Biomedical language and biomedical knowledge: Emerging synergies

Patrick Lambrix
Linköpings universitet

Questions

- what do you think are the main obstacles in bridging "text" with "knowledge"?
- what is your research/development plan in near future?
- what technologies do you think are lacking and would be crucial for future progress?
- what will be accomplished in five years in our field (text mining, text and knowledge management in the bio-medical domain)?
- what kinds of research do you think most promising (and should be promoted) and what kinds of research do you think least promising?

Questions

- what do you think are the main obstacles in bridging "text" with "knowledge"?

PEOPLE

Developers of ontologies and applications:

We did not find an ontology that exactly fits our needs, so we developed a completely new one.

We use another ontology, but we redefined some terms.

We introduced/implemented our ontology in our own format.

Developers of standards:

From slide by Lena Strömbäck

Name	Ver.	Year	Defined by	Purpose	Tools	Data
SBML	2	2003	Systems Biology Workbench development group.	A computer-readable format for representing models of biochemical reaction networks.	Many tools available.	Data available from many databases, for instance, KEGG and Reactome.
PSI MI	2.5	2005	Proteomics Standards Initiative.	A standard for data representation for protein-protein interaction to facilitate data comparison, exchange and verification.	Tools for viewing and analysis.	Datasets available from many sources, for instace IntAct, DIP and MINT.
BioPAX	2	2005	The BioPAX group.	A collaborative effort to create a data exchange format for biological pathway data.	Existing tools for OWL such as Protégé.	Datasets available from Reactome.
CellML	1.1	2002	University of Auckland and Physiome Sciences, Inc.	Support the definition of models of cellular and subcellular processes.	Tools for publication, visualization, creation and simulation.	CellML Model Repository (~240 models).
CML	2.2	2003	Peter Murray-Rust, Henry S. Rzepa.	Interchange of chemical information over the Internet and other networks.	Molecular browsers, editors.	BioCYC.
EMBLxml	1.0	2005	EBI	More stability and fine-grained modelling of nucleotide sequence information.	API support in BioJavaX.	EMBL.
INSDseq	1.4	2005	International Nucleotide Sequence Database Collaboration.	The purpose of INSDSeq is to provide a near-uniform representation for sequence records.	API support in BioJavaX.	EMBL, DBJ and GenBank.
Seqentry	n/a	n/a	NCBI	NCBI uses ASN.1 for the storage and retrieval of data such as nucleotide and protein sequences. Data encoded in ASN.1 can be transferred to XML.	SRI's BioWarehouse and ProteinStructureFactory's ORFer.	Entrez.
BSML	3.1	2002	Labbook.com.	Facilitate the interchange of data for more efficient communication within the	Labbook's Genomic Browser and Sequence Viewer. Converters.	Previously provided by EMBL.

Questions

- what is your research/development plan in near future?

Finding overlap
- Alignment of ontologies
- Alignment of standards

#	CellML: component	SBML: Reaction	PSI MI: Interaction	#
1			imexID	1
2	name	id	id	2
3		name		3
4			xref	4
5	variable			5
6	reaction			6
7		sboTerm	interactiontype	7
8			experimentList	8
	variable-ref	reactant	participantList	
		product		
9		modifier		9
10		id	id	10
11		name	names	11
12			experimental-role	12
13	role	sboTerm	biological-role	13
14			participantidentification	14
15			experimentalpreparation	15
16			confidencelist	16
17	direction			17
18	delta_variable			18
19	stoichiometry			19
20		stoichiometry		20
21		kineticLaw	inferredInteractionlist	21
22			participants	22
23			modelled	23
24			confidencelist	24
25	reversible	reversible		25
26		fast		26

Often (not always) there are good reasons for introducing new ontologies etc.

Needed: Clear understanding of the … of the standards and ontologies

- *Why*: motivation, how to be used
- *What*: what is represented, modeled, described
- *How*: representation language, modeling strategy
- *When*: time/versions
- *Connections*

Questions

- what technologies do you think are lacking and would be crucial for future progress?

Methodologies for creating standards and ontologies

- ☐ Document why, what, how, when, connections
- ☐ Reuse existing standards and ontologies (when possible)

OBO Foundry

1. open and available
2. common shared syntax (how)
3. unique identifier space   (how, connections)
4. procedures for identifying distinct successive versions (when)
5. clearly specified and clearly delineated content (what)
6. textual definitions for all terms (what)
7. Use relations from OBO Relation Ontology (how)
8. well documented
9. plurality of independent users
10. developed collaboratively with other OBO Foundry members (connections)

Many smaller ontologies

- ☐ based on expertise, application, …
- ☐ more complex modeling
- ☐ may need different representation formalisms
- ☐ reasoning?
- ☐ well-defined overlap, 'glue'