

Scientific Modeling and Resource Discovery in Scientific Workflows[†]

Maliha Aziz

Scientific Data management Lab

Computational Biosciences

May 28, 2008

[†] This work was partially supported by the National Science Foundation¹⁷ (grants IIS 0431174, IIS 0551444, and IIS 0612273). Any opinion, finding, and conclusion or recommendation expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Outline

- Terminology
- Introduction
 - An Example Scenario in OMICS
 - Bioinformatics Resources
 - Resource representation
 - Implementing and Executing Scientific Workflows
- Scientific Modeling
- Resource Discovery

Terminology

- GO – Gene Ontology
- PO – Protein Ontology
- XML – Extended Markup Language
- RDF – Resource Definition Framework
- RDFS – RDF Schema
- OWL – Web Ontology Language
- WSDL – Web Service Description Language
- OWL-S – Ontology for describing Semantic Web Services
- WSDL-S – WSDL extended with semantic annotations.
- SAWSDL – Semantic Annotations for WSDL
- BPEL – Business Process and Execution Language
- XScufl – an XML dialect of the Simple Conceptual Unified Flow Language
- BPEL4WS – Business Process Execution Language for Web Services
- BPML – Business Process Modeling Language

Introduction

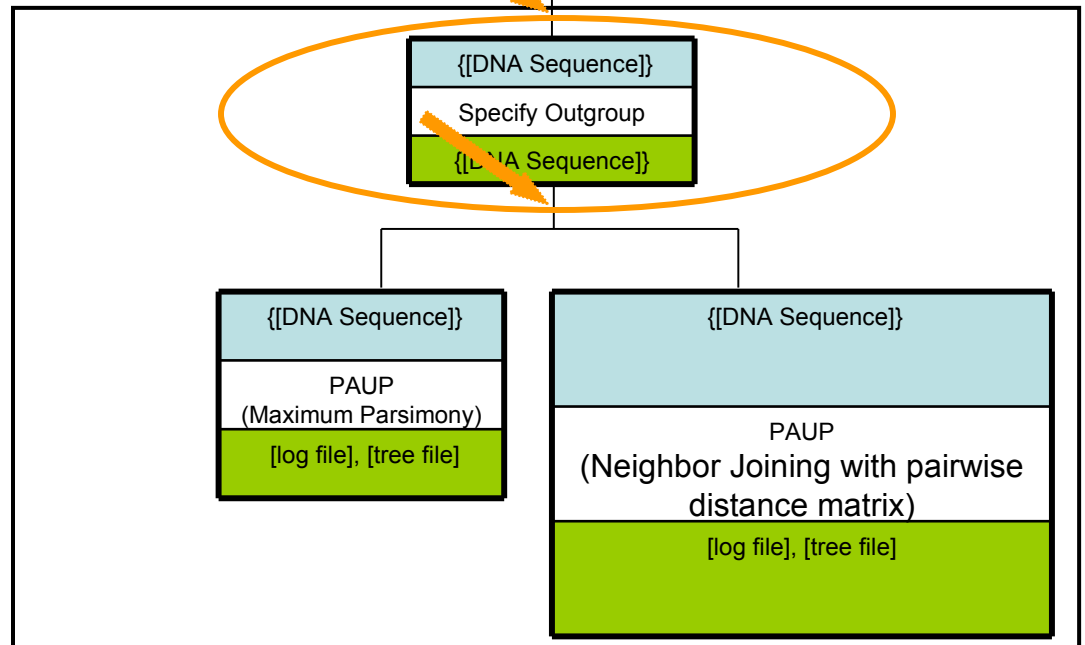
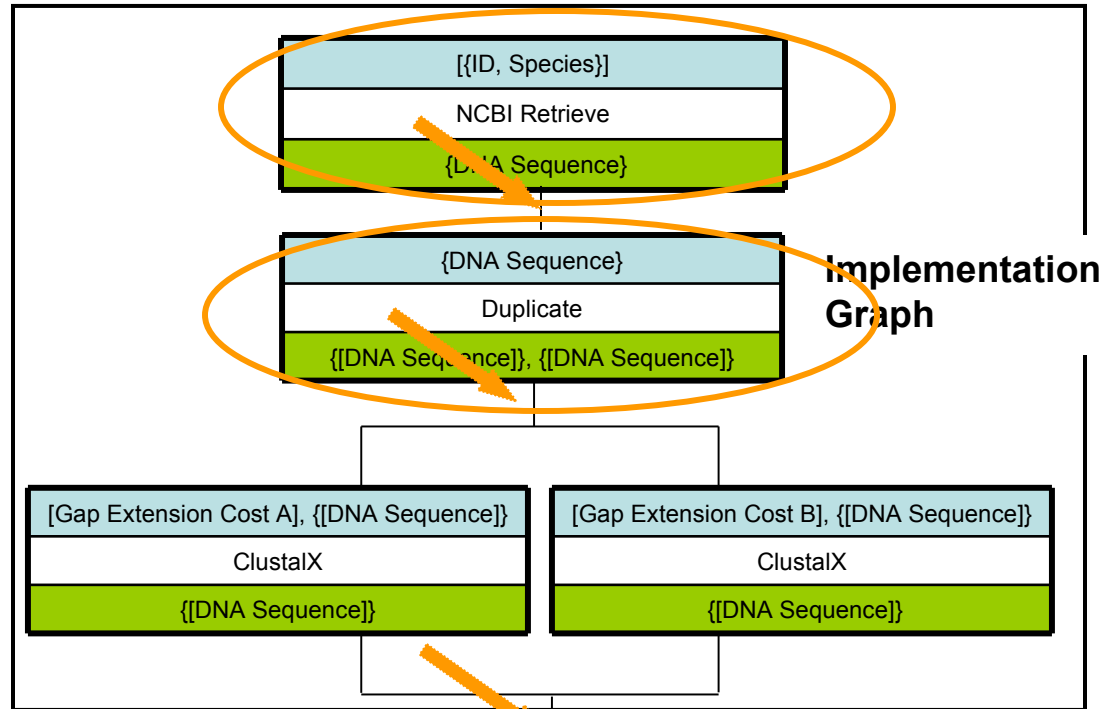
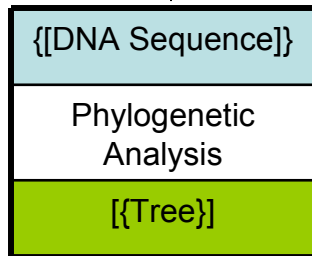
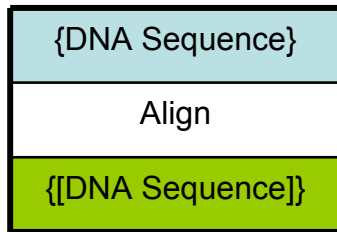
- Scientific approaches for experimentation
 - Hypothesis centric
 - Data centric

An Example Scenario in OMICS

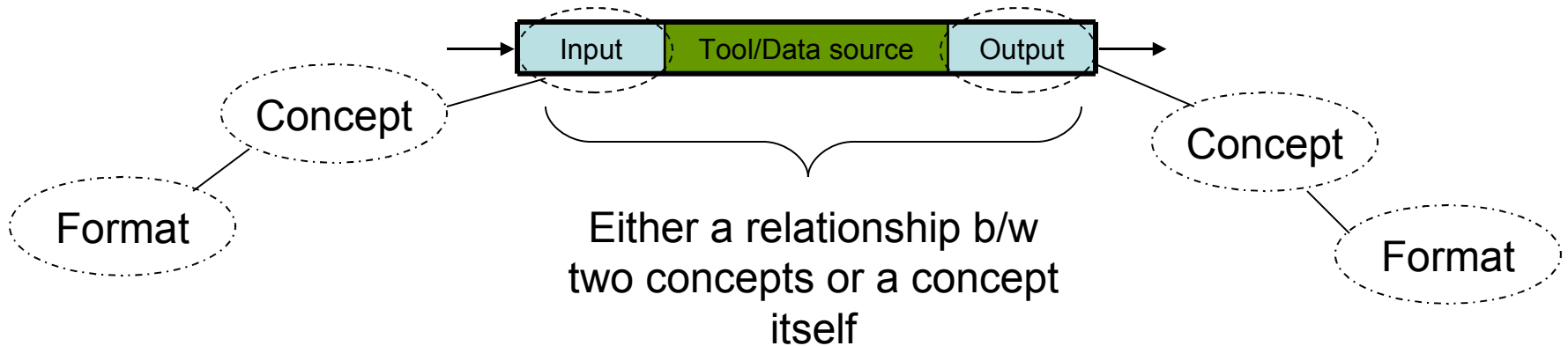
Phylogenetic Analysis

- Retrieve DNA sequences belonging to different species. Align and perform phylogenetic analysis[†].
- Some simple questions that arise
 - How is the experiment to be conducted? [Scientific Modeling](#)
 - Which data sources and tools should I use? [Resource Discovery](#)
 - What would be the input/output formats?
 - How do I connect the output of one resource to the input of the other? [Resource Composition](#)

Design Graph



Bioinformatics Resources



Resource Types:

- Data sources
- Tools
- Ontologies (Def): An ontology is a specification of a conceptualization[†].
- Concepts derived from various domain ontologies e.g., GO, PO.
- Need for placement resources and their inputs/outputs with respect to an ontology.
- Increases resource
 - Visibility
 - Understandability and
 - Reusability

[†] <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>

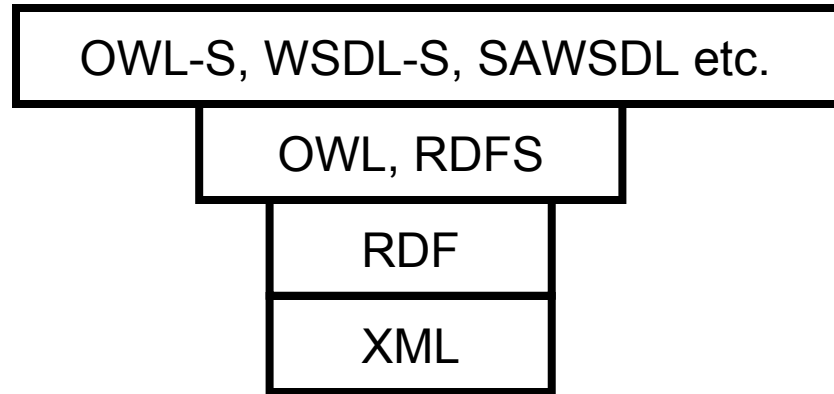
Resource Representation

- To be meaningful data/tool need to be annotated.
- Information about data/tool → ‘meta data’.
- Data and resources need to be structured.
- Structure provided by XML
 - Allows message level interoperability.
 - Machine processable language.
- Some Minimum Information (MI) Standards for structuring and annotating data used in omics[†]
 - CIMR
 - MIACA
 - MIAME
 - MIAME/Env
 - MIAME/Nutr
 - MIAPE
 - MIARE
 - MIFlowCyt
 - MIGS
 - MISFISHIE
 - MIAME/Plant
 - MIAME/Tox
 - MIMPP
 - MIMIX
 - MIQAS
 - MIRIAM

Some Markup languages used in omics

- SBML
- PDBML
- HUP-ML
- AGML
- mzXML
- MAGE-ML

Resource Representation



- Simple XML has limitations.
- Need for layer at conceptual level → introduction of various standards.
- Further automation achieved by making tools available as web services.
- Provides
 - Advanced Automation.
 - Application Integration.

Implementing and Executing Scientific Workflows

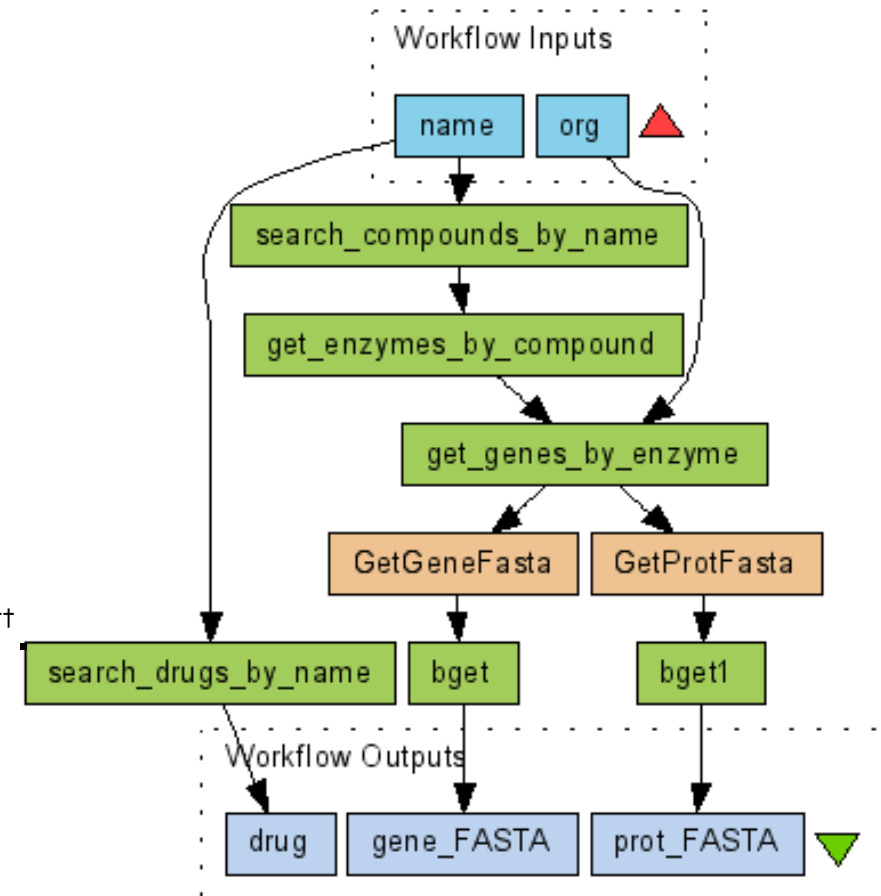
- Simple Scripts written in Perl, Python or Matlab insufficient.
- Need for Workflow Systems.
- Features[†]:
 - Cyclic graph model
 - GUI
 - Tools in the form of services
 - Typed input/output ports
- Languages : XSCUFL, BPEL, BPEL4WS, BPML.
- Example Workflow systems
 - Taverna
 - Kepler
 - Ergatis
 - Wild fire
 - Pegasys
 - Biopipe
 - And many more

[†] Fox, G. C. and Gannon, D. 2006. Special Issue: Workflow in Grid Systems: Editorials. *Concurr. Comput. : Pract. Exper.* 18, 10 (Aug. 2006), 1009-1019.

Implementing and Executing Scientific Workflows

Example workflow in Taverna

For a given drug retrieve drug KEGG ID[†] together with DNA and amino acid sequences in FASTA format of the proteins involved in the metabolism of the given drug^{††}



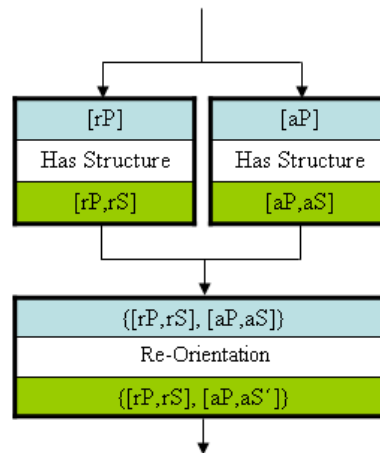
[†] Kyoto Encyclopedia of Genes and Genomes (KEGG)

^{††} http://bioinformatics.istge.it:8080/biowep/details.jsp?wfv_id=23

Scientific Modeling

- A scientific protocol is constructed in two phases[†]
 - design
 - implementation
- **Aim of the protocol:** Superpose proteins structures evaluated at the residue level, when both sequences and structures have diverged.
- GAPS^{††} (Gaussian Based Alignment of Protein Structures) algorithm performs the main superposition task.
- Use of ProtocolDB^{†††} workflow model for design and implementation.
- Concepts derived from domain ontology for proteins i.e., Protein Ontology (PO).

Design Phase



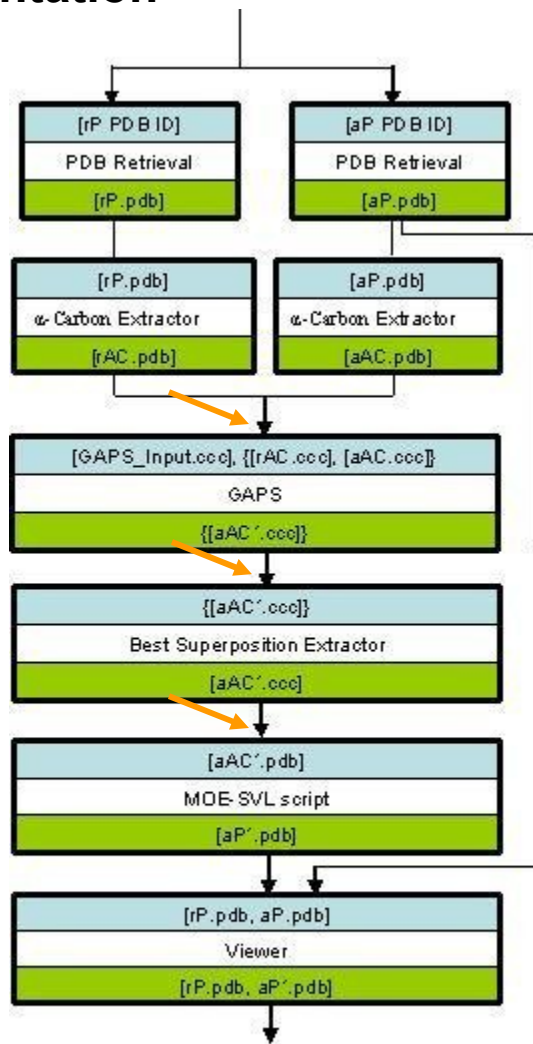
† N Kwasnikowska, Y Chen, Z Lacroix: Modeling and storing scientific protocols. In: OTM Workshops (1). (2006) 730–739

†† Provided by Dr Jordi Mestres - Chemogenomics Laboratory, Research Unit on Biomedical Informatics Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain

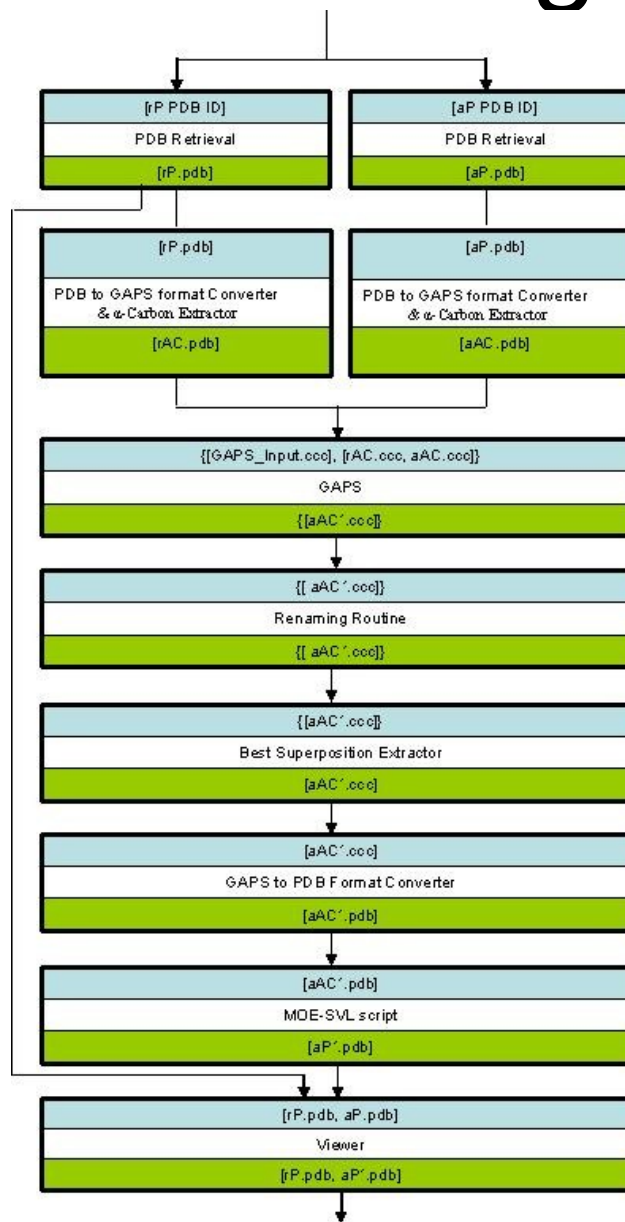
††† M Kinsy, Z Lacroix, C Legendre, P Włodarczyk, "ProtocolDB: Storing scientific protocols with a domain ontology," in International Workshop on Web Data Integration and Mining in Life; LNCS, 2007, pp. 17-28.

Scientific Modeling

Implementation Phase 1

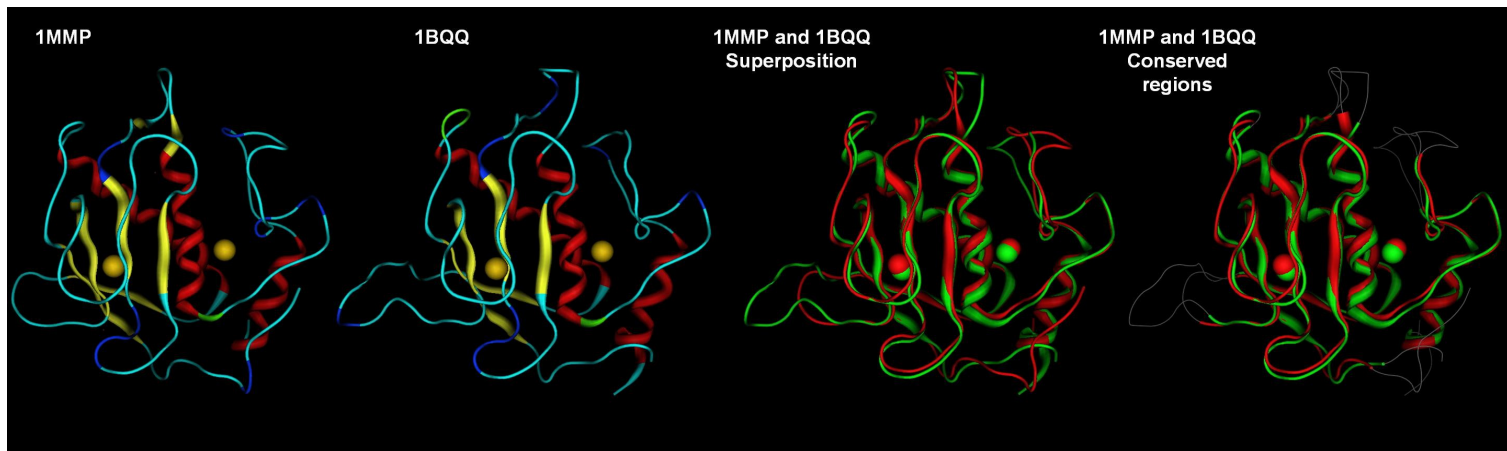


Implementation Phase 2



Results

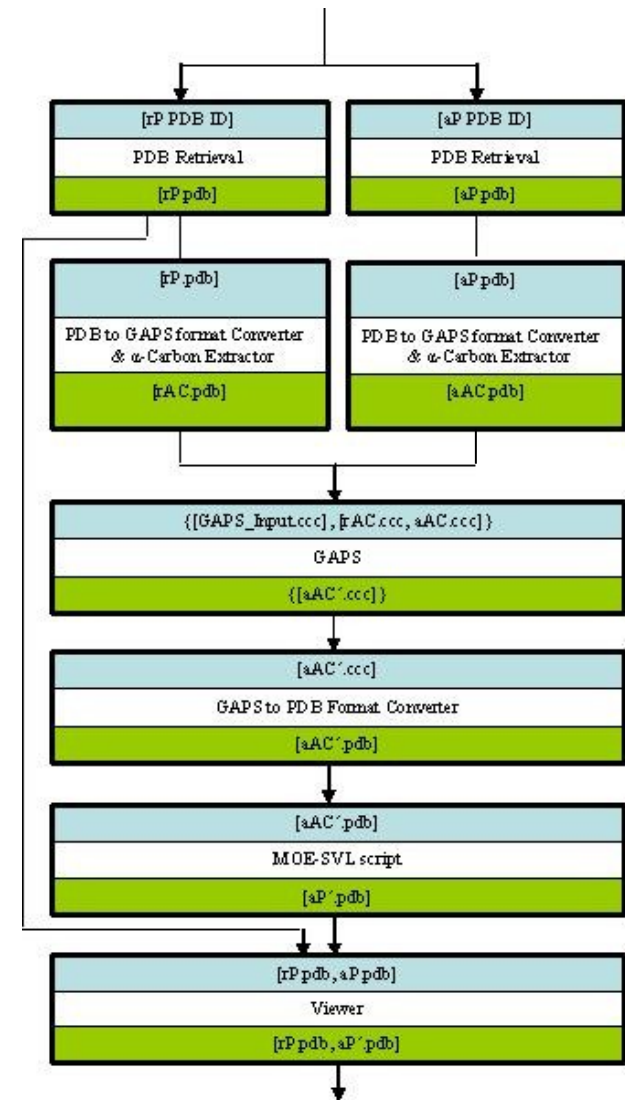
- First implementation of protocol done with Shell Scripting.
- Protocol executed on two proteolytic enzymes structures
 - Matrix Metalloproteinases (MMP) such as gelatinase A (MMP-7, PDB ID: 1MMP) membrane-type MMP-1 (MT1-MMP, PDB ID: 1BQQ).
- Reference Protein: 1MMP
- Adapting Protein: 1BQQ
- Execution Time: ≈ 7 seconds[†].
- 1MMP and 1BQQ are close homologs \rightarrow best superpositioning result is very accurate.



Limitations & Improvements

Performance limitations:

- Manual steps
 - retrieval of files from the PDB repository.
 - cleaning of files to remove any excess chains of the protein structures.
- A more sophisticated version of the protocol.
- User provided an interface to view all the orientations of the adapting protein.
- Availability of a browsing function enables the scientist to select the best orientations of the adapting protein based on other considerations including structural ones.



Resource Discovery

- Resources can be accessed
 - Screen scraping
 - Web crawlers/ spiders
 - Semantic Markups at website
 - Advertising in a semantic enable service registry e.g., BioMoby central or Grimoires
- Usually Resource Discovery
 - Is driven by format.
 - Is not dependant on domain ontology.
 - Does not consider tool efficiency, scientific aim implementation.

BioMoby[†]

- Initiated in 2001.
- Format and an open access repository.
- Resource providers register services.
- Users may select the resources they need.
- Includes various ontologies.
- Contains 1,611 services registered in BioMoby Central^{††}.
- 84 different authorities^{††}.

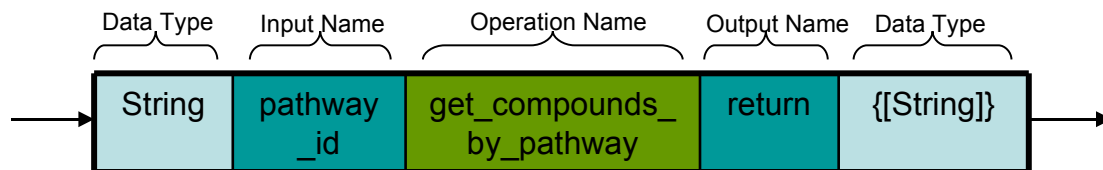
[†] M. D. Wilkinson and M. Links, "BioMOBY: An open source biological web services proposal," Brief Bioinform, vol. 3, no. 4, pp. 331–341, 2002.

^{††} As of April, 2008

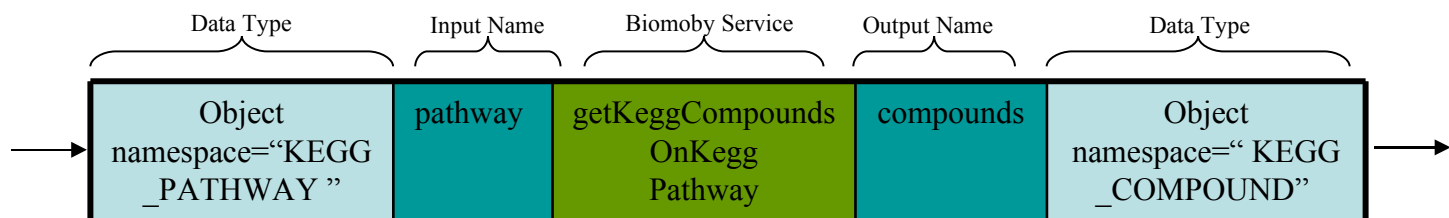
Resource Discovery

- BioMoby differs from WSDL
- Uses an ontology
 - 1) to provide a semantic annotation to the resources.
 - 2) to classify resources with respect to a conceptual hierarchy.

Simple KEGG[†] Web Service



KEGG Service in BioMoby



In our effort we wish to,

- Perform automatic mapping of resource registries in a domain ontology.

Our focus:

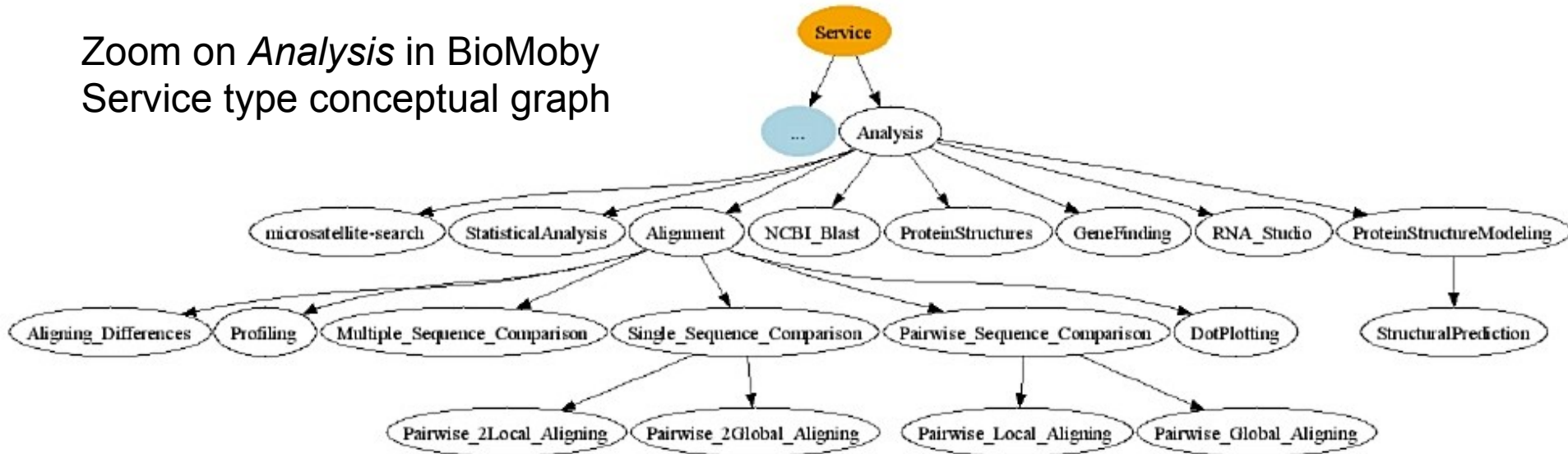
- BioMoby repository.

We analyze the

- Semantic layer of BioMoby services.
- Propose algorithm to map BioMoby services to a domain ontology.

BioMoby Service type Conceptual Graph

Zoom on *Analysis* in BioMoby
Service type conceptual graph



- Tree of root 'Service'
- Total No. of Nodes: 114
- Depth: 5
- Stored in a SQL relational database the Moby Central Registry[†].
- Access is provided through the Moby Central API^{††}.

[†] M. Wilkinson, D. Gessler, A. Farmer, and L. Stein, "The biomoby project explores open-source, simple, extensible protocols for enabling biological database interoperability," in Proceedings of the Virtual Conference on Genomics and Bioinformatics, 2003, pp. 17–27. [Online]. Available: www.virtualgenomics.org

^{††} http://biomoby.open-bio.org/CVS_CONTENT/moby-live/Docs/MOBYSAPI/index_API.html

BioMoby Service type Conceptual Graph

- 5 primitive types[†].
- Expected to cluster most bioinformatics resources,
 - Analysis
 - Parsing
 - Registration
 - Retrieval
 - Resolution

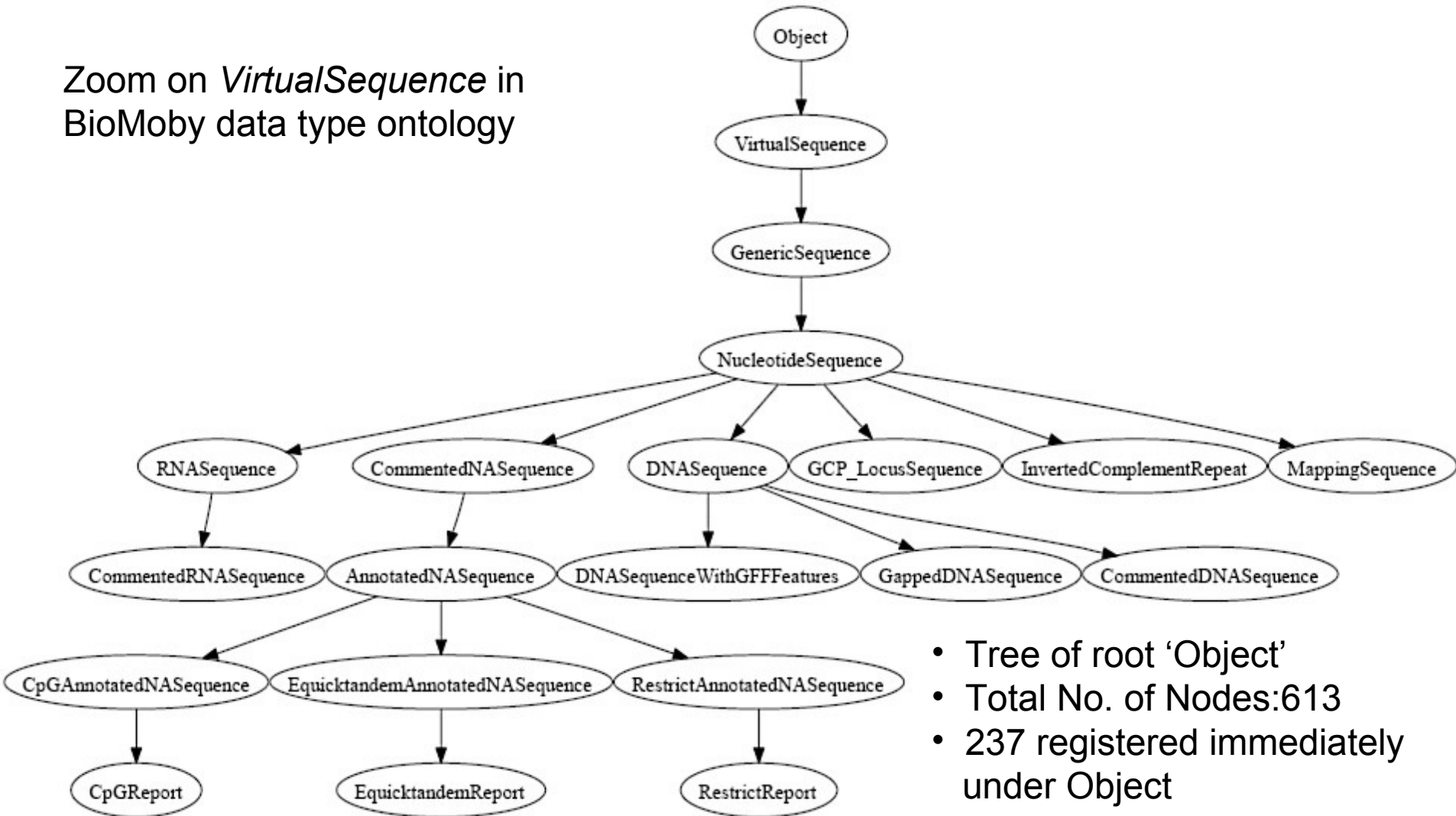
TABLE I
SERVICE TYPE ONTOLOGY

Service Type	Number of Children	Depth	Number of Services
Analysis	20	4	290
Bioinformatics	53	5	248
Conversion	3	2	67
Edit	0	1	202
Parsing	0	1	34
Registration	0	1	0
Rendering	0	1	3
Resolution	0	1	0
Retrieval	3	2	555
Testing	2	2	18

- We identified syntactic and semantic inconsistencies.
- Structural problems in the ontology
 - redundant types.
 - structural classification lacks homogeneity depending on the type.
 - lack of consistency in languages used to name the types, services, and for their description.
- Problem in instances registration under each type.

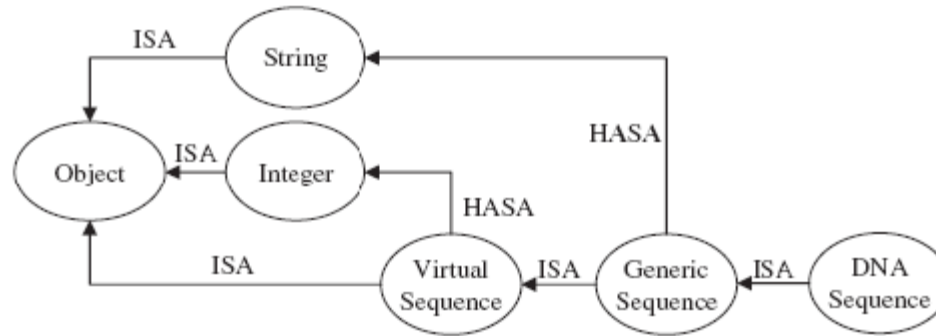
BioMoby Data type Conceptual Graph

Zoom on *VirtualSequence* in BioMoby data type ontology



- Tree of root 'Object'
- Total No. of Nodes:613
- 237 registered immediately under Object
- Depth: 8

BioMoby Data type Conceptual Graph



- The design of the datatype ontology in BioMoby is inspired by the GO ontology.
- Two types of relationships
 - ‘is-a’ depicting subclass.
 - ‘has’ or ‘has-a’ for complex data types depicting container ship.
- Problems:
 - The same ontology contains concepts and formats.
 - All formats must be organized in a tree regardless of their complexity.
 - Lacks consistency.
 - Graph is format driven.
 - The higher nodes should be more conceptual than the leaves but are not.

Extracting a Domain Ontology

- Our approach
 - 1) Retrieve the BioMoby registry from BioMoby Central;
 - 2) Extract all the BioMoby services along with their inputs and outputs using the BioMoby APIs;
 - 3) Extract the service type ontology;
 - 4) Extract the data type ontology;
 - 5) Remove database access (registered under Retrieval and Database);
 - 6) Remove meaningless datatypes;
 - 7) Generate the ontology.

- Pre-Processing Result:
 - Removed 1,104 services out of 1611 services.
 - Removed 20 datatypes from the 613 datatypes.
- For ontology generation remaining services along with their inputs and outputs is considered.

TABLE III
LOW LEVEL DATA TYPES

Irrelevant Data Types	No. of Tools Having these as Inputs/Outputs
User	7
Float	4
text-formatted	45
text-plain	16
ArrayFloat	4
Array	0
Rules	1
Boolean	2
text plain	105
text-html	11
ArrayXYData	3
DateTime	1
text formatted	1
String	28
text-xml	8
EmailAddress	5
simple key value pair	103
Integer	8
Object	214
Zip_Encoded	2

Results

- Running time: ≈ 40 seconds[†].
- 116 concepts selected by the algorithm.
- We classify them in super-class/subclass manner.
- We identify four scientific objects:
 - sequence
 - protein
 - an object for the (semantic) glyco-bio informatics object
 - RNA structure

TABLE II
ALGORITHM EXECUTION - TOOLS, DATATYPES, SERVICETYPES
STATISTICS

	Input	Pre-Processing	Output
Number of Tools	1,611	507	507
Number of Datatypes	613	583	116
Number of Service Types	115	14	109

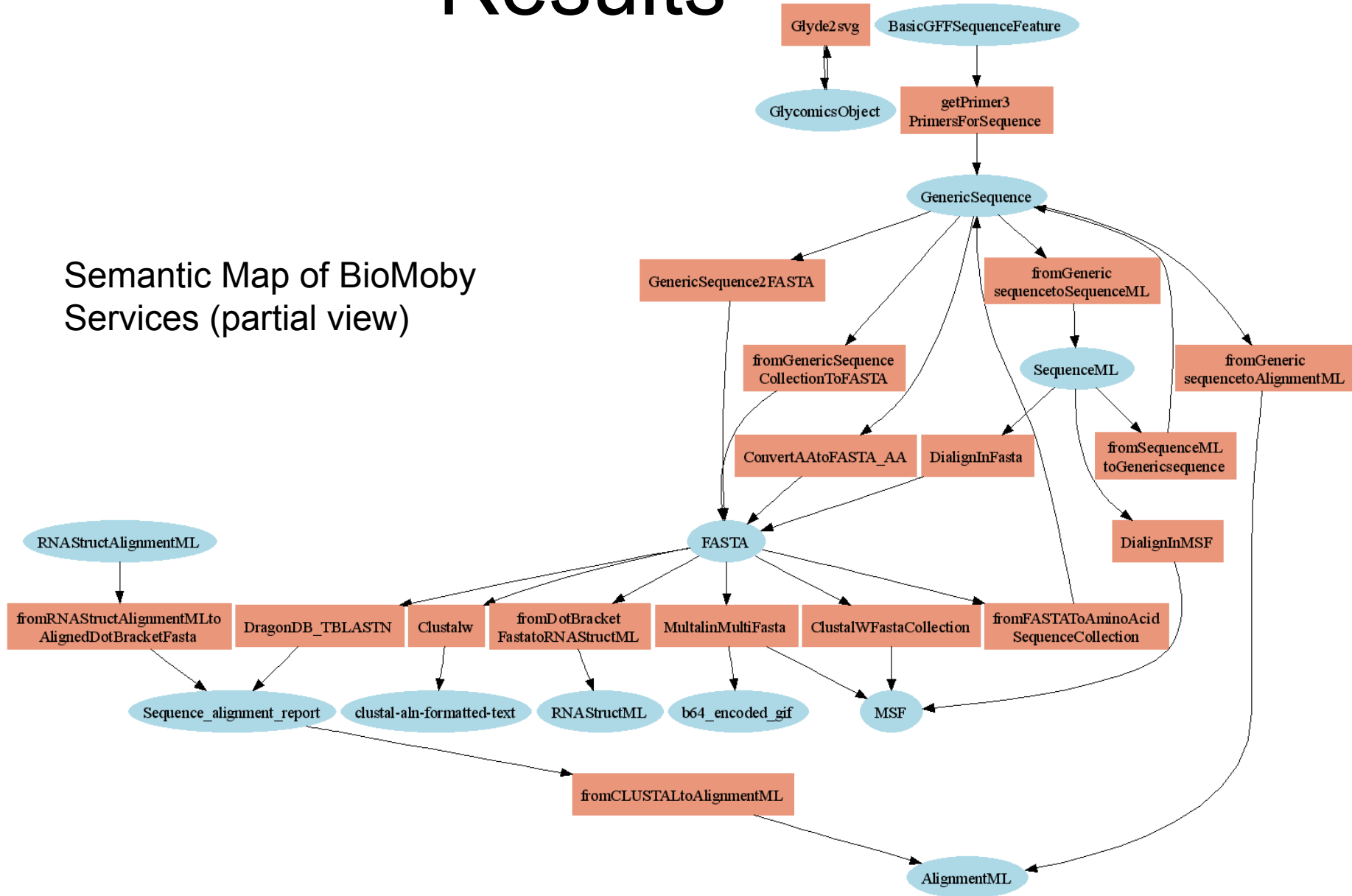
Analyzing the generated concepts

- *Un-Informative and un-Informed Concepts*: Examples, b64_encoded_jpeg, Jackknifernon.
- *Biomoby Ontology Discrepancies* - Example FASTA and FASTA_Text. -should not have been generated as concepts at all.
- *Too Deep a Curation by Our Algorithm* - Selecting highest ancestor as a concept.
- *Implicit Information* - missing intermediate nodes. Example, PepNovoReport.

[†] Processor model: Intel(R) Core(TM)2 CPU 2.13GHz, 2GB of RAM

Results

Semantic Map of BioMoby Services (partial view)



Conclusions and Future work

- Challenges faced by the algorithm
 - generation of too many concepts that do not have any subclasses.
 - overcoming the orthogonal motivation or resource discovery and composition.
- Need for rules and axioms for concept generation.
- As a result, decrease the level of curation and allow expansion of the conceptual mapping.

Acknowledgement

- Dr. Zoé Lacroix
- Dr. Nathalie Meurice