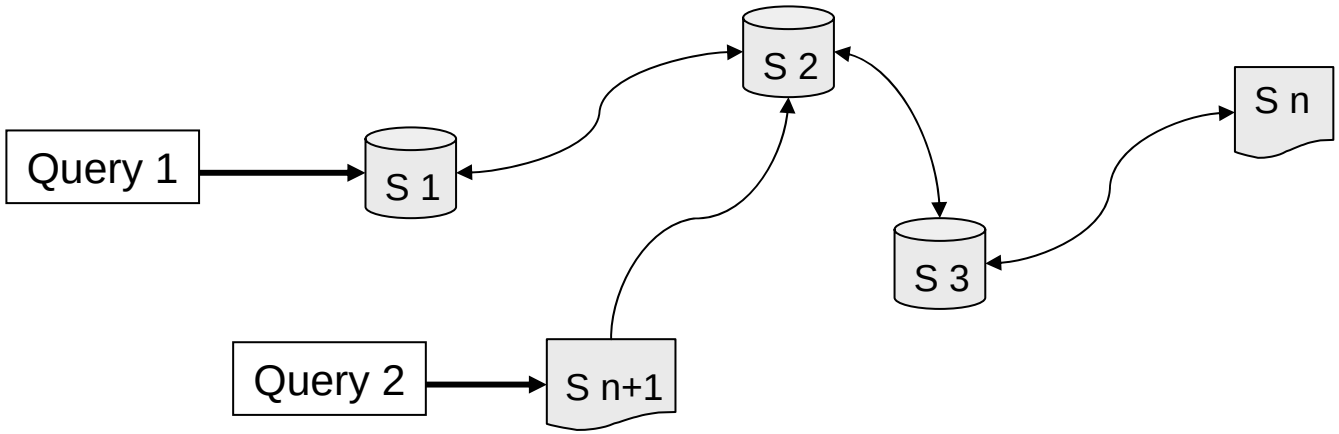
Typical Integration Approaches

- Types: Virtual (e.g. FDBS, mediator), physical integration (e.g. data warehouse)
- Resource-intensive construction and maintenance of
  - Application-specific global schema and
  - Schema mappings between each source and global schema



Mapping-based Data Integration

- Mapping-based integration (peer data mgmt.)
  - No global schema
  - Utilization of bidirectional connections between heterogeneous sources, e.g. based on existing instance correspondences (mapping)
  - Easy link up of new sources (incl. local sources)
  - Querying selected sources and propagating it to relevant neighbors



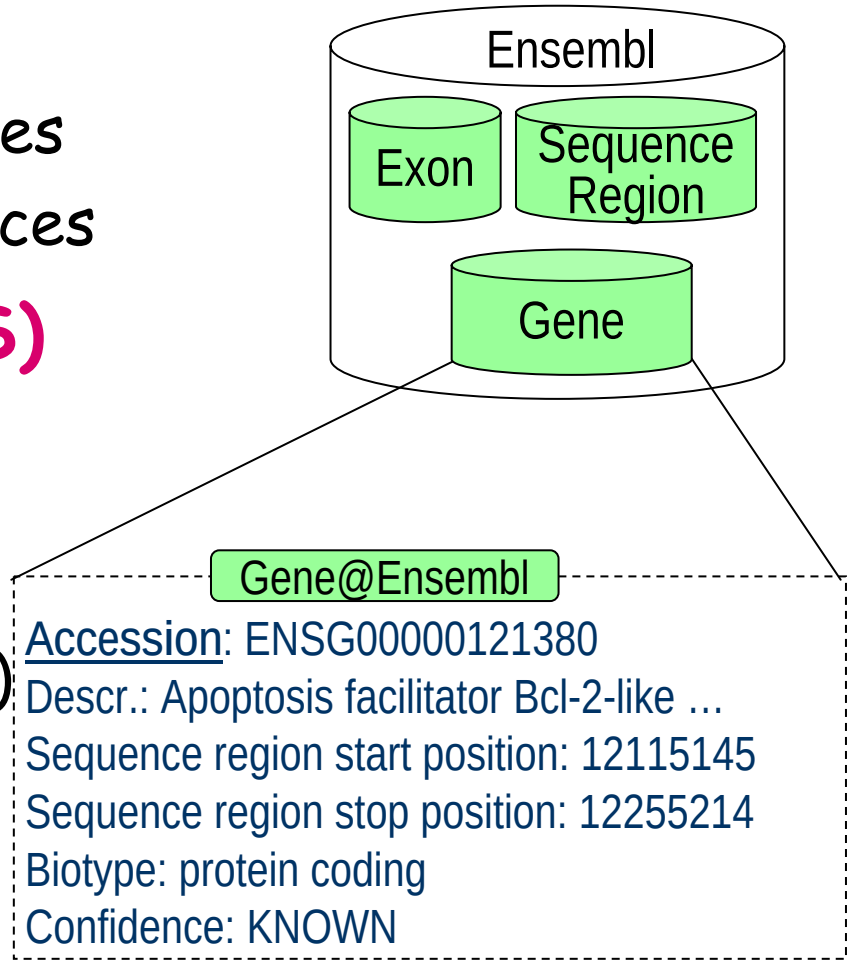
BioFuice

- **BioFuice: Bioinformatics information fusion utilizing instance correspondences and peer mappings**
- Basis: iFuice approach*
  - Bottom up integration
  - **High-level operators**
- P2P-like infrastructure
  - Mappings between autonomous data sources (peers)
  - Mapping: Set of instance correspondences
  - Simple integration of new sources
- Mediator
  - Controlling of mapping- and operator execution
  - Utilization of application specific semantic domain model

* Rahm, Thor, Aumüller, Do, Golovin, Kirsten: iFuice - Information fusion utilizing instance correspondences and peer mappings. Proc. of WebDB, Baltimore, 2005

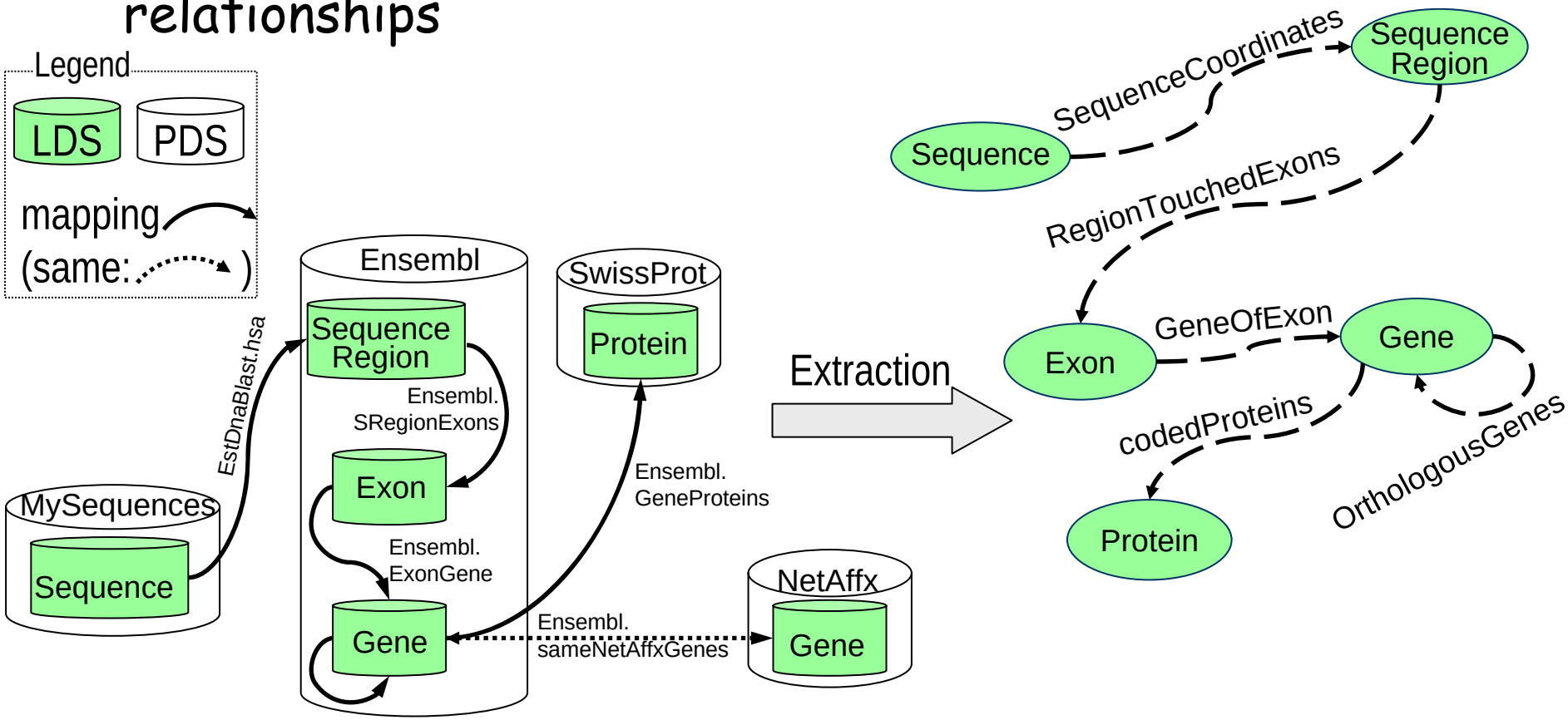
Data Sources

- **Physical data source (PDS)**
  - Public, private and local data (gene list, …), ontologies
  - Subsumes logical data sources
- **Logical data source (LDS)**
  - Refers to an object type and a physical data source, e.g. Gene@Ensembl
  - Contains objects(-instance)
- Object instances
  - Set of relevant attributes
  - One id attribute



Metadata Models

- Used by mediator for mapping/operator execution
- **Domain model** indicates available object types and relationships



Source mapping model **Domain model**

Operators

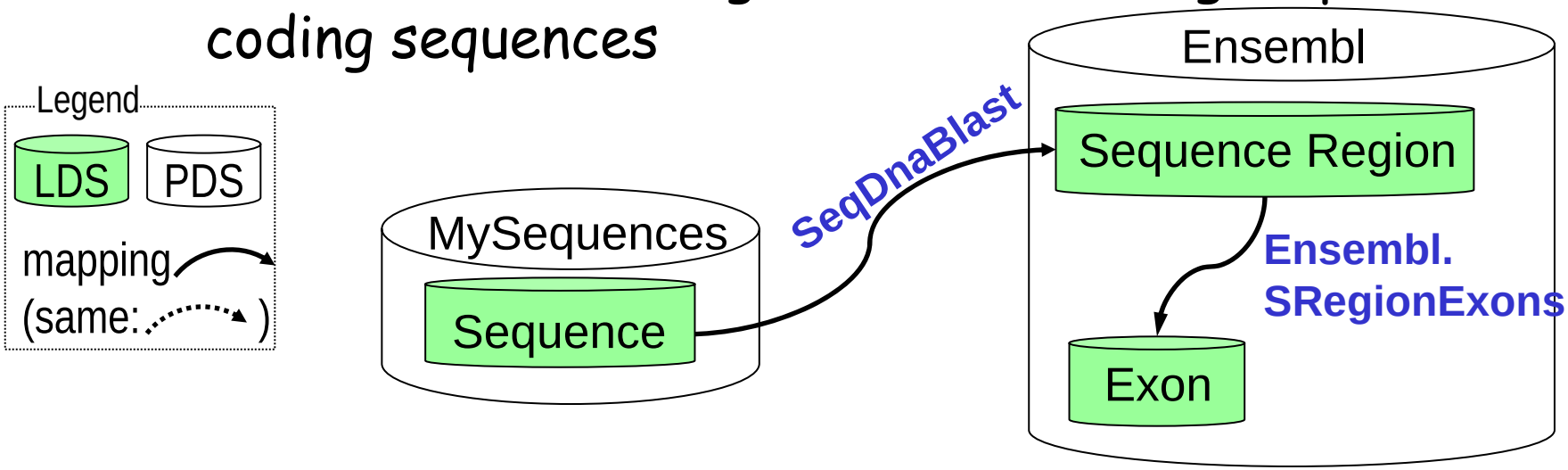
- **Set oriented operators**
  - Input: Set of objects/mappings
    + parameters / query conditions
  - Output: Set of resulting objects

⇒ Combination of operators within scripts for workflow-like execution

- Selected operators:
  - Single source: **queryInstances, searchInstances, …**
  - Navigation: **traverse, map, compose, …**
  - Navigation + aggregation: **aggregate, aggregateTraverse, …**
  - Generic: **diff, union, intersect, …**

Script Example

- Script to solve motivating scenario
  - Three classes: unaligned s., non-coding s., protein coding sequences

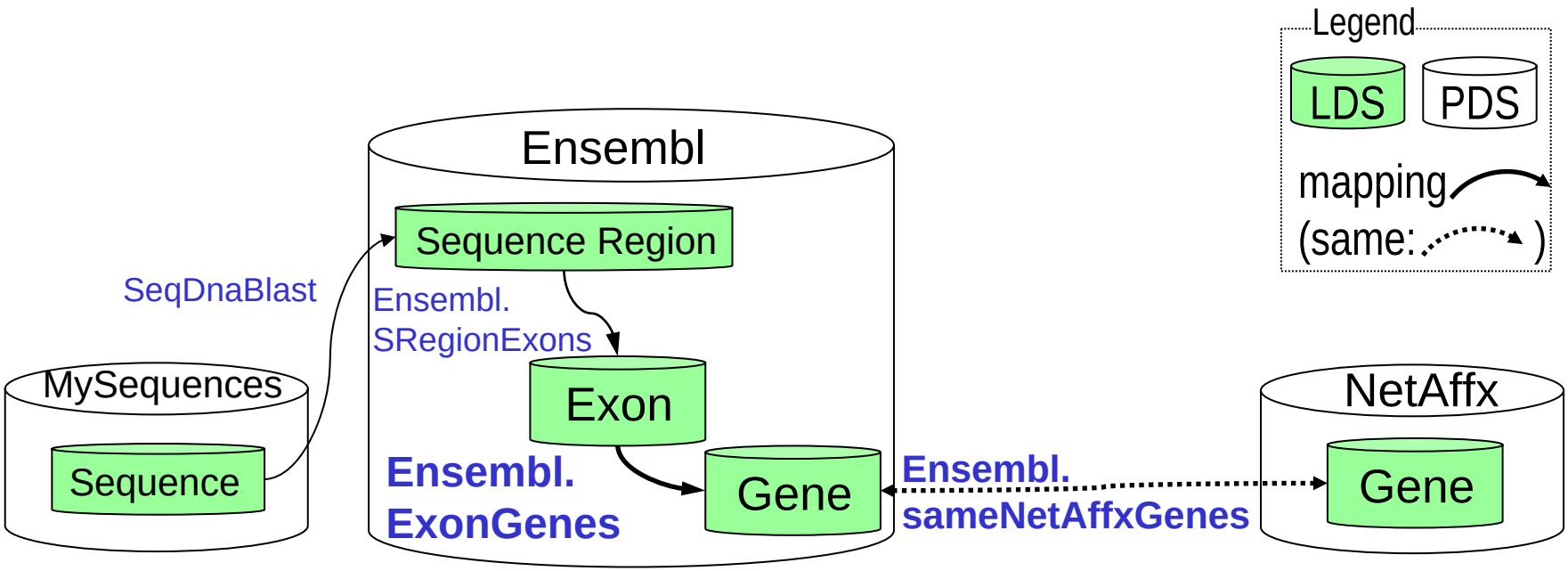



```
$alignedSeqMR := map( MySequences, { SeqDnaBlast } );
$codingSeqMR  := compose( $alignedSeqMR, { Ensembl.SRegionExons } );

$unalignedSeqOI    := diff ( MySequences, domain ( $alignedSeqMR ));
$protCodingSeqOI   := domain ( $codingSeqMR );
$nonCodingSeqOI    := diff ( domain ( $alignedSeqMR ) , $protCodingSeqOI );
```

Aggregation

- Associate and fuse genes of different sources, e.g. for Ensembl and NetAffx




```
$GeneOI := traverse ( range ($codingSeqMR ), {Ensembl.ExonGenes});
$fusedGeneAO  := aggregateSame ( $GeneOI, NetAffx );
```

Aggregation cont.

Overview – Aggregated Objects (i.e. Genes)

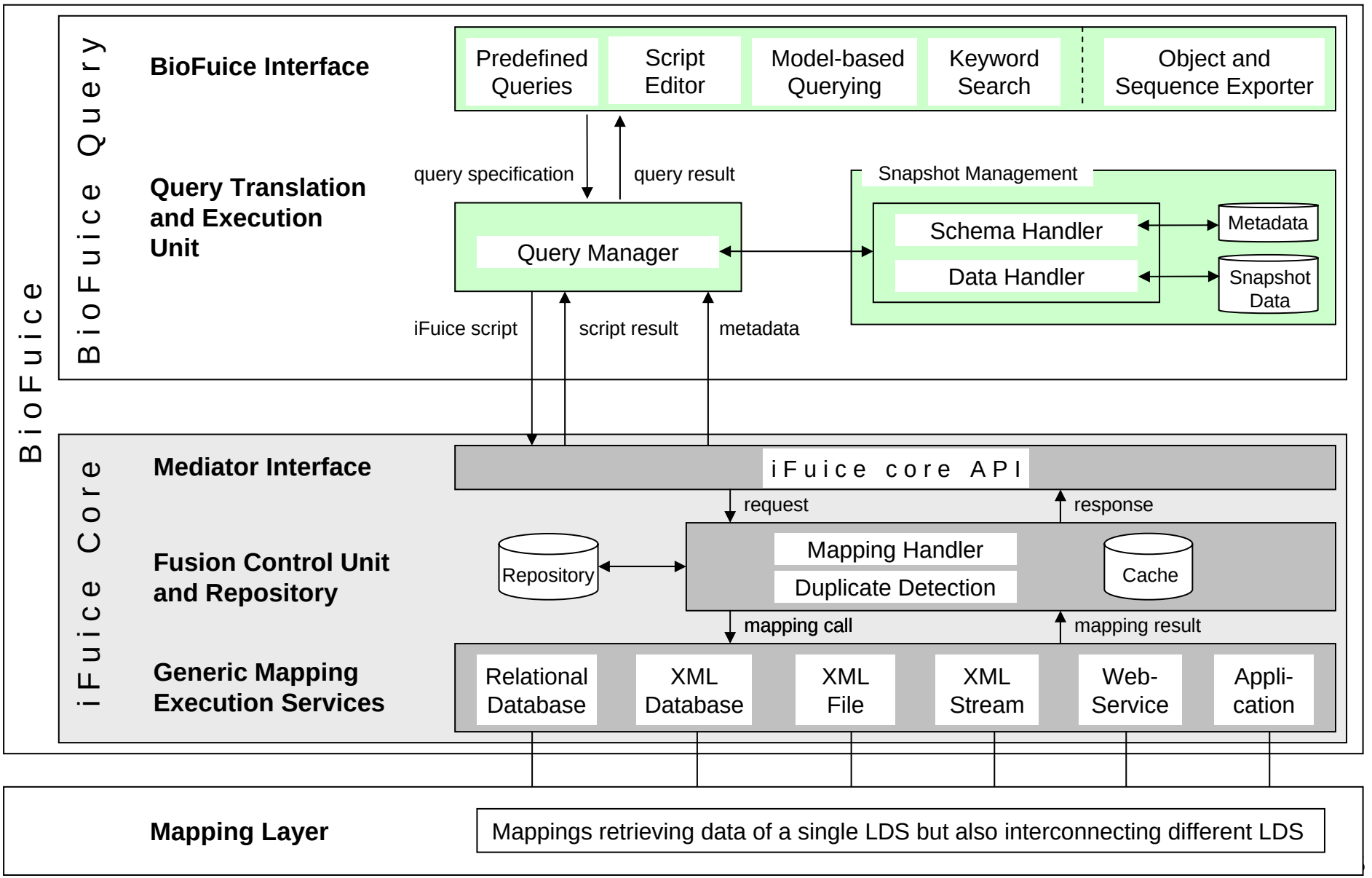
No.	Logical source	Item
1	Gene@{Ensembl,NetAffx}	ENSG00000170581,205170_at
2	Gene@{Ensembl,NetAffx}	ENSG00000166888,2[illegible]s_at
3	Gene@{Ensembl,NetAffx}	ENSG00000173757,1555086_[illegible]05026_at,212549_at,212550_at
4	Gene@{Ensembl,NetAffx}	ENSG00000126561,203010_at
5	Gene@{Ensembl}	ENSG0000016861
6	Gene@{Ensembl}	ENSG00000115415
7	Gene@{Ensembl}	ENSG00000138378

Details – Attributes

		Attribute value
1	Ensembl.accession	ENSG00000170581
2	Ensembl.status	KNOWN
3	Ensembl.source	ensembl
4	Ensembl.bioType	protein_coding
5	Ensembl.strand	-1
6	Ensembl.version	2
7	Ensembl.taxon	9606
8	Ensembl.chromosome	12
9	Ensembl.chromosomeStop	55040176
10	Ensembl.chromosomeStart	55021651
11	Ensembl.description	Signal transducer and act[illegible]ator [illegible]ranscription 2 (p113). [Source:Uniprot/SWISSP…
12	NetAffx.accession	205170_at
13	NetAffx.name	signal transducer and act[illegible]ator of transcription 2, 113kDa

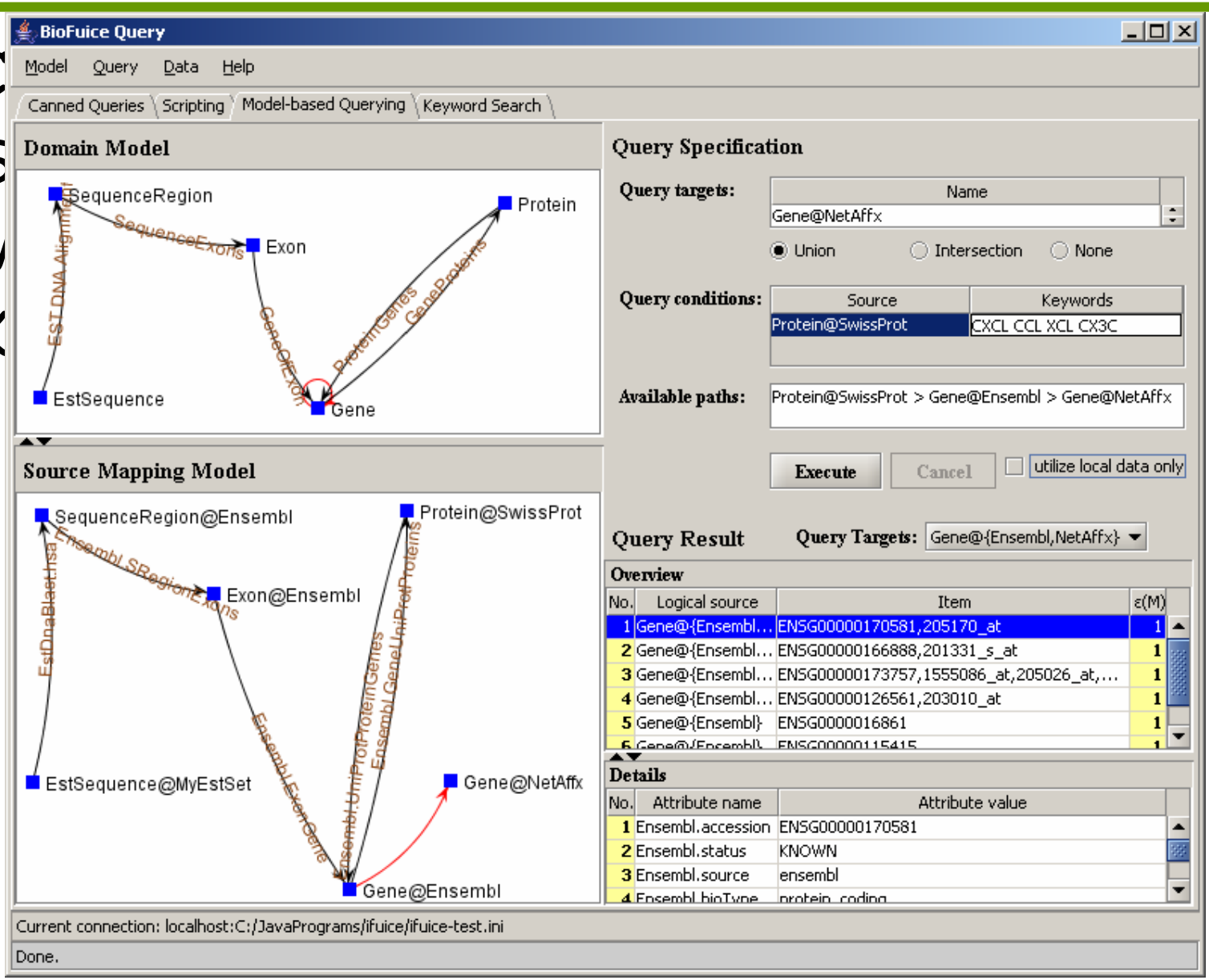
Attribute fusion

System Architecture



Query Processing

- C
- S
- M
- K



Current BioFuice Applications

- Gene expression analysis
  - Sources: Various publicly available + private lists of objects (genes, proteins, …)
  - Find genes of interest to focus the microarray analysis
  - Interpretation and validation of found gene sets
- Analysis of large protein interaction networks
  - Sources: DIP, MINT, BIND
  - Goal: Find network properties to characterize the interplay between behavior, structure and function
- Analysis of non-coding RNA
  - Associate private RNA lists to annotations in Ensembl and GeneOntology
  - Goal: Determination of secondary structure and function

Conclusions & Future Work

- Bioinformatics as complex domain, many sources & mappings
- BioFuice
  - P2P-like infrastructure to integrate data of different heterogeneous sources
  - Domain model using semantic object and mapping types
  - Different operators for query and mapping execution
  - Several applications: Expression ~, protein interaction ~ and non-coding RNA analysis
- Future work:
  - Integration of different analysis applications to create complex analysis workflows (analysis pipelines)

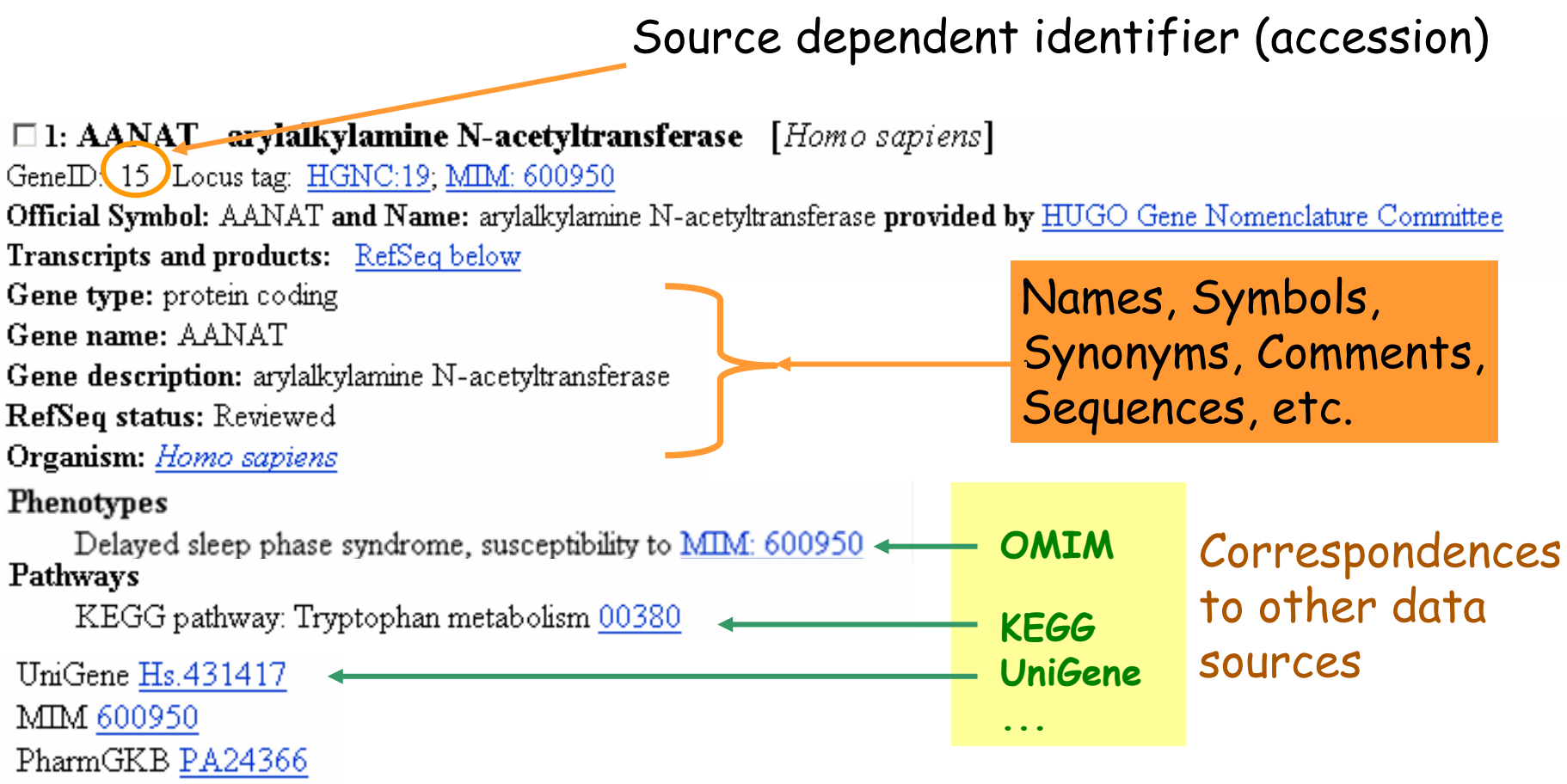
Acknowledgements

- DBS Group, Univ. Leipzig
  - Andreas Thor
  - David Aumüller
  - Nick Golovin
- SAP Research, Dresden
  - Hong-Hai Do
- BioInf, Univ. Leipzig
  - Peter Stadler
  - Claudia Fried
  - Manja Lindemeyer
- TBI, Univ. Vienna
  - Andrea Tanzer
- MPI Mathe. in Sciences, Leipzig
  - Jürgen Jost
  - Anirban Banerjee

Further Information

http://dbs.uni-leipzig.de
http://www.izbi.de

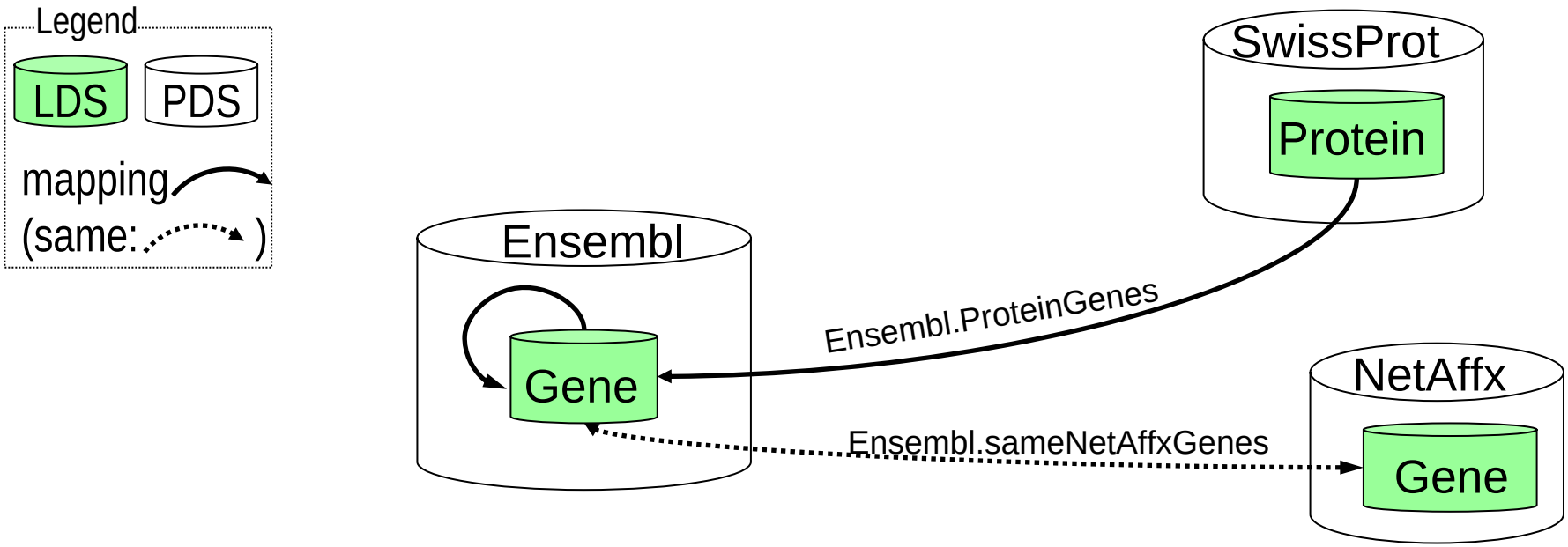
Example: Web-based Source NCBI Entrez



Simple Script Example

- Goal: Return all genes of NetAffx which are associated with Chemokine* proteins

```
$Proteins:=searchInstances(Protein@SwissProt,"CXCL CCL XCL CX3C");
$Genes:=traverse($Proteins, {Ensembl.ProtGenes, Ensembl.sameNetAffxGenes});
```



*Tanaka et al.: Chemokines in tumor progression and metastasis. Cancer Science, 96(6): 317-322, 2005