Automated Annotation-Based Bio-Ontology Alignment with Structural Validation

Cliff A Joslyn[*], **Bob Baddeley**[*], **Judith Blake**[†], **Carol Bult**[†], **Mary Dolan**[†], **Rick Riensche**[*], **Karin Rodland**[*], **Antonio Sanfilippo**[*], **Amanda White**[*]
[*]**Pacific Northwest National Laboratory, Richland, WA, USA**
[†]**The Jackson Laboratory, Bar Harbor, ME, USA**

Nature Precedings : doi:10.1038/npre.2009.3518.1 : Posted 30 Jul 2009

Abstract

We outline the structure of an automated process to both align multiple bio-ontologies in terms of their genomic co-annotations, and then to measure the structural quality of that alignment. We illustrate the method with a genomic analysis of 70 genes implicated in lung disease against the Gene Ontology.

Introduction

Ontologies are commonly aligned based on similar annotations[3, 7], requiring validation of the quality of the induced alignment. In this short paper we make describe an approach to automated annotation-based bio-ontology alignment combined with subsequent measurement of the quality of those alignments. We do so using an example from lung disease genomics.

We begin with a list of 70 genes implicated in lung diseases. These are annotated to the Biological Process (BP) and Molecular Function (MF) branches of the Gene Ontology (GO[4]). The Cross Ontology Analytics tool (XOA, http://xoa.pnl.gov,[10, 12]) is then used to generate proximities between pairs of nodes in the BP and MF branches. The XOA scoring allows generation of putative alignments between BP and MF nodes, and then Joslyn *et al.*'s order-theoretical approach[6] is used to measure the structural quality of the generated alignments.

Lung Disease Genomics

The impact of genomics to study classes of diseases has yet to be fully realized. Research about lung diseases, focused on the cancers and other pathologies of specific tissue types, will benefit from systems analysis of cellular pathways and processes implicated in the presentation of disease states[9]. Genomic and proteomic analysis via ontological representations of gene product location and function has enabled the construction of predictive functional networks awaiting experimental validation[5].

We identified a set of 70 genes through our work in lung development and disease to evaluate the contribution of ontological alignments to further refined experimental hypotheses. We identified these 70 genes through expression analysis of mouse lung samples representing different developmental stages. The gene list is available at `ftp://ftp.informatics.jax.org/pub/curatorwork/ICBO09/lung_dev_genes.txt`. This defined set was chosen to be representative of molecular systems implicated in lung development and function.

The 2/26/09 version of mouse annotations (`ftp://ftp.informatics.jax.org/pub/reports/gene_association.mgi`) yields 1937 lines of GO annotations, including 424 distinct BP annotations, of which 388 were experimental annotations from mouse systems. There were 80 distinct MF annotations, 40 with experimental support. Overall, there are 62 genes with experimental BP annotations and 50 genes with experimental MF annotations. 48 genes, included in the results of the previous sentence, had both MF and BP experimental annotations.

Alignment Generation

XOA automatically generates links between BP and MF nodes based on their common annotations. Information theoretical approaches[8] are effective within one hierarchy. But because they require that similarity between two GO codes be computed in terms of the informational content of the most immediately dominating parent GO code, they cannot link GO codes across distinct gene subontologies. The vector space model approach obviates this limitation by computing the similarity between two GO codes as the cosine of vectors that encode the gene annotation associated with the two GO codes[1]. XOA combines these two approaches by turning relational links across GO codes into hierarchical links[12].

We model semantic hierarchies as finite, bounded, partially ordered sets (posets) $\mathcal{P} = \langle P, \leq \rangle$[2], with nodes $a \in P$ as ontology concepts related by `is-a` links through $\leq$. The XOA similarity between the GO node $a \in P$ and the GO node $a' \in P'$ is then

$$XOA(a, a') := \max \left(\max_{b \in P} (\mathrm{sim}(a, b) \cos(a', b)), \max_{b' \in P'} (\mathrm{sim}(a', b') \cos(a, b')) \right),$$

where $\cos(a, a')$ denotes the cosine measure[11] between

GO nodes $a \in P, a' \in P'$ in the GO node × gene annotation matrix, and $\mathrm{sim}(a, b)$ denotes Resnik's information theoretical similarity measure[8] between GO nodes $a, b \in \mathcal{P}$. An XOA analysis of the GO nodes annotated to our 70 test genes reveals 1970 BP-MF pairs $\vec{l} := \langle a, a' \rangle$ which are significant, with $p \leq 5\%$. Each such pair of **anchors** is a potential link between BP and MF.

An ontology **alignment** is a mapping $f : \mathcal{P} \to \mathcal{P}'$ taking anchors $a \in P$ in a semantic hierarchy $\mathcal{P} = \langle P, \leq \rangle$ to those $a' \in P'$ in another $\mathcal{P}' = \langle P', \leq' \rangle$. But a BP node $a \in P$ which has a high XOA score with an MF node $a' \in P'$ is also likely to have a high XOA score with other MF nodes $b' \in P'$. The complete set of 1970 links $\vec{l}$ yields a many-to-many alignment relation $F \subseteq P \times P'$. We need an alignment function $f : \mathcal{P} \to \mathcal{P}'$ with all left anchors appearing only once, so we sort the links by XOA to select the highest-scoring links $\langle a, a' \rangle$ where a or a' appears. These 36 one-to-one links are shown in Table 1.

Alignment Evaluation

We measure the structural properties of f shown in Table 1 (see[6] for more information). But our primary criterion is that f should not distort the metric relations of concepts, taking nodes that are close together and making them farther apart, or *vice versa*.

For two ontology nodes $a, b \in \mathcal{P}$, their **lower distance** is $d_l(a, b) := |{\downarrow} a| + |{\downarrow} b| - 2 \max_{c \in a \wedge b} |{\downarrow} c|$, where ${\downarrow} x = \{y \leq x\}$ is the set of all descendants of x, and $a \wedge b$ is the set of greatest lower bounds (glbs) below a and b. If a and b lack a glb, we assume a bottom node $0 \in P$ which is below all the leaves. The dual **upper distance** $d_u(a, b) = |{\uparrow} a| + |{\uparrow} b| - 2 \max_{c \in a \vee b} |{\uparrow} c|$ is also available, where ${\uparrow} x = \{y \geq x\}$ is the set of all ancestors of x, and $a \vee b$ is the set of least upper bounds (least common subsumers). Upper distance may appear more natural, but is not generally preferable for technical reasons related to the desire for e.g. siblings deep in the hierarchy to be closer together than siblings high in the hierarchy. While in general it may be preferable to use both in combination, in this paper we use lower distance only.

We can measure the change in distance between $a, b \in P$ induced by f as the **distance discrepancy**

$$\delta(a, b) := |\bar{d}_l(a, b) - \bar{d}_l(f(a), f(b))|,$$

where $\bar{d}_l(a, b) := \frac{d_l(a,b)}{\mathrm{diam}_d(\mathcal{P})} \in [0, 1]$ is the normalized lower distance between a and b in $\mathcal{P}$ given the diameter $\mathrm{diam}_d(\mathcal{P}) := \max_{a,b \in P} d(a, b)$. In this case, we have diam(BP) = 14659, diam(MF)= 8260. Finally, we can measure the entire amount of distance discrepancy at a node $a \in P$ compared to all the other anchors $b \in P$ by summing

$$\delta_f(a) := \sum_{b \in P} \delta(a, b) = \sum_{b \in P} |\bar{d}_l(a, b) - \bar{d}_l(f(a), f(b))|.$$

Note that we use δ_f to indicate that this is an overall discrepancy of a with respect to the entire alignment f. Also note that since f is one-to-one, it is invertible, so $\forall a \in P, \delta_f(a) = \delta_f(f(a))$ and $\forall a' \in P', \delta_f(a') = \delta_f(f^{-1}(a'))$. Thus we can denote $\delta_f(\vec{l}) = \delta_f(a)$ for $\vec{l} = \langle a, f(a) \rangle$, which is also shown in Table 1.

Discussion and Further Work

Fig. 1 shows an abstract representation of a portion of the GO involving the top four scoring XOA links and the top two δ_f links. In general, we are pleased with the quality of the links provided by the XOA scores coupled with the one-to-one link filtering. It is a good sign that the nodes that did come up as significant are ones that make sense in the light of the gene list context (development). With one exception, the top 6 to 8 linked nodes represent molecules and processes associated with cell motility and with known regulators of cellular differentiation, such as the hedgehog signaling pathway. The frequency of nodes associated with motility underscore the importance of cellular migration during differentiation.

The distribution of XOA vs. δ_f is shown in Fig. 2. It can be seen that the XOA scoring method produces a strong alignment, with links having generally low δ_f scores. There are two exceptions which deserve further study to improve the analysis:

BP:GO:0007154 cell communication
MF:GO:0000062 acyl-CoA binding

BP:GO:0000187 activation of MAPK activity
MF:GO:0004672 protein kinase activity

To interpret this, for a given one-to-one link $\vec{l} = \langle a, f(a) \rangle$ between a BP node a and MF node $f(a)$, the XOA score measures the co-annotation of a and $f(a)$, while the δ_f score meaures the distance of $\vec{l}$ from all the other links in virtue of f, that is, the distance of a from all other BP anchors b, and dually the distance of $f(a)$ from all other MF anchors $f(b)$.

The lower distance $d_l(a, b)$ involves the numbers of nodes below a, b, and both of them. Thus from Fig. 1 we can see that both "BP:GO:0007154 cell communication" and "MF:GO:0004672 protein kinase activity" have unusually many nodes below them (341 and 105 respectively). This makes them effectively "far away" from the other nodes in BP and MF, while their corresponding anchor in the other

XOA	δ_f	BP Node	MF Node
10.14	0.070	GO:0006637 acyl-CoA metabolic process	GO:0016290 palmitoyl-CoA hydrolase activity
9.85	0.071	GO:0032927 positive regulation of activin receptor signaling pathway	GO:0050431 transforming growth factor beta binding
9.57	0.072	GO:0050677 positive regulation of urothelial cell proliferation	GO:0042056 chemoattractant activity
9.13	0.072	GO:0007228 positive regulation of hh target transcription factor activity	GO:0005113 patched binding
8.66	0.071	GO:0045723 positive regulation of fatty acid biosynthetic process	GO:0008009 chemokine activity
8.53	0.082	GO:0035023 regulation of Rho protein signal transduction	GO:0005099 Ras GTPase activator activity
8.00	0.079	GO:0048010 vascular endothelial growth factor receptor signaling pathway	GO:0005172 vascular endothelial growth factor receptor binding
7.51	0.087	GO:0050674 urothelial cell proliferation	GO:0005104 fibroblast growth factor receptor binding
7.44	0.076	GO:0016049 cell growth	GO:0005160 transforming growth factor beta receptor binding
7.41	0.233	GO:0048678 response to axon injury	GO:0019899 enzyme binding
7.39	0.103	GO:0007178 transmembrane receptor protein serine/threonine kinase signaling pathway	GO:0004702 receptor signaling protein serine/threonine kinase activity
7.33	0.115	GO:0033144 negative regulation of steroid hormone receptor signaling pathway	GO:0003690 double-stranded DNA binding
6.72	0.177	GO:0009967 positive regulation of signal transduction	GO:0048185 activin binding
6.52	0.080	GO:0007169 transmembrane receptor protein tyrosine kinase signaling pathway	GO:0004714 transmembrane receptor protein tyrosine kinase activity
6.42	0.080	GO:0014044 Schwann cell development	GO:0004675 transmembrane receptor protein serine/threonine kinase activity
6.40	0.103	GO:0045941 positive regulation of transcription	GO:0003713 transcription coactivator activity
6.33	0.075	GO:0048012 hepatocyte growth factor receptor signaling pathway	GO:0005017 platelet-derived growth factor receptor activity
6.31	0.089	GO:0045893 positive regulation of transcription DNA-dependent	GO:0016563 transcription activator activity
6.27	0.071	GO:0042993 positive regulation of transcription factor import into nucleus	GO:0015460 transport accessory protein activity
6.20	0.183	GO:0001558 regulation of cell growth	GO:0019838 growth factor binding
6.08	0.072	GO:0007171 activation of transmembrane receptor protein tyrosine kinase activity	GO:0005161 platelet-derived growth factor receptor binding
6.05	0.071	GO:0030949 positive regulation of vascular endothelial growth factor receptor signaling pathway	GO:0005111 type 2 fibroblast growth factor receptor binding
5.75	0.070	GO:0006919 caspase activation	GO:0019834 phospholipase A2 inhibitor activity
5.57	0.078	GO:0048706 embryonic skeletal development	GO:0005024 transforming growth factor beta receptor activity
5.50	0.415	GO:0000187 activation of MAPK activity	GO:0004672 protein kinase activity
5.46	0.101	GO:0006816 calcium ion transport	GO:0005262 calcium channel activity
5.36	0.776	GO:0007154 cell communication	GO:0000062 acyl-CoA binding
5.30	0.144	GO:0006468 protein amino acid phosphorylation	GO:0004674 protein serine/threonine kinase activity
5.21	0.072	GO:0051795 positive regulation of catagen	GO:0001540 beta-amyloid binding
5.19	0.093	GO:0016481 negative regulation of transcription	GO:0016564 transcription repressor activity
5.17	0.070	GO:0051450 myoblast proliferation	GO:0005021 vascular endothelial growth factor receptor activity
5.04	0.072	GO:0050890 cognition	GO:0019855 calcium channel inhibitor activity
5.01	0.073	GO:0000122 negative regulation of transcription from RNA polymerase II promoter	GO:0003702 RNA polymerase II transcription factor activity
4.85	0.072	GO:0007184 SMAD protein nuclear translocation	GO:0046332 SMAD binding
4.84	0.071	GO:0001707 mesoderm formation	GO:0045545 syndecan binding

Table 1: One-to-one alignment links $\vec{l} = \langle a, f(a) \rangle$ for $p \geq 5\%$, sorted down by XOA score, and showing $\delta_f(\vec{l})$. Underlined links are illustrated in Fig. 1.

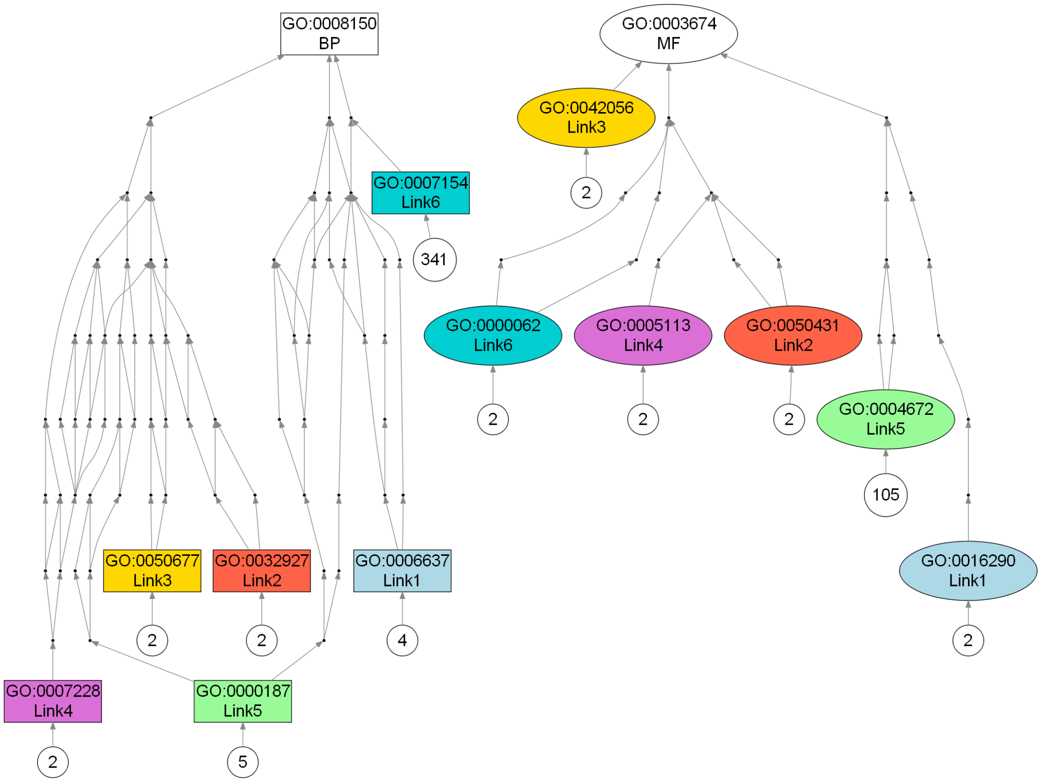

Figure 1: The portion of the BP (left) and MF (right) GO branches involving the top four XOA and top two δ_f links. Only the anchors are shown with their GO IDs (see Table 1 for descriptions). Matching nodes are indicated by color and link numbers. Ancestors are shown, up to the BP or MF root, but all interior nodes are collapsed. Below each anchor is the number of descendant nodes. There are no common nodes below any pair of anchors.

ontology is close to its comrades. This is clear in Fig. 1, and thus our method identifies these links which are clearly significant by XOA, but also distant from the other links.

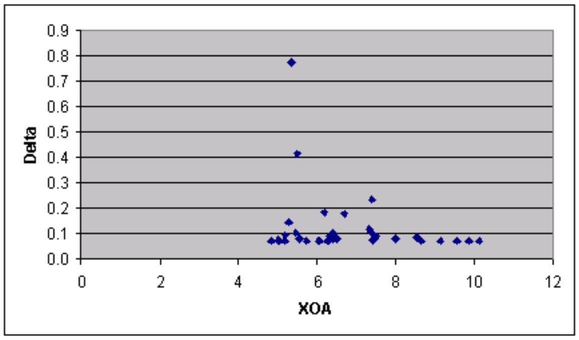

Figure 2: $XOA(a, f(a))$ vs. $\delta_f(a)$.

δ_f provides a measure only about ontology *structure*, and there may be reasons in ontology design or annotation for high δ_f to be preferable, e.g. if it were important that annotations be made high in the structure in some cases. The results would also be different if there were common nodes below pairs of anchors, which is entirely possible in the GO DAG structure with multiple inheritance, especially if the anchors were higher. Finally, note that the number of descendants is correlated both with level in the GO, and the information content (probability) of a node used in semantic similarity calculation. These correlations need to be explored in future work. Further work for a full paper includes:

- There are potential difficulties of mixing experimental and inferential annotations, as reported here, these should be analyzed separately.
- The analytical pipeline needs to be tested for sensitivity at multiple points, especially the filtering to one-to-one links: it is likely that there are links which re-use an anchor which have only a slightly different XOA score, but would produce a preferable mapping according to δ_f. Additionally, the alignment measurement method[6] originally was designed to work on many-to-many alignment relations $F \subseteq P \times P'$, so extensions in this direction may be desirable.
- We have begun analysis on the distribution of δ_f as a function of p-value cutoff.
- Other aspects of the alignment measurement methodology[6] need to be incorporated, including: reconciling the use of upper distance together with lower distance; and the additional use of an **order discrepancy measure**, which rather than being sensitive to the *distances* between links, measures *order violations* (e.g. mapping siblings to parent-child links) implied by an alignment.

Acknowledgements

This work supported under NIH/NINDS grant R01NS057484-03 and NIH/NHGRI grant HG002273.

References

1. Bodenreider OM, Aubry M, Burgun A. “Non-lexical approaches to identifying associative relations in the gene ontology”, *Pacific Symp. Biocomputing*. 2005; 104-115
2. Davey BA, Priestly HA. *Introduction to Lattices and Order*, Cambridge: Cambridge UP; 1990
3. Jérôme E, Shvaiko, P. *Ontology Matching*, Hiedelberg: Springer-Verlag; 2007
4. Gene Ontology Consortium. “The Gene Ontology: Tool For the Unification of Biology”. *Nature Genetics*. 2000; 25(1):25-29
5. Guan Y *et al.* “A Genomewide functional network for the laboratory mouse”. *PLoS Comutational Biology*. 4 e1000165 [PMID: 18818725]. 2008
6. Joslyn CA, Donaldson A, Paulson P: (2008) “Evaluating the Structural Quality of Semantic Hierarchy Alignments”, *Int. Semantic Web Conf. (ISWC 08)*. Available from: http://dblp.uni-trier.de/db/conf/semweb/iswc2008p.html#JoslynDP08
7. Kirsten T, Andreas T, Rahm E: (2007) “Instance-Based Matching of Large Life Science Ontologies”, in S Cohen-Boulakia, V Tannen ed. *DILS 2007, Lecture Notes in Bioinformatics*, 4544. p. 172-187. Heidelberg: Springer-Verlag.
8. Lord PW, Stevens R, Brass, A, Goble CA. “Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship Between Sequence and Annotation”. *Bioinformatics*. 2003; 10:1275-1283
9. Raj JU *et al.* “Genomics and proteomics of lung disease: conference summary”. *Am. J. Physiol Lung Cell Mol Physiol*. 2007; 293:L45-L51, PMID: 17468134
10. Riensche RM, Baddeley BL, Sanfilippo A, Posse C, Goplan B. “XOA: Web-Enabled Cross-Ontological Analytics”. *IEEE Congress on Services*. 2007.
11. Salton GA, Wong A, Yang CS “A Vector space model for automatic indexing”. CACM. 1975; 18(11):613-620.
12. Sanfilippo A, Posse C, Gopalan B, Riensche R, Beagley N, Baddeley B, Tratz S, Gregory M “Combining Hierarchical and Associative Gene Ontology Relations With Textual Evidence in Estimating Gene and Gene Product Similarity”. *IEEE Trans. Nanobio.*. 2007; 6(1):51-59