Additional Gene Ontology structure for improved biological reasoning

Simen Myhre¹,†, Henrik Tveit²,†, Torulf Mollestad²,³ and Astrid Lægreid¹,*

¹Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO-7489 Trondheim, Norway, ²Department of Computer and Information Science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway and ³SAS Institute AS Norway, Pb 2666 Solli, NO-0203 Oslo, Norway

Received on April 7, 2006; revised on May 30, 2006; accepted on June 12, 2006
Advance Access publication June 20, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The Gene Ontology (GO) is a widely used terminology for gene product characterization in, for example, interpretation of biology underlying microarray experiments. The current GO defines term relationships within each of the independent subontologies: molecular function, biological process and cellular component. However, it is evident that there also exist biological relationships between terms of different subontologies. Our aim was to connect the three subontologies to enable GO to cover more biological knowledge, enable a more consistent use of GO and provide new opportunities for biological reasoning.

Results: We propose a new structure, the Second Gene Ontology Layer, capturing biological relations not directly reflected in the present ontology structure. Given molecular functions, these paths identify biological processes where the molecular functions are involved and cellular components where they are active. The current Second Layer contains 6271 validated paths, covering 54% of the molecular functions of GO and can be used to render existing gene annotation sets more complete and consistent. Applying Second Layer paths to a set of 4223 human genes, increased biological process annotations by 24% compared to publicly available annotations and reproduced 30% of them.

Availability: The Second GO is publicly available through the GO Annotation Toolbox (GOAT.no); http://www.g oat.no
Contact: astrid.lagreid@ntnu.no

1 INTRODUCTION

High-throughput biomedical experiments, such as microarray-based gene expression measurements (Schena et al., 1995), generate huge amounts of data. Efficient analysis of the biology behind such data requires standardized vocabularies and ontologies for describing genes, gene products and their functions. The Gene Ontology (GO) (http://www.geneontology.org) has grown in size and popularity to become the standard terminology for describing the function of genes and gene products across species. The potential of GO in biological reasoning is now increasingly recognized by researchers in biomedicine and computing. Thousands of genes and their products have been annotated using GO (http://www.geneontology.org/GO.current.annotations.shtml) and a large number of tools have been developed to utilize this structured vocabulary (Gene Ontology Consortium: GO tools; http://www.geneontology.org/GO.tools.shtml). It should be noted that while GO annotations are often associated with genes, the GO terms always describe the functions of the gene products, i.e. proteins and/or RNAs. Thus, a statement that a gene has a given function described by a GO term should be understood to mean that the gene’s product has this function.

GO consists of three independent subontologies; molecular function describing activities, such as catalytic or binding activities at the molecular level, biological process in which the gene product is active and that is accomplished by one or more ordered assemblies of molecular functions, and cellular component where the molecular function is executed (e.g. see Fig. 1). Each subontology organizes its terms as a directed acyclic graph (DAG), essentially a hierarchical structure, allowing multiple parent terms for each child term. The relationship of child to parent can be either ‘is a’ or ‘part of’. The lower in the DAG a term is located, the more specific it is. This enables the biomedical researcher to find the best-fitting uniform descriptors for the functions of a given gene product according to available information. When annotating a given gene or gene product using GO, the GO DAG displays the way in which annotations within a subontology compositionally relate to each other, and to annotations of other genes. Thus, annotation of a gene product with a GO-term indicates that this gene product is strongly related to all other gene products annotated with the same GO-term as well as to gene products annotated with ancestors or descendants of that GO-term. This knowledge inherent in the GO structure is today used by a number of existing tools using GO (http://www.geneontology.org/GO.tools.shtml).

Although the three GO subontologies are structured as three independent ontologies, it is evident that there exist biological relations between the three subontologies. For example, gene products with the molecular function microtubule motor activity participate in the biological process microtubule based movement, and are active in the cellular component microtubule. Thus, it is necessary to take all three subontologies into account when annotating genes, in order to achieve consistent information in the annotation sets. The Gene Ontology Consortium acknowledges the existence of such intersubontology relationships (The Gene Ontology Consortium,
2001), but the present GO-structure does not reflect them. Yeh et al. (2003) examined the present GO structure using Knowledge-Based Management System and inter alia suggested cross-subontological relationships. Kumar and colleagues used annotation patterns to indicate possible relationships in a data exploration fashion (Kumar et al., 2004). Cross-subontology patterns in existing GO-annotations have also been used for prediction of gene function (King et al., 2003). Lægreid et al. (2003) predicted biological processes based on gene expression profiles. However, no reported initiatives have focused entirely on an inter-subontology model; starting by constructing new inter-subontology GO structure, populating the structure with validated paths in order to prove the practical use of such a model, and providing it as an addition to GO ready-for-use.

In this work, we present a new concept that we name the Second Gene Ontology Layer. It defines relationships between the three subontologies of the GO, linking molecular functions to biological processes and to cellular components, respectively. The inter-subontology relationships are inspired by, but different from, the intra-subontology relationships of the present GO DAG and are therefore organized in an additional structural layer utilizing the existing GO-terms and complementing the present GO DAG structure. Since finding all paths of the proposed structure is a comprehensive task, we applied three methods that generated path candidates to start the population of the Second Layer. All generated path candidates were validated by biological experts before they were included in the model. Thus, the high standard of biological information found in the present GO-structure was maintained. To indicate the value of the Second Gene Ontology Layer, we complemented a set of publicly available annotations. Hence, we believe that this addition to GO provides new information that allows for more advanced reasoning with biomedical functional genomics data and can help ensure consistency in annotation sets.

2 METHODS

2.1 Constructing the Second Gene Ontology Layer

We modeled the Second Layer such that it reflects biological relationships between terms of different GO subontologies. Molecular function is the entity of the subontologies most directly connected to a given gene product. Therefore the paths in Second Layer are directed from molecular function to biological process and from molecular function to cellular component. Paths leading from molecular functions to biological processes are of the type ‘X molecular function is involved in Y biological process’. Paths from molecular functions to cellular components are of the type ‘X molecular function acts in Z cellular component’ (see Fig. 2). For instance, any gene annotated with the

Fig. 1. A section of the Gene Ontology and its three subontologies: molecular function, biological process and cellular component. The GO terms are related through a directed acyclic graph where each path denotes either a ‘is a’ or ‘part of’ relationship. The DAG allows a term to inherit features from multiple parents (example shown in grey). The true path rule states, ‘the pathway from a child term all the way up to its top-level parent(s) must always be true’ (Gene Ontology Consortium, 2001).

Fig. 2. The Second Layer models biological relationships between terms of different subontologies. The new relationships, is involved in and acts in, reflects other biological phenomena than the original GO paths is a and part of and are organized in their own path layer co-existing with the original structure.
molecular function microtubule motor activity is involved in the biological process microtubule-based movement. However, not all genes involved in microtubule-based movement are necessarily microtubule motor proteins. Likewise, a microtubule motor protein is active in the cellular component microtubule, but not all genes annotated with the latter component are microtubule motor proteins.

The proposed new relations were not added directly to the existing ‘is a’ and ‘part of’ paths. Instead, we suggest organizing these new relations in their own ontology layer structure. Thus, paths that represent different biological phenomena are separated and both old and new structures are kept manageable and practical to use. Despite different paths and relationship-structures, the Second Layer utilizes the original GO terms and complements the existing GO paths.

2.2 Populating the Second Gene Ontology Layer

The GO subontologies were downloaded from the Gene Ontology Consortium web page (http://www.geneontology.org; version dated September, 2005).

Both the number of possible inter-subontology term combinations and the number of actual relationships between the subontologies are huge. Hence, finding all correct paths is a non-trivial task, especially if performed manually. To overcome this, we sought strategies that would reduce the number of possibilities to a manageable set of more likely path candidates. Any such method would do, however, they would all require a manual validation before adding paths to the model. Even with semi-automating methods, we could not make the population of the Second Layer complete within our scope. Therefore, we aimed at making the first version of a more compact and efficient model available for public use.

In our approach, we utilized a third in order to capture paths related to metabolic pathways. Because we have an underlying model aligned to GO, we used the properties of the Second Layer and GO to spawn even more paths from the validated, generated paths.

In the first path candidate generating approach, we utilized the fact that many molecular functions and their related biological processes have similar names. For example, molecular functions of the form ‘W inhibitor/transporter/regulator activity’ (W being a biological term) often have related biological processes named ‘W inhibition/transport/regulation’, as is the case with the molecular function spermine transporter activity and its related biological process spermine transport. When we searched for molecular function terms containing wildcard terms such as ‘factor’, ‘inhibitor’, ‘transporter’, ‘regulator’, for each returned molecular function term (e.g. ‘spermine transporter activity’), we copied the words, W, ‘(spermine)’ to the left of the wildcard (‘transporter’) and performed a new search with W (‘spermine’) as the search term. The second search would return biological processes that might be related to the molecular function in question (e.g. ‘spermine transport’). This approach is similar to the lexical approaches used in (Ogren et al., 2004; Bodenreider et al., 2005). But here, as we were to add paths to a formal validated model, experts in biology validated whether the returned biological processes should form paths of the type ‘W inhibitor/transporter/regulator activity is involved in W inhibition/transport/regulation’.

After these paths were established, each of them was used to spawn a number of new paths utilizing the existing GO DAG structure. This was based on the so called ‘true path rule’ (Gene Ontology Consortium, 2001) (Fig. 1), whose rationale is that all terms of the molecular function subontology inherit all the features of their parents in addition to their more specific features. Thus, all children of a given molecular function should be involved in the same biological processes, and execute their function in the same cellular compartments, as their parents. We explicitly modelled this using paths like ‘X_molecular function is involved in Y Biological process’ to spawn new paths like ‘Z_molecular function is involved in Y Biological process’, where X is the parent of Z. For example, the path ‘translation termination factor activity is involved in translational termination’ spawns ‘translation release factor activity is involved in translational termination’ because translation termination factor activity is the GO-parent of translation release factor activity.

New rules were spawned for every descendant of X as defined by the GO DAG. We performed the same procedure for relationships between molecular functions and cellular components. Spawning the validated paths from the term search generated 1867 paths from molecular function to biological process, and 541 paths between molecular function and cellular component.

In the second strategy, we utilized relationships between terms from the three GO subontologies as implicitly indicated in publicly available gene annotations. This was based on the assumption that some of the pairs of GO-terms often annotated to the same gene would reflect an actual interdependence between the given molecular function and the biological process or cellular component. To reduce the number of relationships to be manually validated, we employed a pre-screening step leaving only GO-term pairs with a statistically high-potential relationship. To effectively perform the filtering, we applied the well-known data mining technique called association analysis. Association rule mining aims at the discovery of relationships between items in a dataset. Based on Agrawal and Srikant (1994), we present a short formal description of this technique. Let $I = \{i_1, i_2, \ldots, i_k\}$ be a set of binary attributes called items. A transaction $T$ with a transaction identifier $T_D$ in a database $D$ of transactions consists of an item set such that $T \subseteq I$. Let $X$ be an item set of $|X|$ items such that $X \subseteq I$. A transaction $T$ satisfies $X$ if $X \subseteq T$. An association rule is an implication of the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The percentages of transactions in $D$ that satisfy both $X$ and $Y$ are known as the Support $\%$ of the rule $X \Rightarrow Y$ (and of the rule $Y \Rightarrow X$). The percentage of transactions in $D$ satisfying $X$ that also satisfy $Y$ is known as the Confidence $\%$ of the rule $X \Rightarrow Y$ (not the same as for $Y \Rightarrow X$). Having manually set threshold values minsupport and minconfidence for support and confidence, respectively, the rule $X \Rightarrow Y$ is considered interesting if $\% > \text{minsupport}$ and $\% > \text{minconfidence}$. We downloaded GO annotation sets from the Gene Ontology Consortium website (http://www.geneontology.org/GO.current.annotations.shtml) and used each set as a separate input for the SAS Enterprise Miner’s Association analysis software (http://www.sas.com). We used the gene identifier as $T_D$ and the GO terms as items. We limited the analysis to rules of the format $X \Rightarrow Y$ Biological process with $|X| = 1$ and $|Y| = 1$. As opposed to earlier studies (Bodenreider et al., 2005; Kumar et al., 2004), here, the threshold values had little impact and were mainly used to keep the rule set to a manageable validation size. King et al. (2005) utilized inter alia decision trees to generate probabilistic models. This technique can produce similar output as association analysis as the resulting tree can be split into several rules. However, our association rules differed from the rules from King’s decision tree as the latter produced rules with both positive and negative associations where $|X| > 1$. This is preferable in a probabilistic model as more information is included during the prediction process. In contrast, we intended to build a more formal model from small, validated building blocks. Hence, simple positive binary relations were sufficient for our purpose as long as the quality of the final term associations was ensured by a comprehensive manual inspection. Biology experts validated the returned rules and kept rules reflecting true inter-subontology dependencies only. Analogous to the procedure used in the strategy based on term similarity (above) the validated association paths were used to spawn new paths from all molecular function child-terms, resulting in 2269 paths between molecular function and biological process.

In our third approach, we utilized metabolic pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG). The KEGG pathways describe metabolic processes and the enzymes catalysing them using the Enzyme Nomenclature (http://www.chem.qmw.ac.uk/iubmb/enzyme/), also known as Enzyme Commission numbers (EC-numbers). We focused on the schemas categorized under ‘metabolism’ in the KEGG-database. Of a total
of 142 metabolism schemas, we used 109 that corresponded directly to GO-defined biological processes. For example, pyruvate metabolism in KEGG corresponds to the biological process pyruvate metabolism in GO. The Second Layer paths were generated using the official translations between EC-numbers and molecular functions in GO (http://www.geneontology.org/external2go/ec2go). For example, the EC enzyme Malate dehydrogenase (EC 1.1.1.37) has the corresponding GO molecular function l-Malate dehydrogenase activity. Thus, for each of the 109 biological processes, we generated a table listing the involved molecular functions as according to KEGG. This approach generated 2949 Second Layer paths from molecular function to biological process.

Finally, we merged the sets of paths found by the three strategies and removed duplicates and general paths for which a more specific path existed (<2%). The three strategies combined, populated the Second Gene Ontology Layer with 6721 validated paths.

3 RESULTS

3.1 Populating the Second Gene Ontology Layer

Our main focus in the development of the Second Gene Ontology Layer was to create a framework for connecting the three GO subontologies. With a model in place, we wanted to test the practical use of the model (see Methods for details).

As of today the Second Layer consists of 6721 paths defining relationships between molecular function and biological process and between molecular function and cellular component. Of these, 6180 (92%) are paths from molecular function to biological process, while the remaining 541 (8%) are paths from molecular function to cellular component (e.g. see Fig. 3). These paths cover a large proportion of the entire GO structure. Table 1 shows that for more than half of all terms in the molecular function subontology, we have identified at least one relation to one or more terms in the other two subontologies (Table 1).

3.2 Application of the Second Gene Ontology Layer

The performance of applying Second Gene Ontology Layer paths on a published dataset was assessed with respect to accuracy and coverage as well as the ability to improve completeness and consistency of GO annotations. Publicly available GO-annotations for genes represented by the probes in microarray analysis of the fibroblast cell cycle transcriptional programme (Cho et al., 2001) were retrieved using the Annotation database of the Norwegian Microarray Consortium (http://www.genetools.no, data from the GO database site: http://www.godatabase.org/dev/database/; dated January 2005). A total of 25,965 annotations distributed over all three subontologies were found for 4223 of the 4905 genes represented on the microarrays. We first removed duplicates and annotations using obsolete GO terms as well as annotations for which a more specific annotation was present for the same gene. This resulted in a final set of 21,664 publicly available annotations for the 4223 genes.

Starting with the 7774 molecular function annotations contained in the publicly available annotations, Second Layer paths were used to generate biological process and cellular component annotations. Given a gene, G, and its set of annotations, A, including the molecular function annotation F, we added the biological process, P, to A if a Second Layer path, ‘F is involved in P’, could be found. Similarly, ‘F acts in C’ would add the cellular component annotation C to A. In the resulting set of downloaded and produced annotations, duplicates and less specific annotations were removed.

A total of 5607 annotations were generated—4651 for biological process and 956 for cellular component (Table 2). More than half of the biological process and cellular component annotations present in the publicly available dataset were reflected in the generated set either directly or indirectly through matches for their ancestors or descendants in the GO DAG. This is a strong indication that Second Layer generated annotations identify relevant biological information for the gene products in question.

A large fraction of the generated annotations were new annotations as they were neither present in the publicly available annotations nor among their ancestors or descendants (in the GO DAG). Of the 4223 genes, 1326 (31%) received at least one new biological process annotation, while 290 (7%) genes received at least one new cellular component annotation. An annotation is considered new if it has no ancestor or descendant in the GO DAG already annotated.
Descendants of publicly available annotations included 138 (14%) of total annotations, 1125 (20%) of generated new annotations, 1130 (24%) of generated annotations replacing publicly available annotations, and 979 (21%) of exact match to publicly available annotations.

Table 3. Second Layer complementation of annotations for 4223 genes in the dataset from Cho et al. (2001) 

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Cellular component</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publicly available annotations</td>
<td>8764</td>
<td>5126</td>
</tr>
<tr>
<td>After</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publicly available annotations not replaced</td>
<td>6204</td>
<td>4465</td>
</tr>
<tr>
<td>Generated annotations replaced publicly available annotations (†)</td>
<td>2607</td>
<td>664</td>
</tr>
<tr>
<td>Generated new annotations</td>
<td>2085</td>
<td>301</td>
</tr>
<tr>
<td>Combined annotations</td>
<td>10 896</td>
<td>5430</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Cellular component</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+24%)</td>
<td>(+6%)</td>
<td>(+18%)</td>
</tr>
</tbody>
</table>

The 7774 molecular function GO-terms annotated to 4223 genes analysed in the fibroblast serum response (Cho et al., 2001) were complemented using the ANNEX tool in the GO Annotation Toolbox (http://www.goat.no).

Table 4. Validation of Second Layer annotated genes in the dataset from Cho et al. (2001) 

<table>
<thead>
<tr>
<th>C</th>
<th>UCLP</th>
<th>UCWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological process</td>
<td>408 (91%)</td>
<td>39 (9%)</td>
</tr>
<tr>
<td>Cellular component</td>
<td>47 (94%)</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Total</td>
<td>455 (91%)</td>
<td>41 (8%)</td>
</tr>
</tbody>
</table>

C (Certain annotation) annotation is certain for the respective gene, UCLP (Uncertain annotation Logic Path) annotation is uncertain for the respective gene however the Second layer path is correct. UCWP (Uncertain annotation Wrong Path) Second Layer generated annotation is uncertain for the respective gene due to a wrong Second Layer path. The erroneous paths that generated the three UCWP annotations were removed from the Second Layer.

The 7774 molecular function GO-terms annotated to 4223 genes resulted in 24% new biological process annotations and 6% new cellular component annotations compared with the publicly available annotation set (Table 3). This clearly demonstrates the potential of the Second layer not only to reproduce known annotations but also to complement and enrich publicly available annotation sets when making them more consistent.

Biological knowledge content of annotations increases with their specificity as represented by the location of the annotation in the DAG. Relative to the top subontology terms biological process and cellular component, Second Layer generated annotations for the gene set above were at an average of level 7.3 and 4.8, respectively. Publicly available annotations for the same set were at an average of level 7.5 and 4.6. Thus, Second Layer generated annotations are at a similar level of specificity, and thus have similar biological knowledge content as publicly available annotations.

In order to validate the new knowledge added by the Second Layer paths, biological experts randomly selected 500 (450 biological processes and 50 cellular components) annotations from those classified as new and those replacing publicly available annotations in the combined set (see Table 3). The biologists assessed these annotations by using information from Swiss-Prot (http://www.ebi.ac.uk/swissprot/), Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) and literature (Table 4). Totally 91% of 450 biological process annotations and 94% of the 50 cellular component annotations were confirmed from database and literature information (Table 4). Nine percent of the biological process annotations were not in agreement with database and literature information even though the Second Layer paths applied were reconfirmed. For each of these annotations, we found no information supporting the publicly available molecular function annotation that the generated biological process annotation stemmed from. This indicates that the initial molecular function annotation may be uncertain or incorrect. Of the 39 biological process annotations not confirmed, 30 (77%) were generated from molecular functions with evidence code IEA (Inferred by Electronic Annotation). In comparison, 143 (35%) of the 408 confirmed annotations were generated from molecular functions with evidence code IEA. This indicates that <10% of new annotations generated using Second Layer paths are uncertain, and that the main contributors for uncertain annotations are molecular function annotations with evidence code IEA.

Our results clearly indicate that a Second Gene Ontology Layer can be used to generate relevant biological process and cellular component annotations based on molecular function annotations. It is therefore of interest to consider to which extent Second Layer could be used as an annotation tool which allows the annotator to focus on generating an exhaustive list of molecular function annotations. From these molecular functions, the Second Layer paths would contribute with biological processes and cellular component annotations as described above. To investigate this potential we can take the list of annotations for the 4223 genes as an example. Of the 13 890 publicly available annotations within the biological process and cellular component subontologies, 3221 (23%) were matched by annotations generated by Second Layer paths. However, comparing the unmatched annotations with the biological processes and cellular components involved in one or more paths, we see...
that 8792 (82%) of the remaining 10,669 unmatched annotations used GO terms from the biological process and cellular component subontologies that were represented in the Second Layer by paths using: (1) the exactly same GO term (39%), (2) one of the GO term’s descendants in the GO DAG (38%), or (3) one of its ancestors (23%). Of these 8753 (>99%) had more than one relevant Second Layer path. This shows that the current Second Layer contains paths representing many of the biological process and cellular component terms commonly used in GO-annotations.

The Gene Ontology Consortium (2000) defines a molecular function as the biochemical activity of a gene product, a biological process as assemblies of molecular functions, and cellular component as the place where the activity is performed. Hence, the molecular functions of a gene can be regarded as the atomic entity for describing gene characteristics. In accordance with this, the example of Second Layer application suggested here is based on the assumption that biological process and cellular component annotations can be deduced from molecular function annotations. We realize that mapping and validating all paths from the molecular functions to biological processes and cellular components, respectively, is a major effort and a non-trivial task. However, a fully developed Second Layer would enable annotators to focus on a gene’s molecular functions. By supplying process and component annotations that are biological necessities and consequences of the annotated molecular functions, the Second Layer paths complete the annotation sets and make them consistent.

4 DISCUSSION
The GO is continuously improved by additions of new terms and by redefinitions of existing terms (Lewis, 2004; The Gene Ontology Consortium, 2004). In addition to these refinements, GO should be enriched with new term relationships beyond the existing ‘is a’ and ‘part of’ relationships. For example, the formal definitions of path types between the subontologies proposed here enhance GO by reflecting biological concepts neither present in GO today, nor captured in the OBO Relation Ontology (Smith et al., 2005; OBO Relationship Ontology, http://obo.sourceforge.net/relationship). In constructing the Second Gene Ontology Layer, we focused on defining directed paths from molecular function to biological process and from molecular function to cellular component. The two new relationships, is involved in and acts in, were inspired by the simplicity and intuitiveness of the present GO relationships ‘is a’ and ‘part of’ and are comparable with those of Yeh et al. (2003). All paths of both the present GO and the Second Layer must be interpreted by considering both the relation-type and the involved terms’ descriptions. Many GO terms are themselves quite detailed in the descriptions of their roles as the Consortium direct curators to ‘aim to be reasonably descriptive, even at the risk of some verbal redundancy’ (http://www.geneontology.org/GO.usage.shtml).

Modelling of genome-wide gene expression data often considers broad categories of genes in terms of their function as described by GO, like in models predicting biological process annotations based on temporal gene expression profiles (Lægreid et al., 2003). However, noticeable inconsistency and incompleteness in available annotations have been documented (Dolan et al., 2005), and this inconsistency and incompleteness can impact and reduce the sensitivity and accuracy of such models. The lack of consistency can also be exemplified by our search in the July 12, 2005 version of the Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) which revealed that of all 77 human genes annotated with the molecular function microtubule motor activity, 22 were not annotated with the biological process microtubule-based movement and 26 were not annotated with the cellular component microtubule. This example indicates a considerable variability in annotation practice underlying current GO annotations in public databases and illustrates how the benefit from using a structured vocabulary is reduced due to lack of a standardized way to use it.

Applying the Second Layer to existing gene annotations may enhance annotations to become both more complete and more consistent by adding new biological process and cellular component annotations to genes based on their molecular function annotations and by adding the same biological process and cellular component annotations to all genes with the same molecular functions. This may improve genome-wide models and thus aid the biomedical researcher in developing model-based hypotheses. Since the Second Layer contributes annotations such that every gene to a specific molecular function term also is annotated to the related biological processes and cellular components, some of the existing variations in annotation practice underlying publicly available annotations may be normalized. Thus, Second Layer may improve robustness of models that make use of biological background information in the form of GO annotations.

King and colleagues (King et al., 2003) also supplemented annotation sets based on existing annotations. However, the method produced predictions based on a probabilistic model, rather than annotations generated from a validated biological model. A formal model, like the Second Layer, gives more credibility to the conclusions drawn from it (e.g. annotations), as all relations are explicitly defined. Furthermore, such constructed paths have applications beyond complementing annotation sets, as, for example, reasoning with biological data.

4.1 Future challenges and concluding remarks
The Second Gene Ontology Layer is an effort to model inter-subontological relationships found in nature. We have started populating the proposed structure with 6721 validated paths, but acknowledge that we have not identified all possible relations between the subontologies and that this is a comprehensive task. In the future, all path generating methods, not only the three used here, should be utilized. Still, in its current state, the Second Layer has practical applications giving promising results by making annotation sets more specific, consistent and complete.

We also would like to study how the biological process and cellular component branches relate to each other. Such an investigation could use the proposed relationships as a starting point; possibly by pairing ‘is involved in’ and ‘acts in’ paths. We recognize that such relations can be complex considering that one biological process can span several cellular components. Furthermore, additional structures could model even more biological concepts, e.g. the sequence of functions acting within processes.

We believe that the addition of the Second Gene Ontology Layer to the original GO represents an important step towards a complete framework for biological modelling (e.g. see Fig. 4). By defining additional background knowledge in the form of relations between...
the three subontologies of the present GO structure, Second Layer provides a validated biological model that may improve interpretation of experimental data. Hopefully, the Second Gene Ontology Layer will assist in, and encourage the pursuit of, defining all pathways of the cell and their interactions—one major challenge in systems biology.

4.2 The Second Layer is available through www.goat.no

The Second Layer network is available for download from the GO Annotation Toolbox (GOAT.no) at http://www.goat.no and will be updated regularly. GOAT.no also provides ANNEX, a tool based on the Second Layer, which expands your annotation sets as described above. The Second Layer and the ANNEX-function are also integrated in Blast2GO (http://www.blast2go.de).

ACKNOWLEDGEMENTS

The authors would like to thank Torgeir Hvidsten, Jan Komorowski and Anne-Lise Børresen-Dale for valuable discussions and reviews of this work. The authors are also thankful for the software provided by SAS Institute Norway. Funding to pay the Open Access publication charges for this article was provided by Norwegian University of Science and Technology.

Conflict of Interest: none declared.

REFERENCES


