## **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**



#### **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



## 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: <u>Bio</u>informatics information <u>f</u>usion <u>u</u>tilizing instance <u>c</u>orrespondences and <u>peer</u> 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**



#### Metadata Models

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



#### **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
     Ensembl



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

\$unalignedSeqOl := diff ( MySequences, domain ( \$alignedSeqMR ));
\$protCodingSeqOl := domain ( \$codingSeqMR );
\$nonCodingSeqOl := diff ( domain ( \$alignedSeqMR ) , \$protCodingSeqOl );



 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.

O	verview – Aggre	egated	Objects (i.e. Ge	nes)			
L	No. Logical sou	irce		Item			
	1 Gene@{Ensembl,N	etAffx} E	ENSG00000170581,205170_at				
	2 Gene@{Ensembl,No	etAffx} E	ENSG00000166888,2				
	3 Gene@{Ensembl,Network	etAffx} E	ENSG00000173757,15550869	5026_at,2125	i49_at,212550_at		
	4 Gene@{Ensembl,Network	etAffx} E	ENSG00000126561,203010_at				
	5 Gene@{Ensembl}	E	ENSG0000016861				
	6 Gene@{Ensembl}	E	ENSG00000115415				
	7 Gene@{Ensembl}	E	EN5G00000138378				
De	tails – Attributes	5	All A	ute value			
	1 Ensembl.accession	ENSG	00000170581				
2Ensembl.statusKNO3Ensembl.sourceense4Ensembl.bioTypeprote5Ensembl.strand-16Ensembl.version27Ensembl.taxon96068Ensembl.chromosome12		KNOW					
		ensem	nbl				
		protei	cein_coding				
		-1			Attribute fueion		
		2			Allibule lusion		
		9606					
		12					
	9 Ensembl.chromosomeStop	55040	40176 21651				
	10 Ensembl.chromosomeStar	t 55021					
	11 Ensembl.description Signal transducer and art ator anscription 2 (p113). [Source:Uniprot/				<ol><li>Source:Uniprot/SWISSP</li></ol>		
	12 NetAffx.accession	20517	70_at				
13 NetAffx.name signal transducer and acceptor of transcription 2, 113kDa					kDa		

#### **System Architecture**



Mapping Layer

Mappings retrieving data of a single LDS but also interconnecting different LDS

#### **Query Processing**

Model Query Data Help						
Canned Queries Scripting Model-based Querying Keyword Search						
Domain Model	Query Specification					
-	Query targets: Name	Name				
Pro	Gene@NetAffx					
Exon	Union      Intersection      None	Э				
	Query conditions: Source Keyword	5				
	Protein@SwissProt CXCL CCL XCL CX3	C				
	Auxilable and by Ducksin@SuriseDuck > Cons@Encemble > Cons	NAL-1				
Estsequence Gene	Available pains: Protein@SwissProt > Gene@Ensembl > Gene@	pined				
Source Mapping Model	Emerute Consol Utilize loca	l dat				
	Execute					
SequenceRegion@Ensembl	Prot Query Result Query Targets: Gene@{Ensembl,NetAff:	<} ◄				
SequenceRegion@Ensembl	Prot Query Result Query Targets: Gene@{Ensembl,NetAff: Overview	<} ▼				
SequenceRegion@Ensembl	Prot Query Result Query Targets: Gene@{Ensembl,NetAff: Overview No. Logical source Item IGene@/Ensemble_ENSG00000170581_205170_at	<} -				
SequenceRegion@Ensembl	Prot Query Result Query Targets: Gene@{Ensembl,NetAff: Overview No. Logical source Item 1 Gene@{Ensembl EN5G00000170581,205170_at 2 Gene@{Ensembl EN5G00000166888,201331 s at	<} •				
SequenceRegion@Ensembl	Overview         Item         Item           1         Gene@{Ensembl         ENSG00000170581,205170_at         2           2         Gene@{Ensembl         ENSG00000166888,201331_s_at         3           3         Gene@{Ensembl         ENSG00000173757,1555086_at,205026_at,	<} •				
SequenceRegion@Ensembl	Overview         Item           1         Gene@{Ensembl         ENSG0000170581,205170_at           2         Gene@{Ensembl         ENSG0000170581,205170_at           3         Gene@{Ensembl         ENSG0000173757,1555086_at,205026_at,           4         Gene@{Ensembl         ENSG0000126561,203010_at	<} -				
SequenceRegion@Ensembl	Overview         Exercise           1         Gene@{Ensembl         ENSG0000170581,205170_at           2         Gene@{Ensembl         ENSG0000170581,205170_at           3         Gene@{Ensembl         ENSG0000170581,205170_at           4         Gene@{Ensembl         ENSG0000173757,1555086_at,205026_at,           4         Gene@{Ensembl         ENSG0000126561,203010_at           5         Gene@{Ensembl         ENSG000016861	<} •				
SequenceRegion@Ensembl	Exercise         Catter         Exercise           Prot         Query Result         Query Targets:         Gene@{Ensembl,NetAff;           Overview         Image: Catter         Image: Catter         Image: Catter           I Gene@{Ensembl         ENSG00000170581,205170_at         Image: Catter         Image: Catter           2 Gene@{Ensembl         ENSG00000170581,205170_at         Image: Catter         Image: Catter         Image: Catter           3 Gene@{Ensembl         ENSG00000170581,205170_at         Image: Catter         Image: Ca	<} •				
SequenceRegion@Ensembl	Exercise         Catter         Exercise           Prot         Query Result         Query Targets:         Gene@{Ensembl,NetAff:           Overview         Item         IGene@{Ensembl         ENSG00000170581,205170_at           I Gene@{Ensembl         ENSG00000170581,205170_at         Gene@{Ensembl           2 Gene@{Ensembl         ENSG00000166888,201331_s_at         Gene@{Ensembl           3 Gene@{Ensembl         ENSG00000126561,203010_at         SGene@{Ensembl           5 Gene@{Ensembl         ENSG00000156561,203010_at         SGene@{Ensembl           FNSG00000115415         ENSG00000115415         Details	<} •				
SequenceRegion@Ensembl	Prot       Query Result       Query Targets:       Gene@{Ensembl,NetAff:         Overview       Item         1       Gene@{Ensembl       ENSG00000170581,205170_at         2       Gene@{Ensembl       ENSG00000166888,201331_s_at         3       Gene@{Ensembl       ENSG00000173757,1555086_at,205026_at,         4       Gene@{Ensembl       ENSG00000126561,203010_at         5       Gene@{Ensembl       ENSG0000015861         Come@{Ensembl       ENSG0000115415         Details       No.       Attribute name         No.       Attribute name       Attribute value	<} -				
EstSequence@MyEstSet	Prot       Query Result       Query Targets:       Gene@{Ensembl,NetAff:         Overview       Item       1       Gene@{Ensembl       ENSG00000170581,205170_at         2       Gene@{Ensembl       ENSG00000170581,205170_at       3         3       Gene@{Ensembl       ENSG00000173757,1555086_at,205026_at,         4       Gene@{Ensembl       ENSG0000126561,203010_at         5       Gene@{Ensembl       ENSG0000115415         Ottails       ENSG00000170581         Xaffx       No.       Attribute name         1       Ensembl.accession       ENSG0000170581	<} •				
SequenceRegion@Ensembl	Exercise       Catter       Exercise         Prot       Query Result       Query Targets:       Gene@{Ensembl,NetAff;         Overview       Item       Item         1       Gene@{Ensembl       ENSG00000170581,205170_at         2       Gene@{Ensembl       ENSG00000166888,201331_s_at         3       Gene@{Ensembl       ENSG00000126561,203010_at         4       Gene@{Ensembl       ENSG00000115651         5       Gene@{Ensembl       ENSG00000115415         Details       Mo.       Attribute name       Attribute value         1       Ensembl.accession       ENSG0000170581         2       Ensembl.accession       ENSG00000170581	<} <b>▼</b>				

#### **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. ViennaAndrea Tanzer
- MPI Mathe. in Sciences, Leipzig
  - Jürgen Jost
  - Anirban Banerjee

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

#### **Example: Web-based Source NCBI Entrez**

Source dependent identifier (accession)

□ 1: AANAT <del>a: y</del> ialkylamine N-acetyltransferase	[Homo sapiens	2]			
GeneID: 15 Locus tag: <u>HGNC:19</u> ; <u>MIM: 600950</u>		-			
Official Symbol: AANAT and Name: arylalkylamine N-acetyl	ltransferase <b>provide</b>	d by <u>HUGO Gen</u>	e Nomenclature Committee		
Transcripts and products: <u>RefSeq below</u>					
Gene type: protein coding		Names, Symbols,			
Gene name: AANAT	Synonyme Commente				
Gene description: arylalkylamine N-acetyltransferase 💦 🏲					
RefSeq status: Reviewed		Sequences, etc.			
Organism: <u>Homo sapiens</u>					
Phenotypes					
Delayed sleep phase syndrome, susceptibility to $\underline{MIM}$ Pathways	: 600950 🗸	OMIM	Correspondences		
KEGG pathway: Tryptophan metabolism 00380 🛛 🛶		KEGG	to other data		
UniGene <u>Hs.431417</u>		UniGene	sources		
MIM <u>600950</u>		•••			
PharmGKB PA24366					

#### **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