technische universität
dortmund

Department of Computer Science
Chair for Programming Systems

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Anna-Lena Lamprecht, Stefan Naujokat, Bernhard Steffen, Tiziana Margaria

Constraint-Guided Workflow Composition Based on the EDAM Ontology

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Outline

- Constraint-Guided Workflow Composition Based on the EDAM Ontology.
- Constraint-Guided Workflow Composition Based on the EDAM Ontology.
- Constraint-Guided Workflow Composition Based on the EDAM Ontology.

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

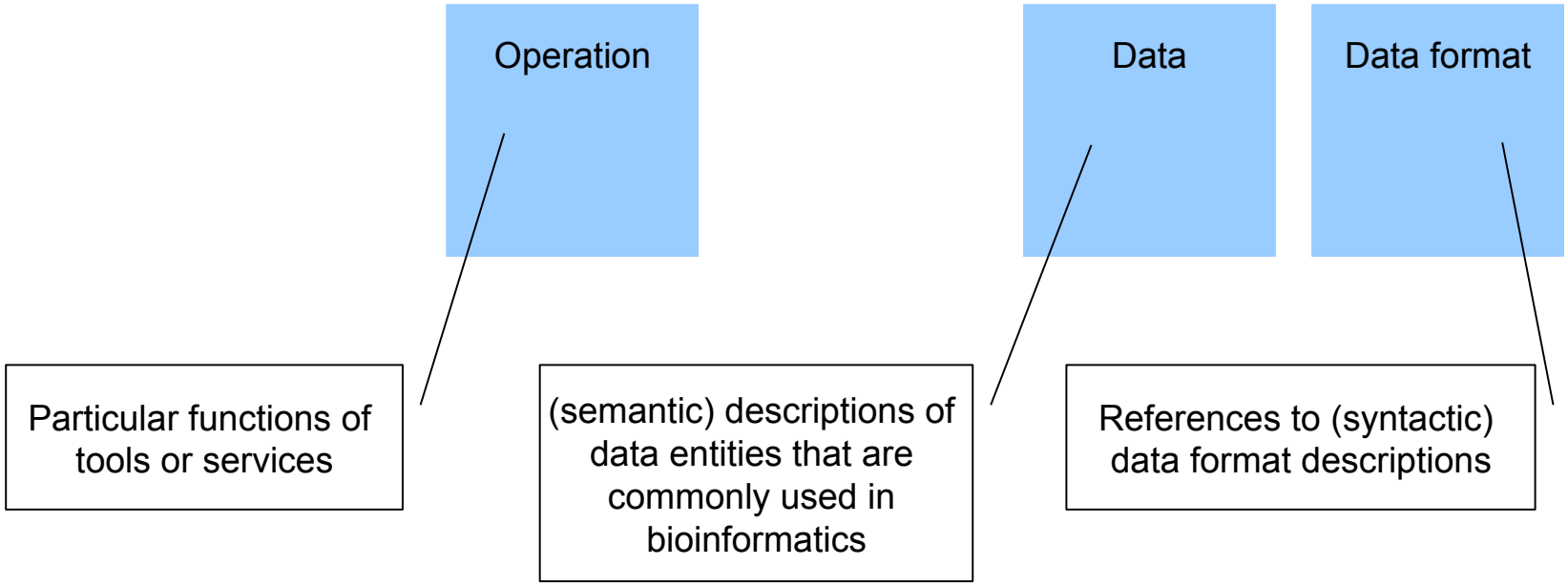
Constraint-Guided Workflow Composition Based on the EDAM Ontology

- EDAM = EMBRACE Data and Methods Ontology
- Vocabulary of terms and relations that can be used for annotating services

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

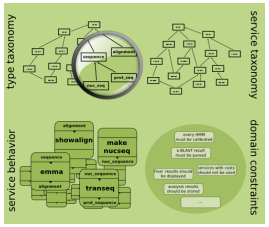
- EDAM = EMBRACE Data and Methods Ontology
- Vocabulary of terms and relations that can be used for annotating services



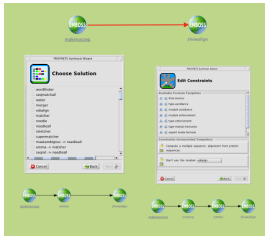
Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

- PROPHETS plugin for jABC/Bio-jETI
  (Process Realization and Optimization Platform using a Human-readable Expression of Temporal-logic Synthesis)



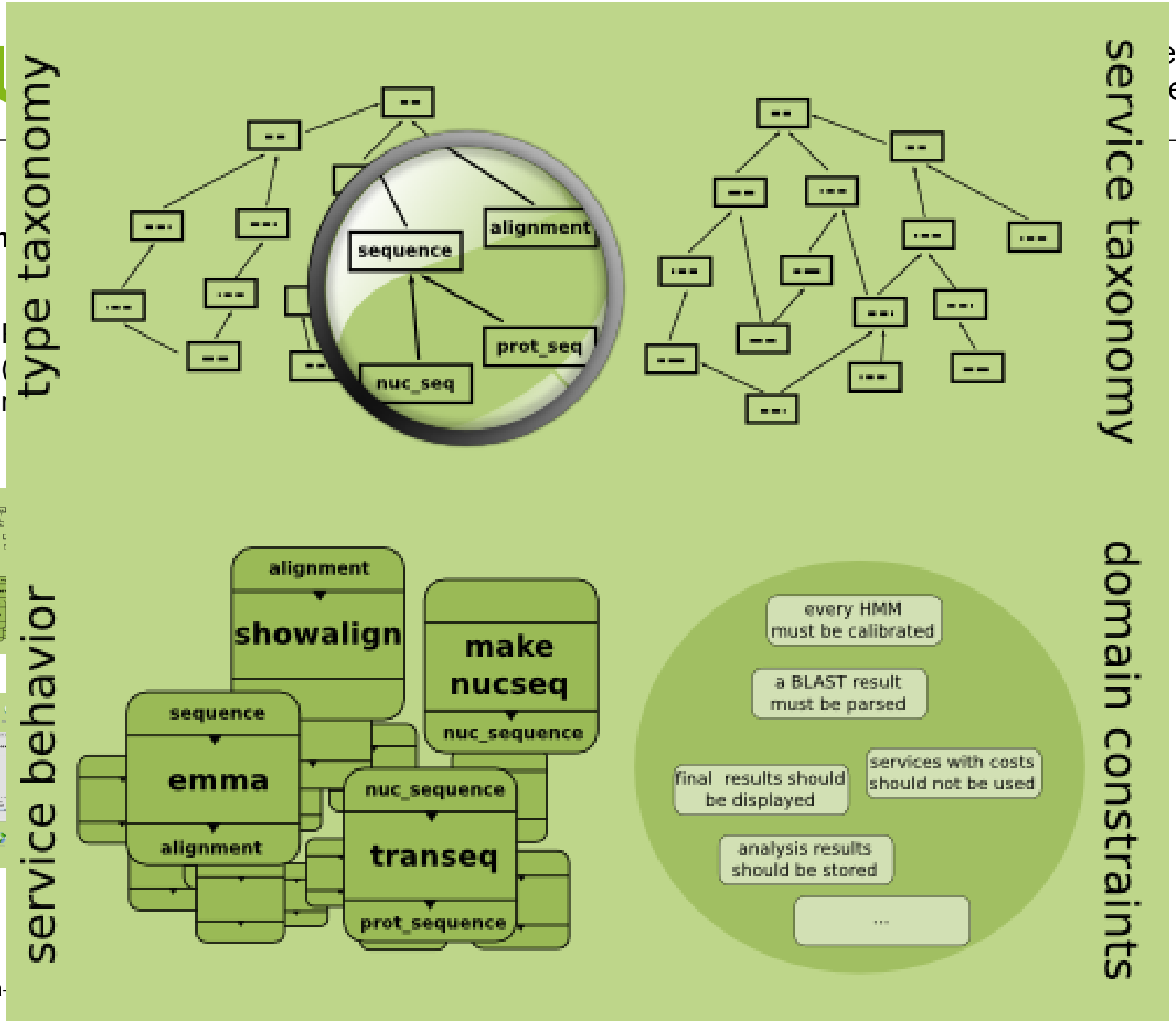
- Domain Modeling:
  - Descriptions of service behavior
  - Taxonomic classifications of services and data types
  - Domain-specific constraints



- Workflow Design:
  - Loose specification
  - Problem-specific constraints
  - Selection and refinement of solutions

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

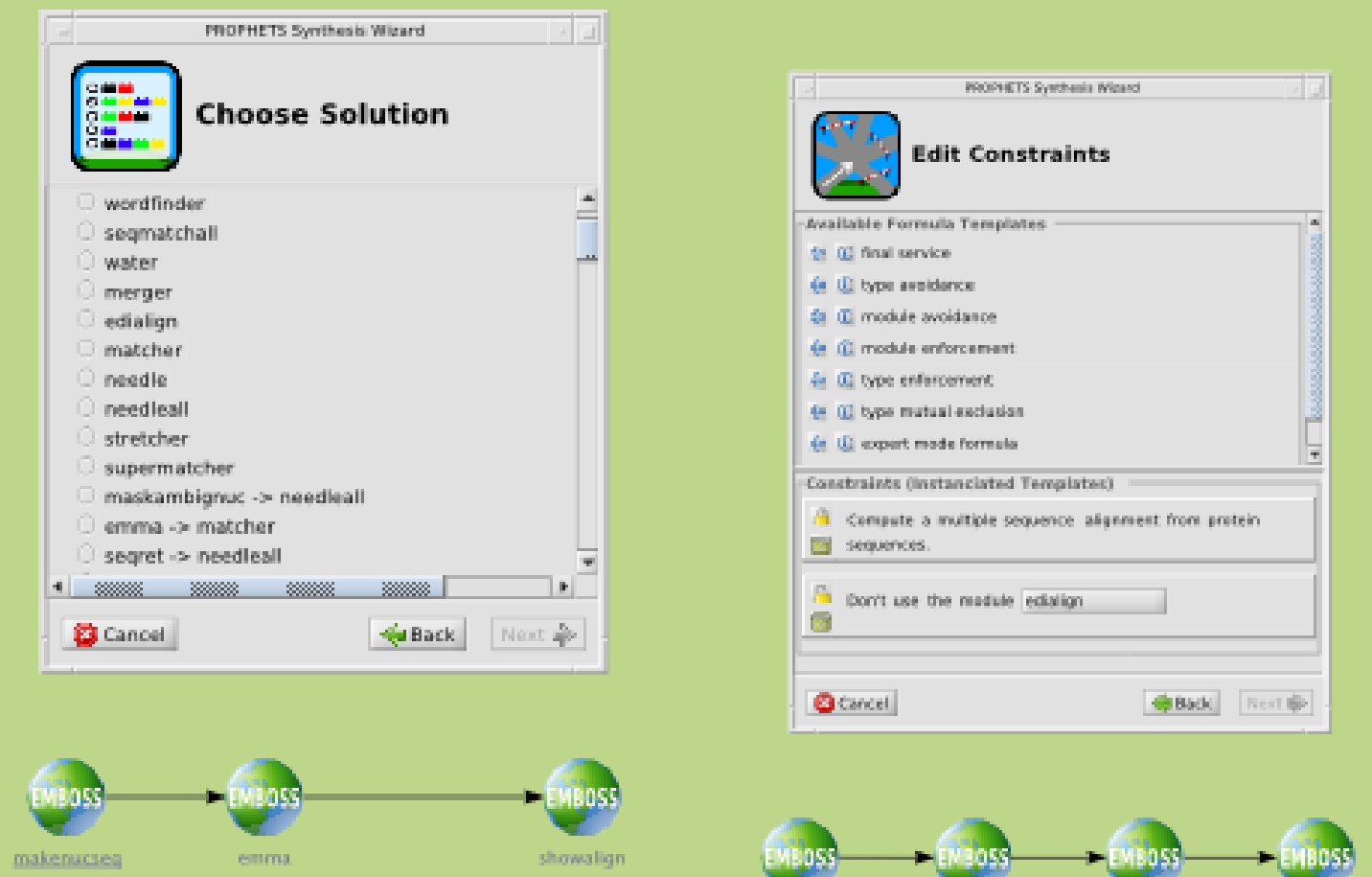
Con



Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Con

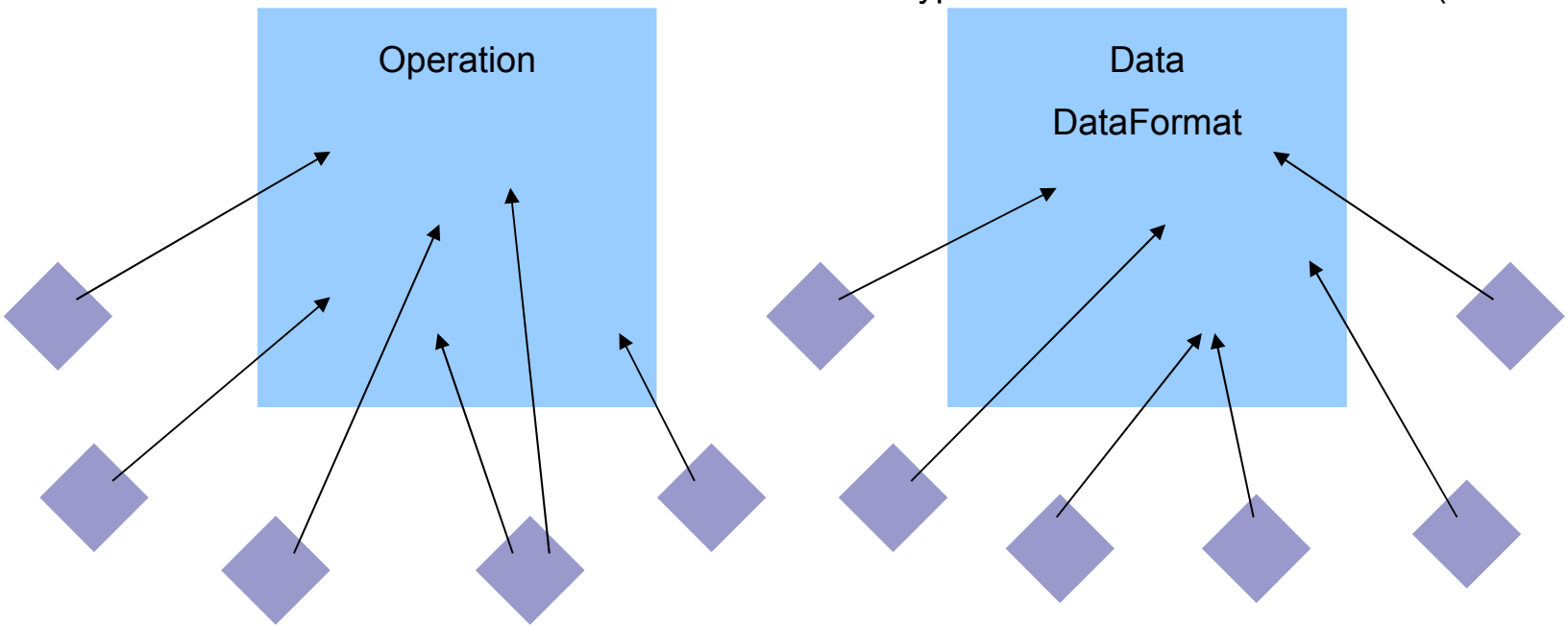
-



Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided **Workflow Composition** Based on the EDAM Ontology

- Domain Modeling:
  - EDAM as background knowledge → skeletal service and type taxonomies (classes)
  - Set of services → services and data types sorted into the taxonomies (instances)

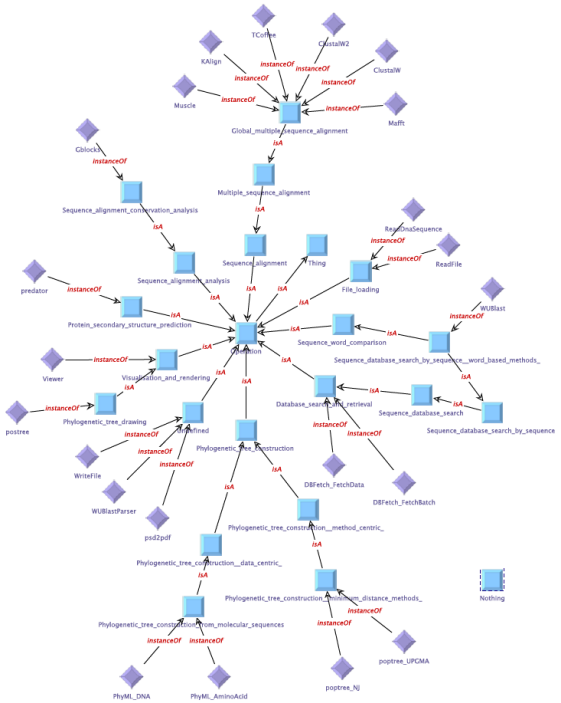


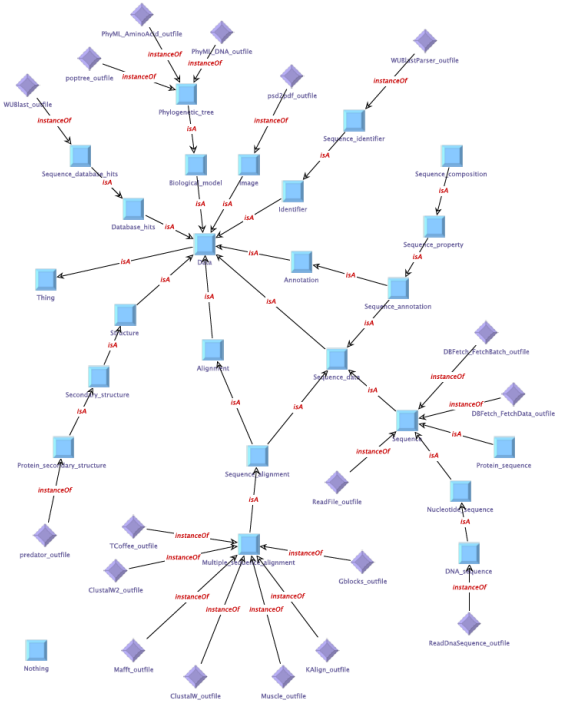
Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided **Workflow Composition** Based on the EDAM Ontology

- Example: Domain model based on EDAM and some selected services.

Service	Behavior
ClustalW *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
ClustalW2 *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
DBFetch_FetchBatch *Database search and retrieval*	In: *Sequence identifier* Out: *Sequence*
DBFetch_FetchData *Database search and retrieval*	In: *Sequence identifier* Out: *Sequence*
Gblocks *Sequence alignment conservation analysis*	In: *Multiple sequence alignment* Out: *Multiple sequence alignment*
KAlign *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
Mafft *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
Muscle *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
PhyML_AminoAcid *Phylogenetic tree construction from molecular sequences*	In: *Protein Sequence* Out: *Phylogenetic tree*
PhyML_DNA *Phylogenetic tree construction from molecular sequences*	In: *DNA sequence* Out: *Phylogenetic tree*
poptree_NJ *Phylogenetic tree construction (minimum distance methods)*	In: *Sequence composition* Out: poptree_outfile
poptree_UPGMA *Phylogenetic tree construction (minimum distance methods)*	In: *Sequence composition* Out: poptree_outfile
postree *Phylogenetic tree drawing*	In: poptree_outfile Out: *Phylogenetic tree image*
predator *Protein secondary structure prediction*	In: *Protein sequence* Out: *Protein secondary structure*
ps2pdf	In: *Image* Out: *Image*
ReadFile *File loading*	Out: *Data*
ReadDNASequence *File loading*	Out: *DNA sequence*
TCoffee *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
WriteFile	In: *Data*
WUBlast *Sequence database search by sequence (word-based methods)*	In: *Sequence* Out: *Sequence database hits*
Viewer *Visualisation and rendering*	In: *Data*
WUBlastParser	In: *Sequence database hits* Out: *Sequence identifier*



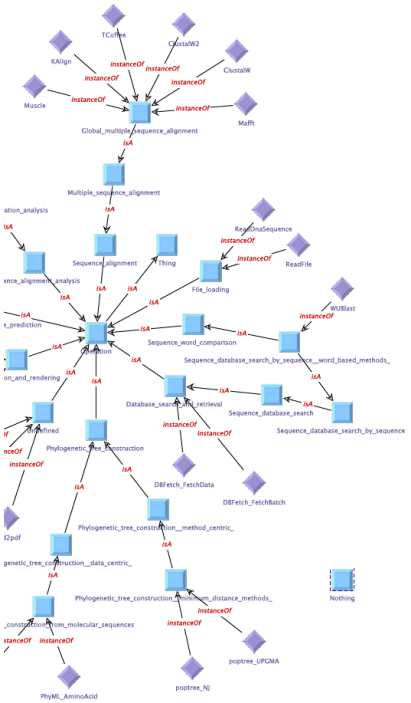


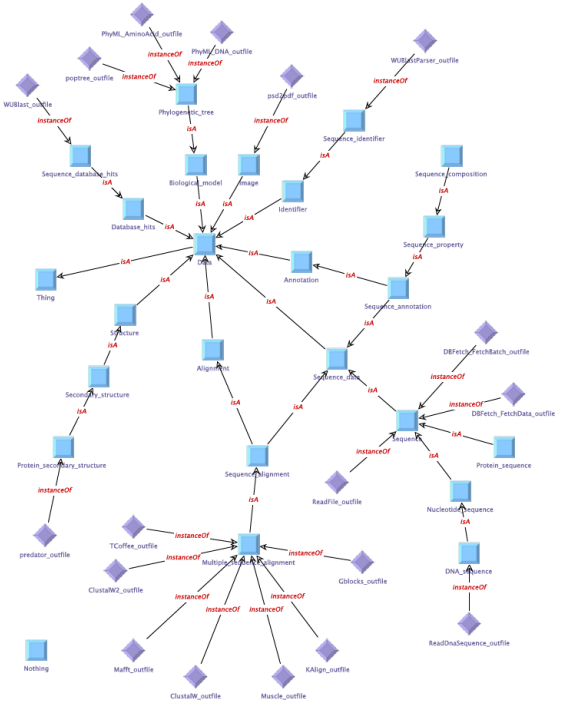
NaturePrecedings - doi:10.1038/npre.2010.5331.1 Posted 15 Dec 2010

Service	Behavior
ClustalW *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
ClustalW2 *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
DBFetch_FetchBatch *Database search and retrieval*	In: *Sequence identifier* Out: *Sequence*
DBFetch_FetchData *Database search and retrieval*	In: *Sequence identifier* Out: *Sequence*
Gblocks *Sequence alignment conservation analysis*	In: *Multiple sequence alignment* Out: *Multiple sequence alignment*
KAlign *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
Mafft *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
Muscle *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
PhyML_AminoAcid *Phylogenetic tree construction from molecular sequences*	In: *Protein Sequence* Out: *Phylogenetic tree*
PhyML_DNA *Phylogenetic tree construction from molecular sequences*	In: *DNA sequence* Out: *Phylogenetic tree*
poptree_NJ *Phylogenetic tree construction (minimum distance methods)*	In: *Sequence composition* Out: poptree_outfile
poptree_UPGMA *Phylogenetic tree construction (minimum distance methods)*	In: *Sequence composition* Out: poptree_outfile
poptree *Phylogenetic tree drawing*	In: poptree_outfile Out: *Phylogenetic tree image*
predator *Protein secondary structure prediction*	In: *Protein sequence* Out: *Protein secondary structure*
ps2pdf	In: *Image* Out: *Image*
ReadFile *File loading*	Out: *Data*
ReadDNASequence *File loading*	Out: *DNA sequence*
TCoffee *Global multiple sequence alignment*	In: *Sequence* Out: *Multiple sequence alignment*
WriteFile	In: *Data*
WUBlast *Sequence database search by sequence (word-based methods)*	In: *Sequence* Out: *Sequence database hits*
Viewer *Visualisation and rendering*	In: *Data*
WUBlastParser	In: *Sequence database hits* Out: *Sequence identifier*

mposition Based on the EDAM Ontology

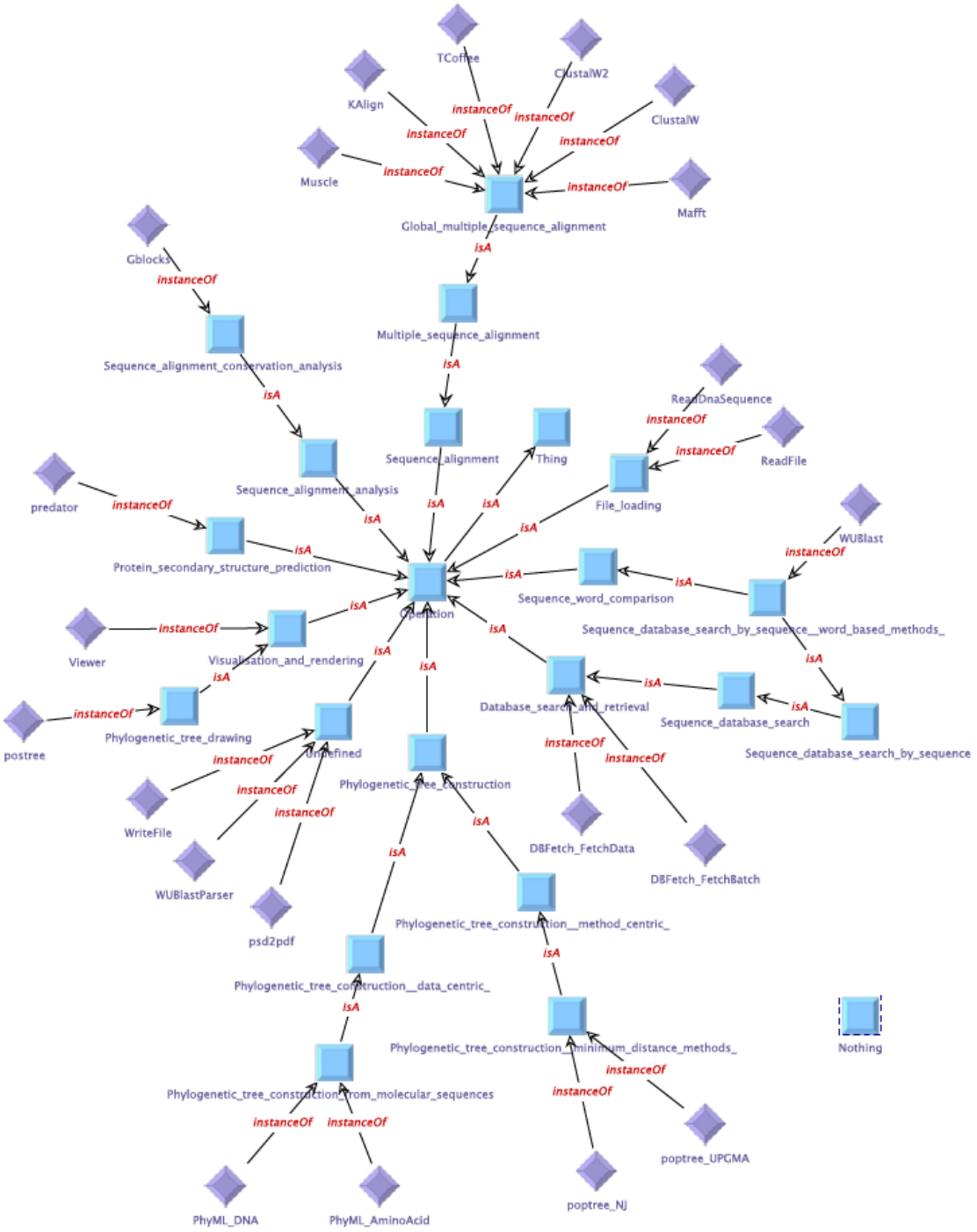
)AM and some selected services.





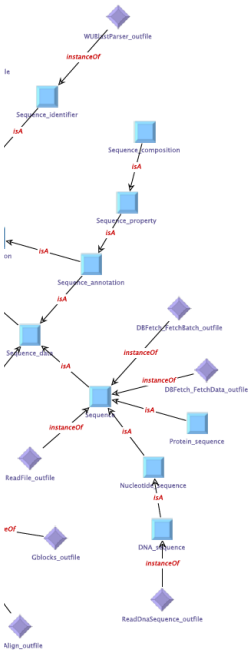
Service
ClustalW *Global multiple sequence alignme*
ClustalW2 *Global multiple sequence alignme*
DBFetch_FetchBatch *Database search and retrieval*
DBFetch_FetchData *Database search and retrieval*
Gblocks *Sequence alignment conservation*
KAlign *Global multiple sequence alignme*
Mafft *Global multiple sequence alignme*
Muscle *Global multiple sequence alignme*
PhyML_AminoAcid *Phylogenetic tree construction from molecular sequences*
PhyML_DNA *Phylogenetic tree construction from molecular sequences*
poptree_NJ *Phylogenetic tree construction (minimum distance methods)*
poptree_UPGMA *Phylogenetic tree construction (minimum distance methods)*
postree *Phylogenetic tree drawing*
predator *Protein secondary structure pred*
psd2pdf
ReadFile *File loading*
ReadDNASequence *File loading*
TCoffee *Global multiple sequence alignme*
WriteFile
WUBlast *Sequence database search by seq (word-based methods)*
Viewer *Visualisation and rendering*
WUBlastParser

Nature Preceedings : doi:10.1038/npre.2010.5039.1 : Posted 15 Dec 2010



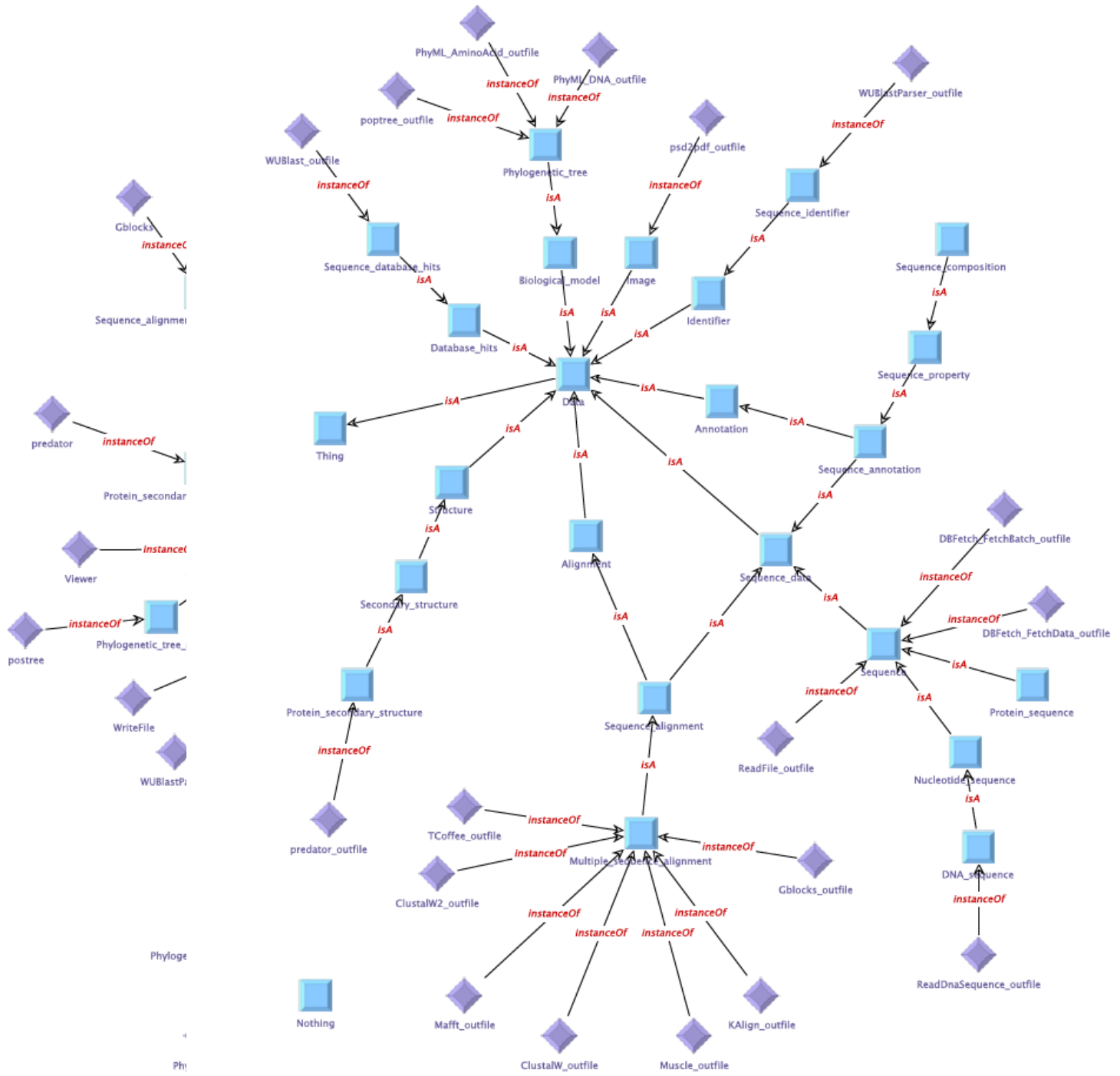
puter Science
ning Systems

ogy



Service
ClustalW *Global multiple sequence alignme*
ClustalW2 *Global multiple sequence alignme*
DBFetch_FetchBatch *Database search and retrieval*
DBFetch_FetchData *Database search and retrieval*
Gblocks *Sequence alignment conservation*
KAlign *Global multiple sequence alignme*
Mafft *Global multiple sequence alignme*
Muscle *Global multiple sequence alignme*
PhyML_AminoAcid *Phylogenetic tree construction from molecular sequences*
PhyML_DNA *Phylogenetic tree construction from molecular sequences*
ptree_NJ *Phylogenetic tree construction (minimum distance methods)*
ptree_UPGMA *Phylogenetic tree construction (minimum distance methods)*
ptree *Phylogenetic tree drawing*
predator *Protein secondary structure pred*
psd2pdf
ReadFile *File loading*
ReadDNASequence *File loading*
TCoffee *Global multiple sequence alignme*
WriteFile
WUBlast *Sequence database search by sequ (word-based methods)*
Viewer *Visualisation and rendering*
WUBlastParser

Nature Precedings : doi:10.1038/npre.2010.5239.1 : Posted 15 Dec 2010

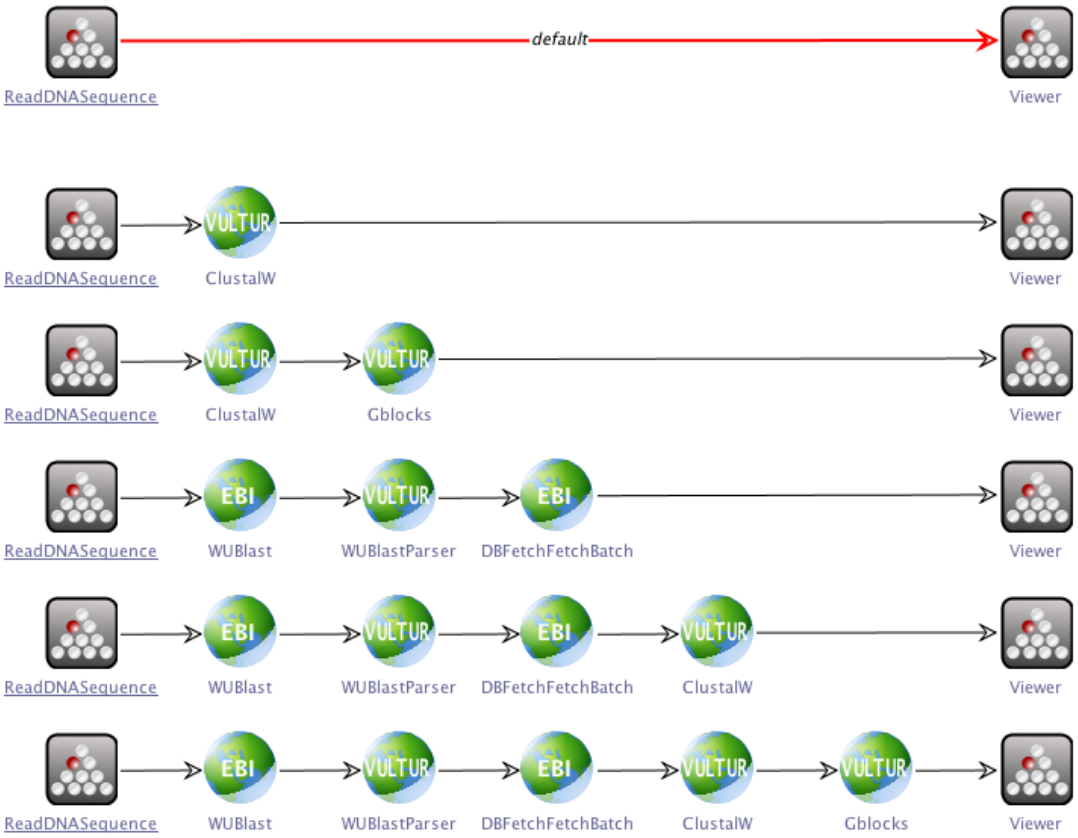


Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided **Workflow Composition** Based on the EDAM Ontology



Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

- Note: for performance reasons limited to a search depth of 5
  (results can be obtained within a couple of seconds on a standard laptop computer)

- **default** synthesis configuration: 264,118 solutions
- **plus** permutation filtering: 5,325 solutions
- **plus** pipelining behaviour: 2,269 solutions

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

- Observation 1: services that make no contribution (e.g. ReadFile, WriteFile)
  → Constraint 1: Avoid use of „ReadFile“, avoid use of „WriteFile“, …

- Observation 2: services that make no progress (e.g. redundant call of Gblocks)
  → Constraint 2: Do not use Gblocks redundantly.

- Observation 3: „dead“ functionality (e.g. BLAST result that is never parsed)
  → Constraint 3: If BLAST is called, a BLAST parser must be used subsequently.

- Observation 4: not the envisaged analysis
  → Constraint 4: Use a *sequence database search by sequence* and after that a *multiple sequence alignment*.
  → Constraint 4': Use a *phylogenetic tree construction* service.

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Constraint-Guided Workflow Composition Based on the EDAM Ontology

- Results (summarized)

Constraints	Visited nodes	Solutions
none	34,026	2,269
1	1,139	55
2	82,343	2,194
3	132,809	1,916
4	436,102	471
4’	129,200	406
1, 2	1,103	49
1, 3	3,123	52
1, 4	8,309	24
1, 4’	2,336	1
2, 3	138,137	1,847
2, 4	443,860	459
2, 4’	181,365	394
3, 4	910,672	138
3, 4’	277,239	359
4, 4’	847,845	18

Constraints	Visited nodes	Solutions
1, 2, 3	9,603	31
1, 2, 4	8,057	24
1, 2, 4’	2,084	1
1, 3, 4	28,545	24
1, 3, 4’	18,699	0
1, 4, 4’	15,919	0
2, 3, 4	919,162	138
2, 3, 4’	284,463	347
2, 4, 4’	859,047	18
3, 4, 4’	1,752,153	0
1, 2, 3, 4	28,545	24
1, 2, 3, 4’	2,084	1
1, 2, 4, 4’	15,235	0
1, 3, 4, 4’	54,711	0
2, 3, 4, 4’	1,764,843	0
all	54,027	0

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

Conclusions

- Summary:
  - EDAM as background knowledge ensures that *possible* workflows are found.
  - Guiding synthesis to actually *desired* solutions requires more knowledge.

- Future work:
  - Proceed to greater search depths.
  - Integrate BioCatalogue services and annotations.
  - Identify general domain-specific knowledge beyond EDAM.

Nature Precedings : doi:10.1038/npre.2010.5397.1 : Posted 15 Dec 2010

The end

Thank you for your interest!