

Formal Ontology and the Gene Ontology

Jennifer Williams

Ontology Works, Inc.

williams@ontologyworks.com

Gene Ontology

- <http://www.geneontology.org/>
- consortium of model organism groups
- about 11,000 terms

Goal: “to provide a set of structured vocabularies for specific biological domains that can be used to describe gene products in any organism”, for the purpose of “facilitat[ing] communication between people and organizations”

Ref: [GO Consortium 2001]

GO Consortium wants:

- Tools to enforce semantic consistency
- Ability to constrain statements according to species
e.g. “glycosome can be part of a cell
for Kinetoplastidae”
- Extend vocabulary to cover “cell types”, “tissues”,
other useful terms.

Ref: [GO Consortium 2001]

Bioinformatics community wants to know:

What would it take to get GO to the point where it could support automated bioinformatics applications?

Could ontologies help?

GONG project: <http://gong.man.ac.uk>
using DAML+OIL & manual enrichment of GO semantics

Why ontologies?

- When theories are expressed in formal language, computers can help enforce internal consistency and data consistency [Karp 2001]
- Ontologies recognized as basis for semantic integration [Schulze-Kremer 1998]
- The more domain knowledge encapsulated in the ontology, the less required for correctly making use of it

ontologies and Ontology

- **ontologies**: logical theories used to talk about a subject domain, usually intended for computation
- **Ontology**: the philosophical discipline concerned with the study of the categories of what exists.
- **Formal Ontology**: that part of Ontology concerned with the most general principles, applicable to *any* subject matter

Why Formal Ontology?

When expressed according to principles of Formal Ontology (e.g., [Guarino 2000]) ontologies can adhere to *domain-independent* organizing principles

- Universal vs. Particular
- Continuant vs. Occurrent

Additional tools for organization & expressivity:

- higher-order properties

Adhering to formal standards should enhance ability to create/integrate ontologies from different users/domains?

Ingredients of Formal Ontology

- **Categories**
 - Universal/Particular
 - Continuant/Occurrent
 - Property/Relation (higher-order properties)
- **Organizing principles**
 - Identity
 - Dependence
 - Unity
 - Modal and temporal rigidity
 - Granularity (Bittner & Smith, 2001)
 - Constitution

Described by an ontology

Properties
("classes")

Kinds of things found in the world.
Examples: cells, species, diseases,
metabolism.

Relations

Ways that properties interact.
Examples: signals, inhibits, is-
between.

Sentences

Constraints on these properties
and relations.

Sentences

Sentences constrain the **relations**
between **properties**

“The **photosynthetic reactions** **take place** *only* in
the **thylakoid**”

“An **axon** is **part of** a *neuron cell*”

Our approach

1. Induce the intended semantics of GO
2. Make ontological decisions:
 - Universal vs. Particular
 - Continuant vs. Occurrent
 - higher-order properties
 - isa and part-of
3. Suggest future directions

What's in GO?*

Three DAGs:

- cellular component (~ 1100 terms)
- molecular function (~ 5100 terms)
- biological process (~ 5100 terms)

Two relationships:

- *part-of*
- *isa*

* July 2002 XML release

Basic ontological distinctions

- Universals vs. Particulars
- Continuants vs. Occurrents

Universal vs. Particular

Universal = "class"

Particular = "individual"

"Dog"

(*collection* of things with dog-distinguishing characteristics)

vs.

"My dog Spot"

(an *instance* that belongs in the Dog collection)

In biological ontologies we don't see many particulars...
(liver cell #124xy?)

Instantiation of Universals

Mouse

inst

Mouse579

Universals can be instantiated,
or exemplified, by objects.

Particulars cannot.

All GO terms can be instantiated, therefore are Universals

Continuant vs. Occurrent

Continuant: things that *exist in time*
(everyday objects, e.g., a protein)

Occurrent: things that *occur or unfold through time*
(*processes and events*, e.g., a chemical reaction)

GO terms: continuant/occurrent?

Cellular Component: “part of a cell” .

Examples: “double layer of molecules” ,
“cytoplasmic bridge” , “rigid... envelope”

Continuant

Molecular Function: “The action characteristic of a gene product” .

Examples: “retards or prevents” ,
“neutralizes” , “mediate”

Occurrent

Biological Process: “A phenomenon marked by changes”

Examples: “actions” , “reactions” ,
“activity” , “process” , “sequence of events”

Occurrent

Relating continuants and occurrents

Continuants (e.g. Cellular Components) are *participants* in the Occurrents (e.g. Biological Processes) in which they act, not *part-of*.

- Flagella are participants in cellular motility processes.
- Ubiquitin ligase complex is a participant in protein ubiquitylation.

GO Relations

- semantics of isa
- semantics of part-of
- non-rooted terms

Semantics of “isa”

Does “isa” indicate subsumption or instantiation?

Subsumption:

- (necessary) specialization relation between properties
- e.g. *mammal* subsumes *horse*

Instantiation:

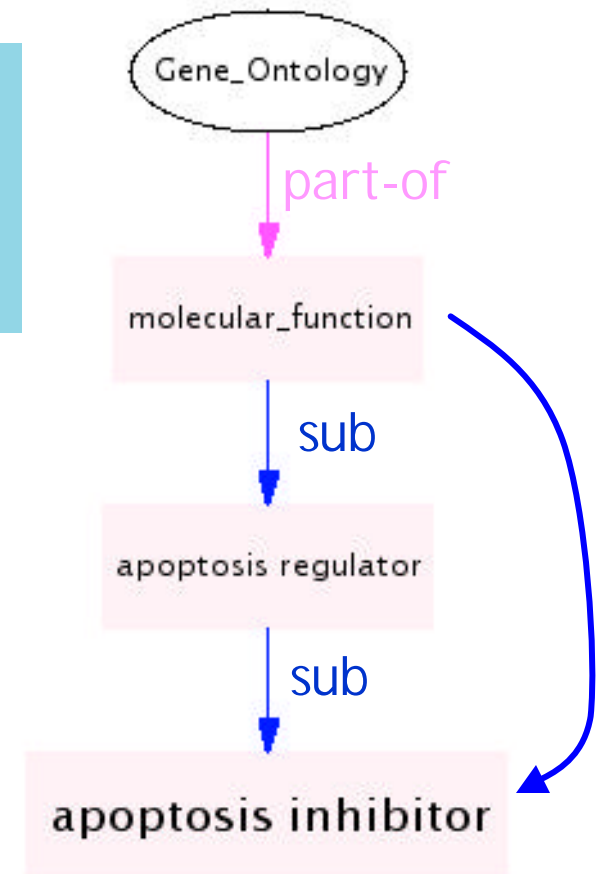
- exemplification relation between property & entity
- e.g. *Murphy* instantiates *horse*

GO's "isa" is subsumption

GO terms represent properties,
(a kind of Universal) therefore
"isa" must indicate subsumption.

This is consistent
with its usage:

Subsumption is transitive



Semantics of “part-of”

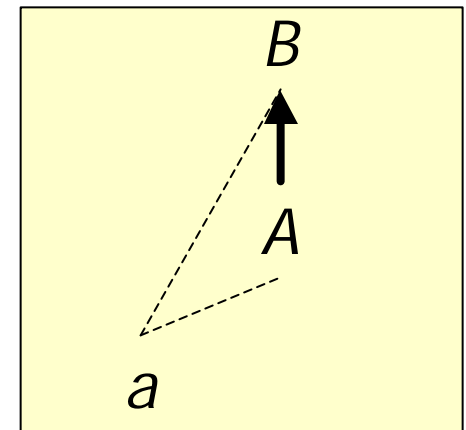
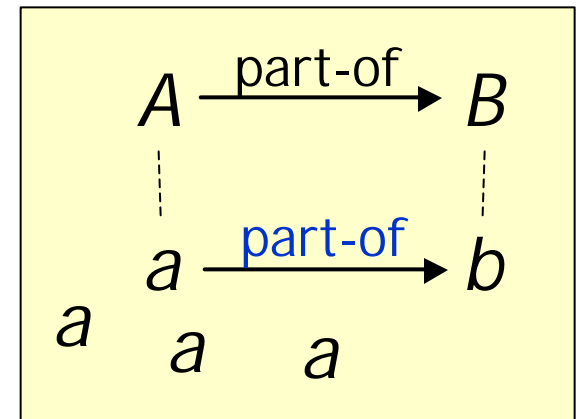
“can be a part of, not is always a part of”

$\diamond (\exists x, y) (A(x) \wedge B(y) \wedge \text{part-of}(x, y))$

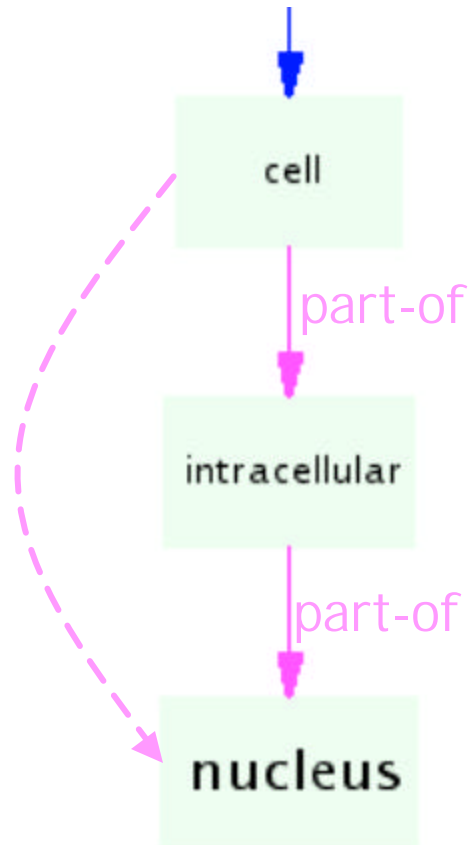
“can”

individual-level parthood relation


This is different from subsumption, where if A is subsumed by B , then every instance of A is **necessarily** an instance of B



part-of is transitive



“part-of” : variety of usages

- steps-in-a-process: synaptic transmission is part-of transmission of nerve impulse
- physical-part-of: inner membrane is part-of membrane
- functional-part-of: casein kinase II catalyst is part-of casein kinase II
-  conceptual-part-of: the term ‘molecular function’ is part-of the gene ontology

More flavors of part-of

- *membrane part-of cell*
 - "a membrane is part of (every) cell"
 - ⇒ **all** cells have-part membrane
- *flagellum part-of cell*
 - "a flagellum is part of (some) cells"
 - ⇒ **some** cells have-part flagellum
- *replication fork part-of cell*
 - "a replication fork is part of the nucleoplasm (only) during certain times of the cell cycle"
 - ⇒ **all** nucleoplasm have a RF during... and not otherwise

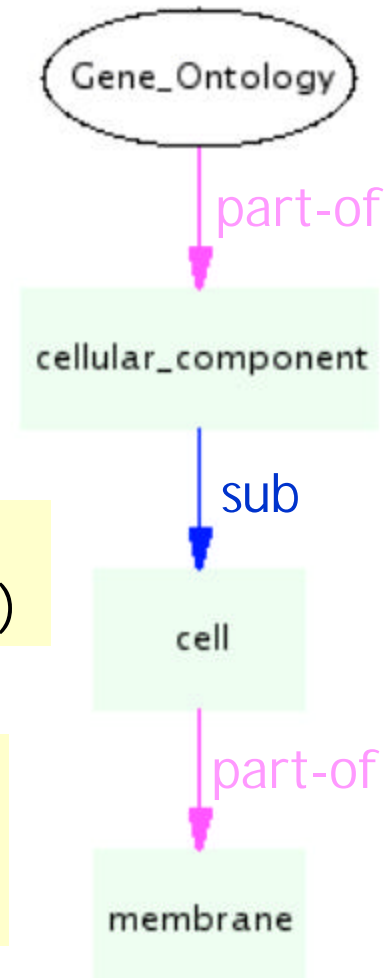
Non-rooted terms

“membrane” is part-of “cell”, but is not subsumed by anything...

What is a membrane?

(there are ~ 1200 of these non-rooted terms in the July '02 GO)

As part of the cellular-component taxonomy, can be inferred to be subsumed by cellular-component.



Higher-order properties

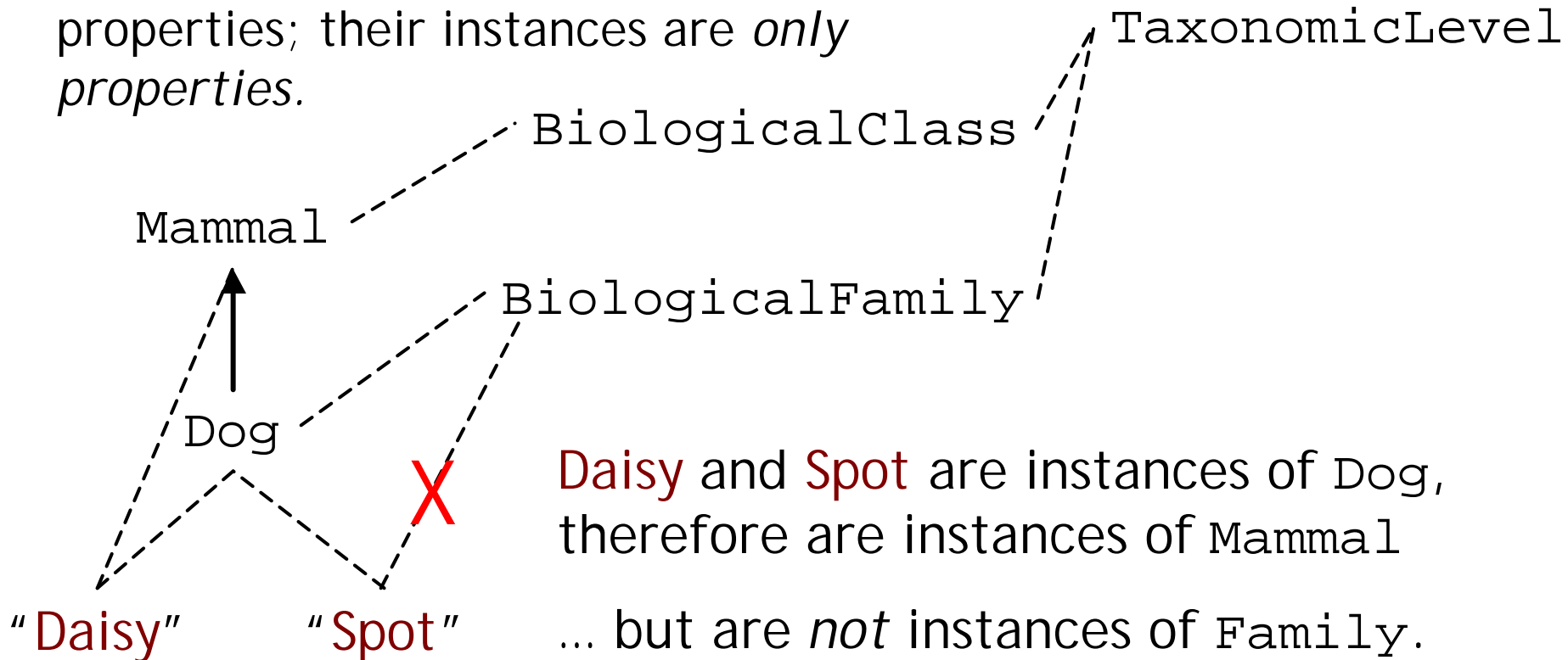
- Different kinds of properties and relations?
- Then we need to talk about them
- We need a language to do this

What are higher-order properties?

- Just as we can use **properties** and **relations** to describe individuals...
 - **Cell**(**cell#2341**)
- ...**higher-order properties and relations** describe properties and relations
 - **Property**(**Cell**)
 - **Relation**(**subsumes**)
 - **subsumes**(**Cell**, **NerveCell**)
- we need logical machinery to do this

Capturing higher-order properties

Class and Family are higher-order properties; their instances are *only properties*.



chemical kinds (e.g. ethers vs. esters)

GO: implied higher-order properties

GO uses only two relations, “part-of” and “isa” – but GO’s structure implies instantiation of higher-order properties:

cellular component kind

cellular component

qwert

yuiop

Do we know anything about *qwert* and *yuiop*?

We know they are **kinds of cellular component**... maybe *membrane*, *cell*...

Property-level relationships

- Modeling where processes take place:

processLocation

(biological-process-kind x cellular-component-kind)

processLocation (photosynthesis x thylakoid)

- What kinds of components are present in plant cells that are absent in animal cells?

More Utility for Instantiation

- Instantiation of Instance Data
- Instantiation of Meta-Information

Instantiation of Instance Data

For an application to work over instance data...need ability to say this (*particular*) is one of those (Universals).

- “*This sample of* spleen cells was obtained from Mouse #1234”
- “We observed *a case of* cell adhesion”

Instantiation: Meta-information

Meta-information on fact assertions:

“Jen says 14C is a radioisotope”

fact-source (sentence x person)

fact-source ((Radioisotope 14C) x Jen)

Formal Higher-Order Properties

Distinguishing fundamental formal differences between properties

Formal differences?

- Identity
- Rigidity
- Dependence

A quick example using **these** (there are more)...

Analysis of *Dog*

Identity: Can I distinguish my dog from others? *Yes*

Rigidity: Must Spot always be a dog? *Yes*

Dependence: Is the fact of Spot's dog-ness dependent on anything else? *No*

Analysis of *Pet*

Identity: Can I distinguish my pet from others? *Yes*

Rigidity: Must Spot always be my pet? *No*

Dependence: Is the pet-ness of Spot dependent on anything else? *Yes (me, the owner)*

Dog vs. Pet in Formal Terms

Dog - is distinguishable ("has identity")
- is unvarying ("is rigid")
- is independent

Pet - has identity
- is not rigid (*pet-ness* can come & go)
- is dependent (on having an owner)

- We call *Dog*-like things (instances of) Type
- We call *Pet*-like things (instances of) Material Role

Formal Properties in GO

Cellular Components e.g. a nucleus

- distinguishable (seq of DNA it contains?)
- rigid (once a nucleus, stays a nucleus)
- independent

Formal Properties in GO

Molecular Functions and Biological Processes

- occurments have identity (same participants, same spatiotemporal extent, same occurrence)
- occurments are rigid because when they stop “occurring” they stop existing

Cellular Components, Molecular Functions, and Biological Processes are instances of Type.

Molecular Function

Molecular function: Type or Role?

- *anti-coagulant* "**substance**... retards or prevents coagulation"
- *enzyme* "**substance**... that catalyzes"
- *structural molecule* "The **action** of a molecule that contributes to the structural integrity..."

"...the GO term describes the chemical reaction carried out by the enzyme and is not a reference to the enzyme molecule itself." [GO Consortium 2001]

Future Directions

- Checking rules of taxonomy traversal against intended semantics
- Ontologizing “annotation” relation(s)
- Exposing relations between GO taxonomies

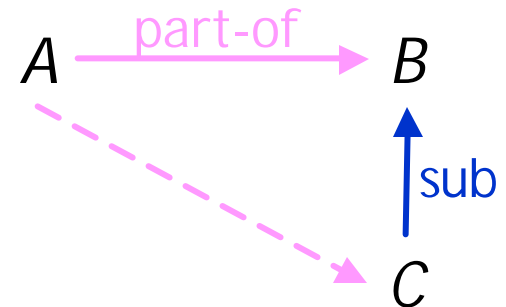
Taxonomy traversal and semantics

GO Usage guide:

If A is part of B

and C is an instance of B

is A a part of C? — YES



If Axon is part-of Cell
and Ascus is subsumed by Cell
then is Axon a part-of Ascus?

Glycosome... only part-of *Kinetoplastidae*
Viral Tegument... only part-of viral particles

Improving consistency

Some help from additional taxonomy of cell/tissue types:

"Axon is part-of Neuron (a kind of cell)"

Lots of useful concepts: *red-blood-cell, cardiac-cell*

But don't want to populate a kinds-of-cell taxonomy with loosely related concepts:

Kinetoplastida-cell, Mammalian-cell, Chordate-cell.

You want the actual zoological taxonomy itself, to be used in constraining statements:

"Glycosome is part of Cell for Kinetoplastida."

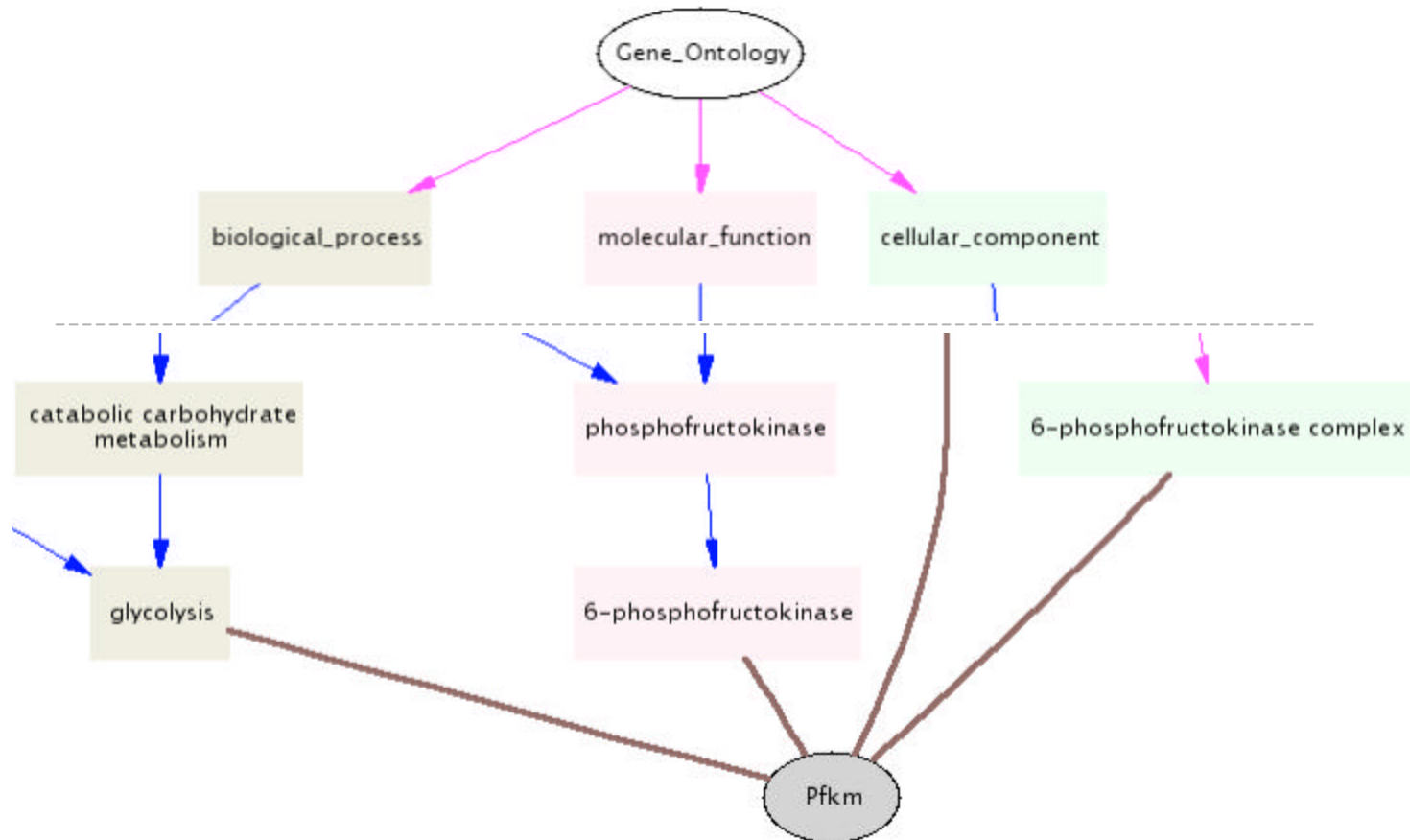
More useful taxonomies

- biological substances:
 - macromolecule subsumes polysaccharide
 - polysaccharide subsumes chitinTherefore, “chitin catabolism” is subsumed by
“macromolecule metabolism”

The more taxonomies are integrated, the more such inferences can replace manual connection-making.

Annotation

Currently annotation of a gene product is being used for “is involved in somehow”:



Specializing annotation relations

- Gene product-to-cellular component:
active-in-component
- Gene product-to-molecular function:
performs-functional-action
- Gene product-to-biological process:
participant-in-process

Connections between taxonomies

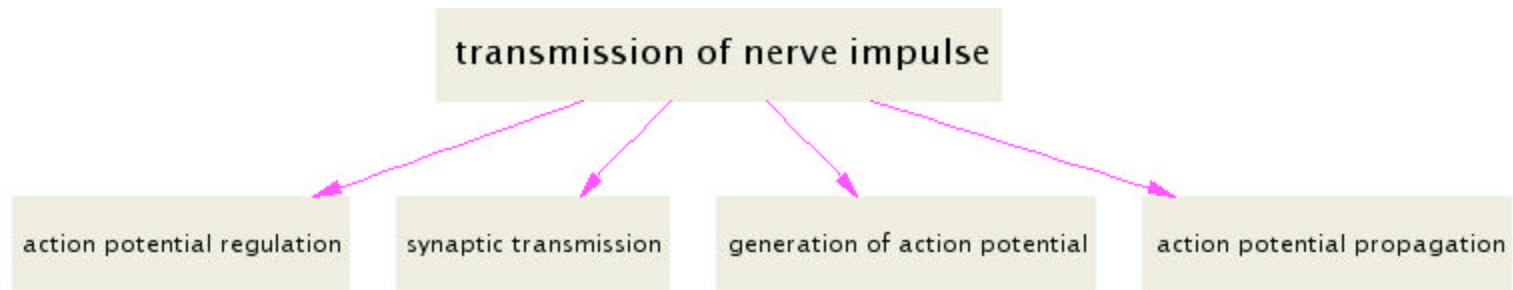
GO Usage Guide: “A biological process is a biological goal that requires more than one function.”

- Molecular function – biological process:
apoptosis inhibitor ... apoptosis inhibition
- Cellular component – biological process
flagellum ... cellular motility

And temporal constraints

“Nucleolus reappearance is associated with *telophase* of mitosis”

Ordering steps-in-a-process:



Thank you!

Olivier Bodenreider, NLM (Washington, DC)

GenNav GO browser: <http://etbsun2.nlm.nih.gov:8000/>

AmiGO browser team

Bill Andersen, Brian Peterson, Joshua Engel (Ontology Works, Inc.)



GO References

GenNav GO browser <http://etbsun2.nlm.nih.gov:8000/>

Bodenreider O. Experiences in visualizing and navigating biomedical ontologies and knowledge bases. Proceedings of the ISMB'2002 SIG meeting "Bio-ontologies" 2002:29-32.

GO General Documentation:

<http://www.geneontology.org/doc/GO.doc.html>

GO Usage Guide:

<http://www.geneontology.org/doc/GO.usage.html>

Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**: 25-29

Gene Ontology Consortium. 2001. Creating the Gene Ontology Resource: Design and Implementation. *Genome Res.* **11**: 1425-1433

GONG: [HTTP://GONG.MAN.AC.UK/](http://GONG.MAN.AC.UK/)

AmiGO: <http://www.godatabase.org/cgi-bin/go.cgi>

Formal Ontology, KR Refs

Guarino N, Welty C. 2000. Ontological Analysis of Taxonomic Relationships. In *Proceedings of ER-2000: The 19th International Conference on Conceptual Modeling*, Laender A, Story V (eds). Springer-Verlag LNCS.

ontology papers:

<http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/OntologyPapers.html>

methodology papers:

<http://www.ladseb.pd.cnr.it/infor/ontology/methodology.html>

Karp PD. 2001. Pathway Databases: A Case Study in Computational Symbolic Theories. *Science* **293**: 2040-2044

Loux MJ. 1998. *Metaphysics: a contemporary introduction*. Routledge: New York.

Schulze-Kremer S. 1998. Ontologies for Molecular Biology. In *Proceedings of the Third Pacific Symposium on Biocomputing*. AAAI Press.