Poset Ontologies and Concept Lattices as Semantic Hierarchies

Cliff Joslyn

Computer and Computational Sciences
Los Alamos National Laboratory
joslyn@lanl.gov

Abstract. We describe some aspects of our research in relational knowledge discovery and combinatorial scientific computing [11], with special emphasis on the relation to the research portfolio of the conceptual structures community. We have recently been developing [10, 12] a combinatorial approach to the management and analysis of large ontologies such as the Gene Ontology (GO) [6]. Our approach depends on casting the GO as a labeled partially ordered set (poset) [16], and then using scores based on pseudo-distance measures which we have developed to categorize lists of labels (in the case of the GO, genes and gene products) concerning their clustering and depth within the GO. We hold that such taxonomic semantic hierarchies serve as the core conceptual structures underlying all ontological databases, and through this work we have developed a number of what we believe to be both fundamental and novel ideas about treating such large posets as data objects, in particular the nature of *distance* in such structures, and the nature of *level* as an *interval-valued* property. After laying out this basic framework, we can then bring these ideas to a particular kind of poset, namely the concept lattice [5]. Considering a concept lattice as a poset, we are then prepared to develop techniques for anomaly detection in relational data by measuring the relative level of concepts vs. their cardinalities.

1 Introduction

Semantic hierarchies are ubiquitous, not just in formal semantic structures like conceptual graphs, ontologies, and concept lattices (CLs), but also in metamodeling environments, object-oriented typing architectures, and even natural and computational linguistics. We are concerned with the fundamental nature of semantic hierarchies, and report on the current state of our work here.

We open with some discussion of the POSet Ontology Categorizer (POSOC), which was the motivation for the beginning of this work. POSOC was in turn motivated by the need for biologists to use algorithmic tools to navigate the Gene Ontology (GO), the best example of the vast, novel conceptual structures which the genomic revolution has thrust into the world only very recently: very large, taxonomically organized, hierarchical data objects as specialized databases.

Our view is that semantic hierarchies naturally live within the theory of partially-ordered sets (posets), and POSOC was developed on that basis. After

reviewing POSOC's foundations, including some elementary partially ordered set (poset) theory, we then move on to consider general semantic hierarchies, and thus arrive at our core points: explicating our new conceptualizations of level and distance in posets as a vector-valued quantity of the height and width of the neighborhood (defined in a particular way) of a collection of poset nodes.

We conclude with some speculations about the use of such concepts in lattices, particularly CLs, conceived of as proto- or putative ontologies generated in the context of available relational knowledge, in our case, protein-ligand binding.

2 The POSet Ontology Categorizer (POSOC)

The computational biology revolution has produced a proliferation of large databases of genomic information. A premier example is the Gene Ontology (GO)[1] [6], a large ($> 16,000$ node), standardized knowledge structure consisting of three branches: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Each branch is organized as a taxonomy of nodes which represent different categories of genomic characteristics. Once a gene is sufficiently characterized, it can be attached to the appropriate node, as shown in Fig. 1 [6].

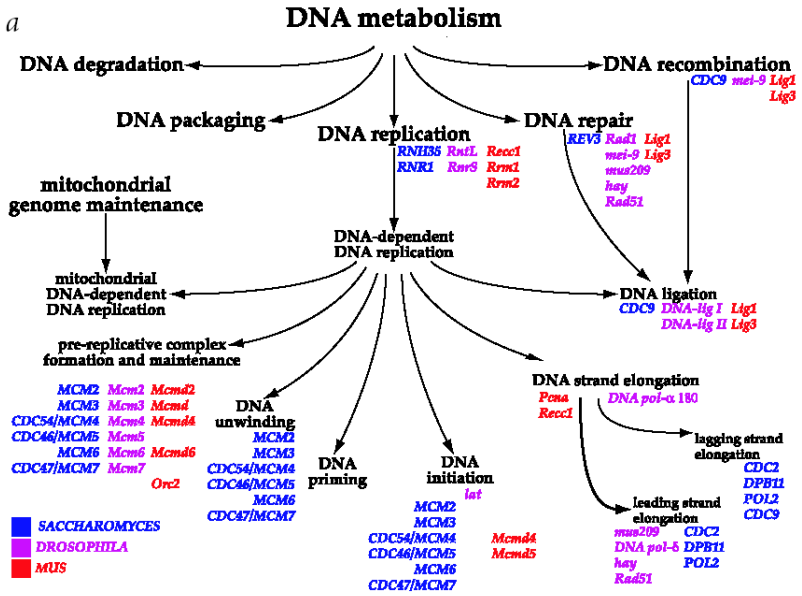

Fig. 1. A portion of the BP branch of the GO [6]. GO nodes in the hierarchy have genes from three species annotated below them.

We have been working on the **categorization** task in the GO, where following a gene expression experiment involving high throughput microarrays or

[1] `http://www.geneontology.org`

Affymetrix gene chips, a biomedical researcher is confronted with a list of a few hundred to a thousand genes, from which she will need to extract useful information about the various types of biological processes that were affected in the experiment. The researcher then wants to take the names of these genes which have been annotated to the GO and gain an understanding of their overall function by examining their distribution through the GO: are they localized, grouped in distinct areas, or spread uniformly? Manual approaches and existing software are inadequate to answer this question over hundreds of proteins and more than $16,000$ GO nodes, and thus an algorithmic approach is necessary.

At its core, the GO is a hierarchy of semantic categories. So to approach this problem, we have needed to address a number of fundamental questions about the nature of such hierarchies, modeled as partially ordered sets (posets), to provide algorithmically determined numerical scoring of the nodes in the GO with respect to the genes of interest. We produce a ranked list of appropriate summarizing nodes within the GO, which act as functional hypotheses about the characteristics of the genes expressed.

POSOC has been developed over the past year [9–12] by researchers at the Los Alamos National Laboratory (LANL) and Procter & Gamble Corp. (P&G), and is currently in use by staff scientists at P&G[2]. In addition, extensions of POSOC to handle textually-based queries have been used recently in a submission by LANL for the BioCreative challenge[3] for automated annotation [18].

2.1 Posets

We first introduce some elementary ideas from the theory of **finite partially ordered sets** (posets). This is mostly standard and elementary [16], but in some cases novel (to our knowledge), at least in terms of notation and perspective.

A finite poset is a structure $\mathcal{P} = \langle P, \leq \rangle$ where P is a finite set and $\leq \subseteq P^2$ is a reflexive, anti-symmetric, transitive binary relation on P. Posets are the most general combinatorial structures admitting to description in terms of *levels*, in our case, levels of semantic generality. While more specific than directed graphs or networks (every poset is a digraph with no cycles), they are more general than trees or lattices (every tree and lattice is a poset), in that collections of nodes can have multiple parents.

The GO is notably a directed acyclic graph (DAG), as is evident in Fig. 1, and every DAG determines both a unique poset and a unique Hasse diagram, in which all transitive links have been removed[4]. In a poset, two nodes $p, q \in P$ are **comparable**, denoted $p \sim q$, if either $p \leq q$ or $p \leq q$; a **chain** $C \subseteq P$ is a collection of comparable nodes; and the **height** $\mathcal{H}(\mathcal{P})$ is the size of the largest chain. Similarly, two nodes $p, q \in P$ are **non-comparable** if $p \not\sim q$, an **anti-chain** is a collection of non-comparable nodes, and the **width** $\mathcal{W}(\mathcal{P})$ is the size of the largest anti-chain. For any node $p \in P$, its **ideal** is $\downarrow p := \{q \in P : q \leq p\}$,

[2] As previously reported [9, 10, 12], POSOC was originally targeted specifically at the GO, and was thus called the Gene Ontology Categorizer (GOC). GOC has now been generalized to deal with any poset ontology, and is thus now called POSOC here.

[3] `http://www.mitre.org/public/biocreative`

[4] I.e., if both $a \leq b$ and $b \leq c$ are included, then if $a \leq c$ is present, it is removed.

its **filter** is $\uparrow p := \{q \in P : q \geq p\}$, and its **hourglass** is $\Xi(p) := \uparrow p \cup \downarrow p$. Define these concepts over a collection of nodes $Q \subseteq P$ similarly:

$$\downarrow Q := \bigcup_{p \in Q} \downarrow p, \quad \uparrow Q := \bigcup_{p \in Q} \uparrow p, \quad \Xi(Q) := \bigcup_{p \in Q} \Xi(p).$$

For any subset $Q \subseteq P$, a node $p \in Q$ is **maximal** in Q if $\nexists q \in Q, q > p$. Let $\mathrm{Max}(Q)$ be the set of all maximal nodes in Q, noting that $\mathrm{Max}(Q)$ must be non-empty if Q is non-empty. Define the set of all minimal nodes $\mathrm{Min}(Q)$ dually. For any two nodes $p, q \in P$ the set $\uparrow p \cap \uparrow q$ is their "joint filter" in some sense, and $p \vee q := \mathrm{Min}(\uparrow p \cap \uparrow q)$ are their **joins**. For a collection of nodes $Q \subseteq P$, let

$$\bigvee Q := \mathrm{Min}\left(\bigcap_{p \in Q} \uparrow p\right).$$

Lower bounds and meets $\wedge$ are defined dually. Note that posets are distinguished from lattices in that $p \vee q \subseteq P$ is not a single node, and is not guaranteed to exist, but is rather an arbitrary, possibly empty, subset of nodes.

If there exists a node $1 \in P$ such that $\mathrm{Max}(P) = \bigvee P = \{1\}$, then we say that $\mathcal{P}$ is **upper-bounded**, and dually for $0 \in P$. If either there is no unique upper or lower bound $0, 1 \in P$, then we can create them easily by constructing the **closure** of $\mathcal{P}$ as $\bar{\mathcal{P}} := \langle P \cup \{0, 1\}, \bar{\leq} \rangle$, where $\forall p, q \in P, p \bar{\leq} q \leftrightarrow p \leq q$, and $\forall p \in P, 0 \bar{\leq} p \bar{\leq} 1$. Most of our results below require either an upper-, lower-, or totally bounded poset. We will presume that when $\mathcal{P}$ is not naturally so bounded, its closure $\bar{\mathcal{P}}$ is available in this way.

For two comparable nodes $p \leq q$, all the nodes "between" them is the **interval** $[p, q] = \{t : p \leq t \leq q\} = \uparrow p \cap \downarrow q$. For comparable subsets $P_1, P_2 \subseteq P$ with $P_1 \leq P_2$ (so that $\forall p \in P_1, q \in P_2, p \leq q$), their interval $[P_1, P_2]$ is

$$[P_1, P_2] := \bigcup_{\langle p_1, p_2 \rangle \in P_1 \times P_2} [p_1, p_2].$$

For two comparable nodes $p \leq q$, the interval $[p, q]$ is equivalent to the set of all chains between p and p, denoted $\mathcal{C}(p, q)$. The **vector of chain lengths** $\boldsymbol{h}(p, q) := \langle |C(p, q)| \rangle$ is the collection of the lengths of all these chains, and finally the minimal and maximum chain lengths between p and q respectively are $h_*(p, q) := \min_{C \in \mathcal{C}(p,q)} |C|$ and $h^*(p, q) := \max_{C \in \mathcal{C}(p,q)} |C|$[5].

The Hasse diagram of an example of a poset on a set of nodes $P = \{1, A, B, \ldots, K\}$ is shown in Fig. 2. Note the inherently two-dimensional structure displayed by division into levels: while nodes can be re-drawn left to right (width) as convenient, vertically it's crucial that higher nodes be placed above lower ones (height).

2.2 The GO as a Labeled Poset

The GO has measurable poset properties, as shown in Tab. 1 and Fig. 3 (GO for September, 2003). The height parameter shows that the GO is properly seen

[5] Here we assume the Hasse diagram, otherwise $p \leq q \rightarrow h_*(p, q) = 1$.

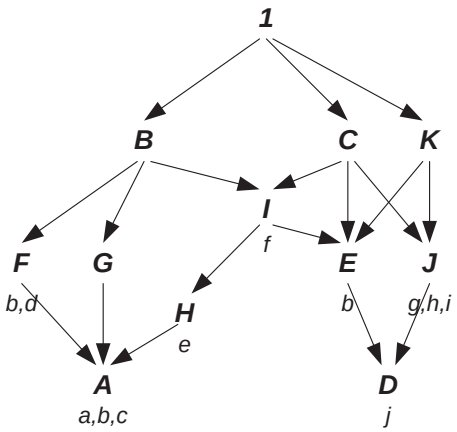

Fig. 2. An example of a labeled poset.

as a structure divided into levels, 15 for BP and 13 for MF and CC. It branches out quickly and broadly, with twice as many nodes (10.6K) being "terminal" leaves compared to interior nodes (only 5.4K). Calculating the width of a poset is still daunting algorithmically, so the width shown here is only a lower-bound estimate. Thus the structure is at least three orders of magnitude wider than it is high. Fig. 3 shows the distribution (on a log scale) of the number of parents and children per node. Note that a few nodes have hundreds of children, and a substantial quantity have at least two parents, some as many as four or five.

	Nodes	Leaves	Interior	Edges	$\mathcal{H}$	$\mathcal{W}$
MF	7.0K	5.6K	1.3K	8.1K	13	$\geq$ 3.5K
BP	7.7K	4.1K	3.6K	11.8K	15	$\geq$ 2.9K
CC	1.3K	0.9K	0.4K	1.7K	13	$\geq$ 0.4K
GO	16.0K	10.6K	5.4K	21.5K	16	$\geq$ 5.9K

Table 1. Poset statistics of the GO.

We can then define a structure $\mathcal{O} := \langle \mathcal{P}, X, F \rangle$ as a **POSet Ontology (POSO)**, where X is a finite, non-empty set of **labels**, and $F\colon X \mapsto 2^P$ is a function mapping each label $x \in X$ to a collection of nodes $F(x) \subseteq P$. In Fig. 2, we have $X = \{a, b, \ldots, j\}$, and e.g. $F(b) = \{A, E, F\}$. In the case of the GO, then P is the collection of GO nodes, $\leq$ is the ordering relations present in the GO, and X is the set of genes annotated to the GO, as illustrated in Fig. 1.

2.3 POSOC Methodology

We can now pose the categorization problem in the context of the example in Fig. 2: given a particular set of genes of interest cast as a query, say $Y =$

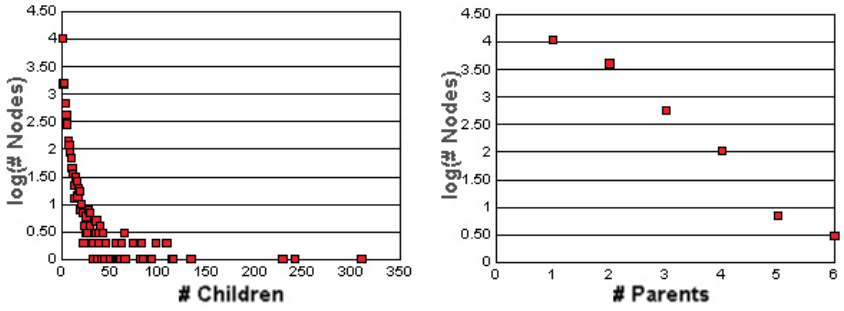

Fig. 3. Distribution of number of children (left) and parents (right) per node.

$\{c, e, i\} \subseteq X$, what node(s) in P best summarize that set? One answer is C, since it "covers" all three genes, and does so in the most specific way. The node 1 also covers the genes, but would not be favored since it's a more general category. But it can also be argued that H is a good answer, since, while it only covers c and e, it does so more specifically than C does. We will see that this interplay between "coverage" and "specificity" will be central to the methodology developed.

To proceed, we need the concept of a **pseudo-distance** as a function $\delta\colon P^2 \mapsto \mathbb{R}$ where $\forall p \leq q \in P, h_*(p,q) \leq \delta(p,q) \leq h^*(p,q)$; and a **normalized distance** as $\bar{\delta} := \delta/\mathcal{H}(\mathcal{P})$. Current pseudo-distances implemented in POSOC include: the **minimum path length** $\delta_m := h_*$; the **maximum path length** $\delta_x := h^*$; the **average of extreme path lengths** $\delta_{ax}(p,q) := \frac{h_*(p,q)+h^*(p,q)}{2}$; and the **average of all path lengths** $\delta_{ap}(p,p') := \frac{\sum_{h \in \boldsymbol{h}(p,q)} h}{|\boldsymbol{h}|}$. Other candidate pseudo-distances are in exploration.

Given a pseudo-distance and a set of nodes of interest $Y \subseteq X$, we then want to develop a **scoring function** $S_Y(p)$ which returns the weighted rank of a node $p \in P$ based on requested nodes Y. We actually use two kinds of scores, an **unnormalized score** $S_Y\colon P \mapsto \mathbb{R}^+$ which returns an "absolute" number, and a **normalized score** $\hat{S}_Y\colon P \mapsto [0,1]$ which returns a "relative" number. We allow the user to choose the relative value placed on coverage vs. specificity by introducing a parameter $s \in \{\ldots -1, 0, 1, 2, 3, \ldots\}$, where low s emphasizes coverages, and high s emphasizes specificity. The scoring function can use either the unnormalized distance δ, or the normalized $\bar{\delta}$. Letting $r = 2^s$, we have the four scoring functions shown in Tab. 2.

We then find non-comparable nodes within the ranked-list to serve as "cluster heads". The resulting clusters are at different depths in $\mathcal{P}$: while "headed" by non-comparable nodes, their contents (the collection of their descendants in $\mathcal{P}$) can overlap. Cluster heads which are non-comparable to all other cluster heads of lower rank are called "primary", and those above some previously identified cluster head "secondary".

Output for the example in Fig. 2 is shown in Tab. 3, for query $Y = \{c, e, i\}$, specificity values $s = -1, 1$, and 3, doubly-normalized score $\hat{\bar{S}}$, and pseudo-

Distance	Score	
	Unnormalized	Normalized
Unnormalized	$S_Y(p) \coloneqq \sum_{x \in Y} \sum_{p' \in F(x): p' \leq p} \left(\delta^r(p', p) + 1\right)^{-1}$	$\hat{S}_Y(p) \coloneqq \frac{S_Y(p)}{\sum_{x \in Y} \lvert F(x) \rvert}$
Normalized	$\bar{S}_Y(p) \coloneqq \sum_{x \in Y} \sum_{p' \in F(x): p' \leq p} \left(1 - \bar{\delta}(p', p)\right)^r$	$\hat{\bar{S}}_Y(p) \coloneqq \frac{S_Y(p)}{\sum_{x \in Y} \lvert F(x) \rvert}$

Table 2. Scoring functions.

distance δ_m. Cluster heads are shown in bold, and secondaries are labeled with $*$. Inspection reveals desirable results: for low specificity, C is the preferred primary cluster, with 1 a secondary; for high specificity, H and J are preferred (J specifically covers i), with C as the next-ranked secondary.

	$s = -1$		$s = 1$		$s = 3$	
Rank	$\hat{\bar{S}}_Y(p)$	p	$\hat{\bar{S}}_Y(p)$	p	$\hat{\bar{S}}_Y(p)$	p
1	0.7672	**C**	0.5467	**H**	0.3893	**H**
2	0.6798	**1***	0.3867	**C***	0.3333	A;**J**
3	0.6315	H	0.3333	A;I;J		
4	0.5563	I			0.0617	**C***
5	0.5164	B			0.0615	I
6	0.3333	A;J	0.2400	**B***	0.0559	F;G;K
7			0.2267	**1***		
8	0.2981	F;G;K	0.2133	F;G;K		
9					0.0112	B
10					0.0060	**1**

Table 3. POSOC output for example in Fig. 2 for query $Y = \{c, e, i\}$.

POSOC was validated by a highly experienced molecular immunologist who had no prior knowledge of the POSOC to assess its utility and accuracy [12]. It was also validated formally by comparing POSOCs annotations to a collection of independent annotations of collections of GO nodes (corresponding to our lists of target genes) available through the InterPro project[6], which catalogs assignments of protein families, domains, and functional sites to GO IDs [12].

As noted, we are in the process of generalizing POSOC's implementation to target any POSO, not just the GO. Current targets include the Enzyme Commission (EC) database[7] and the MEdical Subject Headings (MESH) ontology[8].

3 Requirements for Working with Semantic Hierarchies

While modern bio-ontologies take many forms, an adequate overall description is of a taxonomically organized data object over which automated inference and

[6] `http://www.ebi.ac.uk/interpro`
[7] `http://www.biochem.ucl.ac.uk/bsm/enzymes`
[8] `http://www.nlm.nih.gov/mesh/meshhome.html`

reasoning (for example using description logics [2]) is performed. Leading research in ontologies tends to focus on logical properties, inference, and search. Our view is that what has made existing bio-ontologies such as the GO so successful are their attributes as hierarchical, taxonomic, categorizations of biological objects, coming closer to being specially structured databases.

Moreover, these attributes are fundamental to other aspects: it is clear that large taxonomically-organized database can be very useful without an inference engine, but the converse is not so evident. Indeed, semantic hierarchies are truly ubiquitous. Even a casual observation reveals them at the foundations of knowledge architectures such as conceptual graphs [17], as object-oriented data types [14], in CLs and related work [5], and even in verb type hierarchies from cognitive linguistics [4]. And yet there seems to be little attention paid to the need for algorithmic approaches to their representation, analysis, navigation, manipulation, and measurement, or even their generic properties as formal structures.

While there are no doubt many reasons for this, these likely include the relatively later development of poset theory as compared to lattices and networks (the first serious textbook appeared in 2003 [16]), and especially the novel appearance of these large, taxonomically organized knowledge objects which now *require* this kind of computer-scientific approach.

So we are motivated to continue in a number of directions:

- First, we have found our pseudo-distances δ lacking, as they are only available between comparable nodes. We are thus seeking to generalize this idea to a more inclusive measures of distance, size, level, etc.
- There are many more tasks which need to be addressed within the overall poset ontology world than the categorization task. Examples include:

  Matching: How do we match two parts of a poset ontology? This arises, for example, in both the BioCreative task and the InterPro validation of POSOC, where POSOC has provided certain answers, and we wish to compare those to some "correct" answer provided by someone else. This can be formalized as follows: assume a poset $\mathcal{P} = \langle P, \leq \rangle$, with $P_1, P_2 \subseteq P$, inducing the sub-posets $\mathcal{P}_1 = \langle P_1, \leq|_{P_1} \rangle$ and $\mathcal{P}_2 = \langle P_2, \leq|_{P_2} \rangle$. How can we then measure the similarity of $\mathcal{P}_1$ and $\mathcal{P}_2$?

  Comparison: Assume now that we have two different orderings available on the same underlying set, for example ontologies constructing by different teams of researchers. How can we compare their similarity? This can also be formalized as assuming $\mathcal{P}_1 := \langle P, \leq_1 \rangle$ and $\mathcal{P}_2 := \langle P, \leq_2 \rangle$, then how can we measure the similarity of $\mathcal{P}_1, \mathcal{P}_2$?

  Merger: Finally we have the most general formulation of the problem, assuming two complete different ontologies $\mathcal{P}_1 := \langle P_1, \leq_1 \rangle$ and $\mathcal{P}_2 := \langle P_2, \leq_2 \rangle$. How can we hope to measure their similarity, and ultimately find ways to merge them together into some new poset $\mathcal{P}$ on $P_1 \cup P_2$?

  The general situation is illustrated in Fig. 4, where the EC and the GO are shown as posets on different underlying sets P, but with the same set of labels X. This common labeling can also be used as a source of comparison information, showing, for example, similarity between nodes A, G, F of GO

and E, J of EC in virtue of the annotation of genes b, g, h, i, some of which are analogous, and some of which (e.g. i) are not.

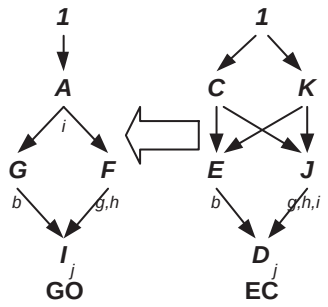

Fig. 4. Cartoon of the general ontology matching problem between the EC and GO.

- We are also interested in considering CLs as semantic hierarchies (see Sec. 5), and using formal measures of level and distance in them to induce hypotheses about both extractable knowledge and potential anomalies in data sets.
- Indeed, this general class of problems arises in a number of more specialized lattices and posets, for example posets of system reconstruction hypotheses in multi-dimensional statistical analysis [8, 13, 15] and classes of random sets in generalized information theory [7].

4 Measures in Semantic Hierarchies

In all these instances, what is required are much better conceptualizations of measures in posets. Our thoughts extend to two important concepts: a general, interval-valued concept of vertical **level** or **rank** within a poset; and a general, vector-valued concept of overall **distance** between two arbitrary nodes.

The scope for this paper allows only a partial formal development. Here we introduce a few suggestive definitions and results, and refer the reader to future work for a detailed development, including more proofs of the basic results.

4.1 Interval-Valued Poset Rank

Rank as a measure of the vertical "level" of a node is an important combinatorial concept [1, 3], but usually used only in more constrained combinatorial structures such as lattices or so-called Jordan-Dedekind, or JD, posets[9]. We have found [16] rank to be defined in posets in a lower-bounded way:

$$r_*(p) := \begin{cases} 0,\ p \in \mathrm{Min}(P) \\ n,\ p \in \mathrm{Min}\left(P - \{q : r_*(q) < n\}\right) \end{cases}$$

[9] Those where all chains between comparable nodes have the same length.

We use r_* suggestively, as its dual function is readily available:

$$r^*(p) := \begin{cases} 0,\ p \in \mathrm{Max}(P) \\ n,\ p \in \mathrm{Max}\left(P - \{q : r^*(q) < n\}\right) \end{cases}$$

And so an **interval rank** function can be easily identified as $R\colon P \mapsto \mathcal{D}(\mathcal{H}(\mathcal{P}))$, with $R(p) := [r_*(p), \mathcal{H}(\mathcal{P}) - r^*(p)]$[10], where $\mathcal{D}(n) := \{[x, y] : x, y \in \mathcal{I}^+, 0 \leq x \leq y \leq n\}$ is the set of all integer intervals for $n \in \mathcal{I}^+$. An example of a bounded poset equipped with its interval rank function is shown in Fig. 5, so that $R(D) = [1, 2], R(K) = [1, 4]$.

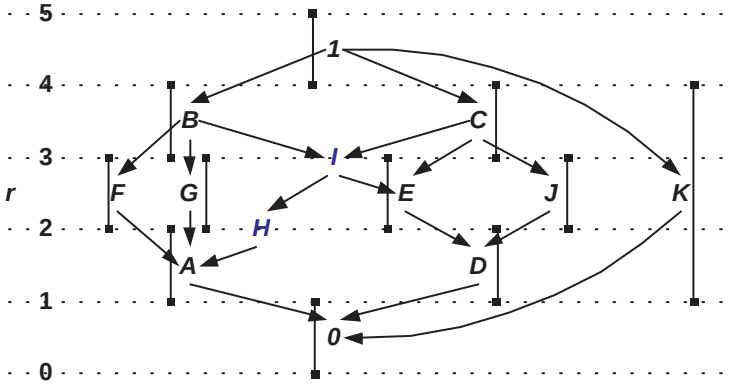

Fig. 5. A bounded poset equipped with its interval rank.

To conceptualize the interval rank, first, $\mathcal{P}$ can possess at most $\mathcal{H}(\mathcal{P})$ "levels". So for any node $p \in P$, it's level has to be no less than as "high up" from the bottom as it is, but no more than how "far down" from the top it is.

Theorem 1. $R(p) = [h_*(0, p), \mathcal{H}(\mathcal{P}) - h_*(p, 1)]$

Proof. Since $R(p) = [r_*(p), \mathcal{H}(\mathcal{P}) - r^*(p)]$, it is sufficient to show that $r_*(p) = h_*(0, p)$, and $r^*(p) = h_*(p, 1)$. For $p = 0$, clearly $r_*(0) = 0 = h_*(0, 0)$. So if $p \in \mathrm{Min}(P - \mathrm{Min}(P))$, then $r_*(p) = 1$, and also $h_*(0, p) = 1$. Indeed, at each step $\mathrm{Min}\left(P - \{q : r_*(q) < n\}\right) = \{q : h_*(0, q) = n\}$, and thus r_* partitions P into bands of lower rank n with $h_* = n$. Thus $\forall p \in P, r_*(p) = h_*(0, p)$. $r^*(p) = h_*(p, 1)$ follows by a dual argument.

R induces an order mapping $R\colon \mathcal{P} \mapsto \langle \mathcal{D}, \preceq \rangle$ from $\mathcal{P}$ to the set of integer intervals $\mathcal{D}$, where $\preceq$ is an interval order on $\mathcal{D}$. Whether R is order preserving

[10] Note that if we instead use $R(p) := [\mathcal{H}(\mathcal{P}) - r_*(p), r^*(p)]$, the "top" and "bottom" as in Fig. 5 are simply reversed.

depends on the interval order used. For two integer intervals $I = [I_*, I^*], J = [J_*, J^*] \in \mathcal{D}$, consider the following weak and strong interval orders:

$$I \preceq_w J := I_* \leq J_* \text{ and } I^* \leq J^*, \qquad I \preceq_s J := I^* \leq J_*.$$

Theorem 2. *R is order preserving from $\mathcal{P}$ to $\langle \mathcal{D}, \preceq_w \rangle$, but not to $\langle \mathcal{D}, \preceq_s \rangle$.*

Proof. Let $p \leq q$. Then $h_*(0,p) \leq h_*(0,q)$, and $h_*(q,1) \leq h_*(p,1)$. Thus we have $R(p) = [h_*(0,p), \mathcal{H}(\mathcal{P}) - h_*(p,1)] \preceq_w R(q) = [h_*(0,q), \mathcal{H}(\mathcal{P}) - h_*(q,1)]$ directly, but it might be that $\mathcal{H}(\mathcal{P}) - h_*(p,1) \leq h_*(0,q)$ or not, and so it could be that $R(p)$ and $R(q)$ are non-comparable in $\preceq_s$.

We also have the following unproved conjecture about how scalar-valued rank arises as a special case of our interval-valued rank.

Conjecture 1. Assume a fully bounded poset $\mathcal{P}$, and a node $p \in P$ with $r_*(p) = r^*(p)$ so that $R(p) = [r, r]$ for some specific $r \in \mathcal{I}$. Then $\forall C \in \mathcal{C}(0,1), p \in C$ iff C is maximal in the sense of $|C|$. Moreover, $\forall p \in P, R(p) = [r, r]$ iff $\mathcal{P}$ is JD.

For example, in Fig. 5, we have $R(H) = [2, 2] = 2$, and H is only on a maximal chain $0 \leq A \leq H \leq I \leq B \leq 1$.

4.2 Vector-Valued Poset Distance

In conjunction with our new sense of "vertical distance" in posets, we also wish to have a general sense of distance which captures the horizontal component as well. Towards that end, for some collection of nodes $Q \in P$, including both comparable and non-comparable pairs, we need to characterize the nodes "between" them in some sense. We characterize this as the **neighborhood** of Q, and our sense of distance is directly related to some measure of the "size" of this region of P. This should be a vector quantity consisting of a horizontal and vertical component, since these concepts are so distinct in posets.

We have some preliminary ideas in this direction, which we report here, although we regret that we haven't yet explored the implications of our definitions deeply yet, nor the relationship to interval-valued rank described above.

Definition 1 (Neighborhoods). *Assume a poset $\mathcal{P}$ and a collection of nodes $Q \in P$. If $\bigvee Q$ exists, then define the* **upper neighborhood** *of Q as the intersection of its filter and the ideal of its lubs:*

$$N^*(Q) := \uparrow Q \cap \downarrow \left(\bigvee Q \right).$$

If $\bigwedge Q$ exists, then define the **lower neighborhood** *of Q dually:*

$$N_*(Q) := \downarrow Q \cap \uparrow \left(\bigwedge Q \right).$$

When both $\bigvee Q$ and $\bigwedge Q$ exist, then define the **neighborhood** *as the intersection of the hourglass and the interval between the lubs and glbs of Q:*

$$N(Q) := \Xi(Q) \cap \left[\bigwedge Q, \bigvee Q \right].$$

In all cases, if $|Q| = 2$ so that $Q = \{p, q\}$, then define for each appropriate form e.g. $N(p,q) := N(Q)$.

Note that $N(Q)$ exists because necessarily $\bigwedge Q \leq \bigvee Q$. A simplified cartoon of the appearance of $N^*(p,q)$ is shown in Fig. **??**.

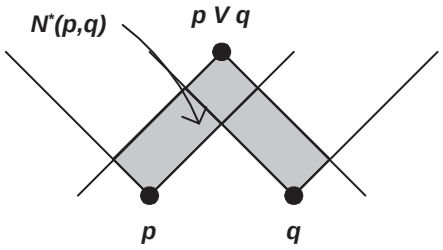

Fig. 6. Cartoon of the upper neighborhood $N^*(p,q)$ (shaded region).

The idea is to say that the nodes in the neighborhood of Q should be "entrained" by both the filter $\uparrow Q$ and ideal $\downarrow Q$ (that is, by the hourglass $\Xi(Q)$), but then also should not be "higher" than the joins $\bigvee Q$, nor "lower" than the meets $\bigwedge Q$; indeed, they should be only those parts of the hourglass between $\bigwedge Q$ and $\bigvee Q$. Thus we have:

Conjecture 2. $N(Q)$ is the set of all chains between $\bigwedge Q$ and $\bigvee Q$ which go through some node of Q.

Chains have no horizontal width, so an easy special case is recovered.

Theorem 3. *If $C = \{p_1, p_2, \ldots, p_n\} \subseteq P$ is a chain with $p_1 \leq p_2 \leq \ldots \leq p_n$, then $N(C) = [p_1, p_n]$.*

Proof. Let $C = \{p_1, p_2, \ldots, p_n\}$ be a chain with $p_1 \leq p_2 \leq \ldots \leq p_n$. Then $\uparrow p_1 \supseteq \uparrow p_i$ for all $2 \leq i \leq n$, and $\downarrow p_n \supseteq \downarrow p_i$ for all $1 \leq i \leq n-1$. Also, $\bigwedge C = p_1, \bigvee C = p_n$ both exist, so that $[\bigwedge C, \bigvee C] = [p_1, p_n]$. Thus we have:

$$\begin{aligned} N(C) &= \Xi(C) \cap \left[\bigwedge C, \bigvee C\right] \\ &= \left(\left(\bigcup_{i=1}^{n} \uparrow p_i\right) \cup \left(\bigcup_{i=1}^{n} \downarrow p_i\right)\right) \cap [p_1, p_n] \\ &= (\uparrow p_1 \cup \downarrow p_n) \cap (\uparrow p_1 \cup \downarrow p_n) = \uparrow p_1 \cup \downarrow p_n = [p_1, p_n]. \end{aligned}$$

Note the trivial corollary that $p \leq q \rightarrow N(p,q) = [p,q]$.

We now define a vector-valued distance in terms of these neighborhood.

Definition 2 (Size and Distance). *Assume a bounded poset $\mathcal{P}$. Then let the vector-valued size of a collection of nodes $Q \subseteq P$ be*

$$\mathbf{D}(Q) := \langle \mathcal{H}(N(Q)), \mathcal{W}(N(Q)) \rangle ,$$

and the vector-valued distance between two nodes $p, q \in P$ be $\mathbf{D}(p,q) := \mathbf{D}(\{p,q\})$.

For examples, consider that in Fig. 2, we have

$$N(B,J) = \Xi(B,J) \cap [B \wedge J, B \vee J] = (P - \{E\}) \cap [D,1] = [D,1],$$
$$\mathbf{D}(B,J) = \langle \mathcal{H}([D,1]), \mathcal{W}([D,1]) \rangle = \langle 5,3 \rangle,$$

and in Fig. 5, we have

$$\begin{aligned} N(J,K) &= \Xi(J,K) \cap [J \wedge K, J \vee K] = (\{0,D,J,C,1\} \cup \{0,K,1\}) \cap [0,1] \\ &= \{0,D,J,C,1,K\} \cap P = \{0,D,J,C,1,K\}, \\ \mathbf{D}(J,K) &= \langle \mathcal{H}(\{0,D,J,C,1,K\}), \mathcal{W}(\{0,D,J,C,1,K\}) \rangle = \langle 5,2 \rangle. \end{aligned}$$

We recover a pseudo-distance easily for the case of comparable nodes.

Theorem 4. *If $p \leq q$, then $\mathbf{D}(p,q) = \langle \delta_x(p,q), 1 \rangle$.*

Proof. If $p \leq q$, then we know from Thm. 3 that $N(p,q) = [p,q]$, and thus $\mathbf{D}(p,q) = \langle \mathcal{H}([p,q]), \mathcal{W}([p,q]) \rangle = \langle h^*(p,q), 0 \rangle = \langle \delta_x(p,q), 1 \rangle$.

In the future, we may recover other pseudo-distances δ if we first restrict our sense of height $\mathcal{H}(\mathcal{P})$ to bounded posets, but then relax it to be the interval $\mathcal{H}(\mathcal{P}) = [h_*(0,1), h^*(0,1)]$ instead of the scalar $\mathcal{H}(\mathcal{P}) = h^*(0,1)$.

5 Distance Measures in Concept Lattices

We recognize formal concept analysis (FCA) as both a foundational tool for the representation of relational information [5], and a way to extract semantic hierarchies from relational data. The trivial observation that every lattice is a poset opens the way to the consideration of the application of our ideas about levels and distances to nodes in CLs.

Space precludes a detailed exposition, instead we refer to the standard references [5]. Instead, we will simply assert the availability of a **context** as a binary relation $R \subseteq X \times Y$ which generates a poset $\mathcal{L} = \langle P, \leq \rangle$, where: $\mathcal{L}$ is actually a lattice, in particular the concept lattice of R; $P \subseteq 2^X \times 2^Y$ is a set of **concepts** generated by R; and $\leq$ is the subset ordering such that $p \leq q := p = \langle A_1, B_1 \rangle, q = \langle A_2, B_2 \rangle$, with $A_1 \subseteq A_2$ and $B_1 \supseteq B_2$, where $A_1, A_2 \subseteq X, B_1, B_2 \subseteq Y$.

Lattices as Special Posets: When a poset $\mathcal{P}$ is a lattice (recalling that P here is finite), we always have that $\forall Q \subseteq P, \bigvee Q$ and $\bigwedge Q$ exist, and furthermore as unique members of P. Thus the formulations of Def. (1) and (2) become largely simplified, for example:

$$N(p,q) = (\uparrow p \cup \uparrow q \cup \downarrow p \cup \downarrow q) \cap (\uparrow(\downarrow p \cap \downarrow q) \cap \downarrow(\uparrow p \cap \uparrow q)).$$

We do not know of the further significance of this at this time, and in particular what the meaning of these kinds of expressions are specifically in the context of $\mathcal{P}$ being a CL $\mathcal{L}$. Indeed, we are suspicious that we are probably recapitulating or generalizing known results from lattice theory [3].

Ontology Induction: One of the great challenges in ontology work is the ability to create ontologies from other information sources such as relational or statistical data. CLs provide such an opportunity. In our case, we are working with molecular biologists and machine learning researchers who are creating relational knowledge bases of the interaction between a set of proteins $R = \{r_i\}$ and ligands (smaller molecules which bind to them to form biologically active complexes) $L = \{l_j\}$. As illustrated in Fig. 7, this provides a formal context in $R \times L$ relating proteins to those ligands with which they bind, and the resulting CL is a semantic hierarchy categorizing proteins r in the context of those ligands l which they bind, and *vice versa*. In this way, an actual POSO $\mathcal{O}$ is generated, where concepts $P = 2^L$ are collections of ligands, $\leq$ is $\subseteq$ on L, $X = R$, and F is determined by the concept lattice. Thus $\mathcal{O}$ is fodder for our methodology, including POSOC for categorization, but also explorations of mappings from these proto-ontologies to other existing ontologies such as the GO or EC.

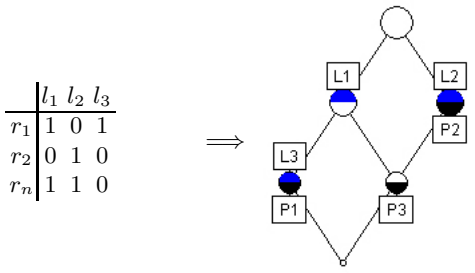

Fig. 7. Mapping a protein-ligand binding relation to its concept lattice proto-ontology.

Anomaly Detection: We conclude with the final direction in which we would like to take this work, namely the use of measures in semantic hierarchies to detect anomalies in relational data as represented in CLs. Simply put, depending on the semantics of the formal context being represented, there may or may not be an expected distribution of nodes in the CL with respect to their cardinalities, that is $|A|$ and $|B|$. In other words, object concepts (where $|A| = 1$) should be "low" in the hierarchy, and attribute concepts (where $|B| = 1$) "high". When this is not the case, it indicates an usual object, attribute, or collection thereof. Much more needs to be explored here, but for now we will leave this as a suggestion for the community to consider further development.

Acknowledgements Many people supported the original development of GOC and POSOC, especially Susan Mniszewski at LANL and Andy Fulmer and Gary Heaton at P&G. Recent support has been from the Protein Function Inference

project at LANL, and in particular from Tom Terwilliger of LANL Biosciences and Michael Wall of LANL Computer Sciences.

References

1. Aigner, M: (1979) *Combinatorial Theory*, Springer-Verlag, Berlin
2. PG Baker, CA Goble, S Bechhofer, N Paton, R Stevens, A Brass: (1999) "An Ontology for Bioninformatics Applications", *Bioinformatics*, v. **15**:6, pp. 510-520
3. Birkhoff, Garrett: (1940) *Lattice Theory*, v. **25**, AMS, Providence RI, 3rd edition
4. Davis, Anthony R: (2000) *Types and Constraints for Lexical Semantics and Linking*, Cambridge UP
5. B Ganter and W Rudolf: (1999) *Formal Concept Analysis*, Springer-Verlag
6. Gene Ontology Consortium: (2000) "Gene Ontology: Tool For the Unification of Biology", *Nature Genetics*, v. **25**:1, pp. 25-29
7. Joslyn, Cliff: (1996) "Aggregation and Completion of Random Sets with Distributional Fuzzy Measures", *Int. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, v. **4**:4, pp. 307-329
8. CA Joslyn and SM Mniszeiski: (2002) "DEEP: Data Exploration through Extension and Projection", LAUR 02-1330, `ftp://wwwc3.lanl.gov/pub/users/joslyn/deep.pdf`
9. CA Joslyn, SM Mniszewski, AW Fulmer, GA Heaton: (2003) "Measures on Ontological Spaces of Biological Function", *Pacific Symp. Biocompuating (PSB 03)*
10. Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy, and Gary Heaton: (2003) "Structural Classification in the Gene Ontology", in: *Proc. 6th Bio-Ontologies Workshop, Intelligent Systems for Molecular Biology (ISMB 03)*
11. Joslyn, Cliff and Mniszewski, Susan: (2004) "Combinatorial Approaches to Bio-Ontology Management with Large Partially Ordered Sets", in: *SIAM Workshop on Combinatorial Scientific Computing (CSC 04)*
12. Joslyn, Cliff; Mniszewski, Susan; Fulmer, Andy, and Gary Heaton: (2004) "The Gene Ontology Categorizer", submitted to the *2004 Conf. on Intelligent Systems for Molecular Biology (ISMB 04)*
13. Klir, George and Doug Elias: (2003) *Architecture of Systems Problem Solving*, Plenum, New York, 2nd edition
14. Knoblock, Todd B and Rehof, Jakob: (2000) "Type Elaboration and Subtype Completion for Java Bytecode", in: *Proc. 27th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*
15. Malvestuto, FM: (1996) "Testing Implication of Hierarchical Log-Linear Models for Probability Distributions", *Statistics and Computing*, v. **6**, pp. 169-176
16. Schröder, Bernd SW: (2003) *Ordered Sets*, Birkhauser, Boston
17. Sowa, John F: (2000) *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks/Cole, Pacific Grove
18. Verspoor, Karin; Simas, Tiago; Joslyn, C, *et al.*: (2004) "Protein Annotation as Term Categorization in the Gene Ontology", submitted to *BioCreative Workshop*