Paul Thomas DRAFT December 29, 2009

LEGO: building biological function modules from molecules to populations

The Gene Ontology development effort has resulted in a very rich ontology. However, annotations remain very simple and limited in their expressiveness, being of the form:

`Gene_product` *`relation`* `GO_term`

This proposal aims to enable much richer annotations. Biological systems are modular at many levels. Within a particular protein domain, there may be multiple different sites that are coupled to each other to perform a particular function, e.g. a catalytic active site and a distinct but coupled (allosteric) binding site that regulates the catalytic activity. At a higher level, there is functional coupling between different domains of the same protein (e.g. the ligand binding domain and protein kinase domain of a transmembrane protein kinase receptor) or between different subunits of a macromolecular complex (e.g. the ribosome). At an even higher level, molecular interactions can define a pathway that can be used or reused (coopted) in multiple different processes (e.g. the ubiquitin-dependent proteolysis pathway or JAK-STAT pathway). The GO comprises an extensive vocabulary for molecular functions, biological processes and cellular components. The goal of GO2 is to define each GO term by a combination of modular terms, and enable extensible representation of this biological modularity: how elemental molecular interactions are combined in different ways to produce compound molecular functions, how molecular functions are combined to produce processes, and how processes are combined to produce larger processes.

GO2 also has a very practical application. Human curators play an essential role in the utility of the GO, by reading papers and manually assigning ontology terms to genes based on these papers. Curation is time-intensive and is the rate-limiting resource for the GO project. The GO2 proposal is specifically aimed at making the most of the curatorial resources by 1) increasing the amount of information captured by the curator in the same amount of time, 2) allowing curators to extend the ontology automatically simply by combining modular terms in an annotation, 3) streamlining and standardizing the curation process by pre-composing modules that either suggest possible additional annotations, or deduce additional annotations from logical definitions.

It is important to note that this proposal is an extension of the current annotation process. *Existing annotations will remain valid*, though in many cases it will obviously be possible to improve the existing annotations by utilizing the new extensions.

There are several critical elements to the proposal:

1. Modularizing molecular function
2. Substrate specificity for molecular function and some biological processes
3. Enabling nested annotations

4. Annotation of classes other than gene products: annotation of molecular functions to biological processes, since functions are executed during processes; annotation of complexes to molecular functions and biological processes; annotation of biological processes to larger biological processes. (This last one is already taking place, but as defined subclasses, not as annotations.)

Within an overall biological process, a gene product can have a biochemical function (which may be compound), a specificity (for each biochemical function) and a molecular role.

An important aspect of this proposal is an extension of the GO that allows the curator to express how a gene product participates in a biological process—i.e. how multiple molecular functions are executed within a biological process, and how sub-processes contribute to larger processes. Function-process links that have been considered in the GO so far are the "easy" ones—namely those that are either so general they cover all specificities, e.g. protease activity (function) and proteolysis (process), or those that have molecular specificity pre-composed into a GO term, e.g. fructose-bisphosphatase aldolase (function) and glycolysis (process).

1. Modularizing molecular function

The current GO molecular function classes can be expressed as a combination of one or more of the following three elements:

1. biochemical function: the type of molecular interaction; the physical mechanistic operation (e.g. kinase)
2. a specificity of function: substrate or molecular interactor; the operand, the thing which is operated upon (e.g. the substrate of adenylate kinase is AMP). This can be taken from a molecule ontology (e.g. PRO, ChEBI) or sequence ontology (SO), and in many cases should correspond to a gene product or gene.
3. molecular role: the reason, or biological "purpose" of the operation as a component within a biological system; the molecular "effect" of the operation

However, these three elements are not clearly separated and combined inconsistently. For example, substrate specificity is included for many metabolic enzymes, but not for enzymes that operate upon proteins. We propose to separate these three elements, and be clear about definitions. First, a molecular function is the function performed by a stable tertiary (gene product) or quaternary structure (complex) due to *direct physical interactions*, and in the case of compound functions, occurring within a very short timeframe. Each elemental biochemical function has a molecular specificity—we propose to capture this using one or more slots. Molecular role and biochemical function will be separate lineages within the ontology, which are combined using Boolean operators (most commonly AND). A

molecular role is the purpose the function fulfills, while a biochemical function is the physical mechanism by which the role is carried out.

1.1. Biochemical function and specificity

1.1.1 Elemental biochemical functions

We first define "elemental biochemical function" as a biochemical function that is performed by the stable tertiary or quaternary structure. There are only a few main classes of elemental function:
Binding activity
Catalysis activity
Electron carrier activity
Photon capture activity

Each elemental function has a "specificity slot", the thing it operates upon. Specificity is generally a specific molecule (e.g. specific protein or small molecule or ion) but it can also be a class of molecule, or a gene, or an electron or photon. The combination of function and specificity is often sufficient to determine the product of the operation. For instance, the product of stable binding between two proteins is a specific complex; the product of a protein kinase function on a particular protein is a phosphorylated protein. In some cases, it is not specific enough, e.g. phosphorylation of a protein can occur on different amino acids, but for now having an operand slot alone might be sufficient. For catalysis we could add slots to define the product more precisely.

1.1.2. Compound biochemical functions

We next define "compound biochemical function" as two or more elemental functions coupled together. For example, consider a receptor protein kinase. In our proposed modular scheme, *receptor* would be its molecular role (see below), and its biochemical function, something like *transduction of binding activity to catalytic activity*, is a compound function that couples two elemental biochemical functions: an extracellular ligand binding activity and an intracellular catalytic activity. In this example, binding positively regulates the catalytic activity. The compound function links together (function-function links) elemental functions that must be performed *together* in order to perform the molecular role.

1.2. Molecular role

The gene product or complex also has a molecular role, i.e. the role it plays as a component of a biological system. It should be possible to express the role in terms of relations to biological processes (or even other molecular functions). An example is the function of copper chaperone. In our proposal, the biochemical function would be *copper ion binding*, while a molecular role is *sequestering of metal ion* (a biological process term).

An important molecular role is to regulate the biochemical function of another tertiary/quaternary structure, by modifying its physical structure through binding

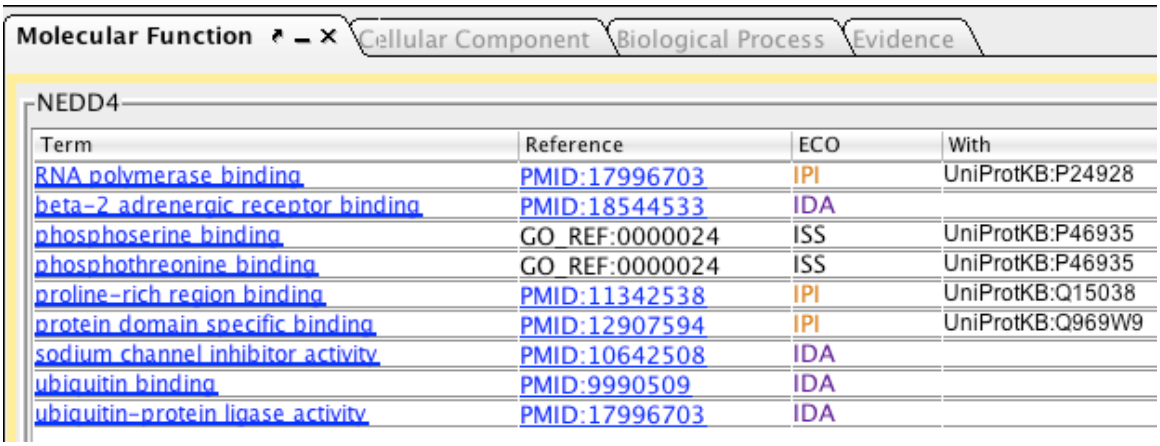
or catalysis (elemental functions). For example the GO function *Enzyme regulator activity* is a protein that binds to another protein (or chemically modifies it using an enzymatic activity), resulting in the regulation of the enzymatic activity of that other protein.

2. Function-process relations, and subprocess-process relations

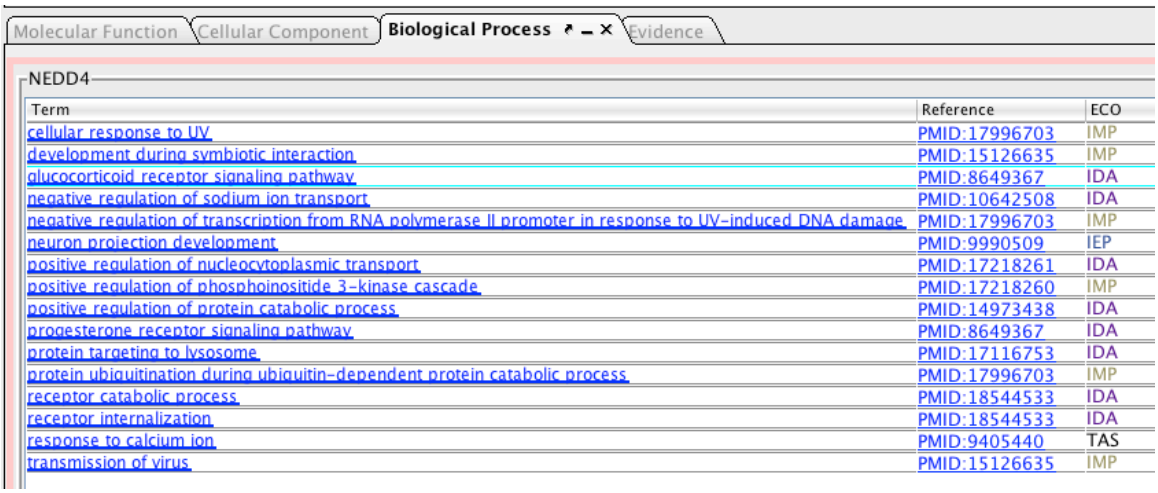
The definition of GO biological process is the execution of one or more molecular functions. Ideally, then, the GO should represent how functions are executed during processes, and how subprocesses are executed within larger processes. Biology is modular, and modules are often reused or "coopted", but with different specificities of one or more of the participating molecular functions, or under different conditions (e.g. different cell types). However, the current GO annotation process does not capture all of the information necessary to represent these modular relationships. GO annotations currently represent the biological processes that genes and gene products participate in, but not which functions (including specificities) are actually executed during these processes. In addition, the GO captures some subprocess-process relations, but these are not implemented as annotations, but rather as part_of relations that are not supported by literature references. The new process would address these shortcomings by allowing **nesting within annotations**: molecular function annotations can be nested inside biological process annotations, and biological process annotations can be nested inside other biological process annotations.

3. Examples of how this proposal will address curator needs

The figure below shows the molecular function and biological process annotations for NEDD4. Curators have done an excellent job of representing the functions of NEDD4 as well as possible given the constraints of the current annotation process. However, with just a few extensions of the current process, they would have been able to express the biology much more accurately.



Term	Reference	ECO	With
RNA polymerase binding	PMID:17996703	IPI	UniProtKB:P24928
beta-2 adrenergic receptor binding	PMID:18544533	IDA	
phosphoserine binding	GO_REF:0000024	ISS	UniProtKB:P46935
phosphothreonine binding	GO_REF:0000024	ISS	UniProtKB:P46935
proline-rich region binding	PMID:11342538	IPI	UniProtKB:Q15038
protein domain specific binding	PMID:12907594	IPI	UniProtKB:Q969W9
sodium channel inhibitor activity	PMID:10642508	IDA	
ubiquitin binding	PMID:9990509	IDA	
ubiquitin-protein ligase activity	PMID:17996703	IDA	


Molecular Function | Cellular Component | Biological Process | Evidence

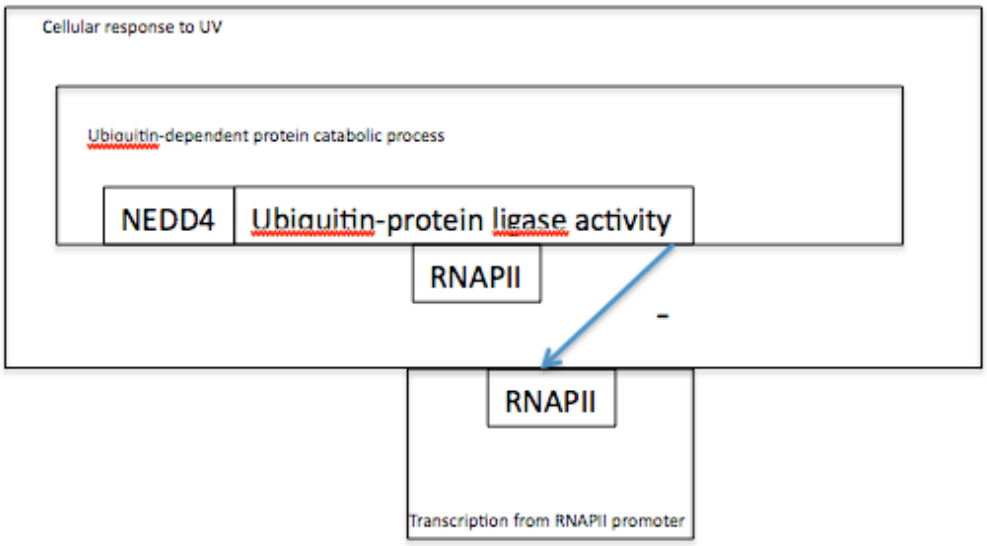
NEDD4

Term	Reference	ECO
cellular response to UV	PMID:17996703	IMP
development during symbiotic interaction	PMID:15126635	IMP
glucocorticoid receptor signaling pathway	PMID:8649367	IDA
negative regulation of sodium ion transport	PMID:10642508	IDA
negative regulation of transcription from RNA polymerase II promoter in response to UV-induced DNA damage	PMID:17996703	IMP
neuron projection development	PMID:9990509	IEP
positive regulation of nucleocytoplasmic transport	PMID:17218261	IDA
positive regulation of phosphoinositide 3-kinase cascade	PMID:17218260	IMP
positive regulation of protein catabolic process	PMID:14973438	IDA
progesterone receptor signaling pathway	PMID:8649367	IDA
protein targeting to lysosome	PMID:17116753	IDA
protein ubiquitination during ubiquitin-dependent protein catabolic process	PMID:17996703	IMP
receptor catabolic process	PMID:18544533	IDA
receptor internalization	PMID:18544533	IDA
response to calcium ion	PMID:9405440	TAS
transmission of virus	PMID:15126635	IMP

NEDD4 has several molecular functions and participates in several biological processes, so there is no way for an end user of these annotations to connect them into the relevant biological pathways. It is not possible from the annotations to disentangle which molecular functions are used in which processes, and how subprocesses are used within larger processes.

Example 1: NEDD4 and response to UV

For instance, one paper (PMID 17996703) demonstrates that NEDD4 *ubiquitin ligase activity*, in the context of a specific subprocess (biological module) *ubiquitin-dependent protein catabolic process*, marks RNAPII for proteasomal degradation, to negatively regulate gene transcription in response to UV damage to DNA. Because of the constraints on the current annotation process, the biological reality cannot be adequately expressed, and the curator has resorted to annotating the gene to numerous apparently separate functions and processes. First, the specificity of the ubiquitin protein ligase function cannot currently be expressed, so the curator has attempted to capture it in a separate annotation, *RNA polymerase binding*, which falls short of the biological reality. Second, this molecular function is not connected to the biological processes within which it is used: *protein ubiquitination during ubiquitin-dependent protein catabolic process,* and *cellular response to UV.* Third, one of these processes is actually used as a sub-process within the other: the entire module of *ubiquitin-dependent protein catabolic process*, not just the protein ubiquitination step, is used as part of the overall *cellular response to UV*. If the annotation protocol were extended as proposed, the curator would capture these using **specificity slots** and **nested annotations** as follows:



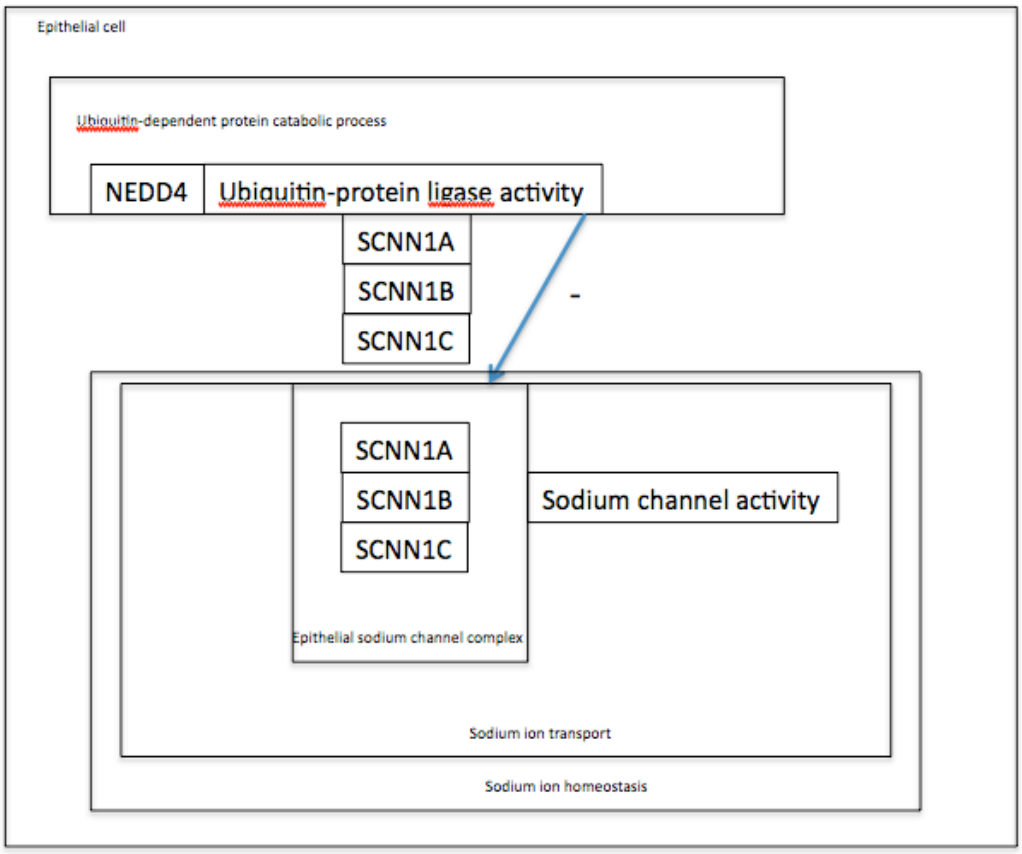
This diagram should be interpreted as follows: each ontology term or molecule is represented by a box. Molecular functions are in larger type, and molecules are given the gene symbol. The gene (product) to the left of the molecular function is the one annotated with the given function; the gene product below it is the target of the function or process; all molecules inside a biological process are annotated with the given process. An arrow indicates regulation (in this example, both the activity and the process regulate the downstream molecule and process). Nesting of one process inside another indicates a process-subprocess relation: in this case, both *ubiquitin-dependent protein catabolic process* (operating on RNAPII) and *negative regulation of transcription from RNAPII promoter* are subprocesses within *cellular response to UV*.

Currently, this would require the explicit construction of a pre-composed GO term to capture the nested biological processes, and this is indeed what the curators requested: *negative regulation of transcription from RNA polymerase II promoter in response to DNA damage during cellular response to UV*. However, even this complex GO term does not capture the use of the ubiquitin-dependent proteolysis module, as opposed to NEDD4 function alone. In our new, proposed process, rather than first creating a new, complex term, , a curator would simply be able to compose this term as a nested annotation, which could be used to create pre-composed GO terms automatically, if desired. Importantly, the nested annotation would also capture which molecular function of NEDD4 (including its specificity) is used within the biological process.

Example 2: Negative regulation of sodium ion transport

In this example, the same process of ubiquitin-dependent proteolysis is used to regulate a different biological process by targeting a different protein. From the abstract:

> The epithelial Na(+) channel (ENaC) regulates Na(+) absorption in epithelial tissues including the lung, colon and sweat gland, and in the distal nephrons of the kidney. When Na(+)-channel function is disrupted, salt and water homoeostasis is affected…. Previously we showed that a proline-rich region of the alpha subunit of the Na(+) channel bound to a protein of 116 kDa from human lung cells. Here we report the identification of this protein as human Nedd4, a ubiquitin-protein ligase that binds to the Na(+)-channel subunits via its WW domains. Further, we show that WW domains 2, 3 and 4 of human Nedd4 bind to the alpha, beta and gamma Na(+)-channel subunits but not to a mutated beta subunit. In addition, when co-expressed in Xenopus oocytes, human Nedd4 down-regulates Na(+)-channel activity.



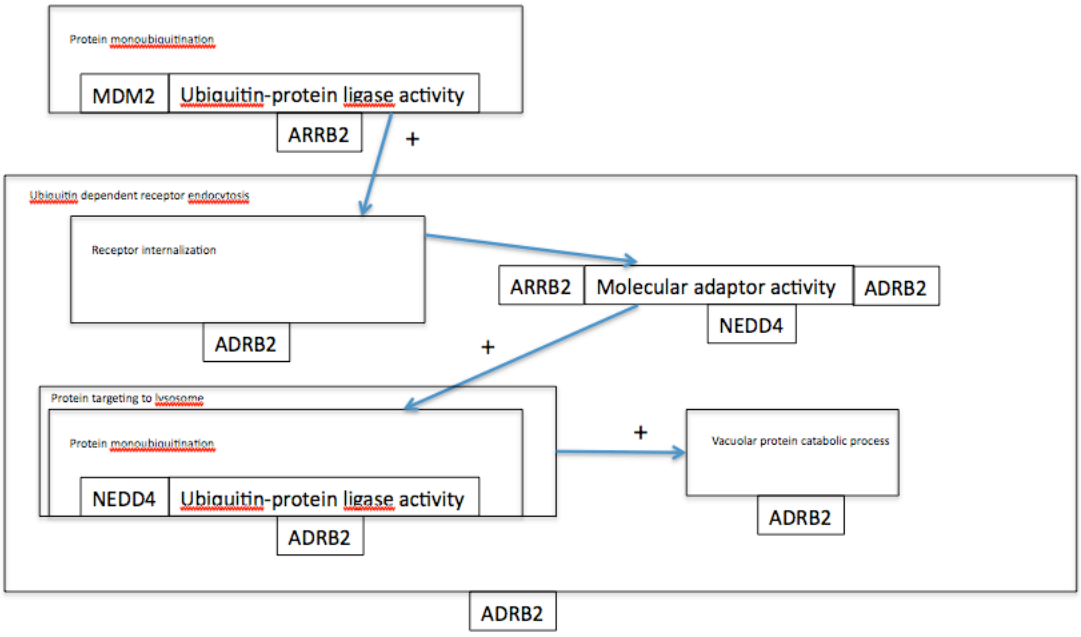
The nested annotations allow the simultaneous annotation of both a molecular function and a biological process as regulating another biological process; this entire super-process can then be annotated as occurring in a particular location. This example also illustrates the annotation of a complex with a molecular function and biological process.

Example 3: NEDD4 and receptor catabolic process

Both the NEDD4 ubiquitin protein ligase function and protein ubiquitination sub-process, with different molecular specificities of course, are used within the context of different larger biological processes. PMID 18544553 is used as evidence for three annotations: beta-2 adrenergic receptor binding (function), receptor internalization (process) and receptor catabolic process. In this case, NEDD4 ubiquitinates (probably monoubiquitination, but requires the same main steps as polyubiquitination above) the beta-2 adrenergic receptor, leading to catabolism (in the lysosome, not the proteasome). Actually, the publication states that the internalization step is mediated by ubiquitination of beta-arrestin by MDM2, and does not involve NEDD4. Both of these steps are part of *ubiquitin-dependent endocytosis* (GO:0070086):

> β-arrestin2 binds at least two E3 ubiquitin ligases, Mdm2 and Nedd4, serving different purposes in β2AR regulation: Mdm2, which mediates β-arrestin ubiquitination (12) and regulates the initial step of receptor endocytosis, and Nedd4, which mediates receptor ubiquitination that targets receptors to lysosomal compartments…We also demonstrate that β-arrestin2 functions as an E3 ubiquitin ligase adaptor to recruit Nedd4 to the activated β2AR.

The findings can be expressed as:

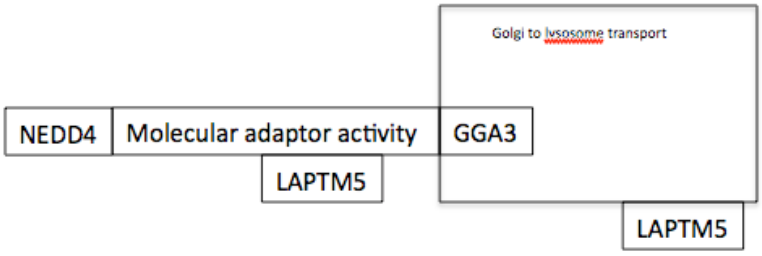


In this example, ubiquitin-dependent endocytosis of ADRB2 has two sub-processes, receptor internalization and protein targeting to lysosome, each of which have a sub-process of protein ubiquitination.

Example 4: NEDD4 and protein targeting to lysosome

NEDD4 does not always function as a ubiquitin protein ligase. Thus, our mechanism for capturing which molecular function is used within a biological process is essential. For instance, NEDD4 is annotated as being involved in *protein targeting to*

lysosome. This paper actually states that NEDD4 binds to a specific protein (LAPTM5), and targets it to the Golgi-to-lysosomal transport machinery by specifically binding a component of that machinery, GGA3. Based on this publication, there is no evidence that NEDD4 is part of a general system for targeting proteins to the lysosome, as implied by the annotation. Rather, NEDD4 uses two coupled *protein binding* functions to target a specific protein (LAPTM5) to another protein, GGA3. Then GGA3, as part of a larger general vesicular transport system, is involved in the actual lysosomal transport process. However, there is currently no way for a curator to capture the specificity of a protein targeting process. In the new process, this would be expressed as:



Note that *molecular adaptor activity* is a molecular role, which is executed in this case by a compound biochemical function, composed of two binding functions (elemental biochemical functions). In general, a biochemical function can be combined modularly with a molecular role. This particular compound function has two specificity slots (one for each binding function): one for the protein "cargo" (LAPTM5) and one for the protein "destination" (GGA3).

Note also that some functions, e.g. protein binding, should automatically result in additional, reciprocal annotations. Because binding is a symmetric relationship, capturing ligand specificity of NEDD4 binding automatically generates binding annotations for its ligands.

4. Effect on curation

In summary, curators will be able to capture more information from each paper, more efficiently and more consistently. They will be able to express how biological systems are constructed in a modular manner from molecular functions and subsystems. Modular annotations can guide ontology development and provide literature evidence for relations in the ontology. Explicit relationships between molecular function and biological process can suggest additional annotations. Templates of subsystems can provide guidance and consistency during curation.

Critically, modular annotations are an extension of the existing annotation process, and all annotations made using the current process will remain valid. For example, a gene product can still be annotated directly to a biological process if its molecular function within the process is unknown. However, one of the main strengths of the proposed new process is that it addresses an important feature of the biological literature: **biology papers are not just about how a gene functions, but how a**

system functions. Curators will be able to annotate systems as well as their parts, resulting in a more complete representation of current biological knowledge.