Utrecht University

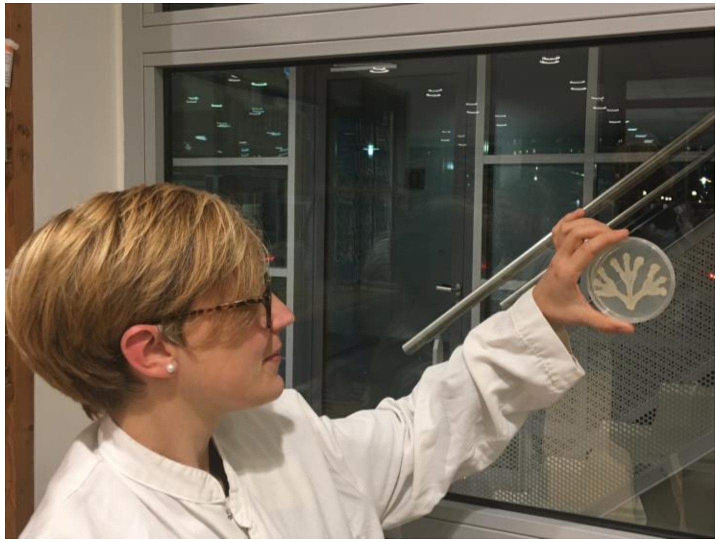
Faculty of Science
Department of Information and Computing Sciences
Software Technology Group

Formal Methods (for Better Software) in Computational Science

Anna-Lena Lamprecht

Workshop "A Research Agenda for Formal Methods in the Netherlands", Leiden

04 September 2018



Who am I?

2002-2007 B.Sc. and M.Sc. in Applied Computer Science (focus bioinformatics) in Göttingen

2007-2012 PhD in Computer Science at TU Dortmund (supervisor Bernhard Steffen)

2012-2015 PostDoc at the University of Potsdam

2015-2017 Research Fellow at Lero Limerick

Since 2017 Assistant Professor and Westerdijk Fellow at Utrecht University

Starting Point

Science across all domains is increasingly data-driven and computational, and thus the **correctness of research software** becomes increasingly critical for the validity of scientific results.

Yet formal methods and software quality assurance in general have not received great attention in this context in the past.

Software Engineering Practices in Science

Different than in industry

Typically no formal development methodology, no proper requirements specifications

Design not treated as a distinct development step

Testing more complicated, as correct results often unknown

Need for scientists to develop software themselves, self-taught programmers

...

Reference: D. Heaton and J. C. Carver. Claims about the use of software engineering practices in science: A systematic literature review. Information and Software Technology, 67:207 – 219, 2015.

"50% of biologists, chemists, ... end up as programmers anyway."
Jan Friso yesterday

Formal Methods for Workflows

Workflows popular in the scientific community

Inherently component-based

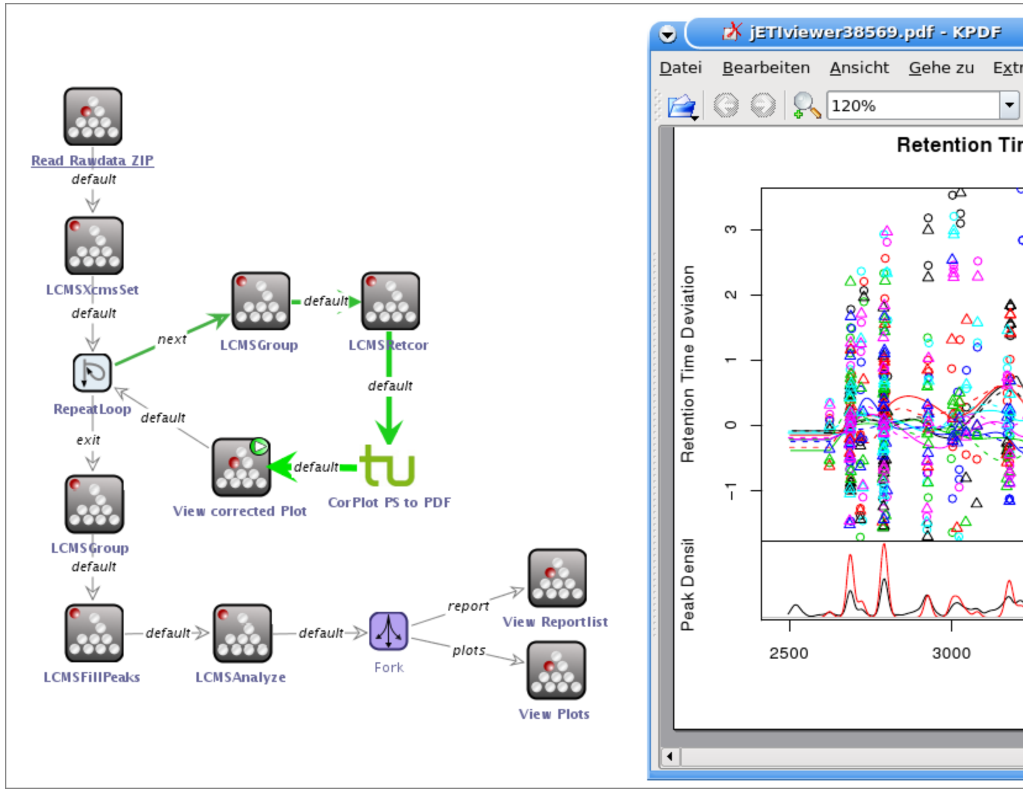
Less error-prone than coding from scratch

(Some) workflow models can be interpreted as transition systems, Kripke structures, …

Model-based formal methods (easily) applicable

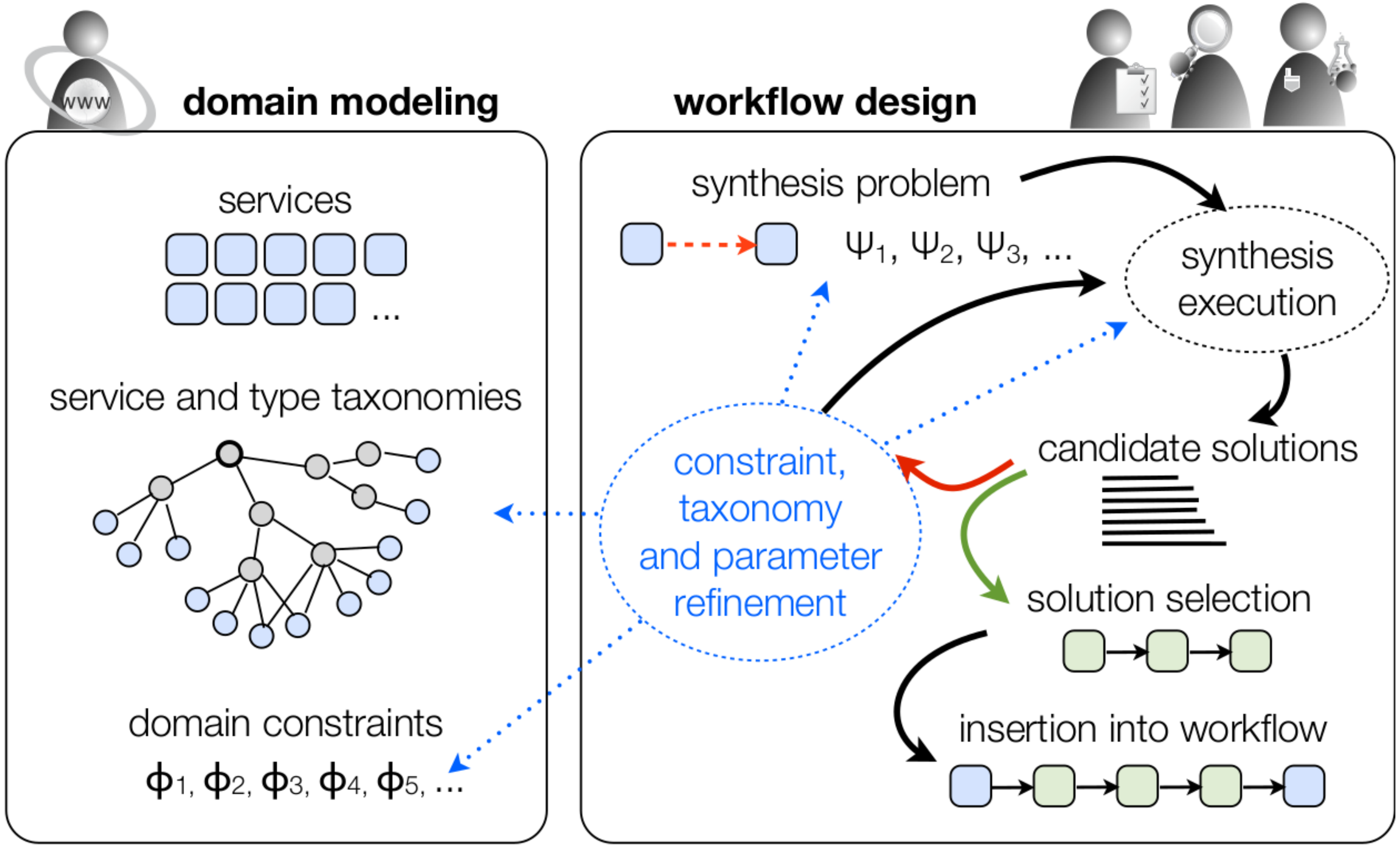
-> Model Checking

-> Model Synthesis



Supporting Process Development in Bio-jETI by Model Checking and Synthesis. 1st Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS 2008), CEUR Workshop Proceedings, Volume 435, 2008.

Synthesis of Scientific Workflows with PROPHETS



Some references:

- Constraint-Guided Workflow Composition Based on the EDAM Ontology. SWAT4LS 2010
- Loose Programming with PROPHETS. FASE 2012
- User-level workflow design. A bioinformatics perspective. Springer LNCS, Volume 8311, 2013
- Constraint-Driven Automatic Geospatial Service Composition: Workflows for the Analysis of Sea-Level Rise Impacts. ICCSA 2016
- Automated workflow composition in mass spectrometry- based proteomics. Bioinformatics, 2018.
- Automated Composition of Scientific Workflows: A Case Study on Geographic Data Manipulation. IEEE eScience 2018

A Recent Case Study

"Automated workflow composition in mass spectrometry-based proteomics"

A collaboration with:

Magnus Palmblad
(microbiologist, UMC Leiden)

Jon Ison
(ontologist, DKU Copenhagen)

Veit Schwämmle
(biostatistician, SDU Odense)

Making use of:

The ms-utils tool collection,

a selection of eliXir tools,

the EMBRACE Data and Methods Ontology (EDAM),

and four real use cases.

ms-utils.org

elixir Tools and Data Services Registry

EDAM
ontology
bioinformatics topics, operations,
data types and formats

Domain Model

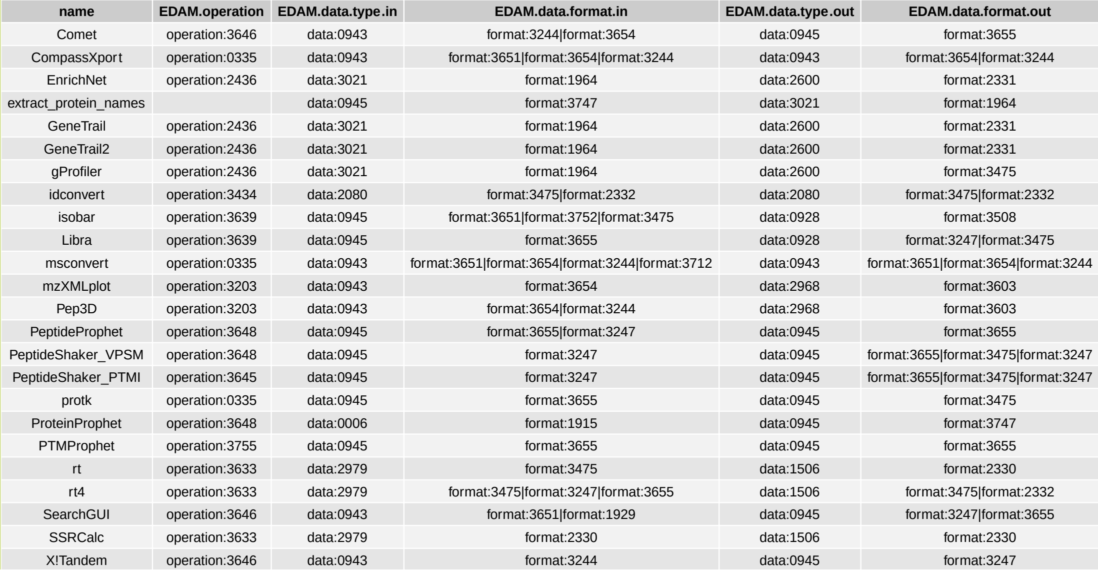
Semantic annotations of tools

from ms-utils.org and
the elixir tools registry

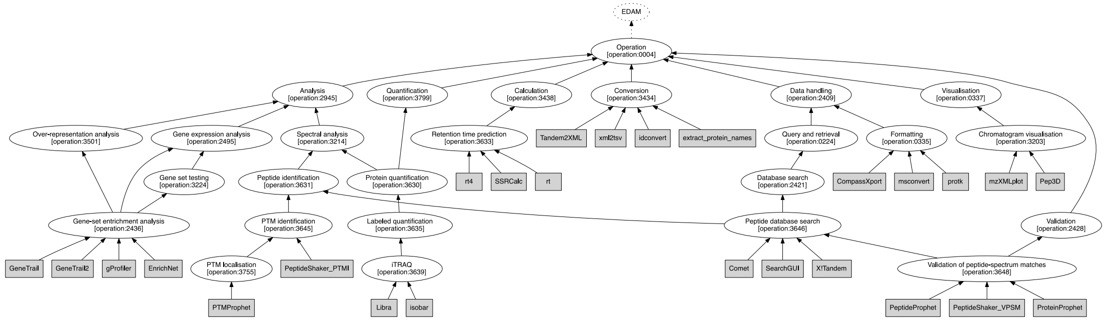
Tool and type taxonomies

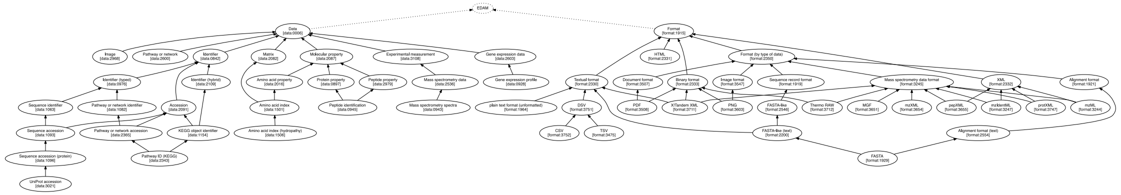
(derived from the EDAM ontology)

EDAM
ontology
bioinformatics topics, operations,
data types and formats



name	EDAM.operation	EDAM.data.type.in	EDAM.data.format.in	EDAM.data.type.out	EDAM.data.format.out
Comet	operation:3646	data:0943	format:3244|format:3654	data:0945	format:3655
CompassXport	operation:0335	data:0943	format:3651|format:3654|format:3244	data:0943	format:3654|format:3244
EnrichNet	operation:2436	data:3021	format:1964	data:2600	format:2331
extract_protein_names		data:0945	format:3747	data:3021	format:1964
GeneTrail	operation:2436	data:3021	format:1964	data:2600	format:2331
GeneTrail2	operation:2436	data:3021	format:1964	data:2600	format:2331
gProfiler	operation:2436	data:3021	format:1964	data:2600	format:3475
idconvert	operation:3434	data:2080	format:3475|format:2332	data:2080	format:3475|format:2332
isobar	operation:3639	data:0945	format:3651|format:3752|format:3475	data:0928	format:3508
Libra	operation:3639	data:0945	format:3655	data:0928	format:3247|format:3475
msconvert	operation:0335	data:0943	format:3651|format:3654|format:3244|format:3712	data:0943	format:3651|format:3654|format:3244
mzXMLplot	operation:3203	data:0943	format:3654	data:2968	format:3603
Pep3D	operation:3203	data:0943	format:3654|format:3244	data:2968	format:3603
PeptideProphet	operation:3648	data:0945	format:3655|format:3247	data:0945	format:3655
PeptideShaker_VPSM	operation:3648	data:0945	format:3247	data:0945	format:3655|format:3475|format:3247
PeptideShaker_PTMI	operation:3645	data:0945	format:3247	data:0945	format:3655|format:3475|format:3247
protk	operation:0335	data:0945	format:3655	data:0945	format:3475
ProteinProphet	operation:3648	data:0006	format:1915	data:0945	format:3747
PTMProphet	operation:3755	data:0945	format:3655	data:0945	format:3655
rt	operation:3633	data:2979	format:3475	data:1506	format:2330
rt4	operation:3633	data:2979	format:3475|format:3247|format:3655	data:1506	format:3475|format:2332
SearchGUI	operation:3646	data:0943	format:3651|format:1929	data:0945	format:3247|format:3655
SSRCalc	operation:3633	data:2979	format:2330	data:1506	format:2330
X!Tandem	operation:3646	data:0943	format:3244	data:0945	format:3247


EDAM
Operation [operation:0004]
Analysis [operation:2945]
Quantification [operation:3799]
Calculation [operation:3438]
Conversion [operation:3434]
Data handling [operation:2409]
Visualisation [operation:0337]
Over-representation analysis [operation:3501]
Gene expression analysis [operation:2495]
Spectral analysis [operation:3214]
Retention time prediction [operation:3633]
Tandem2XML
xml2tsv
idconvert
extract_protein_names
Query and retrieval [operation:0224]
Formatting [operation:0335]
Chromatogram visualisation [operation:3203]
Gene set testing [operation:3224]
Peptide identification [operation:3631]
Protein quantification [operation:3630]
rt4
SSRCalc
rt
Database search [operation:2421]
CompassXport
msconvert
protk
mzXMLplot
Pep3D
Gene-set enrichment analysis [operation:2436]
PTM identification [operation:3645]
Labeled quantification [operation:3635]
Peptide database search [operation:3646]
Validation [operation:2428]
GeneTrail
GeneTrail2
gProfiler
EnrichNet
PTM localisation [operation:3755]
PeptideShaker_PTMI
iTRAQ [operation:3639]
Comet
SearchGUI
X!Tandem
Validation of peptide-spectrum matches [operation:3648]
PTMProphet
Libra
isobar
PeptideProphet
PeptideShaker_VPSM
ProteinProphet


EDAM
Data [data:0006]
Format [format:1915]
Image [data:2968]
Pathway or network [data:2600]
Identifier [data:0842]
Matrix [data:2082]
Molecular property [data:2087]
Experimental measurement [data:3108]
Gene expression data [data:2603]
HTML [format:2331]
Format (by type of data) [format:2350]
Identifier (typed) [data:0976]
Identifier (hybrid) [data:2109]
Amino acid property [data:2016]
Protein property [data:0897]
Peptide property [data:2979]
Mass spectrometry data [data:2536]
Gene expression profile [data:0928]
Textual format [format:2330]
Document format [format:3507]
Binary format [format:2333]
Image format [format:3547]
Sequence record format [format:1919]
Mass spectrometry data format [format:3245]
XML [format:2332]
Alignment format [format:1921]
Sequence identifier [data:1063]
Pathway or network identifier [data:1082]
Accession [data:2091]
Amino acid index [data:1501]
Peptide identification [data:0945]
Mass spectrometry spectra [data:0943]
plain text format (unformatted) [format:1964]
DSV [format:3751]
PDF [format:3508]
X!Tandem XML [format:3711]
PNG [format:3603]
FASTA-like [format:2546]
Thermo RAW [format:3712]
MGF [format:3651]
mzXML [format:3654]
pepXML [format:3655]
mzIdentML [format:3247]
protXML [format:3747]
mzML [format:3244]
Sequence accession [data:1093]
Pathway or network accession [data:2365]
KEGG object identifier [data:1154]
Amino acid index (hydropathy) [data:1506]
CSV [format:3752]
TSV [format:3475]
FASTA-like (text) [format:2200]
Alignment format (text) [format:2554]
Sequence accession (protein) [data:1096]
Pathway ID (KEGG) [data:2343]
FASTA [format:1929]
UniProt accession [data:3021]

Domain Model

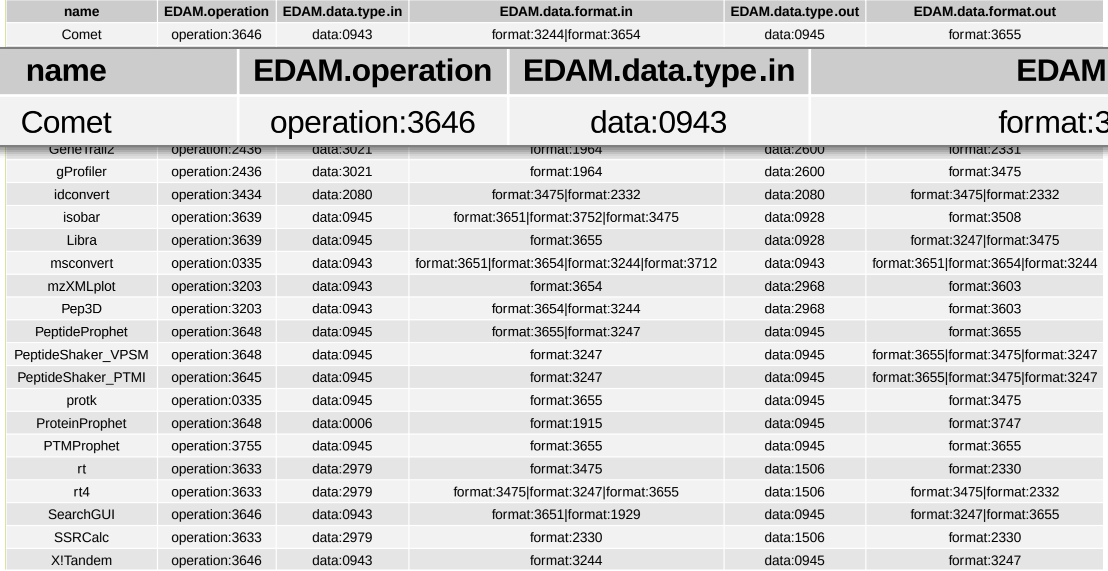
Semantic annotations of tools
from ms-utils.org and
the elixir tools registry

ms-utils.org

elixir Tools and Data Services Registry

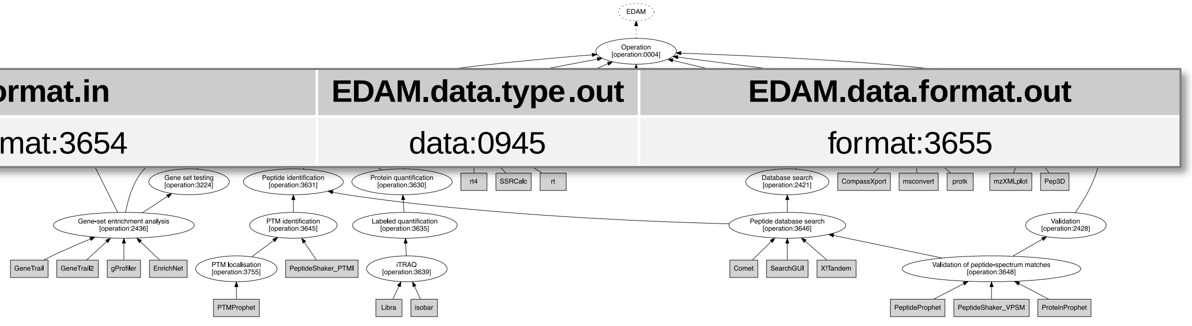
Tool and type taxonomies
(derived from the EDAM ontology)

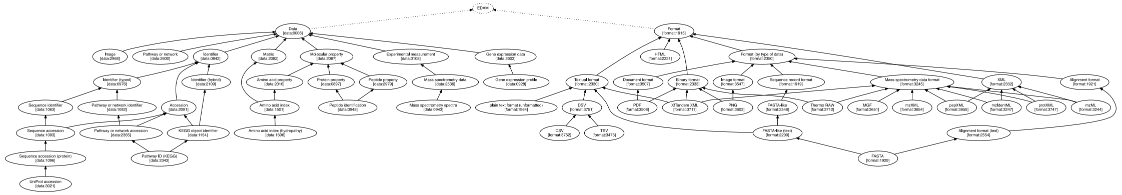
EDAM
ontology
bioinformatics topics, operations, data types and formats



name	EDAM.operation	EDAM.data.type.in	EDAM.data.format.in	EDAM.data.type.out	EDAM.data.format.out
Comet	operation:3646	data:0943	format:3244|format:3654	data:0945	format:3655
GeneTrail2	operation:2436	data:3021	format:1964	data:2600	format:2331
gProfiler	operation:2436	data:3021	format:1964	data:2600	format:3475
idconvert	operation:3434	data:2080	format:3475|format:2332	data:2080	format:3475|format:2332
isobar	operation:3639	data:0945	format:3651|format:3752|format:3475	data:0928	format:3508
Libra	operation:3639	data:0945	format:3655	data:0928	format:3247|format:3475
msconvert	operation:0335	data:0943	format:3651|format:3654|format:3244|format:3712	data:0943	format:3651|format:3654|format:3244
mzXMLplot	operation:3203	data:0943	format:3654	data:2968	format:3603
Pep3D	operation:3203	data:0943	format:3654|format:3244	data:2968	format:3603
PeptideProphet	operation:3648	data:0945	format:3655|format:3247	data:0945	format:3655
PeptideShaker_VPSM	operation:3648	data:0945	format:3247	data:0945	format:3655|format:3475|format:3247
PeptideShaker_PTMI	operation:3645	data:0945	format:3247	data:0945	format:3655|format:3475|format:3247
protk	operation:0335	data:0945	format:3655	data:0945	format:3475
ProteinProphet	operation:3648	data:0006	format:1915	data:0945	format:3747
PTMProphet	operation:3755	data:0945	format:3655	data:0945	format:3655
rt	operation:3633	data:2979	format:3475	data:1506	format:2330
rt4	operation:3633	data:2979	format:3475|format:3247|format:3655	data:1506	format:3475|format:2332
SearchGUI	operation:3646	data:0943	format:3651|format:1929	data:0945	format:3247|format:3655
SSRCalc	operation:3633	data:2979	format:2330	data:1506	format:2330
X!Tandem	operation:3646	data:0943	format:3244	data:0945	format:3247

name	EDAM.operation	EDAM.data.type.in	EDAM.data.format.in	EDAM.data.type.out	EDAM.data.format.out
Comet	operation:3646	data:0943	format:3244|format:3654	data:0945	format:3655


EDAM
Operation [operation:0004]
Gene set testing [operation:3224]
Peptide identification [operation:3631]
Protein quantification [operation:3630]
rt4
SSRCalc
rt
Database search [operation:2421]
CompassXport
msconvert
protk
mzXMLplot
Pep3D
Gene-set enrichment analysis [operation:2436]
PTM identification [operation:3645]
Labeled quantification [operation:3635]
Peptide database search [operation:3646]
Validation [operation:2428]
GeneTrail
GeneTrail2
gProfiler
EnrichNet
PTM localisation [operation:3755]
PeptideShaker_PTMI
iTRAQ [operation:3639]
Comet
SearchGUI
X!Tandem
Validation of peptide-spectrum matches [operation:3648]
PTMProphet
Libra
isobar
PeptideProphet
PeptideShaker_VPSM
ProteinProphet


EDAM
Data [data:0006]
Image [data:2968]
Pathway or network [data:2600]
Identifier [data:0842]
Matrix [data:2082]
Molecular property [data:2087]
Experimental measurement [data:3108]
Gene expression data [data:2603]
Identifier (typed) [data:0976]
Identifier (hybrid) [data:2109]
Amino acid property [data:2016]
Protein property [data:0897]
Peptide property [data:2979]
Mass spectrometry data [data:2536]
Gene expression profile [data:0928]
Sequence identifier [data:1063]
Pathway or network identifier [data:1082]
Accession [data:2091]
Amino acid index [data:1501]
Peptide identification [data:0945]
Mass spectrometry spectra [data:0943]
Sequence accession [data:1093]
Pathway or network accession [data:2365]
KEGG object identifier [data:1154]
Amino acid index (hydropathy) [data:1506]
Sequence accession (protein) [data:1096]
Pathway ID (KEGG) [data:2343]
UniProt accession [data:3021]
Format [format:1915]
HTML [format:2331]
Format (by type of data) [format:2350]
Textual format [format:2330]
Document format [format:3507]
Binary format [format:2333]
Image format [format:3547]
Sequence record format [format:1919]
Mass spectrometry data format [format:3245]
XML [format:2332]
Alignment format [format:1921]
plain text format (unformatted) [format:1964]
DSV [format:3751]
PDF [format:3508]
X!Tandem XML [format:3711]
PNG [format:3603]
FASTA-like [format:2546]
Thermo RAW [format:3712]
MGF [format:3651]
mzXML [format:3654]
pepXML [format:3655]
mzIdentML [format:3247]
protXML [format:3747]
mzML [format:3244]
CSV [format:3752]
TSV [format:3475]
FASTA-like (text) [format:2200]
Alignment format (text) [format:2554]
FASTA [format:1929]

Domain Model

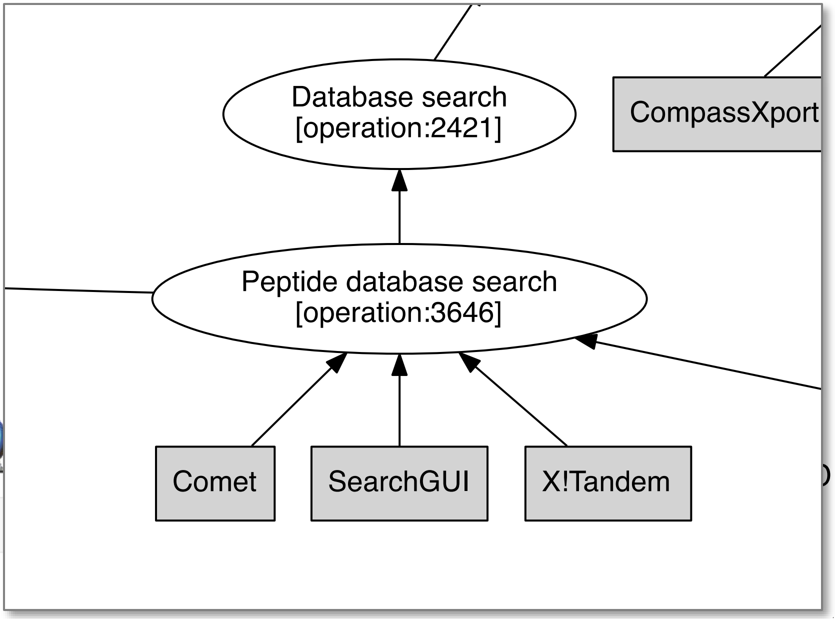
Semantic annotations of tools
from ms-utils.org and
the elixir tools registry

ms-utils.org

elixir Tools and Data Services

ogy)

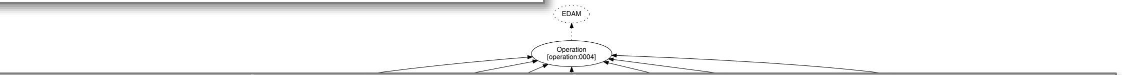
EDAM
ontology
bioinformatics topics, operations, data types and formats

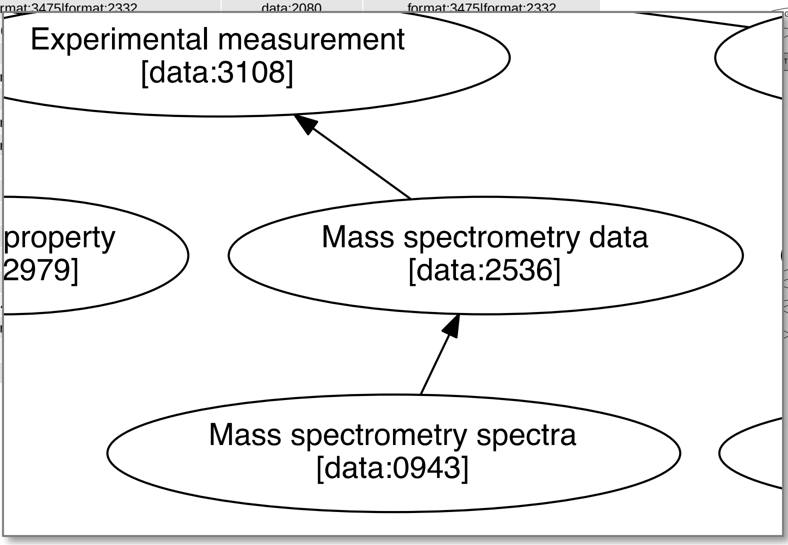


Database search [operation:2421]
CompassXport
Peptide database search [operation:3646]
Comet
SearchGUI
X!Tandem

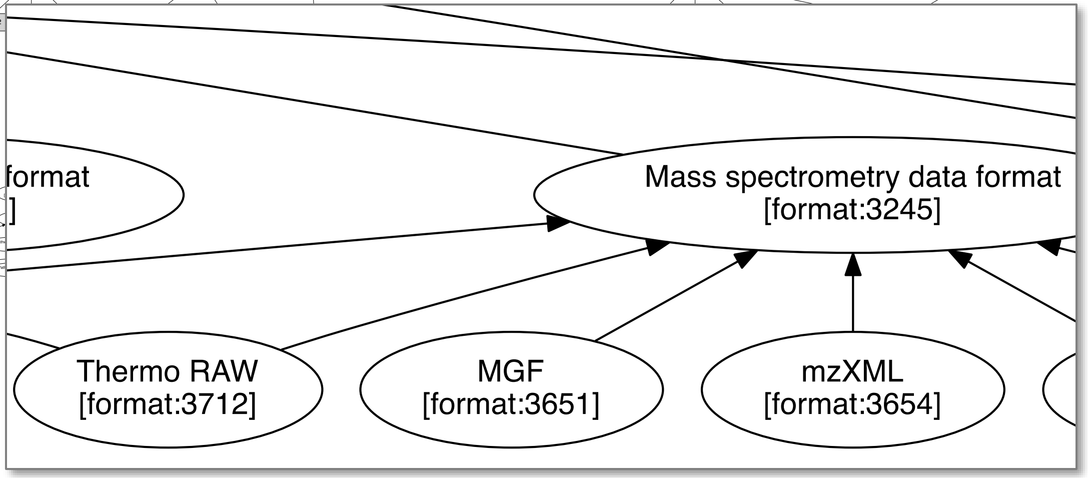
name	EDAM.operation	EDAM.data.type.in	EDAM.data.format.in	EDAM.data.type.out	EDAM.data.format.out
Comet	operation:3646	data:0943	format:3244|format:3654	data:0945	format:3655

name	EDAM.operation	EDAM.data.type.in	EDAM.data.format.in	EDAM.data.type.out	EDAM.data.format.out
Comet	operation:3646	data:0943	format:3244|format:3654	data:0945	format:3655
gProfiler	operation:2436	data:3021	format:1964	data:2600	format:3475
idconvert	operation:3434	data:2080	format:3475|format:2332	data:2080	format:3475|format:2332
isobar	operation:3639	data:0945	format:3		
Libra	operation:3639	data:0945			
msconvert	operation:0335	data:0943	format:3651|for		
mzXMLplot	operation:3203	data:0943			
Pep3D	operation:3203	data:0943	for		
PeptideProphet	operation:3648	data:0945	for		
PeptideShaker_VPSM	operation:3648	data:0945			
PeptideShaker_PTMI	operation:3645	data:0945			
protk	operation:0335	data:0945			
ProteinProphet	operation:3648	data:0006			
PTMProphet	operation:3755	data:0945			
rt	operation:3633	data:2979			
rt4	operation:3633	data:2979	format:3		
SearchGUI	operation:3646	data:0943	for		
SSRCalc	operation:3633	data:2979			
X!Tandem	operation:3646	data:0943			





Experimental measurement [data:3108]
property 2979]
Mass spectrometry data [data:2536]
Mass spectrometry spectra [data:0943]



Workflow Specification

Use Case:

#2

Original Description:

“Next, our hypothetical scientist faces the common task of protein identification and interpretation of lists of identified proteins by enrichment analysis. Our researcher starts again from LC-MS/MS data in the Thermo RAW format. After peptide database search and protein identification, the list of proteins identified via UniProt accessions should be analyzed by gene-set enrichment analysis with respect to KEGG pathways and annotations, reporting KEGG annotations and pathways IDs and associated p- or q-values.”

Workflow Specification

Workflow Input:

Mass spectrometry spectra [data:0943] in Thermo RAW [format:3712]

Workflow Output:

Pathway or Network [data:2600] in any format [format:1915]

Workflow Constraints:

Use gene-set enrichment analysis [operation:2436].

Use Gene-set enrichment analysis [operation:2436] only after peptide identification [operation:3631].

Use ProteinProphet only after PeptideProphet.

Workflow Specification

Workflow Input

Mass spectro
Thermo RAW [format:3712]

Workflow Output:

Pathway or Netw
any format [form]

Workflow Constraints:

Use gene-set enrichm

Use Gene-set enrich
only after peptide ide

Use ProteinProphet only after PeptideProphet.

Start Types:

[“Mass spectrometry spectra”, “Thermo RAW”]

Goal Constraint:

G (X true | (“Pathway or network” & “Format”))

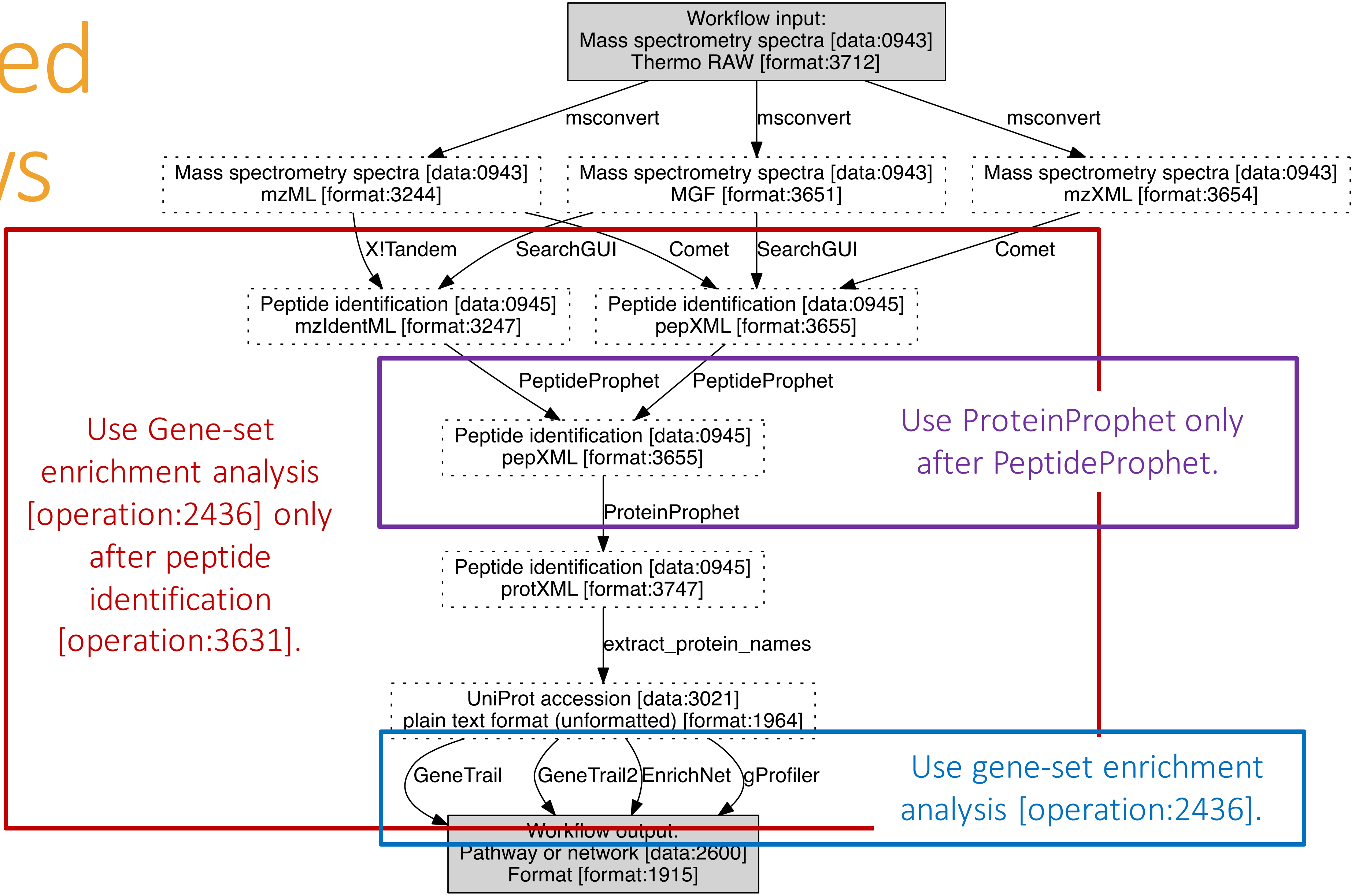
Constraints:

F <“Gene-set enrichment analysis”> true

(~ <“Gene-set enrichment analysis”> true WU <“Peptide identification”> true)

(~ <“ProteinProphet”> true WU <“PeptideProphet”> true)

Synthesized Workflows



Workflow input:
Mass spectrometry spectra [data:0943]
Thermo RAW [format:3712]

msconvert

msconvert

msconvert

Mass spectrometry spectra [data:0943]
mzML [format:3244]

Mass spectrometry spectra [data:0943]
MGF [format:3651]

Mass spectrometry spectra [data:0943]
mzXML [format:3654]

X!Tandem

SearchGUI

Comet

SearchGUI

Comet

Peptide identification [data:0945]
mzIdentML [format:3247]

Peptide identification [data:0945]
pepXML [format:3655]

PeptideProphet

PeptideProphet

Peptide identification [data:0945]
pepXML [format:3655]

ProteinProphet

Peptide identification [data:0945]
protXML [format:3747]

extract_protein_names

UniProt accession [data:3021]
plain text format (unformatted) [format:1964]

GeneTrail

GeneTrail2

EnrichNet

gProfiler

Workflow output:
Pathway or network [data:2600]
Format [format:1915]

Use Gene-set enrichment analysis [operation:2436] only after peptide identification [operation:3631].

Use ProteinProphet only after PeptideProphet.

Use gene-set enrichment analysis [operation:2436].

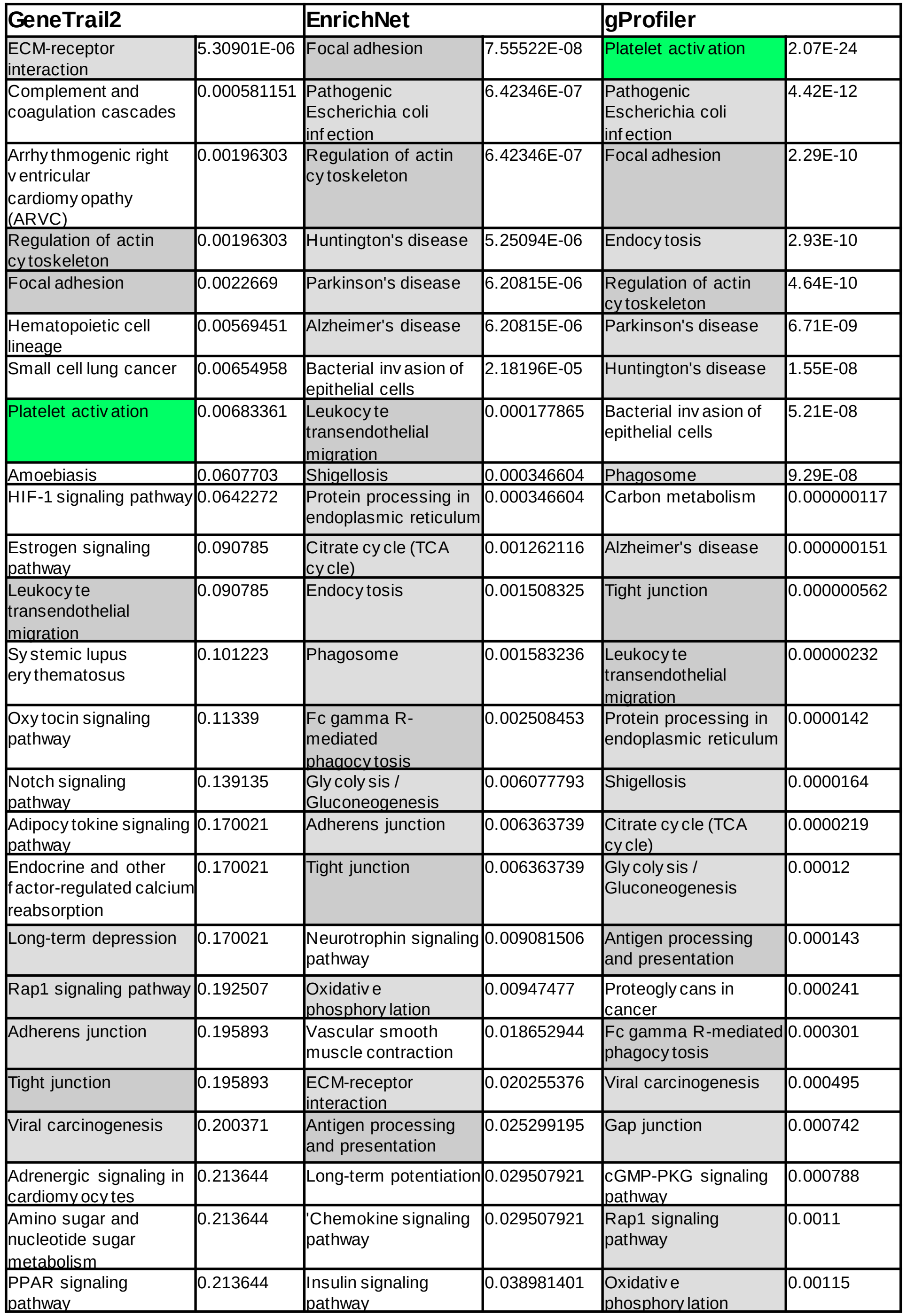
User Perspective: Workflow Evaluation

Select workflows for implementation:

1. *msconvert -> Comet -> PeptideProphet -> ProteinProphet -> extract_protein_names -> GeneTrail2*
2. *msconvert -> Comet -> PeptideProphet -> ProteinProphet -> extract_protein_names -> EnrichNet*
3. *msconvert -> Comet -> PeptideProphet -> ProteinProphet -> extract_protein_names -> gProfiler*

Execute on real data.

Compare results.



GeneTrail2		**EnrichNet**		**gProfiler**	
ECM-receptor interaction	5.30901E-06	Focal adhesion	7.55522E-08	Platelet activation	2.07E-24
Complement and coagulation cascades	0.000581151	Pathogenic Escherichia coli infection	6.42346E-07	Pathogenic Escherichia coli infection	4.42E-12
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.00196303	Regulation of actin cytoskeleton	6.42346E-07	Focal adhesion	2.29E-10
Regulation of actin cytoskeleton	0.00196303	Huntington's disease	5.25094E-06	Endocytosis	2.93E-10
Focal adhesion	0.0022669	Parkinson's disease	6.20815E-06	Regulation of actin cytoskeleton	4.64E-10
Hematopoietic cell lineage	0.00569451	Alzheimer's disease	6.20815E-06	Parkinson's disease	6.71E-09
Small cell lung cancer	0.00654958	Bacterial invasion of epithelial cells	2.18196E-05	Huntington's disease	1.55E-08
Platelet activation	0.00683361	Leukocyte transendothelial migration	0.000177865	Bacterial invasion of epithelial cells	5.21E-08
Amoebiasis	0.0607703	Shigellosis	0.000346604	Phagosome	9.29E-08
HIF-1 signaling pathway	0.0642272	Protein processing in endoplasmic reticulum	0.000346604	Carbon metabolism	0.000000117
Estrogen signaling pathway	0.090785	Citrate cycle (TCA cycle)	0.001262116	Alzheimer's disease	0.000000151
Leukocyte transendothelial migration	0.090785	Endocytosis	0.001508325	Tight junction	0.000000562
Systemic lupus erythematosus	0.101223	Phagosome	0.001583236	Leukocyte transendothelial migration	0.00000232
Oxytocin signaling pathway	0.11339	Fc gamma R-mediated phagocytosis	0.002508453	Protein processing in endoplasmic reticulum	0.0000142
Notch signaling pathway	0.139135	Glycolysis / Gluconeogenesis	0.006077793	Shigellosis	0.0000164
Adipocytokine signaling pathway	0.170021	Adherens junction	0.006363739	Citrate cycle (TCA cycle)	0.0000219
Endocrine and other factor-regulated calcium reabsorption	0.170021	Tight junction	0.006363739	Glycolysis / Gluconeogenesis	0.00012
Long-term depression	0.170021	Neurotrophin signaling pathway	0.009081506	Antigen processing and presentation	0.000143
Rap1 signaling pathway	0.192507	Oxidative phosphorylation	0.00947477	Proteoglycans in cancer	0.000241
Adherens junction	0.195893	Vascular smooth muscle contraction	0.018652944	Fc gamma R-mediated phagocytosis	0.000301
Tight junction	0.195893	ECM-receptor interaction	0.020255376	Viral carcinogenesis	0.000495
Viral carcinogenesis	0.200371	Antigen processing and presentation	0.025299195	Gap junction	0.000742
Adrenergic signaling in cardiomyocytes	0.213644	Long-term potentiation	0.029507921	cGMP-PKG signaling pathway	0.000788
Amino sugar and nucleotide sugar metabolism	0.213644	'Chemokine signaling pathway	0.029507921	Rap1 signaling pathway	0.0011
PPAR signaling pathway	0.213644	Insulin signaling pathway	0.038981401	Oxidative phosphorylation	0.00115

CS Perspective: Synthesis Evaluation

Constraints	Length <= 5	Length <= 6	Length <= 7	Length <= 8
No constraints	12 solutions + 40,484 states 0.223 seconds	156 solutions + 294,972 states 1.114 seconds	1,368 solutions + 2,104,842 states 3.669 seconds	10,452 solutions + 14,878,084 states 14.269 seconds
General domain constraints	12 solutions + 11,463 states 0.279 seconds	108 solutions + 61,185 states 0.903 seconds	692 solutions + 310,766 states 3.558 seconds	3,876 solutions + 1,532,352 states 16.066 seconds
Problem-specific constraints	0 solutions + 652,865 states 6.105 seconds	**20 solutions** **+ 4,643,845 states** **40.376 seconds**	204 solutions + 30,629,133 states 4 min 47 seconds	1,456 solutions + 191,780,655 states 27 min 54 seconds

Executed on a 2015 MacBook Pro, 3.1 GHz Intel Core i7 Processor, 6 GB 1867 MHz DDR3 Memory, OS X El Capitan Version 10.11.6

Challenges / Future Work

- Usability
  - "good" results
  - natural language specification
  - integration into existing eScience infrastructure

- Scalability
  - synthesis performance
  - large-scale comprehensive domain modeling

Formal Methods for Python and R

Dynamically typed languages like Python and GNU R extremely popular

Feel easy-to-use, but...

Runtime errors due to mismatching data types (often difficult to fix)

Type mismatches that remain undetected and cause wrong results (even worse!)

...

(How) can FM help?

Static typing? Domain-specific type systems and error diagnoses? Dependent types?

Partner Communities in NL

General eScience (e.g. eScience Center, NL-RSE)

Life Sciences (e.g. ELIXIR, DTL)

Semantic Web (e.g. Knowledge representation @ VU)