GO Consortium Meeting

St. Croix, USVI

March 31 – April 2, 2006

[Next meeting: November, 2006, Hinxton or Marseille (to be determined)]

GROUP PARTICIPANT LIST

BDGP/SO	John Day-Richter, Karen Eilbeck, Suzanna Lewis, Chris Mungall, Shu ShengQiang
DictyBase	Rex Chisholm
FlyBase	Michael Ashburner, Susan Tweedie
GeneDB Pathogen	Matt Berriman
GeneDB Pombe	Val Wood
GOA	Daniel Barrell, Evelyn Camon
GOEO	Jennifer Clark, Midori Harris, Amelia Ireland, Jane Lomax
MGI	Judy Blake, Alex Diehl, Mary Dolan, Harold Drabkin, David Hill
NCBO	Barry Smith
PAMGO	Candace Collmer, Trudy Torto-Alallibo
Reactome	Lisa Matthews
RGD	Susan Bromberg
SGD / CGD	Mike Cherry, Karen Christie, Stan Dong, Stacia Engel, Eurie Hong, Marek Skrzypek
TAIR	Tanya Berardini, Sue Rhee
TIGR	Linda Hannick, Michelle Gwinn-Giglio
WB	Ranjana Kishore, Kimberly Van Auken
ZFIN	Doug Howe

Friday, March 31, 2006

Principles of Biomedical Ontology Design

Barry Smith, Department of Philosophy, University at Buffalo, National Center for Biomedical Ontology

Barry's talked was divided into five sections, the first four covered general ontological issues, while the fifth was devoted to GO, specifically.

There are many other ontologies besides GO, each with various pros and cons regarding their construction. The least sloppy ontology is the **Foundational Model of Anatomy** (FMA), characterized as follows:

Pros:

- Clear statement of scope, we know what it is: human structural anatomy
- Powerful, proper (formal) treatment of definitions - most important feature
- Single inheritance is_a hierarchy – an objective good
- From the whole organism to the biological molecule

Cons:

- Some unfortunate artifacts in the ontology deriving from its specific computer representation (Protege) – FMA was built manually, but there came a point where it was too big to be maintained manually.

[An ontology should never contain entities dictated by your programmer. Needed to include non-anatomical terms to make it work with the program; this is bad.]

Formal Aristotelian Definitions

An A = Def. a B which Cs

Parent which differentiates in this way.

This is also why single inheritance is good; there is only one place to look.

For example: cell = an anatomical structure which consists of cytoplasm, etc.

There are many circular definitions in GO. These are useless. Not bad, just useless. Every single definition should tell you where in the is_a hierarchy the term belongs. Every definition is an encapsulation; they give you the content you need in a modular form.

An ontology has to be designed both for human beings and computers. Terms used in definitions should be simpler than the term to be defined. Many of GO's existing definitions have this problem.

(FMA 90,000 terms – it would be a nice discipline for the GO to represent terms in this diagram, because this lets you know if you have terms that don't go anywhere.)

The **Gene Ontology** is characterized as follows:

Pros:

- Open source
- Cross-species
- Impressive annotation resource
- Impressive policies for maintenance
- Has recognized the need for reform

Cons:

- Poor formal architecture
- Poor support for automatic reasoning and error-checking
- No cross-ontology relations
- Not (yet) transgranular

Granularity is one very important challenge for bioinformatics. If data comes in granular packages, then ontologies must be granular. GO doesn't give a basis for reasoning in organized granularity. GO can make big strides forward without changing the content, i.e. distinguish cellular from physiological process.

GO deals with definitions in a way that is worse than useless. Logically speaking, they are total nonsense. For example, GO: hemolysis of host red blood cells is defined as:

The processes by which an organism effects hemolysis, the lytic destruction of red blood cells with the release of intracellular hemoglobin, in its host. This sort of definition is worse than circular.

David: What if there is a parent that defined hemolysis?

Barry: This would be fine.

GO is now adopting structured definitions which are built out of genus and differentiae. For example, GO: neuronal cell differentiation – differentiation by which a cell acquires features of a neuron.

Michael raised the issue that, in the past, Barry has criticized the GO because they have complex terms in the definitions. But, some chemical terms are inherently complex, right? Barry's response: Definitions should use terms that are less complex than the terms itself. You're going to have to produce a computer-friendly version of these definitions. If possible, you'll need to produce a human-friendly version, as well.

Judy pointed out that we are also dealing with community vocabularies that were constructed with different concerns. Barry's response: We need to move to a new kind of paradigm. We don't need to allow their terminology to thwart current ontology efforts.

Barry: The problem is that the UMLS accepts any group (community) developed ontologies without worrying about their quality.

Another example, the **National Cancer Institute Thesaurus** (NCIT):

Pros:

- Open source
- Broad coverage
- Some formal structure (OWL-DL)

- OWL-DL: a collection of languages used by WWW, DL maximally expressible formal logic that is still computable
- Has realized the error of its ways (a good ontology needs a more expressive language than DL.)

Cons

- Full of errors (many inherited from UMLS)
- Has verbal definitions
- Has logically incompatible definitions
- Confuses definitions with descriptions

Goals: to make use of current terminology best practices to relate relevant concepts to one another in a formal structure, i.e., to support automatic reasoning.

Of 37,261 nodes, 33,720 remain formally undefined, while about half have verbal definitions, sometimes more than one, e.g., disease progression. This assumes that people already know what is meant. Three verbal definitions are logically incompatible. For example, cancer is defined as a process and an object. Disease definitions treat them as a condition and a process. Like the GO, the definitions here get confused with descriptions.

The NCIT recognizes three classes of plants and three kinds of cells.

How best to deal with this? Barry's response: Generally, use of of_a (use of 'other') is bad practice.

There are three kinds of cells in the NCIT that do not overlap:

- Abnormal cell – top-level class
- Normal cell – is a subclass of 'microanatomy'
- Cell is a subclass of 'Other anatomic concept' (so that cells themselves are concepts)

Neither abnormal or normal cells are types of cells

Another example, the **UMLS Semantic Network**, an upper level ontology for the biomedical domain:

Pros:

- Broad coverage
- No multiple inheritance

Cons:

- Incoherent use of 'conceptual entities'
- Relationship: location_of For example, 'fungus location_of vitamin' – what does this really mean? Every instance of fungus located in some vitamin? Every instance of fungus is located in every vitamin? Should be: every instance of A is such that there is some instance of B.

General Ontological Overview

Good ontologies require a consistent use of terms, supported by logically coherent (non-circular) definitions and a coherent shared treatment of relations in equivalent human-readable and computable formats.

There are **Three Fundamental Dichotomies**:

- Continuants vs. Occurrents
- Dependents vs. Independents
- Types vs. Instances

ONTOLOGIES ARE REPRESENTATIONS OF TYPES, NOT INSTANCES. Types exist to bind different communities, and this is precisely what is missing from the UMLS where the terms are all different and produced by different groups. Types are sometimes called kinds, universals, categories, species, genera, etc.

GO has three ontologies:

- Molecules, cell components, organisms are independent continuants which have functions.
- Functions are dependent continuants that become realized through special sorts of processes we call functionings.
- Processes (occurrents) include: functionings, side-effects, stochastic processes

Continuants (aka endurants) have continuous existence in time. They can gain or lose parts, i.e. preserve their identity through change, but they exist *in toto* whenever they exist at all.

Snapshots of continuants (you, 3D)

Occurrents are never wholly there. They unfold themselves in successive phases and exist only in their phases.

Videos of occurrents (your life, 4D: 3D + time)

How should complexes be treated? There are special problems that arise in the world of molecules…..

Dependent entities require independent continuants as their bearers, e.g., there is no grin without a cat.

Independent continuants are such things as organisms, cells, molecules, and environments.

Dependent continuants are things like qualities, function, or spatial region.

All occurrents are dependent entities. They are dependent upon independents.

The basic ontology has three things:

- Independent continuant: component
- Dependent continuant: function
- Occurrent: process

Two families are occurrents: functioning vs. side-effects, stochastic processes.

Michael asked if it makes sense to have a process of temperature regulation as instantiated in a person? David's response: that would be a great way to define a biological process in the GO. Midori asked if independent continuants can represent entities when something is removed instead of added. Barry's response: Yes, but….

Some dependent continuants are realizable, such as 'expression of a gene', 'applications of a therapy', 'course of a disease', or 'execution of an algorithm.'

Functions vs. functionings. The function of your heart = to pump blood into your body.

The OBO Foundry

There is a movement in the NIH to try to avoid waste of data and to try to encourage reuse of data. There are very generously funded projects to try to serve this need and this is money down the drain. They're not proactive; they accept the data that is thrown at them. You will never make data interoperable unless you actively pursue that.

Old strategy: UMLS – rooted in faithfulness to the ways language is used by different communities. Each community created their own terminology and structure independently.

We need common, enduring ways of organizing biomedical data. We need preestablished, reference ontology upon which groups can draw. This can indeed help make data interoperable.

New strategy: OBO foundry – preemptive regimentation of language, structure, and format. We are making progress on the first two. Draft version can be found on:

http://obofoundry.org/

The goal is a step-by-step evolution. In time, the OBO foundry ontologies will be so recognizably good that groups will enforce use of specific terminology in reporting results. The OBO foundry will be initiated by a subset of ontologies who agree to a core set of principles.

OBO Foundry

- OBO-UBO
- GO
- SO
- RNA ontology
- PATO
- FuGO (Functional Genomics Investigation Ontology)
- Some others

The OBO foundry will consist of two kinds of ontologies: a reference ontology and an application ontology, like NCIT or the FuGO ontology. The reference ontologies will provide a repertoire of database schemas to use.

Criteria for inclusion: the ontologies must be open, must agree to collaborate, must have common formal language, identifier space, versioning, clearly delineated content, textual definitions, well-documented, and a plurality of independent users. Further criteria will be added over time to begin to improve quality.

The main non-trivial step forward is the adoption of methodology of shared, coherent, defined definitions which promotes quality control, guarantees automatic reasoning, and yields direct connection to temporally indexed instance data.

Types and Instances

We've now seen a distinction between types and instances: science text vs. clinical document, man vs. Michael. Instances are not represented in an ontology. We're interested in generalizations. Nevertheless, instances must still be taken into account.

Instances are divided into types which are arranged hierarchically. Once you've got the types in order, the instance becomes less important. But….they should always be in the back of your mind.

Each node in an ontology should consist of a term, an identifier, synonyms, and a definition.

An ontology is a computable representation of biological reality.

When people talk about concepts, they are expressing a fundamental confusion. There are terms in your ontology and types in reality. Nodes are connected by relationships.

We're trying to capture reality; that's why we curate the scientific literature. We want to teach the computer to reason about biological reality like we do. The computer can't read science texts, so the annotator is finding a way to teach the computer how the terms fit the instances. The computer should have up-to-date knowledge.

There are some rules on types. Don't confuse types with words, concepts, ways of getting to know types, etc. Once you have a good word for a type, you should use that term forever more.

John asked what the problem is with concepts. Barry's response: Concepts encourage inward thinking.

is_a and part_of should be used the same way in all ontologies referring to the same types and relations.

There is no type non-mammal, non-membrane, other metalworker in New Zealand.

Ontology of terms is NOT equal to a logic of terms, e.g., there are no conjunctive and disjunctive types

- anatomic structure, system, or substance
- musculoskeletal and connective tissue disorder
- rheumatism, excluding the back

Which types exist in reality is not a function of our knowledge.

Rex asked about the musculoskeletal example. What if there was a disorder that affected both musculoskeletal and connective tissue? Would you add the word 'both' to the term? One point we need to remember is that we wouldn't deny that something exists, but we need to think about how these terms are put together to avoid confusion. The solution here: 'disorder affecting both musculoskeletal and connective tissue.' The word both makes a difference. It's never wrong to be painstakingly literal.

John asked what the basis is for saying something is or is not a type. Precoordination vs. post-coordination. We can observe instances of a type, and we should examine terms in the ontology with and/or to see if they really represent a type. Should we provide an AND search union or intersection? What do people really want?

In the world of instances, there are clear boundaries, but there are also continuums, e.g., temperature, color, bowls, cups. This means that there is a necessary element of conventionality to how you divide the continuum.

Multiple Inheritance

All multiple inheritance can be unpacked into clear, separate hierarchies. There are technologies for normalizing hierarchies, and you can generate any combination of normalized hierarchy.

Using breast cancer as an example: breast cancer can have a parent term neoplasm and a parent term disease of the breast. These could be split into two ontologies: one classified as location, the other as manifestation of disease.

Most of the 'diamonds' can be cleanly unpacked into two different hierarchies, and then one can map between the two; this is the way this should be done.

Problems with multiple inheritance:

- source of errors
- encourages laziness
- serves as obstacles to integration with neighboring ontologies
- hampers use of Aristotelian method of definitions

Compositionality

The meanings of compound terms should be determined by the meanings of the simpler terms.

Common rules allow alignment with other ontologies. There are 15 such rules, which can be sent around, if wanted. If we have rules stated, then it's easier to train, avoid mistakes, and classify. But most of all, if all the ontologies use the same rules, then those ontologies become automatically aligned with each other. The Gene Ontology is useful because lots of people use it. We want lots of people to use the GO and thereby use the cell ontology, the SO, etc.

OBO Relation Ontology

The relation ontology consists of formally defined relationships. An ontology comprises terms with well-defined relationships and good definitions.

is_a

Correct definition of is_a:

> Every instance of A is an instance of B.
>
> A is_a B = def for all x, if x is an instance of A, then x instance of B.

Occurrents: the is_a definition works fine for occurrents.

Continuants: only continuants change. This means that continuants need to take time continuously into account. Every instance of A at time t is an instance of B at time t. This is being a little more careful about time. In the ontology, we're only ever going to say things about time and instances.

part_of

part_of as a relation between types is more problematic than is standardly supposed.

There are two kinds of part_of: relations between types and relations between instances: human heart part_of human and Mary's heart part_of Mary. This is incredibly important if you want to avoid mistakes.

All-some structure: all instances of A are instance-level parts of some instance of B.

This works in the untensed sense of processes. But continuants needs to take some account of time.

How to use the OBO Relation Ontology? The all-some form gives us cascading inferences because if you have all-some form, whichever A you choose as the first term, the instance of B of which it is a part will be included in some C, which will include as part also the A with which you began. The same principle applies to the other relations in the OBO Relation Ontology.

What about something that occurs only sometimes as part_of a process? Barry: That's okay.

There are three kinds of relations: between types, between instances and types, and between instances. You need to keep these three kinds of relations always in mind.

There's no constraint on single or multiple inheritance for part_of relations. You can't define everything. You have to take some terms and relations as primitive.

We now need to deal with continuity. The human body is very highly connected. This means that you have parts which have no joints between them. This means that there are physical boundaries, but that there are also flat boundaries that are not physical boundaries, but boundaries that we create by fiat. There is continuity, attachment, and adjacent. Practically the only things in the body not connected to other things in the body are blood cells.

Sample relationships:

- attached_to
- synapsed_with

There is also attachment, location, and containment. In order to understand containment, you need to understand the different kinds of holes. Containment involves relations to a hole or cavity, e.g., a hole that you dig in the ground has a flat lid, your mouth has a fiat boundary.

This is why you need to distinguish between instances and type relations.

A continuous_with B is different for instance and type.

continuous_with is not always symmetric. Every lymph node is continuous with some lymphatic vessel.

adjacent_to is also not symmetric. This is important because there is an expectation of symmetry.

transformation_of: child become adult, pre-RNA become mature RNA. Always think about the order!

derives_from: zygote derives_from ovum and sperm. Two instances become one instance.

Budding and capture are two other relations that need to be considered. A biological example of capture might be eating (not general agreement about this).

There is a suite of defined relations between types: foundational, spatial, temporal, participation. To be added are: lacks, dependent_on, quality_of, functioning_of.

Alex pointed out that lacks relates an instance and a type, e.g,. this fly lack wings. Is this explicit for this relation? Barry: Yes.

What would be an example of quality_of? Barry: temperature.

We must choose the relations that we can assert.

Comment: There are lots of instances in biology where there are multiple ways to get to a particular state. How do you address that in an ontology? We're going to need a pathway ontology at different levels of granularity. We will need much cruder ontologies for pathways that will take care of every level of granularity.

The Gene Ontology

The Gene Ontology is composed of three ontologies, or so it thinks, with three central questions that need to be addressed: location, function, and process, and three granularities: cellular, molecular, and organ + organism.

GO has cells, but it does not include terms for molecules or organisms within any of its three ontologies, except e.g., GO: xxx host which was a hack. OBO-UBO will provide top level terms, so you would choose the term host from the UBO. Host is kind of a relational term, but UBO has the facility to talk about this.

Instance – a particular entity in spatiotemporal reality.

Type – A general kind instantiated by an open-ended totality of instances which have certain qualities and propensities in common of the sort that can be documented in scientific literature.

Biological process instance – A change or complex of changes on the level of granularity of the cell or organism, mediated by one or more gene products.

Molecular function instance – The propensity of a gene product instance to perform actions, such as catalysis or binding, on the molecular level of granularity.

Molecular function execution instance, aka "functioning": a process instance on the molecular level of granularity that is the result of the action of a gene product instance.

Are the relations between functions and processes a matter of granularity? Molecular functions are defined as the building blocks of biological processes, but you do not assert part_of between ontologies. Michael pointed out that this was a very conscious design choice.

You must get relations between molecular and higher level terms correct.

What does function mean? To say that an entity has a biological function means that it's part of an organism and has a propensity to act reliably to contribute to survival. A better definition would be: function means it's part of an organism and has a disposition to act reliably in such a way as to contribute to the organism's canonical life plan.

Does this exclude the idea of abnormal? What is canonical vs. variance vs. pathological?

There are biological functions and there are molecular functions? Are all molecular functions biological functions?

The function of the heart is to pump blood. But you can have malfunctionings, side-effects, accidents, and background stochastic activity. (Examples?) These things exist on all levels of granularity. If you do not have a prototype of good function, then you do not have a function. Where you have a function, then you have a scale: heart, healthy heart, unhealthy heart.

What about cases like sickle cell, where there is positive selection, but only in some environments? Response: We will need an ontology of biological environments, niches, habitats. A 'reliable' term has built into it the idea of a certain environment. The sickle cell example is really about two different functions: oxygen carrying and malarial resistance.

Can variant be thought of as an intermediate between canonical and pathological? For example, most left lungs have only two lobes, but three lobes is a variant.

How does this relate to instances and types? There are no pathological functions. Malfunctions lead to pathology. We're going to have to recognize variant functions.

Why did we introduce variants? Because functions always come with a scale. It only makes sense to talk about functions with a prototype function. Functions are associated with certain characteristic process shapes. If it's true that there is always a prototypical end to the function, then it follows that there are no bad functions.

Hypothesis: there are no 'bad' functions. It is not the function of an oncogene to cause cancer. Oncogenes were in every case proto-oncogenes with functions of their own. They become oncogenes because of bad (non-prototypical) environments.

Comment that even using the terms oncogene and proto-oncogene involves pathology.

Response: Talking about function is part and parcel of talking about pathologies. Functions are non-pathological.

Is this true for molecules? Yes. Is it true on all levels of granularity? Does it make sense to talk about a pathological molecule? An oncogene would be an example.

Comment: What about hypersensitivity? At the cellular level it results in cell death. This is good for the organism, but is it good for the cell? Response: An immune response is a response at the biological level which includes functions that may be good for the organism, but not good for the cell. Some things may even be good for the population on the whole. But, we need to recognize that there is a huge amount of thought in population genetics and we should be very careful about how we speak about this.

Are there any exceptions to the definition of molecular function? Response: I don't believe there is an exception for molecular functions. They always make a contribution to the canonical life plan.

There is frequent discussion about the use of evidence from pathological molecules to inform what is represented in the GO.

Like the FMA, so the GO is a canonical ontology. That's why thinking about a variance could be important. What does canonical mean? What does normal mean?

The gene ontology is a canonical ontology, a computational representation of the ways in which genes normally function. You need to think carefully about what this means.

The FMA is a canonical representation, a computational representation of types and relations.

<u>Granularity</u>

There are two kinds of causality:

successive causality

Each stage in the history of a disease presupposes the earlier stages

In this case, we need to reason across time, track the order of events in times. We need pathway ontologies on every level of granularity. We especially need these things for the disease level.

simultaneous causality

Illustrated by Boyle's law. Two things happen simultaneously. It's not about events, it's about changes. (compare Boyle's law: a rise in temperature causes a simultaneous increase in pressure.

Networks are continuants. At any given time, there are networks existing in the organism at different level of granularity. Changes in one cause simultaneous changes in all the others.

Generally speaking, when you're dealing with organisms at coarser grains, you're dealing with networks at higher levels. We need ontologies of networks at the molecular and at higher levels, e.g., digestive system – simultaneous causality.

But there is a granularity gulf. The way data is collected, i.e. For most existing data sources, there is a fixed, single granularity. However, many clinical phenomena cross granularity.

The GO consists of three ontologies: MF and BP are dependent, while CC is independent. If we normalize, then we realize that we're missing the independent bearers, such as organism and complex.

Judy raised the issue that we may need separate ontologies for cell and physiological processes. Are many existing terms cross-products of these two? Yes, consider cell differentiation.

But aren't cellular processes dependent upon molecular function? Barry's response: Harping on about granularity. Do the coarser grains as much service as the finer grains. That's why we need a disease ontology.

GO has cellular components, but we've never had anatomy terms.

Normalization of granular levels is key:

molecule, molecular function, molecular process

cellular component, cellular function, cellular process

organism, organism-level biological function, organism-level biological process

What about hosts? Would we need to go up a level? Response: Host is an organism, is this not included in the GO?

Does there need to always be an example, or an instance, of each level? There are likely to be situations where the molecular function and the process are the same thing.

Annotation Methodology

Scientific curators use experimental observations reported in the literature to link gene products with gene ontology terms – actually observe instances. Is it true that they're always looking at typical instances?

The annotations yield a slowly growing map of biological reality. If done properly, this institutes a virtuous cycle, and the bigger and better the ontology becomes.

What we're doing when we're annotating: an experiment is an instance from which we infer facts about types. But, we also learn about the instances acted upon.

The instances described are typical in that sense that there's nothing interfering with them that would mess up the conclusions, i.e. this is not an artifactual experiment. Experimental records document a variety of such instances; they document the existence of real-world molecules that have the potential to execute.   Annotations will help determine what is typical and what is not.

We have a glossary now:

Instance – a particular entity in spatiotemporal reality.

Type – A general kind instantiated by an open-ended totality of instances which share certain qualities and propensities in common of the sort that can be documented in the scientific literature.

Gene product instance – Generated by expression of a DNA sequence, that plays a role.

Biological process instance – A change or complex of changes on the level of granularity of the cell or organism, mediated by one or more gene products.

Cellular component instance –

Molecular function instance – The propensity of a gene product to perform actions, such as catalysis or binding, on the molecular level of granularity.

Types are trivial once you know the instances!

Molecular function execution instance, (aka, "functionings"): a process instance on the molecular level of granularity that is the result of the action of a gene product instance.

Type – a type of molecular function execution instance (aka, a type of functioning).

Should 'activity' be dropped from Molecular Function terms?

Pros:

- functions are never activities (they are propensities, potentials)
- many functions are never realized
- current remedy is ugly, and not universally acceptable structural constituent of bone

Cons:

- much renaming work would be needed to advance clarity

As soon as you try to state carefully what annotators are doing, the activity term messes things up.

Suzi pointed out that this illustrates that additional relationships need to be added in order to go from function to process. Barry's response: One problem that needs to be addressed is that if you have a relationship between molecular function and biological process, then it will look like you are appeasing yourself.

Jane also pointed out that there is a conceptual problem in that there is a mixture of functions and activities in the GO, e.g., catalysis vs. transcription factor.

Rex explained that one of the reasons that 'activity' was added was to distinguish between the gene product and the activity, e.g., DNA polymerase vs. DNA polymerase activity. Barry's response: activities should go into a molecular activity ontology, while functions should go in a molecular function ontology. We need to create a new level of clarity in the way people think and speak.

Can we replace activity with function, since they're all functions when they're under that branch of the tree? If we searched and replaced with function, does this work?

Sometimes the user community mistakes gene products for function. If a potential user of the GO has a protein which has a molecular function characterized as alcohol dehydrogenase, but is not known as such, isn't this confusing to people?

One possible solution would be to have names of molecules classified according to function, delete the word 'activity' and then make sure there are no strange terms. Also, do we want a molecular functioning ontology? Does this parallel a molecular function ontology?

Judy: We cannot change the nomenclature of gene/gene products, this is fixed.

Alex: I am worried about removing 'activity' since it is commonly used.

John: This is just lazy grammar, 'activity' is what it does, not what is has, can be changed.

David: This is more common in the biochemical world, than anywhere else.

ACTION ITEM #1: Very seriously consider removing the word 'activity' from the molecular function terms and consider renaming the molecular function ontology.

Principles for Building Biomedical Ontologies : A GO Perspective

David Hill, MGI

David's talk was centered around the idea of taking the principles that Barry talked about and discussing how we've applied them in GO. Most of the big arguments that GO has had addresses issues that Barry raised.

The principles for building a good biomedical ontology are as follows:

- Univocity – word means the same thing
- Positivity – not a membrane is not a good term
- Objectivity
- Single Inheritance
- Definitions – formal, written definitions
- Basis in Reality
- Types vs. Instances
- Ontology Alignment

The Challenge of Univocity

One of the first problems that GO dealt with is that different people in different communities use words differently. So, how does a computer know what people are talking about? In GO, we dealt with this by creating primary terms and many characterized synonyms.

Another challenge is illustrated by the question, what does a bud mean? The answer is different for vertebrates, plants, and yeasts. This is now the inverse of what we had before: people use the same words to describe different things. So how did GO deal with that? The computer doesn't know the difference, so that was when we decided to create the sensu designations for terms. These describe the term in the case of metazoans, fungi, etc., as the biologists in the field think about it.

Synonyms are incredibly important, as we had to have terms that biologists will search on. One question that arose: how to represent the function of something like a tRNA? It's term: triplet codonamino acid adaptor activity. But no biologist is ever going to search on this term. That's why we created tRNA as a synonym.

What would happen if we removed activity from this term? Not all tRNAs could then be annotated to it.

The Challenge of Positivity

Sometimes absence is a distinction in the biologist's mind. Some organelles have membranes around them, but some, like centrioles, do not. So the GO came up with two types of organelles: membrane and non-membrane bound. Note the logical difference here between 'non-membranous bound organelle' and 'not a membrane bound organelle.' We don't want the latter, as it signifies everything other than what we're talking about. Alex: Biologists would understand 'non-membrane bound organelle'. Rex: A better term is needed.

The Challenge of Objectivity

For some gene products, we have no idea what they do.

Database users want to know if we don't know anything (exhaustiveness with respect to knowledge) so the unknown terms were created. Annotating to these terms means that an annotator looked at all the literature, and we don't know what the function is. But consider 'G-protein coupled receptor, unknown ligand' – there is no difference between this term and the parent term 'G-protein coupled receptor.' In this case, instead of using a term that incorporates 'unknown,' we should annotate to the parent term.

How should we annotate genes that we assume have *some* function, but we don't know what that is? This will be especially critical for annotating the reference genomes.

We will want to annotate to molecular function and use the ND evidence code (see discussion on Sunday).

Single Inheritance

This is something that is going to be really hard for the GO, due to incompleteness in the graph, as well as other reasons. GO has many is_a diamonds. Technically the diamonds are correct, but they could be eliminated. This would involve choosing one term as the primary parent, while the other would be derived.

What do these pairs have in common?

- Locomotory behavior vs. larval behavior
- GTPase regulator activity vs. enzyme activator activity
- Non-membrane bound organelle vs. intracellular organelle

All of these terms differentiate from the parent by a different factor – type of behavior vs. what is behaving. One side is a type of behavior, the other is location.

Conceptually, we can insert an intermediate groups term, a descriptive behavior term, behavior of a thing. This is no better but allows you to figure out why you have this diamond.

Chris: There are more complex diamonds in the GO, where the result is not a cross-product.

Alex asked why you couldn't reason with this type of structure? As a biologist, it seems reasonable.

Could we just annotate to two different terms: larval behavior and locomotory behavior? Does this get users to the same information? Would doing this make it more difficult for annotators to always know all the right terms to select?

Judy pointed out that there is a danger here in having dependency on annotation to represent a term that maybe should be in the ontology. There is also some feeling that annotating to each term separately does not capture all of the same information.

Having Good GO Definitions

There are two types of definitions:

- definition written by a biologist (the one we all see when it's written out)
- definition given by where the term is placed in the graph

We all strive to make the former definitions necessary and sufficient. The latter are also necessary, but they are not often sufficient. The set of necessary conditions is determined by the graph, and the graph is only considered a partial definition dependent upon placement (selecting the appropriate parents) and relationships.

Requests for new terms, thus, need to consider all relationships. For example, a proteasome complex, which may be part_of the cytosol, part_of the endoplasmic reticulum and part_of the nucleus.

Does this mean that logically is has to be a part_of each place all the time? Chris' response: It's valid to not be a part of these parents. But, if it's present in the ER, cytosol, and nucleus and moves between these places, is that still valid? No. It could be wrong if it's not part of each thing. For example, red blood cells wouldn't have the nuclear proteasome complex.

If the same complex is found in two places (no difference in the composition of the complex), are we double annotating if we annotate to both? The feeling is that curators should annotate to both types.

At present, the definition means that if something is a cytosolic proteasome complex, then it is, at some point, part_of the cytosol. Barry stated that is is fine for an is_a relationship, but part_of always means part_of. Chris: We're not making that strong a statement with our part_of complex. Barry: Are you talking about the part_of relationship in the relationship ontology of the GO? Which definition of part_of is used? How do you account for motion? Earlier when this issue was discussed, we said that it had to be part_of sometimes. Is this true for all ontologies? What about the cytokinesis example? Where should we put it—part of mitosis, part of cell cycle – we know that this is not true for all situations all of the time.

The Importance of Relationships

We annotate to regulatory subunits of catalytic activities. But, we didn't know how to express the regulation of that catalytic activity, so we need a new relationship for regulates. This new relationship type would address the idea that this gene product impinges upon, but doesn't actually have the catalytic activity.

There was some discussion of the extent to which the regulates relationship should be used. As an example, suppose that you have a wise father who regulates the behavior of a child who, in turn, regulates the behavior of her pet dog. Does the father also regulate the behavior of the dog via the daughter? What about a kinase regulating a transcription factor that regulates some other activity? Or MAPK signaling pathways – does a MAPKKK regulate a MAPK? There are probably many examples in biochemistry where A regulates B which regulates C, but we only know about A and C. The real issue here, then, is what is the definition of regulation: directly regulates or indirectly regulates?

GO Textual Definitions: we strive to created similarly structured, normalized definitions, e.g., glial cell differentiation.

For process terms, David would like to propose that every process is defined by a beginning and an end. Does this work for all situations? Can we make extremely explicit process definitions? Would this lead to a lot of new child terms? New sensu terms?

Basis in Reality

GO is designed by a consortium and represents a consensus. Large-scale developments are the result of compromise and annotators are constantly looking for examples in the literature.

If we align different ontologies, such as the cell type or chemical ontologies, with GO, then that permits generation of consistent and complete definitions – formal definitions with necessary and sufficient conditions in both human readable and computer readable forms.

Types vs. Instances

What are the instances we are dealing with in our work as ontology builders and scientific curators?

The first question to answer is: What knowledge are we trying to capture? We're interested in understanding how genes contribute to the biology of an organism.

What is meant by gene products? We have gene products types and gene product instances, e.g., shh in this cell in this mouse. The instance is the actual molecule that can be physically isolated and takes up space.

What do the experiments do? Experiments are designed to study the properties of gene products.

How do we represent that accumulated knowledge? We connect what the wet bench biologists see to our understanding of biology in the ontology.

What are the instances? They're in the labs. We use what experimenters report about those instances.

How do we connect instances with knowledge representation in the GO? Some examples:

- A molecular function annotation using IDA Paper reports assaying the activity of retinoid dehydrogenase

Annotation made: retinal dehydrogenase activity

What are the instances? There are gene products instances, molecular function instances revealed by that assay, and instances of molecular function associated with an instance of the enzymes.

Conclusion: If I have this molecule, it has the potential to have this function.

- Using IMP

Gucy2c activity goes away in mutant cells – observation was that when they had this molecule, they had cGMP, when it was gone, they didn't have it.

- Using IGI

Don't have enzymatic activity when animals are doubly mutant for subunits of the enzyme.

Single mutants have activity intermediate between wild-type and double mutants.

The instances are the gene products, hexA and hexB, and the molecular function instances. We say that these molecules have the potential to execute the activity.

Sue commented that the assertion being made in each of these cases is slightly different. For IDA, the annotation is clear, but the latter cases *suggest* that the gene products are involved, but is their role direct or regulatory? Should we use the contributes_to qualifier in these cases?

There is agreement that the evidence code does tell you something about the quality of the inference and what types of caveats a user should expect. But, can you compute that aspect of the evidence code?

Another example is that of molecular function terms, such as protein binding, with the IPI evidence code. What are the instances? Flag-tagged molecules, molecules to which an antibody exists, execution of binding, potential to execute binding, potential to execute specific binding. Some discussion about whether protein binding is a good function term.

Conclusions?

Process annotation using IMP:

- Observation is that in the presence of Shh you get a specific process of heart development. Instances are functional and non-functional molecules of Shh, development of a mouse heart, functioning of a Shh molecule. This process is the result of gene products executing their functions.

Process annotation using IPI:

- IPI is often used to annotate to MF: protein binding
- In this example, a catenin-interacting protein was annotated to BP: Wnt receptor signaling pathway.
- There was no instance of a function during Wnt receptor signaling, so where did the missing information come from?
- An inference was made based upon previous annotations.
- This represents a chain of inferences. How far do we want to take it?

Much discussion ensued on the chain of inferences. Some felt that IPI for protein binding would be the only annotation that could be made. Others felt that the annotation depended on the context of the experiment and what the authors state. If the authors claim that this [participation in a process] is true based upon their IPI experiment, and the experiment is in a peer-reviewed journal then it's okay for an annotator to make that annotation. Otherwise, the correct evidence code would be IC.

In the signaling literature, it is not uncommon to find that authors do co-IP experiments to show that a protein of interest in involved in a particular signaling cascade. But how far do we take the inference? When do we take the inference? Who makes the assertion? Does that matter? Does it matter what the community assumes to be true?

If, for example, protein A bind protein B and protein B is involved in two completely different processes, would you annotate to both? No. If the binding experiment is the only one performed in the paper is that enough? If not, do the other experiments influence the process annotation? If we comprehensively annotated\ genes, can we make the same conclusions as wet bench biologists?

Can we make a rule about if and when we will use IPI for biological process annotations?

A check of the database indicates that almost all of the core databases have used IPI for process annotations.

The conclusion was that we should think about this more and come up with examples of when we would or would not make the inference to inform our decision about making a rule for IPI and process annotations.

This would be an agenda item for the coming GO annotation camp.

Saturday, April 1, 2006

Discussion of Goals in the New Grant Proposal

Aim 1 – Ontology Development

Suzi Lewis, BDGP and David Hill, MGI

What is our aim?

Suzi Lewis, BDGP

As NIH program director Peter Good has stated, the heart of the project is still the ontology; that's the resource that people still use. Thus, GO will maintain comprehensive, logically rigorous, biologically accurate ontologies paying close attention to both content and relationships in these ontologies.

Content Development

David Hill, MGI

Comprehensive annotation can drive ontology development. With this in mind, David presented an example of ontology development using the process of blood pressure regulation, an important and well-funded area of research.

In mid-November, when David began focusing on curation of genes having to do with blood pressure regulation, there were three relevant terms in the GO: regulation of blood pressure, and positive and negative regulation of blood pressure.

To further develop this part of the ontology, he first got a textbook on medical physiology and proposed a basic structure in a SourceForge item that included 43 new terms. Then, the refinement process began…

A textbook was not sufficient for describing all aspects of blood pressure regulation and from reading published papers, he was able to add more terms, as well as synonyms for many terms. Work continued; he read more and more papers and added more and more terms. One interesting aspect that came out of this was that not all new terms added had to do with blood pressure regulation, because the genes involved in regulating blood pressure affect other processes as well, such as water consumption and kidney function. Thus, in the process, the blood pressure node improved and other nodes improved, as well since you're now starting to understand genes as you expand nodes in the graph. In the end, most new terms added were new leaf nodes, with relatively few changes to the actual structure of the graph.

Summary of ontology development steps:

1 Consult a textbook.

2 Identify papers.

3 Read papers.

4 Enter SourceForge items.

5 Modify GO.

6 Curate papers.

Good news: 14 new genes initially annotated to blood pressure regulation. Now, 23 genes annotated, with five genes comprehensively annotated. Along the way, other genes got annotated and not all annotations had to do with blood pressure, which is good for clinicians.

Bad news: There is still an outstanding SourceForge item about the terms, as Amelia found inconsistencies in some terms and definitions. Although these issues are relatively minor, they're still there and need to be addressed and fixed. When proposing new terms, it is important to think as an ontology developer, not as an annotator. Since the issues raised did not get in the way of curation, it was too easy to switch over to the role of an annotator and just go ahead and curate the papers, without tidying up the ontology issues.

How can we prevent this type of short-circuiting? One way would be to have some assignment of responsibility for ontology development. The ontology developers could point out logical issues, while expert curators deal with major logical issues. Minor issues could then be addressed as concrete proposals that could perhaps be presented in a way where the curator could quickly accept or reject the proposal. Any final decisions would then be based on whether the ontology really represents the biology.

Some discussion on this: Does it become the responsibility of a curator to assume the role of an ontology developer when they recognize that new terms are needed? To some extent yes, as we need to keep in mind that ontology development is key to the GO and dependent upon feedback. We have become a more cohesive group and are really at a point where ontology development and annotation intersect, leading to shifting responsibilities. This may mean that we need some changes in how we interact and do things and that we may need to address any problems that exist in how we manage ontology development. We should be creative in how we think about this; maybe the SourceForge method needs improvement, maybe there are other options besides SourceForge?

Harold: When you have the interest and the background, that's when the interest to develop the ontology kicks in.

Lisa: More collaboration between groups is good, we are doing pathways related to human disease.

Rex: We should identify areas of interest or holes in the ontology.

Suzi: We need more structure in how this is done.

There may be some value in alerting people that a particular area of the ontology needs attention. That way, people with expertise in the field can provide insight, since it is not always clear to an annotator when some terms might be missing, especially if the area under development is not congruent with their background and/or interests. It may also be necessary to have someone at each database be responsible for ontology development. Identifying such people would give the whole process more structure. In addition, the editorial office could provide a timeline of projects outlining different areas of development. Annotators could subscribe to specific mailing lists about these issues.

Bottom line: there is shared responsibility for ontology development!!

Editorial office had a poster about time lines of ontology development projects.

(PRE)ACTION ITEM #2: Need to work out the balance of power/responsibility between the GO office and annotator/ontology developers to complete SourceForge items.

David to the GO editorial office: Do you want to be the enforcer?

Jane: We do hound people to finish an item.

Judy: We need a systematic way to bring closure to an item.

Aim 2 – Reference Genomes

Rex Chisholm, dictyBase and Judy Blake, MGI

The second aim of the grant built on the idea of ontology development. In light of the criticisms of the previous proposal, it was important to pay attention to the biology and pay strong attention to how the ontology was used. What is valuable to biologists about GO?

There is an enormous effort to sequence more genomes. There are 100 billion letters of sequence in GenBank representing 160,000 different organisms. The next step, however, is to use that information to understand these organisms and GO has an important role here because some organisms may only have a handful of experiments. In light of this, GO would like to have a set of well defined, well characterized reference genomes that could be used for electronic annotation of new organisms.

What are the characteristics of a reference genome?

1 Needs to have a sequenced genome.

2 Needs to have an active and robust MOD supporting it.

3 Needs to have broad-scale functional genomics projects.

4 Needs to have an adequate research community adding new information.

5 Needs to have a sufficient literature base.

6 The reference genome list needs to be distributed across the tree of life in such a way as to ensure a range of organisms are represented.

So, a list of nine reference genomes was chosen, starting with the two extremes: E. coli and humans. Humans have to be there since that's where interest lies and much funding derives. E. coli is there since it is the largest component of our biomass and there is a proposal for an E. coli database that is planning to use GO in their annotation. Other genomes identified include mouse, fly, worm, S. cerevisiae, Arabidopsis, and Dictyostelium. Those genomes not identified as reference genomes are not considered less significant, but the reference genomes will be the focus of a committed, coordinated annotation effort.

What did we commit to? Curation of the reference genomes will provide broad and deep annotation, with each of the genomes being consulted and agreeing to establish foci of curation. Curation of genes that are relevant to human disease will be emphasized and at a minimum their curation will include: 1) looking at a provided list of genes, 2) using that list to prioritize annotations to determine if you can add information about the function of those genes in your organism, and 3) looking at those genes and also at the human annotations to see if the human annotations can be improved based upon your annotations.

As the reference genomes are annotated, it will be important to provide data for metrics.

It is unlikely that we will have complete annotation for those nine reference genomes, but we want to be able to measure progress. This may require some slight reorganization of how information about GO curation is captured.

We want to establish a GO annotation team for this part of the project. For most of the reference genomes, there already are GO-supported, "embedded" curators. This fact is very relevant to the previous discussion, as these individuals will play a coordination, as well as annotation, role for their respective genomes. These individuals will be funded by GO and will answer more to the GO, as we will need to expand leverage to complete these annotations.

Another role that these individuals will play is in providing outreach and training to other curation efforts. They will agree to participate in annotation consistency exercises to see if curators within their database have some level of consistency in how to think about curating GO. Part of the annotation consistency efforts may include everyone curating a human paper and comparing annotations. They will also participate in annotation camps and workshops, for which there will be a more regular process. There will be regular communication in the form of biweekly (fortnightly) phone conferences. And lastly, there will be a bidirectional process where each database has a responsibility of reviewing the GOA annotations for their organisms, ensuring feedback between GOA and the reference MODs.

This will require a lot of communication with the EBI group about human gene curation. EBI already accepts some human annotation from MGI and other groups and would be happy to take additional annotations. Using EBI's curation tools, annotators can directly annotate to the GOA database. However, there is some concern about other MODs performing human gene curation. How will the outside world view this? We will need to make concrete and tangible progress on this or it will look bad. It might be best to focus on particular genes and show that we've made progress. All MODs have high priorities within their own organisms and so we will need to be realistic and methodical in our approach. The key role of the MODs will be to still do what the MODs do. That's the easy part of this! We're promising to do what we're already doing.

GOA/EBI group annotation tool is online and can be made available to outside groups, e.g. AgBase, _____, Roslin group for chicken

Sue expressed concerns that the annotation of human could be problematic, if we say we will do human, and then don't do it, we will look bad.

MGI already has a list of human disease genes, ~4000, people are already contributing to InParanoid and using it to compare MOD genes to the human list, we can then also track MOD progress on this list of genes.

This new effort will need a lot of coordination and we will need to know how to represent progress. We will generate a list of human disease genes, which will number several thousand. This is not an impossible list to generate. In addition, we can provide a list of InParanoid orthologs to help groups identify the relevant genes to annotate. This leads well into the discussion on co-annotation of the mouse/rat/human genomes…..

Cooperative Annotation - Human/Mouse/Rat Annotation

Judy Blake, MGI

There is a heavy call from the GO user community for human annotation, and it will be beneficial to look at how the mouse/rat/human groups can work collaboratively since in many cases papers report experiments on genes from all three species. Mary Dolan has already done some work on the annotation consistency between mouse and human annotations and will also be adding chicken, fly, and zebrafish to this analysis (all predicated on knowing the orthology). One consideration, though, is that there are different intersections of information based upon the experimental focus on an organism. For example, rat gene products are often studied in the context of neurophysiology.

At present, these groups do things independently and likely have similar but different priority in curation sets. How can these groups exchange information and help each other out? One way would be to create a shared annotation group, which may be a core group of people funded by GO, of human, mouse, and rat annotators that focus on gene products implicated in human disease. This would involve shared curation of the literature, collective work on the same dataset, and an update of each group's editorial tools to accommodate multiple species. These groups would also need to develop shared quality control protocols that would ensure high quality curation. Currently at MGI, they annotate rat gene products as they come to them, and these annotation lines are sent straight to GOA and posted on the MGI ftp site.

Another proposal that relates to David's idea of depth of annotation: when these groups take on a certain area for curation, they also take on coordinated ontology development. The groups would begin by identifying and reading reviews and then using the reviews to improve data sets, triage the primary literature, and improve the ontologies. This would involve identifying all of the organisms and gene products described, and they are working on various computational ways to identify that information.

Overall, these efforts would help the groups to simultaneously curate the same, focused set of genes and allow for development of a working group amongst these three genomes. Many mouse papers have data from rats and humans and it doesn't make sense to have three different MODs look at the same paper.

Curation of rat genes begins with abstracts and then focuses on disease genes, particularly nervous system disease genes. This is fine, because concentrating on disease collectively gets the synergy. It is agreed that it doesn't make sense for people to look at the same genes in mouse and human, but with this new idea, people would be looking at the same group of genes involved in a particular area and this would help by reflecting how people in each field currently think about this topic.

How would the curation be split up? Ideally, by references and by organism. This effort would involve literature-based, shared discussions that could start with a recent review and then choose a focus for annotation.

This work might also lead to productive interactions with Reactome. If out of this work came a list of genes that were well curated in mouse/rat/human, then these genes could be bumped to the top of Reactome's curation priority list and help give them ideas about what pathways to coordinate.

The exact tools and processes for doing this are not all in place yet, but will come via discussion as work progresses and we see growth in each of these areas.

For GOA, this will be different from their normal curation strategy, as they currently fully curate each gene product. However, the hope is that this strategy extends beyond mouse/rat/human. For example, FlyBase could look at their homologs, too. In addition, we could put in place responsibilities so that when you curate a paper, you curate the whole paper and all the information in that paper. Sharing this gene list amongst all the reference genomes will provide the broadest representation of what these genes do in biology.

Reference Genomes

Rex Chisholm, dictyBase

How do we think of metrics for both breadth and depth?

One important point in annotating the reference genomes is to have metrics, coordinated by and agreed upon by the consortium, to assess the breadth and depth of annotation. How do we develop these metrics? There was some discussion of this on the email list while this proposal was being written. In some ways the answer seems obvious, but there are a lot of hidden complexities and different groups think about this in different ways.

Nevertheless, we should all collect the same set of numbers.

What is meant by breadth? A group would have broad annotation when they've annotated all the genes in the genome. There is, however, different kinds of information from different organisms. For example, rat is good for physiology, mouse for disease models, yeast for signaling, etc., and biologists tend to think about a process based upon what they know from a variety of organisms.

For a given organism, we will need to track the number of genes that actually have experimental annotations. In addition, we'll want to know what genes' functions are known mainly by ISS and what percentage of the genome is annotated with ISS exclusively. We'll also want to know what is annotated based on IEA (sometimes this is the only type of annotation known). Tracking these numbers for each reference genome and collectively we allow a way to monitor progress towards breadth. The goal is to do the entire job – a huge task – and we need to be able to show progress along the way.

It would be good to compare the snapshot taken at the time the grant was being written with one taken this summer. It would also be good to add, in the annotation file, the number of annotatable genes and proteins, since this will be critical for assessing metrics. It is not always trivial, however, to know what percentage of genes you can annotate. The cerevisiae genome, for example, has 20-25% of genes with no annotatable information and it has a compact genome with a long history of experimental study! It may still take a future publication to annotate these genes.

In addition to measuring how many annotatable genes and proteins exist for each organism, it might also be helpful to record how much ontology space is covered for each organism. This bears on the second point, which is how to measure depth. If you've annotated every single paper ever published, then you've achieved 100% depth. But that's not realistic, as many papers aren't gene-centric and not all papers have curatable information. So, what should the denominator be here? Is there any way to capture the number of *relevant* papers? Would working with natural language processing groups be useful for gathering this kind of information?

Another measure of depth might include an assessment of how far down any given branch the GO is used, ie how close do you get to the leaves? Shu (FlyBase) has written an algorithm for assessing this. One off-shoot (so to speak) of this would be that lots of annotations to a leaf node would highlight potential new areas for ontology development.

Depth: can be measured as:

% of papers used for curation

average # papers per curated gene

can we capture the # of relevant papers since we know it is a subset of all papers. How?

Problems:

Even if a paper mentions rat, it may not mention any genes

Even if a paper mentions a gene, it may not give you information that is curatable for GO

Some of these types of projects will be very useful to the natural language processing groups; Mike, Judy already have collaborations with such groups.

ACTION ITEM #3: Begin to coordinate processes for reference genomes to start setting priorities and tracking progress. Acquire and distribute lists of genes for curation focus and set-up fortnightly discussions. [Point Person: Rex Chisholm]

We need to develop a list of “disease genes” and process to monitor progress.

We need to refocus and integrate roles of GO funded annotators.

Aim 3 – Outreach

Michael Ashburner, FlyBase and Suzi Lewis, BDGP

We promised a tool that would provide better ways of annotating through the ortholog sets. This might be software that shows a phylogenetic tree and protein alignments and allows curators to click on any gene in the tree to see the gene's annotations. It might also allow for dragging terms from one organism to the other, resulting in ISS annotations. To accomplish this, we will need ortholog sets (from other groups) and protein sets from each of the reference genomes.

Many groups are calculating orthologs – each does this differently and each uses a different data format. It would be really nice to have a common dataset so that we can compare these different methods on one set. It would also be nice to have a common output file. We'd like to be able to add in any group/organism that can provide a protein set (Matt has 40).

Michelle mentioned differences in current annotation of gene models in eukaryotic organisms.

Common Coding Sequence ?not sure of name? – collaboration between 3 groups annotating human.

ACTION ITEM #4: Any other ideas for shared curation software, please forward to Chris Mungall.

To avoid circular annotations, curators would only want to drag and drop experimentally derived annotations, which could be color-coded based upon evidence code. Where would the annotations go afterwards? The tool could generate an annotation file, and GOA already has a tool to do this. One thing to keep in mind, though, is that if anything changes about the *original* annotation, you would want to check the related annotation for continued accuracy.

WORKING GROUP: We should establish a working group for this software so that we can leverage what other groups have done and figure out, technically, how to pull this all together.

Two critical need for this type of software are: 1) reliable, agreed upon protein sets for establishing orthologs, and 2) a common output from the various ortholog groups.

There is pressure to keep the number of genomes annotated up for taxonomic spread, but also to keep it down because of the enormity of work involved.

The GO will support annotation across all organisms, and has been doing so for some time now. TIGR became an official member in GO in 2001, and TAIR joined us before that. As of the time the grant was written, there were 1,867 genome projects, of which 339 were published and in some sense complete. The former number includes over 900 prokaryotic genomes, over 500 eukaryotic genomes, and 26 meta-genomes. A number of additional mammalian genomes have just been funded. Over 100 pathogens are being sequenced at Sanger. Thus, there will be an increasing number of genomes coming in the future.

For the great majority of these genomes, however, there is either no or relatively poor funding for annotation efforts and associated databases. This provides quite a major problem for the GO which has traditionally interacted with database groups and not with individual sequencing projects.

It is in the interests of science and the GO Consortium to reach out to these groups to encourage and support GO annotations across as broad a phylogenetic breadth as possible. To this end, it is important for all of us to assume an ambassadorial role at talks and meetings and this role has been formalized under Jen Clark at the editorial office.

But, we need to take a number of additional steps to go beyond outreach and beyond initial contacts. We need to provide a basic toolkit of training materials for new groups to annotate their organisms.

This would include reference annotations from reference genomes that could be used for transferring annotations. This is one of the reasons for having a broad phylogenetic spread of reference genomes.

In the past, we have also been asked to give advice about tools. People want to know what the best tools are for annotating a genome, but we have been reluctant to recommend one groups' tool over that of another. What we could do, however, is provide the best SOPs for using the existing tools.

There are many tools that use GO for microarray analysis, and when people see the list on the web page, they often don't know where to start. When giving tutorials, we currently use one or two of these tools, but don't necessarily work that closely with developers to give them feedback on the software. It would be very good to begin working with these people behind the scenes to help improve the tools.

It would also be good to have a wiki for each of the tools and to partner with groups that have a vested interest in further developing and improving these tools. Mike suggested that we could partner with someone, e.g. MGED, to work on this. Sue suggested that we could be more proactive on collaborating with these various software developers – they are often looking for nodes that are overrepresented, which of course will be affected by nodes which are underannotated. GO slims are often too broad for some uses, so we may need to help people develop better slims to really do what they need. We also need to encourage developers to incorporate evidence codes, so they don't get ignored by so many people. Mike seconded Sue's comment about some developers ignoring evidence codes entirely, and they are using a gene-association file to predict other annotations, and have no idea whether they are using experimental or IEA annotations.

In summary, we know that some tools don't use the evidence codes, some ignore the WITH column and thus, the NOT qualifier, and that tools using the GO slims may be using GO annotations that are too high in the tree to be really useful. Accordingly, each group should think about what is the right depth of the ontology for their organism.

Has anyone performed an independent analysis of GO tools? Sorin Draghici at Wayne State University has done this. (See his publication in Bioinformatics, Khatri and Draghici 21 (18): 3587.) End Discussion.

Another form of curation outreach can be provided in the mappings of external classification systems to GO, such as ec2go or interpro2go. When using these, though, it is important to consider how extensively they are maintained, since some are updated regularly and some probably haven't been updated in three or four years. Information about frequency of updates should be transparent in the mappings file. Also, we should always be on the lookout for new classification systems to map to GO, such as the KEGG orthology.

Most importantly, however, is that we're going to provide training. We've already had two annotation camps, and are planning a third for this coming July in Palo Alto. We would also like to have one annotation workshop per year, alternating on which side of the Atlantic it's held. We should especially reach out to Asian groups, particularly some groups in Japan who've already sent curators to the 2005 Annotation Camp and are very interested in learning how to do manual annotation .

WORKING GROUP: There is going to be an outreach working group and they will need to come up with a plan for how to encourage and support annotations from a broad spectrum of genome projects. Part of this group's focus should be on writing SOPs for GO curation and working very closely with new groups to ensure regular submission of gene association files. This working group should include the "embedded" GO curators at each of the MODs.

Another issue related to this has to do with freshness, or currency, of annotations, since we know that annotations can go stale for a number of reasons. For reference genomes and well-supported MODs refreshing annotations is an integral part of their job. For single-pass annotations, however, there may not be resources to refresh them, especially if they are performed at a sequencing center that has long since moved on to the next genome. We do have an existing system to do with this in that a gene association file that has not been refreshed within the last twelve month is put into a separate gene association file archive.

Another outreach idea is that of an annual annotation challenge. This might be similar to the CASP (Critical Assessment of Techniques for Protein Structure Prediction) and GASP (Genome Annotation Assessment Project) projects. An annotation challenge might also be a good way to interact with software developers creating tools for GO.

We could give groups a set of genomes, have them run their programs on them, submit resulting GO annotations, and then compare their annotations to known annotations to see how the different programs perform. The winning program could then be run against all of the genomes that need refreshing. The annotation challenge could perhaps also be extended to evaluating tools that assess consistency of annotation within the consortium.

GO should also continue trying to reach out to additional prokaryotic annotations groups, as there are now a number of different groups in many different locations, such as the Joint Genome Institute, Argonne National Labs or the Pasteur Institute, to name a few.

We could also reach out to groups involved in the NIAID program to sequence organisms that are potential bioterrorist agents, which is ongoing at several BRCs (Bioinformatics Resource Centers).

There are some particular issues, such as the operon issue or annotation by pathway hole filling, that arise when annotating prokaryotic genomes that may merit some changes in the GO. Specifically, this may require changes to the evidence codes, and we may need to improve links between function and process. These issues are especially important as annotation of the E. coli genome, one of the reference genomes, becomes more coordinated (see Riley et al, NAR, 2006 Vol. 34, No. 1, 1-9).

Aim 4 – Community Advocacy

Mike Cherry, SGD

The main idea behind community advocacy is that there would be advocates to seek out what is needed by those people that use the GO. As a group, we want to know, and should start thinking about, who else is using the GO and what can we do to meet their needs?

To begin to address this, we may want to designate a publicity person for the GO and have a place on the website where we highlight the use of the GO. This is not simply an issue of interface design, but also really addresses the questions of what is GO and why should I care about it in an effort to draw people in more. We need to recognize that in talking to non-genomics biologists about GO, we may need to use new language and be pro-active about determining exactly what people do and do not know about GO. GO is now a resource that is expected to be there and as such, will be subject to criticism. But, it is our responsibility to get out there and help people who are using it and address issues they might have.

Michelle suggested that we might want to work with someone professional to develop training materials. Candace seconded this and mentioned that there are lots of people at colleges/universities using GO in their courses and training materials would be really helpful.

Perhaps we could collaborate with an educational professional to produce GO training materials. Reactome has had some experience with a group that came in to produce training material that they would then sell as a subscription. The number of people subscribed might help to give some idea of how many people use their resource.

How else to determine who uses GO? We could compile lists of people who mention GO in publications and in talks at meetings. For the former, we could perform a full-text search using “gene ontology.”

It would also be good to foster a user community that gets immediate feedback from GO in the form of quick responses to email queries. We could also try some sort of monthly call-in where people can ask questions or simply lurk and listen.

Could we put into place some mechanism that allows people to tell us what genes they would really like to see annotated?

We could also make some changes to the email lists. Sometimes people post to a GO mailing list, but it’s not to the right list. We should consider consolidating the lists and making it more transparent which list people should post to when they have a question or comment.

Having good training materials on the web is key. Courses on genomics and informatics would use GO if there were training materials available. This would increase student awareness of GO.

Is there any way that we can determine who is downloading files and perhaps get automatic feedback from them when they do so? Tracking downloads, however, would be a low estimate of how many people use GO.

Sue was recently at a Systems Biology meeting, and nearly every talk mentioned GO. Attending meetings may be one way to get a feel who is using it. She also felt that looking for citations in publications is also a decent step.

John: We need to foster the idea that we are responsive to the community. Karen brought up Rama’s comment that we need to consolidate/manage our various email lists: go-database, amigo, go, and make sure that questions don’t get lost, that everything gets answered in a timely fashion.

David: Medium to large scale analyses where they use GO for analysis– we could offer to focus on annotating genes which would be relevant to the analysis and get it done so that they can use it before publication of the paper. Jane suggested that we could have a tracker for this.

What about a monthly newsletter, a wiki, rss feeds? As many of the MODs already send out monthly newsletters, perhaps they could append a new section devoted exclusively to GO. Using the MODs to reach more people was especially helpful for the GO survey done last fall, results of which are included as an appendix to the grant submission.

Also, any courses that people teach, such as those at TIGR, Jackson Labs, Woods Hole, CSHL, could include GO as part of the curriculum.

There will be no new hires for outreach, so existing GO curators should think about how to reach out to their user community. If you have more ideas about how best to do this, please send them to Eurie Hong at SGD who will be coordinating community outreach.

One other point to keep in mind is that a lot of groups doing functional annotation are doing structural annotation, as well. This is partly why we need the Sequence Ontology (SO). SO has become the fourth ontology, and although there are differences between the SO and GO ontologies, they are all supported and needed to do the complexity of gene annotation.

Aim 5 – Organization

Suzi Lewis, BDGP

Three years ago, the aims of the GO grant proposal were to create structured vocabularies, support and promote use of GO in annotation projects, add new MODs to the consortium, build and disseminate informatics resources and tools to support community use of GO vocabularies..

How can we measure progress towards these aims? The number of hits to the web site has gone up, the number of publications mentioning GO has gone up, and the number of links to the GO web site has increased. 17/24 NIH agencies that fund research have supported projects that use GO and there are many hits when searching for GO in Google Scholar. The GO survey netted close to 1500 replies in three weeks, almost all of which were positive. So, where do we go from here?

Looking at the 2006 aims:

1 maintain comprehensive, logically rigorous, biologically accurate ontologies

2 comprehensively annotate nine reference genomes in as complete detail as possible

3 support annotations across all organisms

4 provide annotations and tools to the research community

Given that the funding level is expected to remain the same, how are we going to scale up? The answer is that we need to do a little better, need to get more efficient, and need to be more coordinated. This can be done!

We already have some working groups, such as the AmiGO and OBO-Edit working groups, and these have been successful. But, any new working groups that form will have specific questions to ask, including:

1 Why is this group here?

2 What is the lifespan of this group? (could differ greatly between groups)

3 Who is the group leader?

4 Are we making progress? What are the obstacles to making progress?

5 What are the group's priorities and what are the criteria for setting priorities?

6 What will the group deliver?

7 What are the criteria for membership? Who has a vested interest in this issue?

8 How often should you meet? How should you meet? Via email? Other ways?

Each group can decide what works for them and what communication method will work best, but having metrics is essential. How are you going to measure your progress/success? What tests do you want to build? Different groups will need to communicate with one another. How are you going to share information? The decision making process for all of this still needs to be fleshed out, but efficiency and lack of bureaucracy is key. If there are software issues that come up, please raise them with Chris.

As an example, the reference genome group will have Rex Chisholm in charge, produce annotations and protein sets, will comprise members for each of the nine reference genomes, participate in fortnightly conference calls, and have as their metrics the percentage of the genome that is annotated. This group will work with the computational group and the user community, processes for which will be decided.

Levels of organization in the GOC can be described as follows:

1 GO PIs, Judy, Michael, Mike, Suzi – set priorities, obtain funding.

2 Midori, David, Jen, Rex, Eurie, and Chris – ontology content, annotator-ontology liason, annotation outreach, reference annotation, community advocacy, computational architecture, respectively.

3 Curators, some of whom are directly funded, from MODs: human, mouse, zebrafish, fly, weed, worm, slime mold, budding yeast, microbes – perform annotation

Parting thoughts: This isn't a drastic change from what we're already doing, but it is trying to be more conscious about what we're doing. We don't want things to fall through the cracks. One of the reasons that projects fail is that when there are so many things to do, you pick the easy things over the things you should do. We need to do the hard, important things.

Ontological Content: Relationships and Terms

Missing "is_a"s

Chris Mungall, BDGP

Tangled DAGs and complexity – complexity is increasing exponentially (i.e total number of paths to the root node).

The first topic addresses the issue of missing is_a relationships in the GO. Jane Lomax has worked on filling in missing is_a relationships in the cellular component ontology. At the onset, the process and component ontologies are not is_a complete. In the past, our methodology has been that as long as a term has a parent, we haven't cared so much about the relationship.

Why is there a need to have an is_a parent for each term? Because without them, we're missing terms and when we're missing terms it's much harder to get the ontology to work with other ontologies and tools that are out there. For example, the Protégé ontology editor assumes is_a completeness and so does not work well with GO, as orphan terms appear as root terms. Accurate ontologies also aid accurate searching.

How can we get started at untangling complex DAGs? In our current display, when drawing the DAG, we ignore the relationship type. We label it, but we don't use the information on the label. For example, the Srb-mediator complex in it's current incarnation is displayed multiple times showing every possible way to draw a route to the top.

In Jane's work, she took all of the is_a orphans and either found the correct parent or created the correct parent. There were 277 is_a orphans in the cellular component ontology. In the 'fixed' version of the cellular component ontology, these is_a orphans are gone, which increases the total number of mixed-paths-to-root, actually making it more complex to look at. Some browsers, such as that of the FMA (Foundational Model of Anatomy) have distinct is_a and part_of browsers, which helps to think about the ontology better.

A summary of Jane's work:

old

- 277 is_a orphans/1688 terms
- avg is_a paths to root: 1.4
- avg mixed-paths to root 6.97

new:

- 0 is_ a orphans
- avg is_a paths to root: 3.36
- avg mixed-paths to root: 38.6

This work illustrates how this project needs to coordinate with the AmiGO working group. We want to make this new version live as soon as possible, but the display will need to be addressed. It would be good to let people have a look, since the component ontology will look very different when these changes are implemented.

There are also some terms missing part_of roots. For example, unlocalized complexes don't have part_of roots at the top of the tree. All of these complexes need to be housed somewhere, although some of them are so old that there are no references, no definitions, and no annotations made to them.

Demonstration of new is_a complete cellular component ontology in OBO-Edit

To ensure is_a completeness, a new high level term 'X component' was introduced. For example, cell projection component was introduced to have a complete is_a path for existing terms, such as cell projection membrane. This is analogous to what the FMA, which is is_a complete, did when they invented the term 'cardinal body part' to have a parent term for the 'head', 'trunk', etc. since extremities are not organs. FuGO also needed these types of top level terms.

One issue that still remains is what to have as an is_a parent for cell components, ie

terms that are part_of cell. Would it be okay to have a parent term cell component, even though the whole ontology is called cellular component? Barry Smith recommended that this trick only be used when dealing with entities that are not anything else, since we don't want to introduce multiple inheritance for this purpose. If the cell component term was introduced, we'd have the same number of high level terms and could remove some redundancies.

Chris and Suzi also suggested renaming the ontology. We could refer to a 'cell-level entity' and then also have cell parts. Barry pointed out that the FMA uses part in the name, and uses 'cardinal part' if it's the only way to have an is_a relationship for every term. It is used, however, for entities that are there only when the whole is there. So, we could use the term 'cell part', 'membrane part', etc……

So Jane has done the following:

- created terms like 'cell projection component' to be is_a parent to give the various types of cell projections an is_a parent.
- may have added some is_a parents that are needed
- an issue came up about: she didn't create a term "cell component" which is almost identical to name of the ontology, i.e. "cellular component" so there are still a bunch of things right off the top.

ACTION ITEM #5: Consider what, if any, are the repercussions of renaming the cellular component to cell level entity?

ACTION ITEM #6: Change new terms ending in '…component' to now end in '…part.' [Jane Lomax]

Note that the current version of AmiGO does not have the ability to disentangle paths, nor do other groups browsers.

In the meantime, the MODs will need to address this in their display, AmiGO will need to deal with this. We don't necessarily want to wait for the tools to catch up to implement this, but the displays will explode a bit. Annotators should take a look at this in OBO-Edit to see if they can live with this. There may be some redundancies that could be removed, which would help with the increased paths to root.

OBO-Edit needs a verification system to check for is_a completeness and warn before saving.

ACTION ITEM #7: Jane will send an is_a complete cellular component ontology to the GO list.

There are also lots (i.e., thousands) of missing is_a relationships in the process ontology that will need to be dealt with. This should be a little bit easier since the tools are already dealing with this with respect to cellular component. However, once we make this live, we're committed to maintaining is_a completeness. Any new terms entered should not create new is_a orphans. OBO-Edit has a verification system for this that could be revived and could be tested by the **OBO-Edit WORKING GROUP**.

Fixing the Upper Levels of Biological Process

Chris Mungall, BDGP and Barry Smith

Examining the top levels of the Biological Process ontology, we also find that there are some issues with granularity. We have been talking about diamonds and DAGs and how they're not so good for ontologies. If diamonds are treated correctly they're not such a problem., but diamonds at the upper levels of the ontology *are* more of a problem, especially when they are at the very top. What is the solution?

Some of the high level terms in the BP ontology are: cellular process, physiological process, cellular physiological process, and organismal physiological process. Do these terms define the granular level of the process or of the end result? The definitions seem inconsistent. In some cases, the definitions refer to the goal of the process, not the level at which the process occurs. Further there is a lack of symmetry in these terms and definitions. For example, why is there no term 'organismal process'? And what does physiological process really mean? It is vague and hard to distinguish from biological process, although some dictionaries define it as everything but developmental processes.

We should think of and define a process as something that has a beginning and an end.

But where's the beginning and end of a physiological process? Of metabolism?

There is also an issue with behavior terms. We've got something called behavior which is not considered a physiological process, but 'behavioral physiological' child terms are useful. All of the behavior terms are now 'response to…' terms, but shouldn't response be a physiological process, since the definition is about a set of things that happen in response to something else?

Jen commented that she once looked up "physiology" and it appears to be a catch for anything that wasn't "development" and some other major branch of research at the time the term was introduced; David feels that physiological process is equivalent to biological process; idea of eliminating "physiological process".

Should we just eliminate the term 'physiological process'?

Alternatively, we could add the granularity at the top level of the process ontology itself.

We could have molecular, cellular, and organismal levels. These are disjointed, whereas physiology is not. The advantage is that each is distinguished from a parent in the same way.

So, for example, a term like cardiac contraction would be an is_a term of organism level process, cell contraction would be a part_of that and also an is_a of a cellular level process.

Barry's revision:

Biological process:

- molecular level process
- cellular level process
- organismal level process

Barry: "I do not believe that there is such a thing as "molecular physiology", however there are such things as "molecular biological process".

Rex and David provided examples to the contrary that people do actually talk about "molecular physiology".

What about single-celled organisms? This has been a source of past and current confusion, since in these cases cell and organism are the same thing. From the perspective of the GO, we need to make a decision about what a single-celled organism is.

We may also want the term 'multicellular level process' or 'multicellular organism process', and we need to consider the cases of organisms like Dicty that can exist in both states and some bacterial species that associate but don't meet criteria for being deemed multicellular.

Barry suggested adding two children of cellular level process: single cell process and multicellular process. We could also have single organism process and multiorganism process.

Agreed that we should look at a few branches of the ontology and see how this would work. What will the effects of these changes be? We can sort through this in-house as there may be many merges and changes, but we shoulud keep an eye towards what other ontologies, such as FMA, do.

We also need to consider the implications of this for pathological processes. What are molecular level processes vs cellular processes? For example, if a molecular penetrates a cell, it's a molecular process, but if a cell captures a molecule this would be a cellular process. We could figure out the granularity by determining the agent. The trick is to find an example where there is no apparent classification as one or the other. It should be obvious, though, where the action is – at the molecular or cellular level. Think about viruses attaching to cells, DNA molecules penetrating a hole in the cell, etc.

Also think about how this reorganization would affect development and cell differentiation terms in the GO. Development is an organismal process, but a big part of that, cell differentiation, is a cellular process. Will this be okay? The relationship to the former would be part_of, while the relationship to the latter would be is_a.

ACTION ITEM #8: Take the new molecular/cellular/multicellular arrangement of the biological process ontology, try it out, and see how it works. The WORKING GROUP for this will include: Chris, Jane, David, Alex, Michelle, Rex, and Val. Jane will make a .obo file so that people can look at this. The group will also need to create a document outlining their philosophical approach and the results. Should there be a development site to help with working on this?

David brought up the sticky issue of making developmental process be under 'organismal process' and then you eventually get down to 'cell differentitation' which also needs to be under ' cellular process'.

Problem of development and the child terms of development.

Relations Between Function, Process, and Component

Chris Mungall, BDGP

How can we make links between the GO ontologies? The discussion of this focused on links between molecular function and biological process using histidine degradation as an example. (See Overbeek, et al, NAR 2005 33(17):5691-5702).

Histidine degradation is a complex process with a branched pathway. In GO, it is represented by the BP term 'histidine catabolism.' Each of the reactions in the process are mapped to GO MF terms, such as 'histidine ammonia-lyase' activity. We have some variants for different ways that histidine catabolism takes place in different organisms and could map pathway branches seen in different organisms to GO child terms. Two could be mapped, histidine catabolism to glutamate and formate, and histidine catabolism to glutamate and formamide, but one could not: histidine catabolism to glutamate and formiminotetrahydrofolate is not represented in the GO. Because of this, there is no way to link relevant functions to processes.

Michelle: The microbial community is looking for this type of relationship because reaction steps for each process is how they (TIGR) annotate a bacterial genome. There is clearly lots of curation work that needs to be done. How will the results fit into the current ontological structure?

But, an important issue here is what is a function and what should be in the function ontology? We must settle that first.

We also need to be clear about what we mean when we are talking about types, processes, and functions. Every single term in the GO should represent an actual type in biology, but the converse is not necessarily true.

What are the relations in reality? There are relationships between types in the same ontology and between functions and processes at a given level of granularity.

What are the instances and relations in reality? Some particular gene product has some molecular function instance – has a functioning- which makes up a more complex multi-step process. Examples of type: histidine ammonia lyase function, histidine ammonia lyase reaction, histidine catabolism.

How hard is it going to be to determine when we can make these relations? What are the types and relations in reality? Some reactions will only take place in one pathway.

Barry commented that if this is the case, then we can't assert part_of relations. But can lack of knowledge be used to exclude a part_of relationship?

What about using the relationship has_part? This exists in the OBO relation ontology.

We need a way to distinguish those reactions that always occur as part of 'histidine catabolism' versus those that are sometimes part of histidine catabolism depending upon the organism and maybe sometimes part of other processes. For example, alcohol dehydrogenase functions in many pathways.

Barry's comments: if we have multiple inheritance in part_of parents, then every single instance must have that part. If we have an entity that is sometimes part of one thing and sometimes another, then we cannot assert the part_of relationship.

The part_of definition means all the time; has_part means the parent process necessarily has that function as part of the process.

We need to collect more data and think about how to integrate that data. The data is captured in annotations and so we should be able to use annotations to help predict the relationships between process and function.

These ideas will probably work fine for well-characterized pathways, but what about other pathways, such as cytokinesis, where all of the molecular events are not yet defined? At some point we are going to want to talk about pathways at an organismal level, not just at the molecular level. To do this we will want to consult with the pathway databases and have some level of synchronization with them.

In GO, however, we will still be missing information about the order of events, so we would need to bring in new relations such as 'preceded_by.' Prokaryotic annotators would like to see those relationships brought into GO soon. (Michelle gets complaints from people in her classes that GO doesn't represent pathways.)

So, how would we manifest this in GO? Molecular functions reside in gene products and the function is distinct from the functioning. In reality, there is a certain bit of redundancy here; do we want to manifest this directly in the GO? Do we want to have functioning terms in the process ontology? We know that not all GO function terms have a corresponding functioning term.

Some redundancy already exists in the GO, in cases such as 'iron transport' and 'iron transporter.' We also have some redundancy between component types and function types. This is not necessarily bad, but we should be aware of the reasons why they're there.

The functioning of the gene product is implicit; it exists in reality but not in GO. If we're not going to create functioning terms, then we will need to link the function and process ontologies in another way. We shouldn't use part_of between the ontologies. What about having a functioning_of relationship?

Another proposal would be to remove the function terms and replace them with highly granular process terms [see below].

One point to consider is that every organism may not always have a particular reaction as part of its pathway. This is why we're using the has_part relation and not the part_of relation, but this arrangement might still be problematic for creating GO slims. It would then be important to annotation NOT to the process if you *know* a reaction/enzyme exists but the pathway is broken, which happens a lot in bacteria. Use of the NOT qualifier in these cases would be consistent with the way we normally use it.

Can one function have many types of functionings? There can be many instances of functionings but only one functioning type.

Processing Function Terms

Amelia Ireland, GOEO

In the past there has been a lot of discussion about function terms and a number of term obsoletions. The proposal presented here would move function terms representing steps in a process to the process ontology. For example, many terms under 'protein modification' in the process ontology are just one-step reactions. On the other hand, we have multi-step reactions that appear in the function ontology, e.g., ent-kaurene oxidase activity. We also have processes and functions that appear to be identical, such as histone methylation and histone methyltransferase activity.

We have had a one-step rule for creating functions, but what constitutes a step can vary between functions. Some function terms fit the definition and some do not.

Some discussion as to where the one-step rule is stated. It is in the documentation and has been used by the editorial office for quite some time. When the one-step rule was instated, no one spoke up to disagree. But, perhaps the one-step rule does not clearly state what we want and can be interpreted in different ways.

We seem to agree that a function is an instance of what a gene product can do at the level of the molecule and does not represent just one step in a process. For example, think of the Michaelis-Menten equation which takes into account several steps in an enzymatic reaction, e. g., binding, catalysis. However, even though there are steps, some said that there is still just one thing that the molecule does.

Does the presence or absence of reaction intermediates have any bearing on this discussion? No, since for any enzymatic reaction that occurs there are going to be some intermediate steps.

What about situations where, for a given function, there is one gene product in eukaryotes, but many gene products in prokaryotes? Would we represent the same reaction in different ways in GO? Would there be multiple functions for the prokaryotic gene products and one bundled function for the eukaryotic gene product? But why annotate to one conglomerate term when the current incarnation of GO allows annotators to annotate to more than one term, i.e. one term for each reaction? (See cerevisiae fatty acid synthetase, for example.) This could create problems in term nomenclature.

What about multifunctional proteins? In these cases, curators would annotate to one term if this function represented a cascade of events that won't stop once started. Otherwise, curators should annotate to separate functions.

Amelia commented here that it seems clear that many functions don't fit the definition of "elemental activities describing the actions of a gene product at the molecular level", and that there are "clean" and "dirty" functions in the GO. "Clean" functions include terms like 'arginase activity', 'transaminase activity', etc. and we can make "clean" relations between these functions and the process term 'arginine catabolism to glutatmate.' Another example of a "clean" function would be 'Notch binding' which can be related to 'Notch signaling pathway.'

On the other hand, "dirty" functions are not necessarily steps in a process, but represent a combination of two or more attributes from function, component, process, or something other domain. An example of a "dirty" function is 'transcription factor activity' which is defined as: Any activity required to initiate or regulate transcription; includes the actions of both gene regulatory proteins as well as general transcription factors. Other examples include receptor activity, structural molecule activity, enzyme regulator activity, hormone activity, etc. These "dirty" functions represent a role or a class and cannot be linked to process terms using existing GO relations.

What is the solution? Amelia proposes that we move terms representing events into the process ontology. We could start with catalytic activities and move them under the parent term 'metabolism,' creating terms that correspond to particular EC numbers.

But, if enzyme activities are "clean", why move them to process? Because these functions are events – occurrents - and should be dealt with in a different way to those function terms representing a role or class of gene product.

Using arginine biosynthesis as an example, individual reactions could be organized according to the substances involved. Reactions involving arginine would be is_a children of arginine metabolism. The reactions making up a specific arginine biosynthesis pathway could be made part_of children of the term representing that pathway, e.g. 'arginine biosynthesis from xxx'. Will listing the individual steps as children of arginine metabolism and the relevant pathway be consistent with how biologists think about this?

Another example of moving function terms is illustrated by the binding terms. For example, 'Notch binding' could be a part_of 'Notch signaling pathway.' We could do the same for transporter activity, permease activity, receptor activity, ligand binding during signaling pathway, regulator activity, etc. We could also reinstate some obsolete terms such as cell adhesion molecule as the concept of cell adhesion receptor binding. All moved terms will be given many function- and process-style synonyms.

So, what's left? The Brave New Function World. This world would redefine function in more colloquial terms, consistent with the dictionary definition that defines a function as the purpose that a gene product serves in the normal activity of an organism. We could keep "dirty" terms representing combinations of function, process, or other information in function, and the new function definition would allow us to add useful terms currently not allowed, such as toxin, ligand, other suggestions?

Does this proposal reflect that there are problems with our function terms, or with the definitions? For example, biochemists and cell biologists look at ligands differently. There is a function in the connotation of ligand; was it just poorly defined? Maybe the definitions need to be clarified.

The Brave New Function World would include the following functions:

energy transducer

enzyme

motor

nutrient reservoir

signal

structural molecule

receptor

regulator

transporter

toxin

There is great concern that this will create new problems and confusion, in part because these are names of things, not functions. Barry's suggestion is that perhaps GO should create an ontology of molecular entities so that GO can clarify use of words to describe what a molecule is, what it does, what it could do, etc. But do we want to have a molecular entity ontology and a function ontology?

This discussion seems to be raising a number of separate issues:

- For "clean" functions, how would you show which functions are needed for a particular process?
- Perhaps we need to talk about better upper level organization for function, similar to what we discussed for process?

- There is a lot of messiness in the function ontology, but processing terms to group functions is not right.
- There is an issue with physical entities.

We do have a chance to do this clean, and we don't want to invent a jargon any more than we have to.

We could keep the Brave New Functions, but call them molecular function entities or add 'function' to the names. There has been a long-standing issue with trying to define the intersection between the molecules we can assay and our understanding of their function. This is part of why we added the word 'activity' to the function terms. We've purged anything from the function ontology that sounded like a protein, but we still have these issues.

We need to readdress the one-step rule and correct this, if needed. The prokaryotic community wants an association between function and process, so we're going to have to find a way to do this, but not by moving function terms into process. This seems backwards to people, as the "cleanest" functions, such as enzymatic activities, are the ones that people think biologists would most expect to see in a function ontology.

In summary, Amelia presented the pros and cons to her proposal:

Pros:

- Reactions involving a certain substrate can all be homed underneath one term, xxx metabolism
- Reactions can be linked to pathways
- Greater precision in annotation
- Ontologies more consistent and pure
- Less redundancy between process and function
- GO function closer to colloguial interpretation

Cons:

- Some adjustment needed by those accustomed to seeing binding and enzymatic reactions in function
- May take some time to implement

An alternative solution presented would be to allow complex functions and redefine molecular function to for terms that represent all the biological activities a gene product has. But some molecules have completely separate functions, such as actin which polymerizes into filaments and inhibits DNaseI. Other examples cited include GPCRs and immunoglobulins.

Other thoughts from this discussion:

- Make grouping terms under function that indicates that this function initiates processes.
- Perhaps we need to broaden our definition of entity so that we can annotate not only to the function of gene products, but to complexes, as well, since in some cases, such as RNA polymerase activity, no single gene product has that function. This relates to the issue of granularity of entities.
- Can we establish links between function and complex terms in cellular component?

ACTION ITEM #9: Make actin polymerization a function term.

ACTION ITEM #10: Amelia, Chris, and David (and Barry, if he’s available) should get together and come up with a single proposal for making connections between function and process. This will likely include writing new definitions for function and process.

Two Taxon Option

Jane Lomax, GOEO

This issue arose from the content meeting held at TIGR in November 2005, where a proposal was drafted for a dual annotation system that accommodates annotation of multi-organism processes.

A key component of this proposal is that for multi-organism annotation, two taxa may be placed in the Taxon ID field. The first taxon ID refers to the organism that encodes the gene product, while the second refers to the organism that interacts with it. The two taxa should be pipe-separated.

Dual-taxon annotation will also require adding a lot more terms and so we're considering having a few GO curators visit PAMGO later this year to develop that part of the ontology.

Other plans include writing detailed documentation and guidance for annotators that may need to curate these interactions. There will also be an announcement of this documentation on the annotation mailing list and dual-taxon annotation will be discussed at the upcoming annotation camp. At the moment, though, only TIGR and PAMGO groups are probably doing this.

If the interaction is between two individuals of the same species should annotators still put two taxa in the ID field? Yes.

The issue of dual-taxon annotation started out originally because some bacterial proteins are injected into plants cells, which leads to the 'hypersensitive response'. Currently, this term is a child of 'programmed cell death', but is this the right parentage? We made need to create a separate ontology for these interactions, since 'hypersensitive response' doesn't fit where it currently resides.

It is agreed that the GO recognizes that there are still some big problems in the underlying structure of the ontology with regard to this type of annotation and we may need to recognize a new structure of the process ontology to get this right. Having an upper level term of 'organismal process' may help understand how to place these terms correctly.

> **ACTION ITEM #11:** Add detailed documentation on dual-taxon annotation, announce this to the annotators' mailing list, include the info in annotation camp discussions, and work on developing the ontology. [Candace Collmer, Jane Lomax, Amelia Ireland, others?]

Two-Taxon Annotation – Capturing the Host Side of Interactions

Jane Lomax, GOEO

Given that there is a range of interactions, some harmful, some beneficial, are we annotating these interactions consistently? At present, when an interaction is mutually beneficial, we annotate both organisms. But, if the interaction is pathogenic, we only annotate the agent, not the host.

A new proposal would allow for annotation of host gene products in pathogenic interactions.

There is some opposition to this because, for the host, this represents abnormal processes. For example, would we want to annotate CD4 to HIV binding? Would we want to allow for ISS transference of this kind of annotation? In the case of viruses this could be particularly dangerous as there are many species-specific host-virus interactions.

What about cases of rhizobium binding to plant proteins (plant nodulation)? Does apparent co-evolution, or positive selection from both organisms, matter for this discussion? There may be a grey area between interactions that are called mutualistic

versus interaction that are pathogenic. Is it always clear how to distinguish the two?

Where does normal start and stop? How do you distinguish an agent in the environment that is neutral versus one that causes a pathogenic response? A binding reaction could lead to a pathogenic response; how do you not annotate that activity? Isn't it a normal process to bind something in the environment? We do have MF terms that involve a gene product interacting with DDT, which may not reflect that gene product's normal function.

Does the canonical life plan include interaction with viruses? Yes. But again, the normal function of CD4 is not to bind HIV. Yet it does bind, and not when mutated.

There is some agreement that it is difficult to draw the line between what is solely beneficial and what may sometimes be pathogenic (think about intestinal flora here). In the absence of any real objective dividing line, then at this point, we have to say that capturing these types of annotations is not okay and we continue to only annotate 'normal' processes. But, there are likely many different examples that will need to be considered before taking a definitive stance on this issue.

ACTION ITEM #12: Candace and Trudy will send examples and relevant references to the annotation list as they come up, so that we can consider these on a case-by-case basis. [Candace Collmer, Trudy Torto-Alallibo]

Obsoleting Justification Policy

Judy Blake, MGI

We have been getting a lot of feedback on term obsoletion from the GO user community, for example, GO users at the BRC meetings. It is necessary, therefore, to try to clarify the discussion about complex functions and about how we obsolete terms. The one-step rule for defining functions should have been challenged before becoming embedded into the system.

There are two main issues with the current obsoletion situation:

1 when we obsolete terms

2 the way we handle our systems when we do obsolete terms.

When searching in AmiGO for a term that has been obsoleted, a term that people use biologically, the obsolete terms come to the top of the page. We need to correct this.

The AMIGO WORKING GROUP will address this.

We continue to debate how we work with the rules about changing definitions. We can take an absolutist approach and say that any change requires a new ID, but what we've come to over time is that if we felt a definition didn't quite capture the essence of the term, we would change the definition, but not the ID.

Another issue (see Item #7 in Judy's handout, Obsoletes Redux) is that obsoletion has been used as a shorthand way to get people to change their annotations. But is there a better way to get people to look at their annotations? We could have annotations deprecated until they are re-checked.

It may also be necessary to have a better versioning system for GO terms. It would be possible to granularly track the way terms are changing, and perhaps we could embed the version of the term in the OBO file. One question, though: what constitutes a version-worthy change? OBO-Edit already has a mechanism in place to track obsolete with new replacement, but do people need to track less fundamental changes? SGD has a way for curators to add 'date last updated' tags to their annotations. This is manual and it is always left up to the curator to decide what changes necesitate a change to the update tag.

OBO-Edit 1.2 can assign automatic or manual replacement.

We need to be clear on why we obsolete terms, though. Sometimes we really need to obsolete a term but in other cases we know that there is a concept we want and we wouldn't make the term obsolete except for the fact that we know there are erroneous annotations. Bottom line: we'll still want to have automatic replacement as well as suggestions for having a look at another possible replacement term.

One concern is that our users don't understand why we keep adding and deleting terms such as cytokinesis. Cytokinesis has been obsoleted because we've diddled with the definition, but users don't understand why we're doing this. The key here is that probably the definition hasn't altered the correctness of the term. Another case concerns that of the molecular function plasma protein. We obsoleted this term, but didn't make a term to replace it, such as a component term 'extracellular.' Would this still have been within the scope of the GO, ie is plasma a cellular component?

There seem to be two issues here: 1) There does need to be some mechanism for obsoleting terms and checking annotations. But not any change to the definition warrants obsoletion. Trivial changes don't warrant obsoletion. 2) Annotations drive what the GO is. Operationally, there needs to be a lot closer communication between the editorial office and the annotators. We need to think about our policies for communication.

ACTION ITEM #13: Develop a new policy for communicating about term obsoletions. The person proposing obsoletion should get in touch with the annotating groups (using contact information from the gene association file) informing them that the term is under review while also soliciting input and suggestions on what changes to make. This could be scripted.

But, we still need to have a system for making sure people check their annotations if changes have been made to a term. Perhaps we could use the regular monthly reports to alert people that a term has changed somewhat and ask them to please review their annotations. But how do we enforce review? It could be set up so that until a term is reviewed it will be removed from the file. Filtering would be done by annotation date.

An additional feature built into this script could alert curators when leaves have been added to the original parent annotation so that curators could check and possibly add higher granular annotations. Concerns: Is there a way, without having a gene association file, to find and check for obsolete terms? How do we make term obsoletions obvious to groups that are not part of the consortium?

We may now have a mechanism in place for communicating obsoletions, but we still need criteria for deciding when an annotation is potentially affected by a change in the definition. And, people really need to check their annotations before a change happens.

Should we revisit old obsoletes and for those made obsolete because of annotation issues, should we go back to the old terms and make their IDs secondary IDs?

ACTION ITEM #14: Reinforce the policy that we will no longer obsolete a term just because the definition has changed or because annotations are thought to be bad or incorrect.

ACTION ITEM #15: Explore the proper technical solution for establishing a mechanism to notify GO users when a term has changed, or rather, when we are thinking of changing a term.

This solution should consider:

- Do we need to enforce a way of making sure people make any necessary changes?
- How best to contact people? Monthly reports? Newsletter? Web page?
- Can we adapt the existing system but have a version for ‘proposed changes to a term’s definition’?
- Could we have users subscribe to a mailing list to be alerted when there are changes to terms they’re interested in? We could have a ‘track this term’ feature in AmiGO for communication outside GO.
- Which people should be alerted to obsoletions?

The EC has a procedure whereby, on a periodic basis, they post all proposed changes and have a public comment period.

The editorial office shouldn’t have to pester people to review their annotations; this is something that the PIs should do.

At many past GO meetings, for many past proposals, there were no comments or objections raised. Is silence tacit approval? That seemed to be the case, but the editorial office would like an acknowledgement that the proposed changes are *really* okay with people, since silence does not mean ‘okay.’

Maybe we also need to have a minimal length of time for which a SourceForge item must stay open to allow people time to look things over. With the new grant proposal refocusing ‘embedded’ GO curators efforts, there should be a greater response to potential term changes.

We could also try having a wiki where each group needs to actively check a box, yes or no, by a certain time. We will need to give people a reasonable amount of time to respond, and also perhaps provide a box for comments. A script would remind people, 24 hours before the deadline, if they haven’t responded. Each site must appoint someone to take care of this responsibility. GO PIs will be notified of chronically non-responding groups.

ACTION ITEM #16: Revisit obsolete terms to see which can be merged with current terms. Decide if the IDs of obsoletes could be made secondary IDs to the currently existing term.

ACTION ITEM #17: John Day-Richter will talk to the GO Editorial Office to find the best way to implement these changes and suggestions in OBO-Edit.

ACTION ITEM #18: Add a term creation date to the .obo file.

Annotation Issues

Issues Arising from Annotation Camp

Karen Christie, SGD

There were a number of annotation issues that arose at the 2005 annotation camp that seemed appropriate for discussion by the entire consortium. However, there has been a considerable amount of time (nine months) between the annotation camp and this latest consortium meeting. Since we are about to have another annotation camp, we should work out procedures for how to reconcile annotation issues that arise at camp in a timely fashion.

The upcoming annotation camp will consist of two parts, one of which is internal and will involve representatives from each of the reference genomes and the other of which will consist of outreach and training.

For non-contentious annotation issues, is it okay for camp participants to put concrete proposals directly to the list to be checked and ratified, rather than wait for the next consortium meeting? Such proposals would be specifically about annotation issues; content issues would still need to go through the appropriate channels. The general feeling is that yes, this would be okay, but it would be good to provide a bullet summary of the recommendations in the camp minutes, which were excellent last year.

Also, since we now have group leaders for different aspects of GO, such as AmiGO, content, we should make explicit contact with each of the other groups if issues arise at camp that affect these groups.

The upcoming annotations camp will emphasize annotation consistency. How can we measure and track annotation consistency and quality across groups? This is something that we said we wanted to do in the grant proposal.

Dates for the upcoming annotation camp are tentatively set for July 10 -14, 2006. Camp will be funded, in part, by the Genetics Department at Stanford.

Annotating to Unknown

David Hill, MGI

Proposal: unknown terms in the ontology would be eliminated and annotations to these terms would be removed and gene products reannotated to the root of each branch of the ontology, BP, MF, CC. Everyone agreed that this was okay, but we did not establish a time table for this change.

Annotation of Common Knowledge

Paper Introductions

Karen Christie, SGD

This issue came up at the last annotation camp. Some groups allow curation of information from the introduction of papers where authors cite research papers as references. These groups use TAS for the evidence code in these cases. Other groups require curators to actually look these statements up. At camp, we didn't come up with a recommendation for the best annotation practice for this situation.

It was generally agreed that TAS is not a high quality evidence code and for reference genomes especially, the goal should be for each gene product to be annotated with an experimental evidence code. References cited in the introduction of a paper are not always the most relevant and may not even be from the same species as the gene product being annotated.

Some groups have TAS and have used it for different ways. Rat uses it, pombe uses it as a placeholder until the original paper can be curated, sgd used it to curate information from reviews. TAIR has a large number of TAS annotations and GOA has TAS annotations inherited from Proteome and annotation from Swiss-Prot where annotators weren't distinguishing between author statements and experimental evidence. Should groups retro-curate and remove TAS annotations? Yes, if possible.

ACTION ITEM #19: TAS is no longer considered a useful evidence code and will not be used in any consistency measures of reference genome annotation. Since part of the idea of the reference genomes is to provide a source of IEA annotations for other groups, we strongly encourage reference genome annotators to not use TAS, and instead use experimental evidence codes whenever possible. The GO documentation should also state this in a clear fashion.

Annotation of Common Knowledge

‘Textbook Knowledge’

Karen Christie, SGD

Common knowledge, such as that found in textbooks like Stryer’s Biochemistry text, has, in the past, been annotated using the TAS evidence code. For example, alcohol dehydrogenase has been annoated to CC:cytosol using TAS and Stryer, but this information really isn’t traceable from Stryer to an experimental paper.

This type of TAS annotation predates the IC evidence code which might be a much better way to annotate this kind of information. Consider the case of ribosomal proteins, where there may be strong sequence conservation leading to an ISS annotation to CC:ribosome. In this case, then, an annotation to BP:translation, using the IC evidence code would be appropriate. (Note how IC is being used in this way to link ontologies together.)

What reference should be used? What entity goes in the WITH column? Curators should cite the paper that they’re reading in the reference column and put the GO term that they used to make the connection in the WITH column. If the original ISS annotation was made with an internal db_ref, then that reference would be used for the IC annotation, too.

We should consider establishing criteria for making IC statements, as some inferences may be more explicit than others. This would be a good topic for discussion at annotation camp.

NAS vs Experimental Evidence Codes – Data Not Shown

Harold Drabkin, MGI

This is another issue that came up at annotation camp. When experimental results are reported in a paper but followed with ‘data not shown,’ what is the appropriate evidence

code to use? MGI uses IDA (or whatever is the correct experimental evidence code) because it is often the case that the experimental method is clear and journals would require the data to be shown, if needed. ‘Data not shown’ can often be the result of space limitations for publication.

’Data not shown’ can be subdivided into two types, though: the first being cases where it is clear what assay was used, and the second being cases where it’s not so clear how the data was acquired. Should curators annotate the latter, and if so, using what evidence code? NAS? Should curators contact the authors about this type of data? What about information cited as personal communication?

The general consensus seems to be that curators should use their best judgement in these cases. If you are confident that the author is clearly stating what they did, then use an experimental evidence code. It’s okay to accept that the authors did what they said they did. If you are not confident about the experimental evidence code, however, it’s probably best not to curate the information, as NAS is not a very useful evidence code and we should strive not to use it. Information cited as personal communication should not be annotated.

Same Protein, Different Organisms, Different Strains

Michelle Gwinn-Giglio, TIGR

When annotating different bacterial strains, is it okay to use experimental evidence codes if the actual experiments were performed in a different strain? This question was posed to the mailing list and no real consensus was reached.

Should the annotations be made using the ISS evidence code, or is IDA okay? We know that bacterial strains can be hugely different from one another. What constitutes a strain or species in bacteria?

Michelle thinks that it's okay to transfer the experimental evidence code when the sequence similarity is 100%. Some agree with this, but others disagree and think that ISS or even IC would be the more appropriate evidence code.

People think that this type of annotation is okay, because when we make annotations, we're always making inferences and asserting the typical function of a gene product in that species. However, there are cases where gene products that are 100% identical do not have the same function in different strains.

The feeling is that we do need to be pragmatic, here. The sequencing and the biochemistry of a particular organism may be done on two entirely different strains and we are really annotating the *potential* of a given gene product. Further, it is probably more likely that the distinctions annotators will encounter will lie not in molecular function but in biological process. You may have the same base activity for a gene product, but the process it's involved in might be different.

ACTION ITEM #20: Add to the documentation that it is okay to use experimental evidence codes for identical/similar gene products from different strains of the same species.

Annotation of Gene Products 'Acted Upon'

Michelle Gwinn, TIGR

In the current documentation, we state that we only annotate those gene products that are involved in a process, not those that are acted upon by a process. For example, we annotate gene products that are involved in the process of secretion, but not the product being secreted.

There is some concern that if we follow this argument, we could annotate to genes upon which a transcription factor acts, or that when two proteins bind, that protein A acts upon protein B by binding it. We need to be careful how we go down this route.

There is strong feeling that if we decide to pursue this type of annotation that we don't do so in the context of the current annotation file format. We should, instead, place these annotations in a different file.

There are groups, such as MGI who capture this kind of information in a text/notes field under the category of 'target.' Other groups, such as TAIR, have added new relationships such as 'has_protein_modification_type' that are in TAIR, but are not included in the gene association file sent to GO. This highlights that individual databases can easily store additional information, but can we be consistent about how this information is eventually presented?

But does GO want to support this type of data? At present, the relationship between the genes and GO terms is implicit. If the relationship is made explicit, then it would be okay to capture this type of information. We can easily add more to the same file, or to a new one, where we explicitly state these relationships, since there are groups out there that want to see this type of information. Could this be a way for GO to handle those things that are not judged as 'normal,' such as host-virus interactions?

This problem arises because we have these implicit relationships. However, these relationships are actually subtle that we usually state. For example, the cellular component annotations really represent the end destination for a gene product. We don't annotate every point along its path while it gets to that final destination.

There is a feeling that we could capture this type of information, but that until we make relationships more explicit we probably want to keep these annotations in a separate file. We could add another column to the current gene association file, but users might then have to filter the file to remove this set, and it could create confusion. Also, what are the ramifications for existing GO tools?

John pointed out that we are already discussing creating a more expressive gene association file and recording this type of data could be part of a pilot project related to this.

But, there is still some discussion about whether or not GO really wants to support this type of annotation, since this type of annotation would really expand the role of what GO provides. Is this within our project scope? Some annotators see this as a logical progression of what information users will want to see in GO. For example, if a gene is annotated to MF: tyrosine kinase, the next question is likely, what gene products does it phosphorylate?

General consensus: having these annotations will be useful, but will require more explicit relations than what we are currently making, so we need to decide how we want to do this.

ACTION ITEM #21: Individual groups can collect this [acted_upon] data knowing that, in the future, GO will present this type of relation. Chris, John, Sue, Candace, and Michelle will form a **WORKING GROUP** to come up with a proposal for how to implement this. Note that this type of annotation will not be a core requirement, but that GO will facilitate its display if groups want to do this.

Annotations Inferred from Genetic Context

Michelle Gwinn-Giglio, TIGR

What evidence code could be used for annotations made on the basis of genome context, a situation that arises in frequently in bacteria when evidence for surrounding genes being involved in a process is well-supported, but for other genes , also within the operon, the evidence is less well-supported?

One possible evidence code is IEP, but if the expression data is not shown, this won't work. What about IC? Michelle would like to propose a new evidence code, IGC, Inferred from Genomic Context, to deal with these annotations. The idea here is that you are using the positional context of the gene to make the annotation.

What would curators put in the WITH column for these annotations? The WITH column could have the SO ID for operon, but it is agreed that we would want to capture which operon is being annotated and perhaps even list all genes in this operon, or at least the first and last genes in the operon.

Could annotators make function annotations in these cases, in addition to process annotations? There was general agreement that, for prokaryotes, using operons to make function as well as process annotations is okay, especially since TIGR uses operon position as a contributing factor in the annotations, rather than the sole support for one.

Could this evidence code also be used for eukaryotes in cases of synteny? This might be harder to do, but we do want a way to annotate cases where the evidence based on sequence similarity might not be very strong, but there is other evidence, ie genomic context, that suggests a gene product is involved in a particular process. Matt brought up the example of variant surface proteins for illustration. This issue also came up at annotation camp, where the sequence similarity of flanking genes was very good, but the sequence similarity of the gene to be annotated was not so good. If you just used ISS to annotate the latter, what would you put in the WITH column?

ACTION ITEM #22: Create a new evidence code IGC, Inferred from Genomic Context. The precise definition of this code and procedures for annotation (what to put in the WITH column) will be hashed out and added to the documentation. [Michelle Gwinn and Matt Berriman]

Pseudogene Annotation

Michael Ashburner, FlyBase

There is general and strong agreement that GO will not annotate to pseudogenes as defined in the sequence ontology (SO).

It was generally agreed upon that annotating pseudogenes is wrong and that they should not have a GO annotation, instead they should have a SO annotation. When the pseudogene acquires a function it is no longer a pseudogene and could then gain a GO term.

One issue that arises with this, though, is that MODs need to be careful and consistent with how they define pseudogenes. This is especially true for the reference genomes.

Does a single frameshift constitute a pseudogenes? Probably not, in most cases. There are 12 genes in FlyBase that have a premature stop codon and these are not called pseudogenes as they are known to be functional in other strains and there are no other frameshifting deletions within their sequence.

Looking up Sequence Accession Numbers

Karen Christie, SGD

This is another annotation issue that came out of discussions at last year's annotation camp. For papers that discuss sequence similarity (or show alignments of protein sequences classified in the same family) but do not give accession numbers for the proteins listed (and many papers don't), is it okay for curators to look up the accession IDs to make an ISS annotation? If the paper being annotated doesn't show the experimental evidence for the gene product used to make the ISS annotation, and you don't know if an experiment has actually been done with that gene product, should you make the annotation? Can you check to see if the experiment has been done and then make the annotation?

MGI curators look up accession numbers all the time. They establish the orthology relationship and then look to see if there is a direct experiment for that gene product.

For proteins, they then add the Uniprot accession ID to the WITH column. Ref_seq IDs could be used in the case of ISS annotations based on nucleotide similarity.

Not all databases have been doing this, though. Other groups take the authors' word on the sequence similarity, but don't look up the accession IDs. For these ISS annotations, there is nothing in the WITH column. The ISS evidence code, in fact, actually predates the WITH column.

There is agreement, though, that going forward, we need to fill in the WITH field for ISS annotations and that there must be an experimental evidence code for the gene product cited in the WITH field. This is not stated in the current documentation on ISS.

ACTION ITEM #22: Update the documentation on using the ISS evidence code to emphasize that annotators need to enter something in the WITH field. In the case of gene products, there must be an experimental evidence code for that gene product which supports the annotation, i.e., we don't want to have circular ISS annotations. Uniprot IDs, ref_seq IDs, or individual MOD gene IDs would be okay to use in the WITH column. Old ISS annotations that don't have an entry in the WITH column will not need to be retrofitted immediately.

Usage of HMM Evidence

Michelle Gwinn-Giglio, TIGR

This issue first arose about a year and a half ago and refers to cases where HMM models, or other models such as a neural network model, are used to determine sequence similarity. Can these models be cited in the WITH field? TIGR has been using them in this way from the start, but there seemed to be some concern that this was not okay.

One of the concerns was about whether or not the models and their IDs are stable. The models may change over time, but the ID associated with a given model is stable.

For CBS (Center for Biological Sequence Analysis, Copenhagen) models, the names of the models correspond to their IDs and these models are available to anyone who may want to test their sequences against them.

There is general agreement that it is okay to use the CBS models in the WITH field for ISS annotations. SignalP, another program that uses both an hmm and a neural network model, will required a new, specific abbreviation.

WITH Column Working Group Report

Harold Drabkin, MGI

What entries are acceptable for the WITH column? WITH is an evidence code qualifier that consists of a searchable database ID and that relates to the evidence codes in the following way:

ISS – something similar to the gene product

IMP – could be an allele

IPI – whatever interacts with the gene product

IGI – whatever interacts with the gene

IC – the informative GO term

IEA – under discussion

IDA – under discussion

There was some discussion about whether there is a legitimate entry in the WITH column for IDA annotations. Some curators have suggested that a target or drug could be added to the WITH column for IDA, but the general consensus is that this is not in the spirit of what we want to capture in the WITH field and therefore, we won't allow entries in the WITH column for IDA annotations at this time.

What to put in the IEA WITH field generated much discussion. We want users to be clear on what the WITH field signifies, but at present, a number of different IDs, from different algorithms, are used here, for example, interpro2go, spkw2go, ec2go, etc. Are these IDs sufficient for users to understand the relationship between the gene product and the WITH field?

There is a general consensus that the combination of a reference and an appropriate ID in the WITH column is enough information for users to figure out what the ID means for IEA annotations. Although the IDs may vary, the common relationship between them and the gene product is that the ID is an 'object that the gene product matched when the algorithm was run.'

ACTION ITEM #24: GO will disallow WITH column entries for IDA annotations.

ACTION ITEM #25: Document that WITH column entries are essential for all match-based methods of annotation and that a valid database ID is required for IEA WITH entries. The WITH column won't be mandatory for tools that just predict GO annotations, as the reference entered will describe the tool/algorithm used.

What is inferable from RCA Evidence?

Linda Hannick, TIGR

In some cases, large-scale experiments make statements about function from their experiments, such as physical interactions. Many groups feel comfortable annotating to a process based upon large-scale experiments, but not to molecular function terms.

Should the GO have a policy on what types of annotations can be made using large-scale data and the RCA evidence code?

One specific example that Val put forward to the mailing list concerned a cerevisiae paper where authors made function assertions based upon analysis of large-scale interaction data. The data from this paper was then used to annotate to MF terms.

Subsequently, SGD has reviewed these annotations and removed any annotations to MF terms. But can, and do, we want to make a statement that MF annotations can never be made from computational analysis?

The general consensus is that it is difficult, at this point, to say that curators should never annotate to MF using the RCA evidence code, especially since the body of work that relates to this question is still relatively small.

ACTION ITEM #26: Add more examples of how the RCA evidence code can and should be used for GO annotation based on published literature to date.

Large-Scale vs Small-Scale, but Same Evidence Type

Eurie Hong, SGD

Large and small-scale experiments can, and often do, result in annotations that use the same experimental evidence code, such as IPI or IMP. Should GO somehow try to differentiate the two types of experiment for users?

The discussion was centered around whether large-scale vs small-scale is really the crux of the matter. A large-scale experiment is not synonymous with a poor quality experiment, and likewise, small-scale does not equate with high quality. Is the real issue one of experimental method, rather than scale? Are we trying to inform users about the potential pitfalls of different experimental methods and is that within the scope of GO? Would doing so require expanding the current evidence codes?

ACTION ITEM #27: No conclusion about how to distinguish large- vs small-scale experiments was reached. People are encouraged to keep thinking about this issue which clearly needs more discussion.

GO Reference Collection

Midori Harris, GOEO

The GO reference collection, a collection of descriptions of methods that groups use for ISS, IEA, and ND evidence codes, needs to be more visible and easier to use. This information is immensely useful, but not even all GO Consortium members knew that it existed even though it currently has a home in the GO CVS repository.

The **AMIGO WORKING GROUP** agreed that it is really important to make these references more visible and there is also general agreement that different groups who are using the same process should be citing the same references, thus avoiding duplication in these references.

ACTION ITEM #28: The AMIGO WORKING GROUP will implement a strategy to incorporate and display the contents of the GO references.

ACTION ITEM #29: Existing GO references will be examined to check for and eliminate redundancy. [Midori Harris and Karen Christie]

Ontological Content, continued

The Use of Sensu in the GO

Chris Mungall, BDGP

There are currently two uses of sensu in the GO. The original use of sensu was as a linguistic qualifier or linguistic disambiguator meant to distinguish cases where the same word referred to different types that are not related, for example, ‘bud’ or ‘trichome’. A second use has arisen, though, which is that of a type qualifier. This use is legitimate, but shouldn’t be lumped together with the original use.

Organism type specificity is a genuine challenge for the GO, but sensu has been wrongly recruited to fix this. There are two problems: 1) We have conflated the meaning of sensu resulting in lack of precision, and 2) We have added taxon IDs which isn’t quite right for this use.

The proposed solution is to retain sensu for its original purpose, that of a linguistic qualifier. Its interpretation then becomes: ‘as used in the XYZ’ community. Taxon IDs would not be required, as the use would not be restricted to organism-specific communities. Biochemists or cell biologists working on the same organism may talk about the same term differently.

A second part of this solution is to introduce a new relation for genuine organism-specific terms (contextual parts). This involves the idea of contextual synonyms, exact synonyms with a context qualifier. This allows users to configure particular applications, e.g., a user could configure to use the plant context exact synonyms, but we don't need to be as specific here as the actual taxon ID. The context should be the insect community, plant community, etc, in all places where this occurs. Context, in these cases, referring to sociolinguistic context. Further, the use of sensu would not be inherited as it is right now. There would be no need to carry its use through.

For other situations, we do want to introduce genuinely different biological subtypes.

In this case Chris proposes adding an 'in_organism' relationship. An example would be that of 'thylakoid', where we would have 'thylakoid, in cyanobacteria' as a subtype of 'thylakoid' instead of thylakoid (sensu Cyanobacteria).

We could use the NCBI taxonomy as our organism ontology to make the relationship between the term and the taxon. These could be put in the .obo file, but this would mean that the .obo file would no longer be an insular, standalone file. Alternatively, we could keep the links in a separate file.

What about cases where something is present in many organisms, such as all gram negative bacteria? What would we do then? Would we need to great a new ID? We do allow for combinations in the .obo file, so there are ways to address this. We don want to use valid taxon IDs.

ACTION ITEM #30: Chris Mungall and Jen Clark will discuss the different aspects of changes to our use of sensu, write documentation on this, and implement the new strategy. This change will then be announced to the community.

Demonstration of Cross-Products between GO and CO

Chris Mungall, BDGP

Chris showed a demo of how cross-products terms could be represented in OBO-Edit and in AmiGO using the example of 'larval locomotory behavior,' which exists as a diamond in the current tree. First, he took the term and created a logical, cross-product definition, locomotory behavior during larval stage, using the larval stage definition from the FlyBase anatomy/dev stage ontology. The generic term here is locomotory behavior, and its differentiating characteristic is that it occurs during the larval stage. Having the logical definition makes it possible to disentangle the diamond.

Another example is that of the term 'differentiation,' a term that implicitly refers to the cell-type ontology. Performing a first-pass using the OBO software, we can make cross-product terms such as 'osteoblast differentiation,' meaning: cell differentiation that has_participant osteoblast. We need a better relationship, though, and don't want to use a relationship from the cell ontology because we want a relationship between a cell and a process, ie between a continuant and a process, and the cell ontology relationships are between continuants.

Are the definitions created necessary and sufficient? Yes, the OBO intersection_of tag indicates necessary and sufficient.

Adding these terms will allow for querying GO by cell types, but we will need OBO 1.2 to be able to represent them.

One concern: the example used the term ‘larval locomotory behavior’ but lots of species have larvae. This highlights the need for more general anatomy and developmental stage ontologies and provides one reason why cross products between GO terms and cell type are being tackled first – it’s much less species-centric.

When looking at cross-products in OBO-Edit, annotators can see the creation of new terms based upon is_a relationships in the cell ontology. For example, ‘macrophage cell activation’ would have an is_a child ‘microglial cell activation’. OBO-edit infers that this child term is okay, but curators would be able to check yes or no to confirm that the term makes sense and should remain in the ontology. Reasons for not accepting a term might include: 1) The relationship between the cells is not correct, or 2) The process may not actually occur. OBO-Edit will allow for curators to check the correctness of terms before they get added and will allow for changes to the computable definitions within the cross product window. There will be ways to this in bulk, if needed.

Where do we go from here? We are ready to start putting these terms into the .obo file, and will need to use obo 1.2 format to accommodate these relationship types. We will also need to have a way to hide these from the average user until the world is ready for this.

Slightly off topic discussion ensued about splitting the gene_ontology.obo file into edit and general versions, generally agreed upon. Chris and John agreed to experiment with the edit version.

Slightly off topic discussion ensued about post-composition in the gene-association files, for example a new column in the file format: slots: eg. OBOREL: located_in [MA:liver] OBOREL: has_primary_participant [FBbt: Y_neuron].

ACTION ITEM #31: Chris and John will develop a plan for implementing cross-products between GO terms and the cell-type ontology. Part of this plan will involve splitting the Gene Ontology .obo file into an edit version that would be filtered into a gene ontology .obo file still using obo-edit 1.0. Also, will need to consider what to do with explicitly stated relationships (relevant to earlier discussion of acted_upon) and work out the specifics of what should happen if curators need to provide feedback on relationships within the cell ontology (eg, contact Oliver Hoffman). In parallel, we should also come up with a plan for AmiGO development and user education.

Working Group Reports

The AmiGO Working Group

Eurie Hong, SGD

This presentation addressed who the AmiGO working group is and how they should interact with the rest of the group. The AmiGO working group is open to all consortium members and there is a major domo mailing list for people to sign up for if they would like to participate. (Works the same way as other GO mailing lists.)

Currently, AmiGO operates under a three-month release cycle broken up approximately by : one to two months of developing mock-ups and specifications and one month of a testing cycle to identify bugs and tweak features, followed by production and installation at Stanford. This implies four releases a year, but that number could be higher, depending upon what issues arise during each release cycle.

Questions: At what point should users become involved with AmiGO development? Should we have focus groups for AmiGO prototypes, since there are different types of users? How would we identify users that are not curators?

The consensus seems to be that it's important to get people from outside the consortium to provide feedback on AmiGO development. We could perhaps find these people via SourceForge entries or from a GOC newsletter. We could also tap into other communities that use AmiGO, such as the plant ontology consortium.

If there are proposals for changes, should they be sent out to the consortium for a period of feedback? Agreed that it would be helpful to send an email indicating which issues are being addressed for the next release and which remain open in SourceForge. This would allow people to see the current priorities and make specific comments or requests, if needed.

How should we handle user support? The email addresses off of the AmiGO web pages could go to a more directed group of people to make sure that they go to thegroup that will definitely deal with them.

News on upcoming releases:

Current Release

- Main fix – viewing term siblings and parents
- Ontology filter for terms
- Obsoletes sorted to the bottom of the page

Next Release

- Improved searching and filtering (search box on every page)
- Reorganization of term search results to look more like gene search results
- Try to make it obvious to users where they will end up when they click on something

High Priority Items

- Dealing with tangled paths to root
- Displaying cross-product terms
- Displaying IEA annotations (takes a lot longer to load the database with IEAs)

Final comments: If any MOD users have questions about AmiGO, please send them on to the working group, and please join if you are interested in AmiGO. One note: you don't have to be on the AmiGO mailing list to email the group.

OBO-Edit: The OBO-Edit Working Group

John Day-Richter, BDGP

The OBO-Edit Working Group was formed to address bugs in the OBO-Edit software and to give direction to OBO-Edit development. The working group currently consists of members of every group in the consortium.

A user survey listed several new features as having high priority. First on this list was a user's guide, followed by a basic annotation tool to have some way of associating a gene product ID with a term, OBOL integration, and bug fixing. Other desirable features include usage movies, FAQs, how-tos, and public webinars open to the world (see below). Midori requested a future directions email so that the working group could help prioritize the rest of the user requests.

A user's guide with greater documentation will be released separately from software releases, since in the past, having more documentation wasn't necessarily grounds for issuing a new release. Members of the working group have signed on to co-write and edit various sections of the documentation.

Another goal has been to have a more regular release schedule with input from the working group required about when to make a beta version an official version. The group still needs to work out what criteria need to be met before a release will be deemed official, but getting to another official version of OBO-Edit should be a high priority.

Once an official version exists, however, the official editing software will not change until another official version is released.

Improved communication in the form of a remote tutorial was tried in February, using different technologies like VNC, Gizmo, and IRC text chat, but there were some technical problems with VNC and Gizmo that will need to be addressed for the success of future webinars. Complete transcripts of the IRC commentary from this first training session are coming.

ACTION ITEM #32: To alleviate technical problems with remote tutorials, which are likely more cost-effective than flying everyone to a particular place for training, GO will investigate retaining the services of a company for hosting future webinars. John will investigate various options available to us.

Once a new version of OBO-Edit has been released, should users get it immediately?

In other words, what quality control measures are in place? John recommends that, when a new version is released, annotators use the new version but don't commit their files.

He has written a test suite that must be passed before each new release. If there is a new bug report for a given version, then a test is added which must be passed before the next release is issued. All working group members have a list of things that they must test because John can't run all possible GUI tests. This has helped a lot.

ACTION ITEM #33: The OBO-Edit User's Guide, which is available in OBO-Edit in the docs directory, will also be made available on the GO website. John will talk to Mike Cherry about how to get this done.

ACTION ITEM #34: The OBO-Edit working group should come up with a time line for release of the official version.

Transition to OBO-Edit Version 1.2

Mike Cherry, SGD

We will need to make an announcement to the user community regarding the switch to OBO 1.2. We should pick a date for switching over and then be very deliberate about this change, giving users plenty of time so that they can fix any scripts that use the file.

What is the best vehicle for announcing the change? The GO home page, the GO Friends email list, a monthly GO newsletter (see below)?

When we switch to obo 1.2, are we going to generate both and old and new obo file?

Yes, but perhaps the gene_ontology.obo should be in the new format and the old file should be renamed something else.

We will also want to make announcements when there are other big changes to the file, such as changes to the use of sensu or having links between process and function.

The need to make specific versions of the gene_ontology.obo file was also discussed: Maybe we need to make specific versions, like when there is a substantial change in format/structure, eg., when a new relationship type is brought in, say make version 2 when the 'sensu' changes are made. OBO 1.2 will have the functionality for meta-data (like tracker ids in sourceforge).

How best to communicate changes in the GO? One suggestion was to bundle the monthly release with a report/release notes that would document any changes. Currently, a semi-automated monthly report is generated. These release notes would include the results of the monthly report, along with a human-readable summary that highlights any major changes in the file. For the monthly report, Rama has suggested that SourceForge IDs be added so that changes came be traced back to the original request.

ACTION ITEM #35: The monthly archive of the GO will also include Release Notes. These notes would include the output of the monthly report script, including the relevant SourceForge IDs, as well as human-readable text that summarizes significant additions or changes to the GO file. [GO Editorial Office]

ACTION ITEM #36: Form a NEWSLETTER WORKING GROUP to develop a GO newsletter that will provide a vehicle for making announcements about a number of GO-related issues, such as major changes to GO, our meeting schedule, what decisions were made at meetings, GO workshops and tutorials, and maybe even the new-term-of-the-month or GO tip-of-the-day. Determine the proper target group for the newsletter. The first newsletter should go out before the switch to obo 1.2. [Eurie Hong, Jane Lomax, John Day-Richter, GO PIs, and other to-be-determined volunteers]

Production Priorities

Mike Cherry, SGD

The production priorities for the GO include:

1. Genome protein sets (gp2protein)
2. User support
3. Production systems change
4. Database changes
5. On-the-fly species annotations

Genome protein sets

Genome protein sets (gp2protein files) should include all proteins in the organism, not just the proteins that are annotated. We also want to put together a fasta file from each organism, which is important to have as the input file for InParanoid and other analyses.

What would the defline format be? The standard fasta format.

InParanoid actually wants a one-to-one mapping file of gene to protein, but for the reference genomes we may need to have all ID mappings between genes and protein products, as well as other ID mappings such as Uniprot, IPI, CCD, MOD IDs, GI, protein_id, ref_seq, etc. The protein set should include predicted proteins, but we may want to make a separate file for ncRNAs. How often should these files be updated?

Updates could coincide with each MODs release cycle.

ACTION ITEM #37: Mike will send an email out to let groups know what the genome protein set fasta file format should be, specifically, what IDs should be included.

Mike has also been tracking experimental evidence codes (IMP, IDA, IPI, IGI, and IEP) for each organism, including information about the percentage of genes from each organism that has been annotated to an experimental code. One issue is that we need to figure out a way to report how many genes are actually in each organism.

ACTION ITEM #38: Each MOD should supply the current number of genes for their organism. Rex will then determine exactly what should be included in the gene-association file, ie the total number or the number of genes split out based upon protein coding genes and ncRNA genes.

User Support

In the past, user support has included email lists for users to report problems. Do we need to make changes to where these emails go? Do we want specialized lists or some more generic lists?

Currently there are the following e-mail lists: GO, GO-database, GO-webmaster, GO friends, GO-In, GO-Top.

For users, it is probably easier to have one email address, but we would then need to have a system for replying to the emails, dealing with the problem, and tracking the resolution. In essence, a more formalized system for user support. As always, if a question comes into the go friends email list, then the first person to whom it's relevant should still answer the email.

ACTION ITEM #39: Set up a formalized system for coordinating user help. This will include rotating responsibility for reading the emails and answering or forwarding them to the appropriate person or group. To facilitate dealing with issues raised in the emails, each group should send a list of contacts to the GO Editorial Office. Information about the new user support system will be added to the newsletter. [Mike, Eurie, GO Editorial Office]

Production Systems Changes

The GODB and AmiGO currently run off of SGD servers. GO will be moving to a cluster environment, where there will be multiple database servers and GO HTML servers.

Last year, we started building the GO Lite (minus IEA) database three times a week. This mean that AmiGO is always using 2-4-day old data. More cluster nodes are now on order which will allow for daily updates. Including IEAs, however, would increase the time required to build the database.

We will now have CVS available through the GO web site which will allow users to get back to an old version of GO on the web.

Other potential production changes include: building an updating script so that the database would not always have to be rebuilt from scratch every time and switching the GO database schema to Chado, a GMOD product. AmiGO does not work off of Chado yet, but Chris has written a schema to simulate the GO schema over Chado. Letters of support needed.

We would also like to consider adding history tracking of GO terms and IDs.

And, lastly, we need to consider what types of files need to be archived.

Survey of GO File Downloads (from 44 replies)

The most popular download is the MySQL database. This is followed by the GO flat file (all users currently sitting in this room – no non-members said that they're using the flat file). If we switch to a different database schema, we will need to make sure that we bring the MySQL users along.

On-the-Fly Species Gene Annotation

This is mainly useful for multi-species gene association data sets, such as those that come from GOA or Uniprot.

Daniel and Evelyn commented that this is trivial to maintain and that GOA already has a web interface for that, since people like to download specific data sets. Rolf Apweiler is keen to allow people to pick and choose which data set they want to see.

An on-the-fly viewer could be done at Stanford.

Outstanding Issues

Content issue regarding protein complexes (Midori).

Names and synonyms (Jen).

IPI chain of inferences (David).

This will also be a good issue for annotation camp.

Consortium members are urged to think about annotation issues for the upcoming annotation camp in July.

Final Comments

Judy Blake, MGI

The GO grant has been submitted for renewed funding, but we will be adding supplementary material in May. Any short-term items that may have user impact could be added to the supplementary material.

April 17th is the date set for receiving the first reports from the **WORKING GROUPS**.

Judging from the size of this meeting's agenda, it would be a good idea to have anothermeeting in ~ six months, instead of waiting a whole year. The next GO meeting will thus likely be in November 2006 in either Hinxton or Marseille.

Action Items

ACTION ITEM #1: Very seriously consider removing the word 'activity' from the molecular function terms and consider renaming the molecular function ontology.

(PRE)ACTION ITEM #2: Need to work out the balance of power/responsibility between the GO office and annotator/ontology developers to complete SourceForge items.

ACTION ITEM #3: Begin to coordinate processes for reference genomes to start setting priorities and tracking progress. Acquire and distribute lists of genes for curation focus and set-up fortnightly discussions. [Point Person: Rex Chisholm]

ACTION ITEM #4: Any other ideas for shared curation software, please forward to Chris Mungall.

ACTION ITEM #5: Consider what, if any, are the repercussions of renaming the cellular component to cell level entity?

ACTION ITEM #6: Change new terms ending in '…component' to now end in '…part.' [Jane Lomax]

ACTION ITEM #7: Jane will send an is_a complete cellular component ontology to the GO list.

ACTION ITEM #8: Take the new molecular/cellular/multicellular arrangement of the biological process ontology, try it out, and see how it works. The WORKING GROUP for this will include: Chris, Jane, David, Alex, Michelle, Rex, and Val. Jane will make a .obo file so that people can look at this. The group will also need to create a document outlining their philosophical approach and the results. Should there be a development site to help with working on this?

ACTION ITEM #9: Make actin polymerization a function term.

ACTION ITEM #10: Amelia, Chris, and David (and Barry, if he's available) should get together and come up with a single proposal for making connections between function and process. This will likely include writing new definitions for function and process.

ACTION ITEM #11: Add detailed documentation on dual-taxon annotation, announce this to the annotators' mailing list, include the info in annotation camp discussions, and work on developing the ontology. [Candace Collmer, Jane Lomax, Amelia Ireland, others?]

ACTION ITEM #12: Candace and Trudy will send examples and relevant references to the annotation list as they come up, so that we can consider these on a case-by-case basis. [Candace Collmer, Trudy Torto-Alallibo]

ACTION ITEM #13: Develop a new policy for communicating about term obsoletions. The person proposing obsoletion should get in touch with the annotating groups (using contact information from the gene association file) informing them that the term is under review while also soliciting input and suggestions on what changes to make. This could be scripted.

ACTION ITEM #14: Reinforce the policy that we will no longer obsolete a term just because the definition has changed or because annotations are thought to be bad or incorrect.

ACTION ITEM #15: Explore the proper technical solution for establishing a mechanism to notify GO users when a term has changed, or rather, when we are thinking of changing a term.

ACTION ITEM #16: Revisit obsolete terms to see which can be merged with current terms. Decide if the IDs of obsoletes could be made secondary IDs to the currently existing term.

ACTION ITEM #17: John Day-Richter will talk to the GO Editorial Office to find the best way to implement these changes and suggestions in OBO-Edit.

ACTION ITEM #18: Add a term creation date to the .obo file.

ACTION ITEM #19: TAS is no longer considered a useful evidence code and will not be used in any consistency measures of reference genome annotation. Since part of the idea of the reference genomes is to provide a source of IEA annotations for other groups, we strongly encourage reference genome annotators to not use TAS, and instead use experimental evidence codes whenever possible. The GO documentation should also state this in a clear fashion.

ACTION ITEM #20: Add to the documentation that it is okay to use experimental evidence codes for identical/similar gene products from different strains of the same species.

ACTION ITEM #21: Individual groups can collect this [acted_upon] data knowing that, in the future, GO will present this type of relation. Chris, John, Sue, Candace, and Michelle will form a WORKING GROUP to come up with a proposal for how to implement this. Note that this type of annotation will not be a core requirement, but that GO will facilitate its display if groups want to do this.

ACTION ITEM #22: Create a new evidence code IGC, Inferred from Genomic Context. The precise definition of this code and procedures for annotation (what to put in the WITH column) will be hashed out and added to the documentation. [Michelle Gwinn and Matt Berriman]

ACTION ITEM #22: Update the documentation on using the ISS evidence code to emphasize that annotators need to enter something in the WITH field. In the case of gene products, there must be an experimental evidence code for that gene product which supports the annotation, i.e., we don't want to have circular ISS annotations. Uniprot IDs, ref_seq IDs, or individual MOD gene IDs would be okay to use in the WITH column. Old ISS annotations that don't have an entry in the WITH column will not need to be retrofitted immediately.

ACTION ITEM #24: GO will disallow WITH column entries for IDA annotations.

ACTION ITEM #25: Document that WITH column entries are essential for all match-based methods of annotation and that a valid database ID is required for IEA WITH entries. The WITH column won't be mandatory for tools that just predict GO annotations, as the reference entered will describe the tool/algorithm used.

ACTION ITEM #26: Add more examples of how the RCA evidence code can and should be used for GO annotation based on published literature to date.

ACTION ITEM #27: No conclusion about how to distinguish large- vs small-scale experiments was reached. People are encouraged to keep thinking about this issue which clearly needs more discussion.

ACTION ITEM #28: The AMIGO WORKING GROUP will implement a strategy to incorporate and display the contents of the GO references.

ACTION ITEM #29: Existing GO references will be examined to check for and eliminate redundancy. [Midori Harris and Karen Christie]

ACTION ITEM #30: Chris Mungall and Jen Clark will discuss the different aspects of changes to our use of sensu, write documentation on this, and implement the new strategy. This change will then be announced to the community.

ACTION ITEM #31: Chris and John will develop a plan for implementing cross-products between GO terms and the cell-type ontology. Part of this plan will involve splitting the Gene Ontology .obo file into an edit version that would be filtered into a gene ontology .obo file still using obo-edit 1.0. Also, will need to consider what to do with explicitly stated relationships (relevant to earlier discussion of acted_upon) and work out the specifics of what should happen if curators need to provide feedback on relationships within the cell ontology (eg, contact Oliver Hoffman). In parallel, we should also come up with a plan for AmiGO development and user education.

ACTION ITEM #32: To alleviate technical problems with remote tutorials, which are likely more cost-effective than flying everyone to a particular place for training, GO will investigate retaining the services of a company for hosting future webinars. John will investigate various options available to us.

ACTION ITEM #33: The OBO-Edit User's Guide, which is available in OBO-Edit in the docs directory, will also be made available on the GO website. John will talk to Mike Cherry about how to get this done.

ACTION ITEM #34: The OBO-Edit working group should come up with a time line for release of the official version.

ACTION ITEM #35: The monthly archive of the GO will also include Release Notes. These notes would include the output of the monthly report script, including the relevant SourceForge IDs, as well as human-readable text that summarizes significant additions or changes to the GO file. [GO Editorial Office]

ACTION ITEM #36: Form a NEWSLETTER WORKING GROUP to develop a GO newsletter that will provide a vehicle for making announcements about a number of GO-related issues, such as major changes to GO, our meeting schedule, what decisions were made at meetings, GO workshops and tutorials, and maybe even the new-term-of-the-month or GO tip-of-the-day. Determine the proper target group for the newsletter. The first newsletter should go out before the switch to obo 1.2. [Eurie Hong, Jane Lomax, John Day-Richter, GO PIs, and other to-be-determined volunteers]

ACTION ITEM #37: Mike will send an email out to let groups know what the genome protein set fasta file format should be, specifically, what IDs should be included.

ACTION ITEM #38: Each MOD should supply the current number of genes for their organism. Rex will then determine exactly what should be included in the gene-association file, ie the total number or the number of genes split out based upon protein coding genes and ncRNA genes.

ACTION ITEM #39: Set up a formalized system for coordinating user help. This will include rotating responsibility for reading the emails and answering or forwarding them to the appropriate person or group. To facilitate dealing with issues raised in the emails, each group should send a list of contacts to the GO Editorial Office. Information about the new user support system will be added to the newsletter. [Mike, Eurie, GO Editorial Office]