

Towards an integrated knowledge system for capturing gene expression events

Aravind Venkatesan **Vladimir Mironov** and Martin Kuiper

Department of Biology, NTNU, 7491 Trondheim, Norway

ICBO 2012

Graz, Austria

July 24

Contents

- Motivation
- Objectives
- Design principles
- Implementation
- Future

Working on demand

Cell Cycle Ontology



BioGateway



Gene Expression KB

Requirements

- Catering for various groups of biomedical researchers
- Capability to answer complex real life questions
- Efficient data retrieval
- Easy extensibility
- Compatibility with other resources

Sample Questions

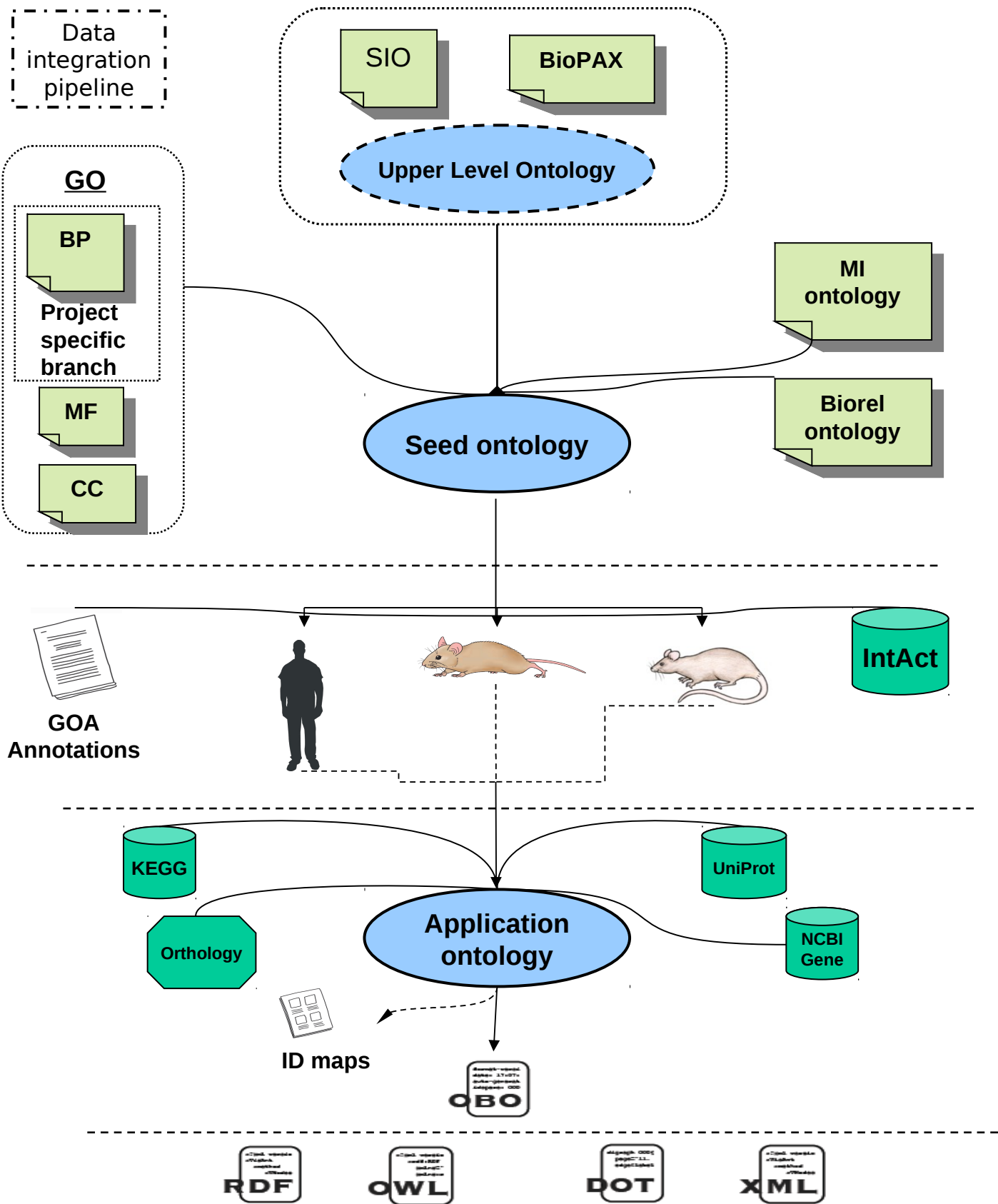
- Retrieve transcription factors.
- Retrieve co-transcription factors.
- Retrieve chromatin modulators.
- Retrieve proteins involved in two processes .
- Retrieve transcription factors that co-operate with a given transcription factor.
- List activators of a given transcription factor.
- List components of a protein complex.
- Retrieve proteins involved in a given pathway.
- Retrieve proteins participating in two pathways.

Design principles

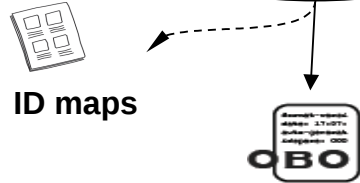
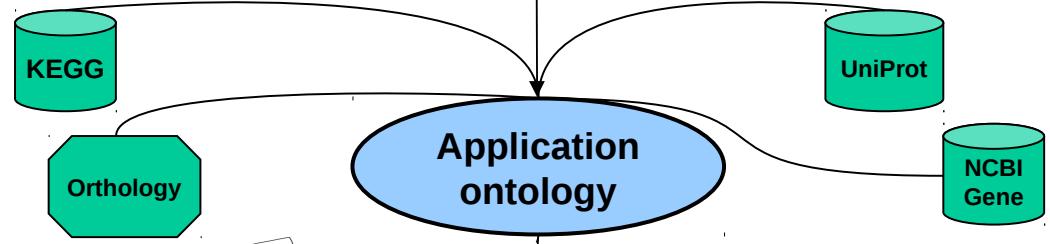
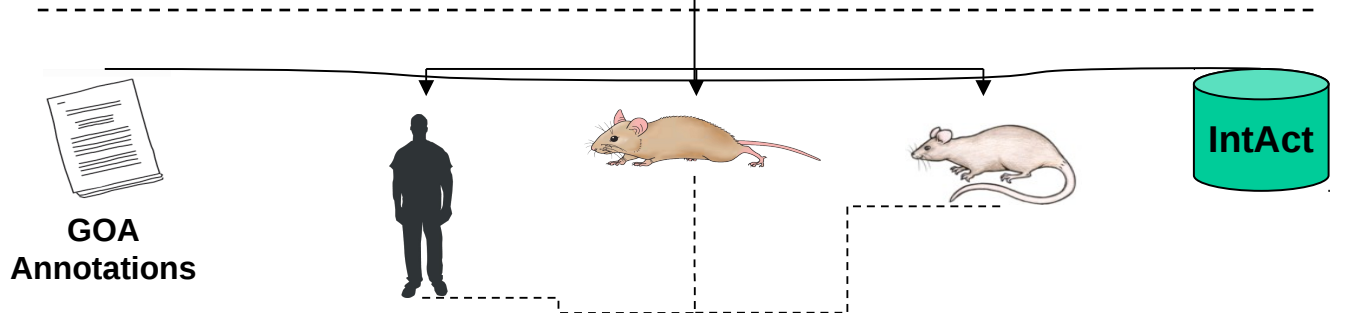
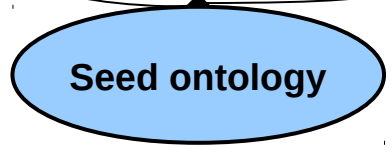
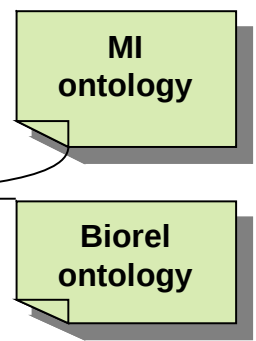
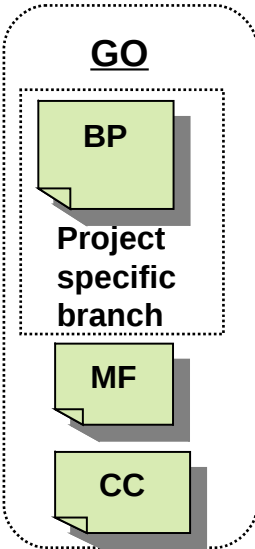
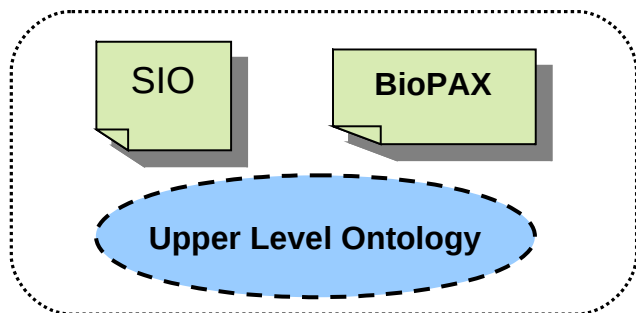
- 'is a' completeness
- 'all-some' semantics
- only classes used for modeling of the domain of discourse
- maximal flexibility both for users and for future extensions

Scope

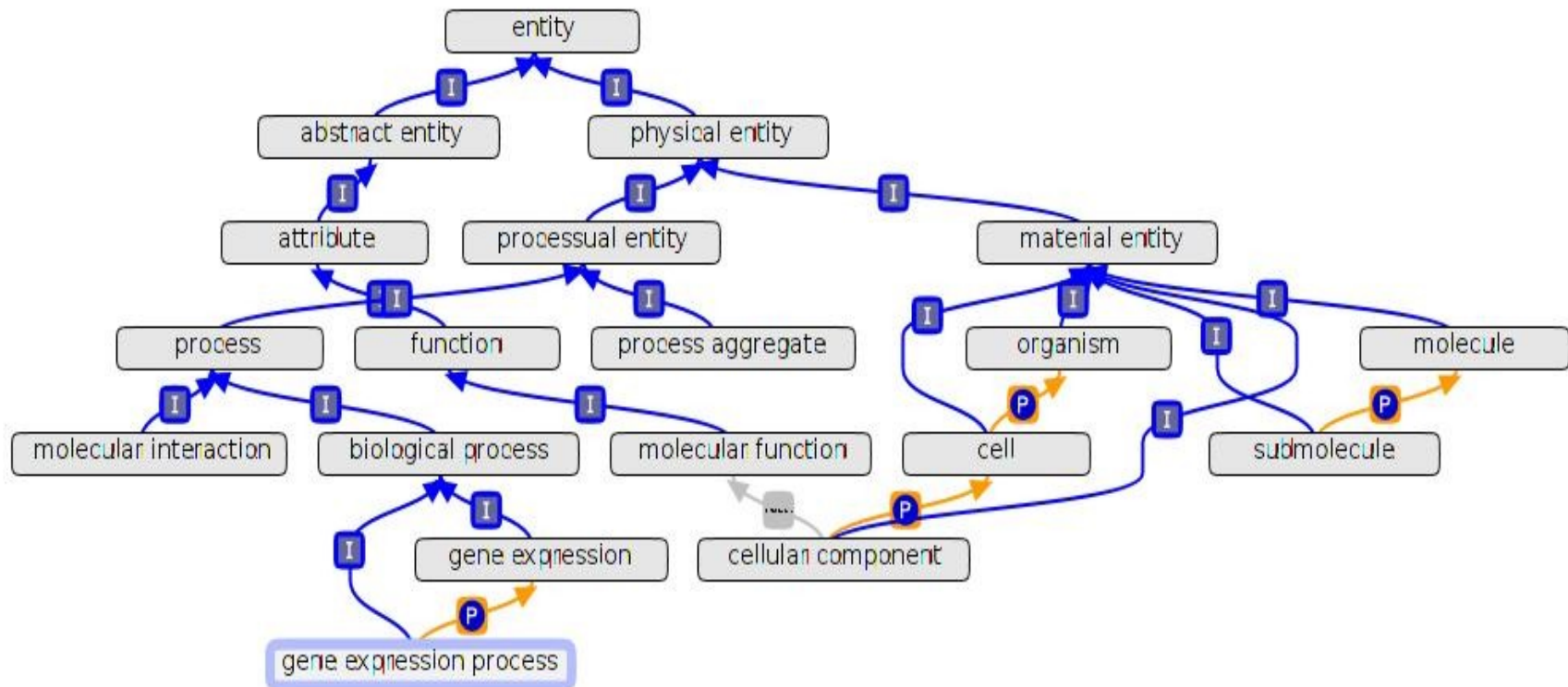
- 'gene expression' (GO:0010467)
 - 'regulation of gene expression' (GO:0010468)
 - 'regulation of transcription, DNA dependent' (GO:0006355)
-
- Human
 - Mouse
 - Rat



Data integration pipeline



Upper Level Ontology



Ontological Overview

	Classes	Relation types	Instances
GeXO	168417	15	0
ReXO	152962	15	0
ReTO	141095	15	0

Guilt by Association

	GOA	IntAct	KEGG	Orthology	Total
GeXO	18419	8528	5433	7300	39680
ReXO	14010	9046	4851	6009	33916
ReTO	12636	9056	4162	5163	31017

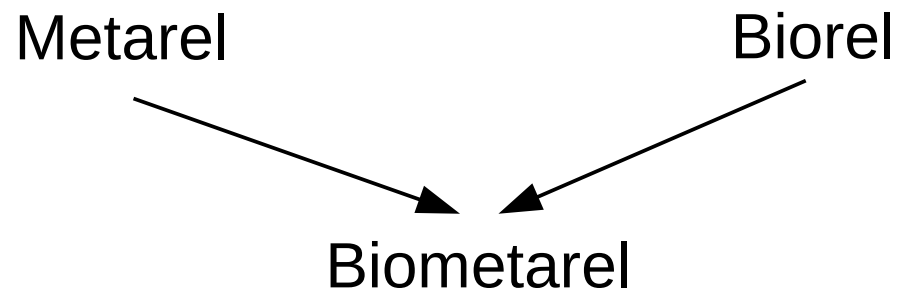
Relations

subject namespace	property	object namespace	GeXO	ReXO	ReTO
UPKB	bearer_of	SSB	60189	56487	53330
NCBIgn	codes_for	UPKB	25285	22588	20794
GO	contains	UPKB	92574	77531	76391
KEGG	has_agent	KEGG	3574	3028	2430
intact	has_agent	UPKB	79633	73177	66824
UPKB	has_function	GO	115548	95925	97407
GO	has_participant	UPKB	39271	24181	20919
UPKB + NCBIgn	has_source	NCBItx	115491	104551	97869
UPKB	member_of	KEGG	11993	10090	8643
GO	negatively_regulates	GO	45	12	6
GO	part_of	GO	1225	1103	1083
SSB	part_of	SSB	4	4	4
GO	positively_regulates	GO	44	10	6
GO	regulates	GO	146	43	20
UPKB	orthologous_to	UPKB	19756	15436	13608
SSB	transformation_of	UPKB	60189	56487	53330

Demo

<http://www.semantic-systems-biology.org/apo/queryingcco/sparql>

Inferencing



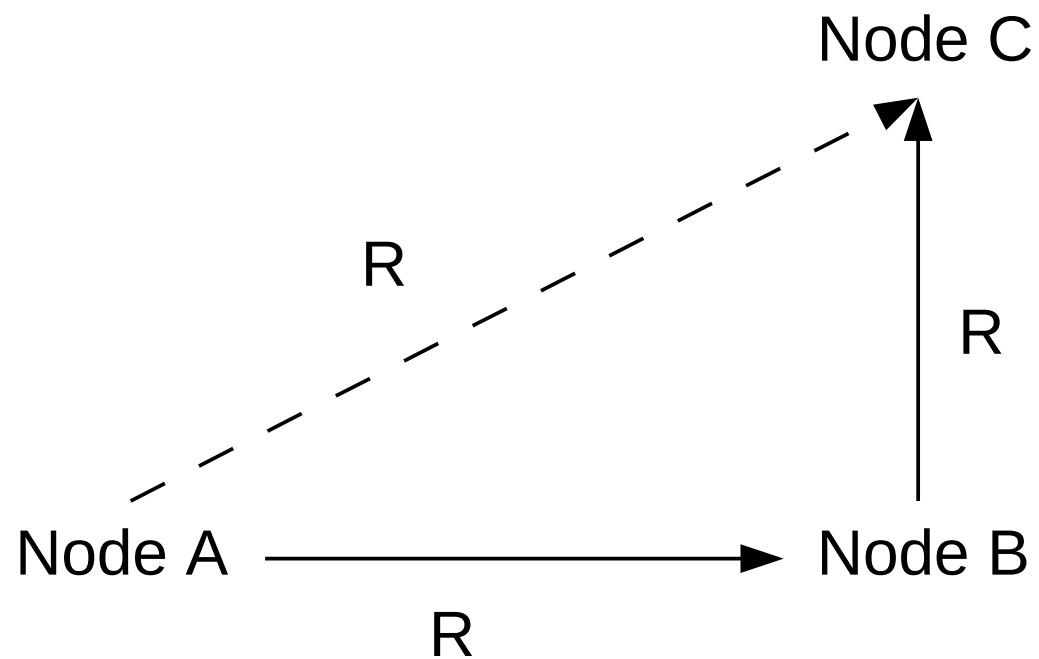
<http://www.semantic-systems-biology.org/metarel>

Ward Blonde

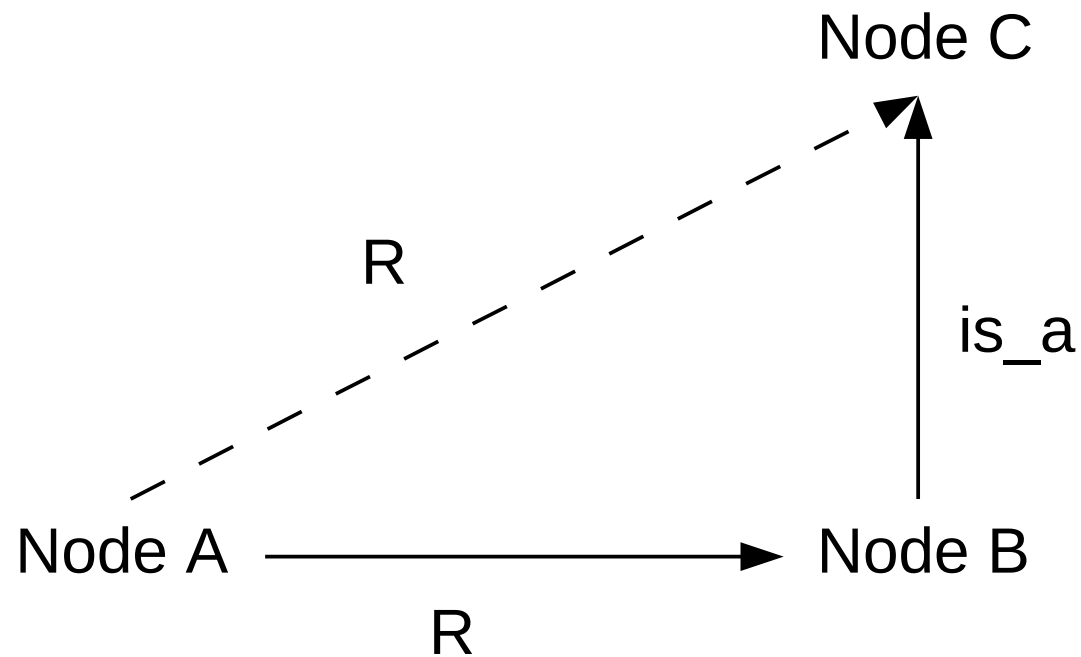
Closures

- Transitivity
- Priority over `is_a`
- Chains
- Reflexivity
- Super-relations

Transitivity

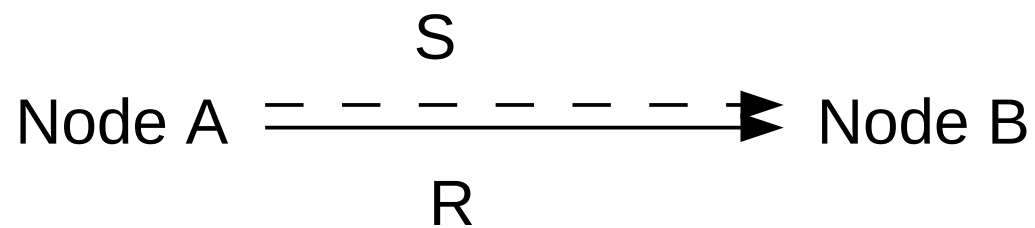


Priority over is_a



Super-relations

R rdfs:subPropertyOf S



Asserted vs Inferred Triples

in millions

	Asserted	Inferred	Total
GeXO	3.3	19.7	23.0
ReXO	3.0	16.9	19.9
ReTO	2.8	16.3	19.1

Target Genes

- 63 projects integrated into GeXKB
- 453 transcription factors
- 4419 target genes
- 37 types of experimental evidence



Future

- GO:0003700 'sequence-specific DNA binding transcription factor activity'
- More target genes
- Expression data for TF knockouts
- Migration to the Linked Data cloud
- Data minig

Thanks for your attention!