Redescription Mining and Applications in Bioinformatics

Naren Ramakrishnan[1] and Mohammed J. Zaki[2]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA
[2]Department of Computer Science, RPI, Troy, NY 12180, USA

January 8, 2009

Abstract

Our ability to interrogate the cell and computationally assimilate its answers is improving at a dramatic pace. For instance, the study of even a focused aspect of cellular activity, such as gene action, now benefits from multiple high-throughput data acquisition technologies such as microarrays, genome-wide deletion screens, and RNAi assays. A critical need is the development of algorithms that can bridge, relate, and unify diverse categories of data descriptors. Redescription mining is such an approach. Given a set of biological objects (e.g., genes, proteins) and a collection of descriptors defined over this set, the goal of redescription mining is to use the given descriptors as a vocabulary and find subsets of data that afford multiple definitions. The premise of redescription mining is that subsets that afford multiple definitions are likely to exhibit concerted behavior and are, hence, interesting. We present algorithms for redescription mining based on formal concept analysis and applications of redescription mining to multiple biological datasets. We demonstrate how redescriptions identify conceptual clusters of data using mutually reinforcing features, without explicit training information.

1 Introduction

Our ability to interrogate the cell and computationally assimilate its answers is improving at a dramatic pace. The transformation of biology into a data-driven science is hence continuing unabated, as we become engulfed in ever-larger quantities of information about genes, proteins, pathways, and even entire processes. For instance, the study of even a focused aspect of cellular activity, such as gene action, now benefits from multiple high-throughput data acquisition technologies such as microarrays [4], genome-wide deletion screens [7], and RNAi assays [14, 15, 16]. Consequently, analysis and mining techniques, especially those that provide data reduction down to manageable quantities, have become a mainstay of computational biology and bioinformatics. From simple clustering of gene expression profiles [10], researchers have begun uncovering networks of concerted (regulatory) activity [20, 26], reconstructing the dynamics of cellular processes [9, 23], and even generating system-wide perspectives on complex diseases such as cancer [25].

The successes at being able to rapidly curate, analyze, and mine biological data obscure a serious problem, namely an overload of vocabularies now available for describing biological entities. For our purposes, a vocabulary is any way to carve up a domain of interest and posit distinctions and equivalences. While one biologist might study stress-responsive genes from the perspective of their transcriptional levels, another might assess downstream effects such as the proteins the genes encode, whereas still others might investigate the phenotypes of deletion mutants. All of these vocabularies offer alternative and mostly complementary (sometimes, contradictory) ways to organize information and each provides a different perspective into the problem being studied. To further knowledge discovery, biologists need tools to help uniformly reason

across vocabularies, integrate multiple forms of characterizing datasets, and situate knowledge gained from one study in terms of others.

The need to bridge diverse biological vocabularies is more than a problem of data reconciliation, it is paramount to providing high-level problem solving functions for the biologist. As a motivating context, consider a biologist desiring to identify a set of *C. elegans* genes to knock-down (via RNAi) in order to confer improved desiccation tolerance in the nematode. Assume the biologist would like to decompose this problem via two lines of reasoning. First, proceeding along a stress response argument, the investigator would like to identify genes that serve as master controls (transcription factors) whose knock-down will cause significant change in downstream gene expression of other genes, leading to a modulation in the desiccation response (positive or negative), culminating in a disruption or delay of any shutdown processes. Second, following a phenotypical argument, efforts would be directed at identifying key physiological indicators of tolerance and adaptation, restate these indicators in terms of pathways that must be activated (or inactivated), and identify genes central to these objectives. To support such lines of reasoning, and integrate their answers, the biologist needs to be able to relate diverse data domains using a uniform analytical methodology. Redescription mining is such an approach. Redescriptions empower the biologist to define his own vocabularies, relate descriptors across them uniformly, and relationally compose sequences of redescriptions to realize complex functions. We will show how redescriptions are not specific to any data acquisition technology, domain of interest, or problem solving scenario. Instead, they can be configured to support a range of analytical functions that straddle vocabularies.

2 Reasoning about sets using redescriptions

As the term indicates, to redescribe something is to describe anew or to express the same concept in a different vocabulary. The input to redescription mining is a set of objects and a collection of subsets defined over this set. It is easiest to first illustrate redescription mining using an everyday, non-biological, example; consider, therefore, the set of ten countries shown in Fig. 1 and its four subsets, each of which denotes a meaningful grouping of countries according to some intensional definition. For instance, the colors (G) green, (R) red, (B) blue, and (Y) yellow (from right, counterclockwise) refer to the sets 'permanent members of the UN security council,' 'countries with a history of communism,' 'countries with land area $> 3,000,000$ square miles,' and 'popular tourist destinations in the Americas (North and South).' We will refer to such sets as *descriptors*. A redescription is a shift-of-vocabulary and the goal of redescription mining is to identify subsets that can be defined in at least two ways using the given descriptors. An example redescription for this dataset is then: 'Countries with land area $> 3,000,000$ square miles outside of the Americas' are the same as 'Permanent members of the UN security council who have a history of communism.' This redescription defines the set {Russia, China}, once by a set intersection of political indicators ($R \cap G$), and again by a set difference involving geographical descriptors ($B - Y$). Notice that neither the set of objects to be redescribed nor the ways in which descriptor expressions should be constructed is input to the algorithm. The underlying premise of redescription analysis is that sets that can indeed be defined in (at least) two ways are likely to exhibit concerted behavior and are, hence, interesting.

What makes redescription mining pertinent to biology is that the domain of discourse, i.e., the descriptors, is defined by the biologist. For instance, descriptors of genes in a given organism can be organized into vocabularies such as cellular location (e.g., 'genes localized in the mitochondrion'), transcriptional activity (e.g., 'genes up-regulated two-fold or more in heat stress'), protein function (e.g., 'genes encoding proteins that form the Immunoglobin complex'), or biological pathway involvement (e.g., 'genes involved in glucose biosynthesis'). More vocabularies can be harnessed from computational studies (e.g., 'genes forming module #124 identified by the Segal et al. algorithm') or literature (e.g., 'genes hypothesized to be involved in desiccation response in the Potts et al. paper'). Redescription mining then constructs set-theoretic expres-

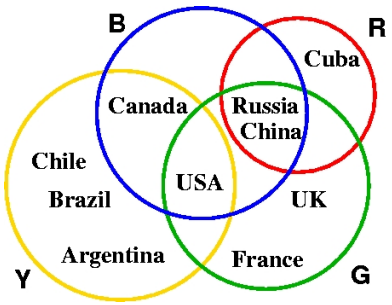

Figure 1: Example input to redescription mining. The expression $B - Y$ can be redescribed into $R \cap G$.

sions that induce the same set of genes in different vocabularies. See Fig. 4, to be described in detail later, for examples of redescriptions from studies on budding yeast *S. cerevisiae*. Redescription 1 in Fig. 4, for instance, restates 'the ORFs negatively expressed in the histone depletion experiment (6 hours)' as those 'negatively expressed two-fold or more in the heat shock (10 minutes) experiment.' Notice that this is an approximate, albeit strong, redescription which holds with Jaccard's coefficient (the ratio of the size of the overlap to the size of the union) 0.78. These ORFs comprise functions related to metabolism, catalytic activity, and their action is localized in the cytoplasm. The Pearson coefficients for these ORFs in the histone depletion experiments match very strongly, showcasing the use of redescription in identifying a concerted set of ORFs. Similarly, it is easy to conceptualize redescription scenarios where the descriptors are defined over proteins, processes, or other domains.

In fact, the importance of 'descriptors' to encode domain specific sets in biology and using them as a starting point for biological data mining has been recognized by other researchers. Segal et al. [25] focus on pre-defined sets of genes and this work defines descriptors based on the results of clustering, on expression in specific cell types, and membership in certain functional categories or pathways. The MSigDB (molecular signatures database) [29] supporting the Gene Set Enrichment Analysis (GSEA) algorithm is another resource that defines gene sets based on pathways, annotations, and similar information. There are many more such methods but essentially all of them are interested in casting interpretations over pre-defined, biologically meaningful, sets. Redescription mining views these databases as the primary resource for mining and reveals inter-dependencies within them.

3 Theory and Algorithms

Formally, the inputs to redescription mining are the universal set of objects (e.g., genes) $G = \{g_1, g_2, \ldots, g_n\}$, and a set $D = \{d_1, d_2, \ldots, d_m\}$ of proper subsets (the descriptors) of G. This information can be summarized in a $n \times m$ binary *dataset matrix* whose rows represent genes, columns represents the descriptors, and the (i, j) entry is 1 if object g_i is a member of descriptor d_j, and 0 otherwise. Typically, descriptors are organized into vocabularies, each of which provides a covering of G. An expression bias (more on this below) dictates allowable set-theoretic constructs involving the descriptors. This setting is similar to one studied by Pu and Mendelzon [21] but there the goal is to find one most concise description of a given set of objects using the vocabulary. The goal in redescription mining is to find equivalence relationships of the form $E \Leftrightarrow F$ that hold at or above a given Jaccard's coefficient θ (i.e., $\frac{|E \cap F|}{|E \cup F|} \geq \theta$), where E and F are expressions in the specified bias comprising the descriptors D. The key property of a redescription, like most data mining patterns, is that it must be falsifiable in *some* interpretation (dataset). Notice that this rules out tautologies, such as $d_i - (d_i - d_j) \Leftrightarrow d_i \cap d_j$, which are true in *all* datasets.

Redescription mining exhibits traits of many other data mining problems such as conceptual clustering, constructive induction, and boolean formula discovery. It is a form of conceptual clustering [11, 17] be-

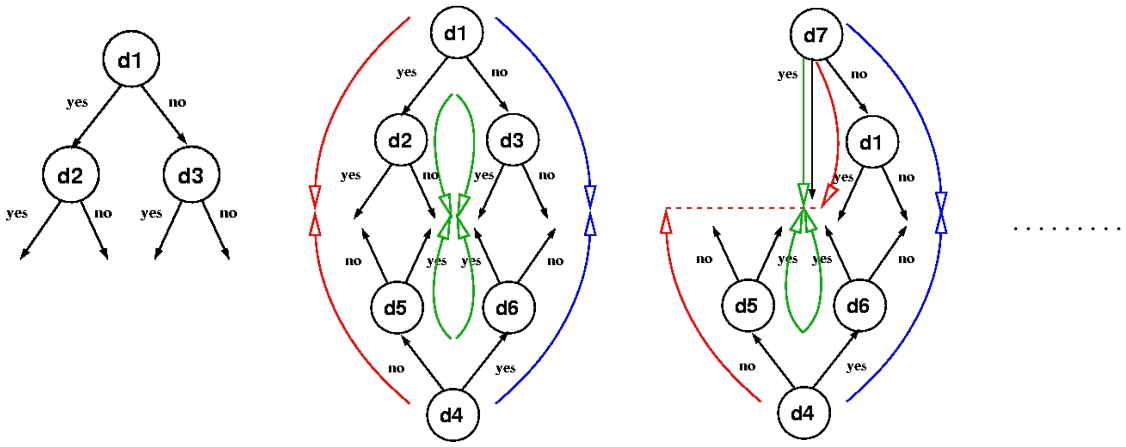

Figure 2: Mining redescriptions using the CARTwheels algorithm. The alternation begins with a tree (first frame) defining set-theoretic expressions to be matched. The bottom tree is then grown to match the top tree (second frame), which is then fixed, and the top tree is re-grown (third frame). Colored arrows indicate the matching paths. Redescriptions corresponding to matching paths at every stage are read off and subjected to evaluation by Jaccard's coefficient. For instance, in the second frame, the matching paths give rise to three redescriptions: $d_1 \cap d_2 \Leftrightarrow G - d_4 - d_5$ from paths on the left (red), $G - d_1 - d_3 \Leftrightarrow d_4 - d_6$ from paths on the right (blue), and $(d_1 - d_2) \cup (d_3 - d_1) \Leftrightarrow (d_4 \cap d_6) \cup (d_5 - d_4)$ from paths in the middle (green).

cause the mined clusters are required to have not just one meaningful description, but two. It is a form of constructive induction since the features important for learning must be automatically constructed from the given vocabulary of descriptors. Finally, since redescriptions are equivalence relationships, the problem of mining redescriptions can be viewed as (unsupervised) boolean formula discovery [6].

3.1 Structure Theory of Redescriptions

In [19], a structure theory of redescriptions is presented that yields both impossibility and strong possibility results. For instance, if the dataset matrix is a truth table, i.e., the number of genes n is 2^m, where m is the number of descriptors, then there can be no redescriptions. This is because the number of subsets of genes (2^n) coincides with the number of possible boolean formulas over m variables (2^{2^m}). Each boolean formula is then in an equivalence class by itself and induces a different subset of objects from other formulas. On the other hand, if the dataset is less than a truth table (i.e., missing one or more rows), then the 'islands' of boolean formulas begin to merge, with each missing row reducing the number of equivalence classes by a factor of two. In such a case, *all* boolean formulas have redescriptions! This dichotomy law is rather disconcerting but holds only under the assumption that the expression bias is general enough to induce all subsets of genes. If we algorithmically restrict the bias, e.g., to conjunctions, or length-limited constructs, then it is not obvious which subsets afford redescriptions, leading to a non-trivial data mining problem. As a result, all algorithms for mining redescriptions focus on a specific bias and mine expressions within that bias.

3.2 Algorithms for Mining Redescriptions

The CARTwheels algorithm [24] mines redescriptions between length-limited disjunctions of conjunctions. The CHARM-L algorithm [33] mines redescriptions between conjunctions (with no restrictions on length). A recently developed extension to CHARM-L (BLOSOM [34]) provides a way to systematically mine complex boolean expressions in various biases, e.g., conjunctions, disjunctions, CNF, or DNF forms. We

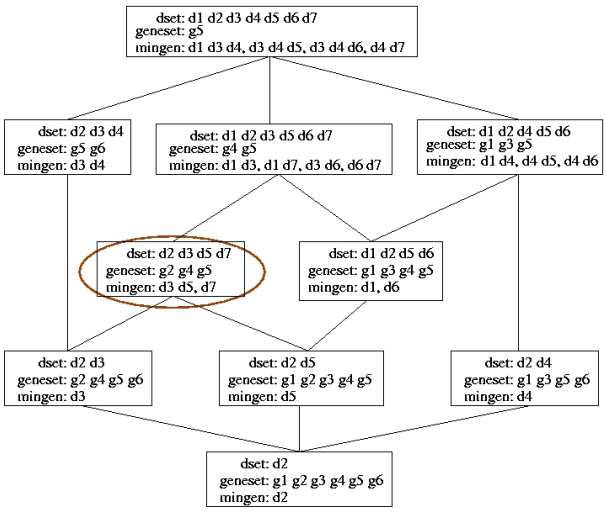

Figure 3: Mining redescriptions using the CHARM-L algorithm. Each node in the lattice denotes a closed set, comprising genes, descriptors (abbreviated as dsets), and their minimal generators. The only redescribable sets are the closed sets; redescriptions for these are obtained by relating their minimal generators. For instance, $d3 \cap d5 \Leftrightarrow d7$ is a redescription because both $d3 \cap d5$ and $d7$ are minimal generators of the closed set circled in the lattice.

highlight the main features of all these algorithms in this section.

CARTwheels: CARTwheels mines redescriptions by exploiting two important properties of binary decision trees [22]. First, if the nodes in such a tree correspond to boolean membership variables of the given descriptors, then we can interpret paths to represent set intersections, differences, or complements; unions of paths would correspond to disjunctions. Second, a partition of paths in the tree corresponds to a partition of objects. These two properties are employed in CARTwheels which grows two trees in opposite directions so that they are joined at the leaves. Essentially, one tree exposes a partition of objects via its choice of subsets and the other tree tries to grow to match this partition using a different choice of subsets. If partition correspondence is established, then paths that join can be read off as redescriptions. CARTwheels explores the space of possible tree matchings via an alternation process (see Fig. 2) whereby trees are repeatedly regrown to match the partitions exposed by the other tree. Notice the implicit restriction of bias to disjunctions of one to three clauses, each involving one to two descriptors (in negated or non-negated form). By suitably configuring this alternation, we can guarantee, with non-zero probability, that any redescription existing in the dataset would be found. Exploration policies must balance the potential of identifying unseen regions of descriptor space against redundancy from re-finding already mined redescriptions.

CHARM-L: CHARM-L, employing the conjunctions bias, adopts a different approach and exploits connections between boolean formulas and closed sets, a concept popular in the association mining community [1, 2]. A closed set is a set of genes together with a set of descriptors such that the conjunction of the given descriptors induces the given set of genes, and no subset of the given descriptors induces the same set of genes. In other words, the gene set and descriptor set are maximal w.r.t. each other and we cannot reduce either of these sets without losing any elements of the other. The closed sets form a lossless representation of the underlying dataset matrix in that the only redescribable sets are the closed sets. Additionally, the redescriptions of a closed set are precisely the non-maximal versions of the given set. CHARM-L further focuses on only the minimal variants, called *minimal generators*. The problem of redescription mining then

reduces to mining closed sets and relating their minimal generators to form redescriptions (see Fig. 3). However, datasets for redescription analysis, when studied in the association mining framework, are very dense. Since a gene participates in either a descriptor or its negation, the underlying dataset matrix is exactly 50% dense (or sparse). We hence cannot rely merely on support pruning as a way to curtail the complexity of data mining. CHARM-L's solution is a constraint-based approach so that the lattice of closed sets is selectively computed around genes (or descriptors) of interest.

BLOSOM: A generalization of CHARM-L is now being developed in the BLOSOM data mining framework [34]. BLOSOM is a framework for mining closed boolean expressions of all forms, and defines closure operators for specific families of expressions such as conjunctions, disjunctions, CNF, and DNF forms. It focuses on mining the minimal boolean expressions that characterize a set of objects (e.g., genes). The main data mining engine is based on a new approach to mine disjunctions (OR-clauses) instead of conjunctions. A number of effective pruning techniques are utilized to effectively search for all the possible frequent boolean expressions in normal form.

Besides their choice of biases, CARTwheels and CHARM-L/BLOSOM approach redescription mining from alternative viewpoints, namely exploratory search versus enumeration and pruning. In this chapter, we show the application of all these algorithms for biological studies.

4 Applications of redescription mining

4.1 Redescriptions for *S. cerevisiae*

We have applied redescriptions to studying descriptors defined over the yeast genome [24, 27], resulting in considerable biological insight. The biologist first na rrows down on a reference set of a few hundred genesand then defines descriptors over this reference set.In [27] we defined the reference set to be the set of 210 'high-expressors' – genes exhibiting more than five-fold change in *some* t ime point (not necessarily all or even a majority) across the yeast desiccation and rehydration time course. The descriptors are drawn from a variety of sources. One vocabulary denotes expression levels in specific microarray measurements taken from Gasch et al. [12] and Wyrick et al. [32]. For instance, 'genes negatively expressed two-fold or below in the 15 minute time point of the 1M sorbitol experiment' is a descriptor in this vocabulary. A second type of vocabulary asserts membership of genes in targeted taxonomic categories of the Gene Ontology (biological processes (GO BIO), cellular components (GO CEL) or molecular functions (GO MOL)). A third class of descriptors is based on clustering time course datasets using a k-means clustering algorithm [28] and using the clusters as descriptors.

The redescriptions presented here, although approximate, have been whetted at a p-value significance level of 0.0001. Essentially, we characterize the distribution of set size overlaps for targeted descriptor cardinalities and reason about the possibility of obtaining the specified Jaccard's coefficient purely by chance.

Fig. 4 depicts some statistically significant redescriptions obtained by CARTwheels, illustrating the diversity of set constructions possible. Of these, redescription R1 has been discussed earlier. R2 relates a k-means cluster to a set difference of two *related* GO cellular component categories. While the 8 ORFs in R2 appear to be part of different response pathways, 5 of these 8 ORFs are similarly regulated according to the work of Segal et al. [26]; these genes relate to the cellular hyperorganization and membrane dynamics in the regulation network.

R3 is a triangle of redescription relationships involving three different experimental comparisons, with 10 ORFs being implicated in all three expressions. From a biological standpoint, this is a very interesting result – the common genes indicate concerted participation across stress conditions; whereas the genes participating in, say, two of the descriptors, but not the third, suggest a careful diversification of functionality.

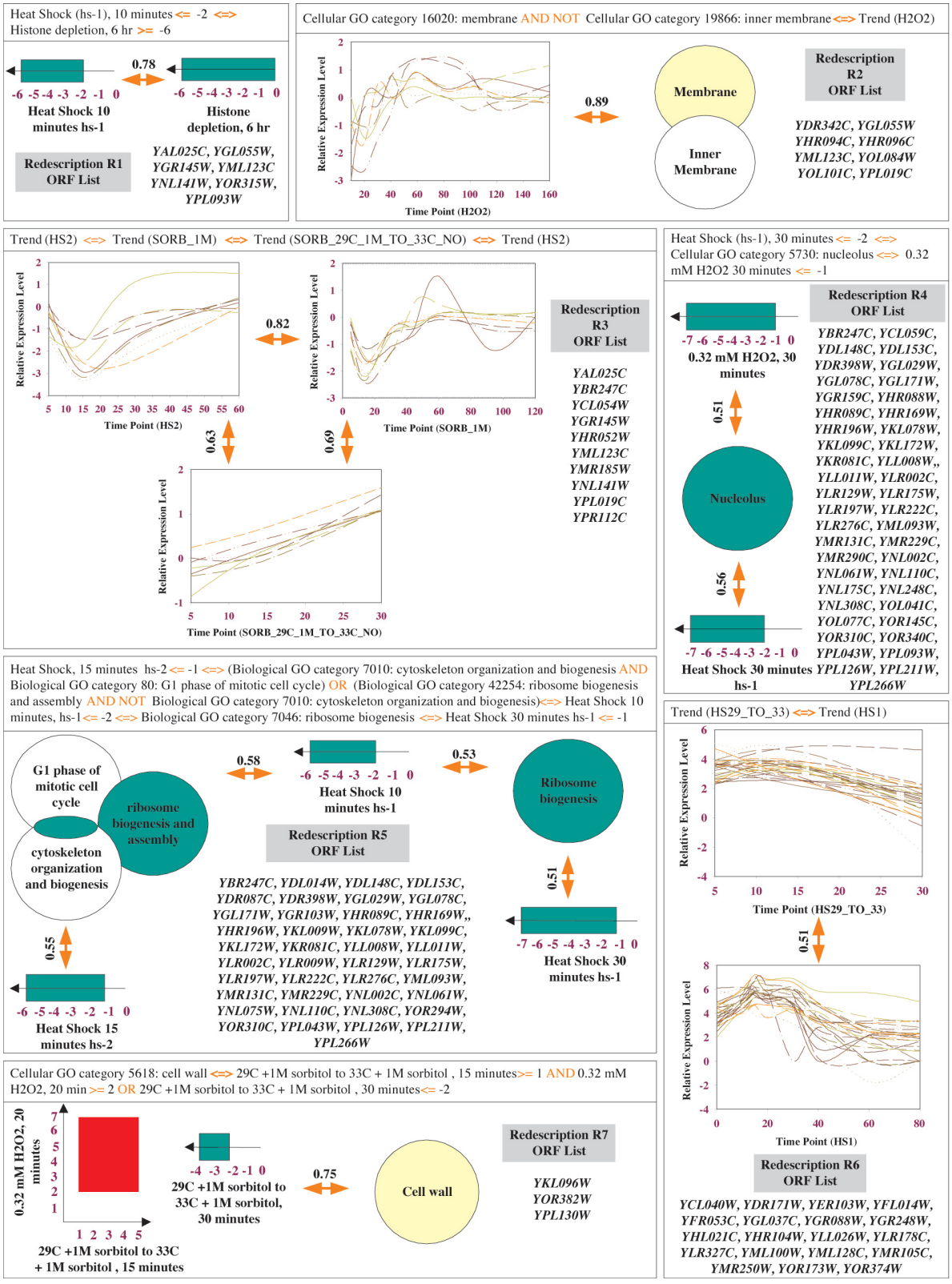

Figure 4: Seven (approximate) redescriptions mined from gene expression studies on *Saccharomyces cerevisiae*. Each box gives a readable statement of the redescription, presents it in graphical form, and identifies the ORFs conforming to the redescription. The Jaccard's coefficient is displayed over the redescription arrow. Notice that some redescriptions (e.g., R7) involve few ORFs, whereas others such as R5 involve larger numbers.

6 of the 10 ORFs are related to cell growth and maintenance. 5 of the 10 ORFs have binding motifs related to the DNA binding protein REB1. The importance of phosphate and ribosomes appears to be salient in this redescription. It is important to note that the circularity of R3 is not directly mined by CARTwheels, but inferred post-hoc from a linear chain.

The theme in R4 is ribosome assembly/biogenesis and RNA processing. R4 is a linear chain comprising two redescriptions, and uses a GO descriptor as an intermediary between two expression-based descriptors. It is also interesting that this redescription involves a set of 45 ORFs!

R5 is an even longer chain involving 41 ORFs that are common to all descriptors. Notice the rather complicated set construct involving a disjunction of a conjunction and a difference, involving three different GO biological categories. Incidentally, this is the most complicated set expression representable in a 2-level tree. Although R3, R4, and R5 are linear chains, CARTwheels is not a story telling algorithm since it cannot find such relationships between user-supplied descriptors. The examples shown here are snapshots from the continuous alternation of CARTwheels.

R6 is a relationship between two k-means clusters, between heat shock stresses. The ORFs participating in R6 demonstrate a clear focus on sugar or sugar phosphate metabolism.

R7 is a redescription relating a disjunction of descriptors to a GO cellular component category. It is also an interesting example of constructive induction, since a rectangular region is mined in a 2D space involving two different experimental comparisons.

The capabilities of the CHARM-L are best illustrated through an interactive scenario where, given constraints, the algorithm reasons about the conditions under which two given descriptors would be the equivalent. In one scenario described in [33], a biologist is exploring descriptors around his favorite gene – YOR374W, an ORF in *S. cerevisiae* that encodes an NAD-dependent aldehyde dehydrogenase (an enzyme—EC 1.2.1.3—that catalyzes the conversion of an aldehyde and NAD+ to a carboxylic acid and NADH), which has been determined to be very highly expressed in time point 20 minutes of the Gasch heat shock condition (more than five-fold). YOR374W (Ald4p) is important from the perspective of metabolism as it provide a means to generate reduced cofactor (NADH) for fueling electron transport and oxidative phosphorylation (ATP synthesis). The biologist is particularly interested in relating two descriptors that Y0R374W participates in. One of them is descriptor d184 that denotes all ORFs that are expressed more than five fold in the above time point; it contains 19 genes. Looking at the nearby time point (15 minutes) the biologist notices that the corresponding descriptor (d183) contains 21 genes, with 18 in common with d184. The Jaccard's coefficient between these descriptors is already high (0.857) but the biologist is curious to determine if there could be an exact redescription by using the GO vocabularies. CHARM-L uncovers the following redescription:

$$d183 - d388 - d460 - d515 \Leftrightarrow d184 - d309$$

In other words, to make d183 equivalent to d184, we need to subtract descriptors d388, d460, and d515 on the left (to remove 3 genes) and subtract descriptor d309 on the right (to remove 1 gene), bringing the commonality to 18, as desired. Here, d388 refers to the GO molecular function category: mannose transporter, d460 refers to the GO cellular component category: external protective structure, and d515 refers to the GO biological process category: fructose metabolism. d309, on the right side, incidentally happens to refer to genes whose molecular function, according to GO, is unknown. The implied message, from the above redescription, is that as we go from time point 15 minutes to time point 20 minutes, genes belonging to the above three categories drop out of the highly expressed (≥ 5 fold) category.

4.2 Expanding redescriptions to uncover pathways

Fig. 5 describes how we can uncover an entire pathway by integrating redescription analysis with domain theories. The source redescription in Fig. 5 (panel A) states that the genes that are onefold or more down-

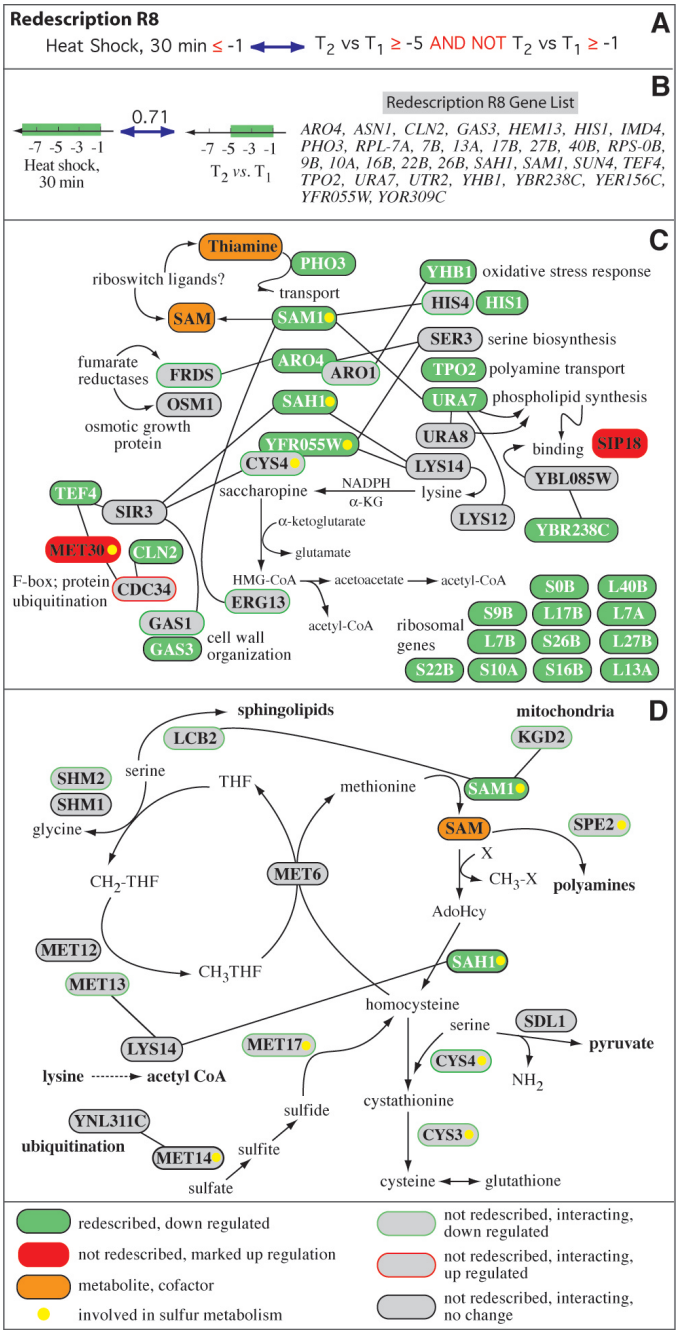

Figure 5: The use of redescriptions to uncover pathways. (A) Statement of redescription relating heat shock to desiccation experiment. (B) Graphical depiction of redescription and the genes identified. (C) Desiccation and heat shock lead to down-regulation of sets of genes with a central function in sulfur metabolism. (D) Functions of genes involved in methyl group transfer and sulfur metabolism.

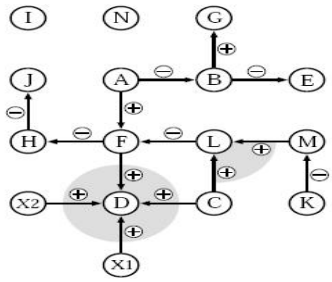

Figure 6: Gene Network

regulated during heat shock (30 min) can be restated as those that are between one-and five fold down-regulated in the desiccation experiment. A conspicuous feature of this redescription (see Fig. 5, panel B) is the presence of three genes involved in sulfur metabolism: SAM1 (S-adenosylmethionine synthetase gene), encoding the protein that synthesizes the potential riboswitch ligand S-adenosylmethionine [8], SAH1 (S-adenosyl-L-homocysteine hydrolase gene), and YFR055W (cystathionine β-lyase gene). Riboswitch ligands such as SAM appear to serve as ancient master control molecules whose concentrations are monitored to ensure homeostasis of a much wider set of metabolic pathways [30, 31], and indeed SAM has recently been implicated in G_1 cell cycle regulation [18].

We can procedurally uncover a pathway from the above redescription as follows. We discard genes encoding for ribosomal activity due to their consistent expression across the range of time courses in the dataset and study the remaining genes and their interacting partners. Sam1p was reported to interact with 13 other proteins [13], and the gene for one of these, URA7, is also present in the redescription. Using each of the genes (and their respective protein interactions culled from [5]), we systematically expand the given genes to form a network of interactions. We then use the primary microarray data to infer possible additional relationships (based on expression correlation). For example, MET30, encoding a cell cycle F-box protein and also involved in sulfur metabolism and protein ubiquination, can directly or indirectly be associated with TEF4 (translation elongation factor EF-1γ gene) and CLN2 (cyclin-dependent protein kinase regulator gene), both of which are present in the redescription. Note that MET30 itself was not present in the redescription. Finally, we improve the network further by incorporating genes from a pathway that show interactions with other genes in the redescription but not with one another (for example, the clustering of HIS4 and HIS1, ARO1 and ARO4, LYS14 and LYS12, and GAS1 and GAS3). The end result is depicted in Fig. 5 (panel C). With the exception of MET30, SIP18, and CDC34, the transcription of each gene in the proposed network was either down-regulated or unchanged, suggesting a central role for sulfur metabolism in desiccation response. In view of the potential role of its protein in phospholipid binding SIP18 is shown close to other genes associated with lipid synthesis and binding (URA7, YBL085W), although it doesn't itself participate in the redescription. Studying this system further, Fig. 5 (panel D) illustrates the functions of genes involved in methyl group transfer and sulfur metabolism. Additional information can be obtained from comparison of Panels C and D; for example, note TPO2 (polyamine transport, Panel C) and role of SAM1 in polyamine synthesis (Panel D); in addition the connectivity of SER3 and ARO4 (serine biosynthesis, Panel C) and the role of serine in lipid biosynthesis (Panel D).

4.3 Modeling gene regulatory networks

A final application of redescription mining is to finding complex gene regulatory networks, which can be represented in a simplified form, as boolean networks [3]. For this purpose, we demonstrate the application of BLOSOM for redescription mining. Consider the network involving 16 genes, taken from [3], shown in Fig. 6.

Here $\oplus$ and $\ominus$ denote gene *activation* and *deactivation*, respv. For example, genes B, E, H, J, and M are expressed if their parents are not expressed. On the other hand G, L, and D express if all of their parents

express. For example, D depends on C, F, $X1$ and $X2$. Note that F expresses if A does, but not L. Finally A, C, I, K, N, $X1$ and $X2$ do not depend on anyone, and can thus be considered as *input* variables for the boolean network. We generated the truth table corresponding to the 7 input genes but BLOSOM was provided the values for all genes, without explicit instruction about which are inputs and which are outputs. This yields a dataset with 128 rows and 16 items (genes). We then ran BLOSOM to discover the boolean expression corresponding to this gene network; we used a minimum support of 100%, since we wish to find expressions that are true for the entire set of assignments. BLOSOM output 65 expressions in 0.36s, which hold true for the entire dataset. After simplification these can be reduced to the equivalent expression, as shown in Table 1.

$\big(\overline{D} \mid (A\ \overline{B}\ C\ E\ F\ \overline{G}\ \overline{H}\ J\ K\ \overline{L}\ \overline{M}\ X1\ X2)\big)$ AND
$\big(\overline{L} \mid (C\ \overline{F}\ H\ \overline{J}\ \overline{K}\ M)\big)$ AND
$\big((\overline{A}\ B\ \overline{E}\ G) \mid \overline{C} \mid D \mid L \mid \overline{X1} \mid \overline{X2}\big)$ AND
$\big((\overline{A}\ B\ \overline{E}\ G) \mid (C\ L) \mid (F\ \overline{H}\ J)\big)$ AND
$\big((\overline{F}\ H\ \overline{J}) \mid (A\ \overline{B}\ \overline{C}\ E\ \overline{G}) \mid (A\ \overline{B}\ E\ \overline{G}\ K\ \overline{M})\big)$

Table 1: Boolean Network Expression

We verified that indeed this expression is true for all the rows in the dataset! It also allows us to reconstruct the boolean gene network shown in Fig. 6. For example, the first component of the expression in the first row $\overline{D} \mid (A\overline{B}CEF\overline{GH}JK\overline{LM}X1X2)$ can be converted into the implication $D \Rightarrow (A\overline{B}CEF\overline{GH}JK\overline{LM}X1X2)$, which means that D depends on the variables on the right hand side (RHS). If, at this point, we supply any partial knowledge about the input variables or of the maximum fan-out of the network, we could project the RHS only on those variables to obtain $(ACKX1X2)$, which happens to be precisely the relationship given in Fig. 6. The second row tells us that L depends on the activation of C and inactivation of K, i.e., $\overline{K}$, if we restrict ourselves to the input variables. Note that C and $\overline{K}$ give the values for the remaining varibles. Note that other dependencies are also included in the mined expression. For example, we find that B and A always have opposite values, and so do B and E, and K and M. G and B always have the same values, and so on. Thus this example shows the power of BLOSOM in mining gene regulatory networks.

4.4 Cross-taxonomic and cross-genomic comparisons using redescriptions

Assume that we are provided with two families of functional annotations or ontologies, E and F, over the same space of objects (e.g., genes). The objective is to conduct an all-pairs redescription study relating categories or concepts between E and F. From the results of such a study, if $e_1 \in E$ is redescribed to $f_2 \in F$ with a very high Jaccard's coefficient, we could help impute annotations and properties typically associated with e_1 to f_2 (and vice versa). The results of such a study can then be used for funtional enrichment of unclassified genes, to analyze the structural consistencies (and inconsistencies) of different ontologies, and in general as an educational tool to communicate similarities and differences across taxonomies. Finally, when the ontologies apply to multiple organisms, we can study the extent to which redescriptions transfer across organisms and whether some organisms have more developed ontologies than others.

We conducted a cross-taxonomic GO comparison study using the GO biological process (GO BIO), GO cellular component (GO CEL), and GO molecular function (GO MOL) assignments available for the *Arabidopsis thaliana*(arabidopsis), *Drosophila melanogaster*(fly), *Homo sapiens*(human), *Mus musculus*(mouse), *Caenorhabditis elegans*(worm) and *Saccharomyces cerevisiae*(yeast) genomes from the GO database website. The GO hierarchy information used for propogation of GO categories up the GO tree was also taken from the same website. For each organism, only those genes were considered that have atleast one GO category, other than the categories for unknown GO BIO, GO MOL and, GO CEL, defined. The summary of the data used is provided in Table 2.

Table 2: Summary of input GO categories for the 6 species considered

	Arabidopsis	Fly	Human	Mouse	Worm	Yeast
Universal set size	13572	8911	23424	25142	11606	5731
Genes with BIO defined	6340	7424	18068	18193	9299	4711
Genes with CEL defined	3114	4131	16002	17362	5179	5713
Genes with MOL defined	12817	7606	21135	21887	8975	5714
BIO categories involved	1043	2493	2837	2774	1361	1691
CEL categories involved	205	530	543	496	261	424
MOL categories involved	1212	2013	2516	2230	1062	1470

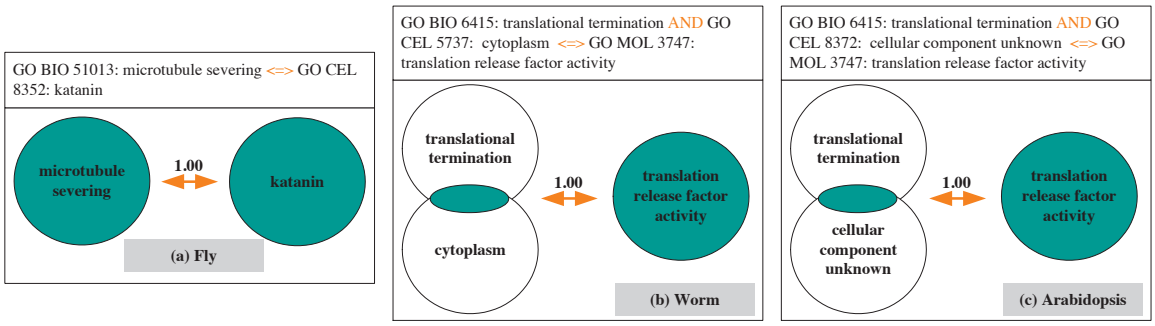

Figure 7: Examples of cross-taxonomic redescription: (a) A redescription between two functional categories for the Fly genome, (b) A redescription involving an intersection between two categories for the Worm genome, (c) A redescription involving the GO CEL category 'celullar component unknown' for the Arabidopsis genome. This redescription relates to the one in (b).

The experiment using the GO assignments for genes in each genome was performed as follows. The universal set of genes was defined as above. Within this universal set, the three families of GO categories were used individually as also in pairs to form input sets of descriptors. In all the runs, the descriptor family for which redescriptions were sought was used to build one-level trees. Two level trees were used for each pair of descriptor families used to construct derived descriptors. Thus, if a redescription was sought between the GO BIO categories on one side and combinations of GO CEL and GO MOL categories on the other, the study was done using a one-level tree for the GO BIO categories and upto a 2-level tree for GO MOL and GO CEL categories. We also restricted all derived descriptors to involve just intersections and differences between descriptors. The support threshold was set at 3 to retain only the most significant redescriptions. The Jaccards coefficient threshold was set at 0.5.

Fig. 7 shows a few examples of redescriptions mined using CARTwheels. Fig. 7(a) shows a simple redescription between a GO BIO (51013) and GO CEL (8352) category. This redescription holds for the Fly genome with Jaccard's coefficient of 1 and involves 4 genes. This type of a redescription can be easily used to relate functional enrichments across different taxonomies as described earlier. Fig. 7(b) shows a redescription involving a derived descriptor formed by the intersection of a GO BIO and a GO CEL category which relates to a GO MOL category. This redescription holds for the Worm genome with Jaccard's coefficient 1 and involves 3 genes. Fig. 7(c) shows a redescription for the Arabidopsis genome where the GO BIO and GO MOL involved are the same as in Fig. 7(b). The difference here is that GO CEL is assigned the GO category 'Cellular category unknown.' This redescription also holds with a Jaccards coefficient of 1 and involve 11 genes. The pair of redescriptions found could potentially be used to better characterize the GO CEL categorization for the genes involved for the Arabidopsis genome.

Table 3 summarizes the number of GO categories for which redescriptions are available and the number of redescriptions mined for each of the six species. In all cases, the use of derived descriptors (intersection and difference based) results in a significant increase in the number of categories involved in at least one

Table 3: Summary of redescriptions obtained for the 6 species. Numbers in bracket indicate number of redescriptions with no derived descriptors

	Arabidopsis	Fly	Human	Mouse	Worm	Yeast
BIO categories involved	259 (244)	169 (138)	389 (915)	375 (314)	257 (239)	260 (224)
CEL categories involved	50 (40)	146 (92)	102 (71)	94 (70)	70 (58)	149 (103)
MOL categories involved	176 (140)	230 (149)	369 (237)	358 (241)	271 (217)	205 (162)
BIO redescriptions	6852 (469)	15483 (324)	43828 (589)	41969 (622)	37174 (713)	23473 (513)
CEL redescriptions	4971 (139)	12567 (207)	22046 (163)	15531 (147)	11280 (178)	43293 (329)
MOL redescriptions	9352 (408)	39788 (363)	68765 (582)	66920 (581)	52445 (711)	20163 (388)

Table 4: Pairwise overlap between redescriptions obtained for the 6 species. The numbers in bracket indicate th e number of distinct descriptors involved.

	Arabidopsis	Fly	Human	Mouse	Worm	Yeast
Arabidopsis	-	2744 (48)	4550 (129)	4383 (120)	4133 (124)	4824 (59)
Fly	2744 (48)	-	13306 (123)	11198 (94)	8237 (96)	5561 (117)
Human	4550 (129)	13306 (123)	-	59674 (475)	29912 (291)	9871 (116)
Mouse	4383 (120)	11198 (94)	59674 (475)	-	26884 (282)	5278 (91)
Worm	4133 (124)	8237 (96)	29912 (291)	26884 (282)	-	4555 (86)
Yeast	4824 (59)	5561 (117)	9871 (116)	5278 (91)	4555 (86)	-

redescription. Also, as is to be expected, a much higher number of redescriptions are found for genomes that have more categories involved with the genes (giving more descriptors to form derived descriptors with). Comparing Table 2 and Table 3, we can conclude that a large proportion of the functional categories have redescriptions associated with them for all genomes.

Redescriptions found in the cross-taxonomic study described above can be used to validate and check the consistency of GO category assignments across different genomes. For this analysis, we conducted a pairwise comparison of redescriptions found for two different species and checked for overlap (same descriptors involved). Importantly, we did not require that the support or Jaccards coefficient be the same for the same redescription across a pair of species. The overlap observed is summarized in Table 4.

The species with large number of redescriptions (mouse and human) have high overlaps between them as also with other species. The fly and arabidopsis redescriptions show the minimum overlap. This is a result of the low number of descriptors available and redescriptions found for these species. As would be expected, arabidopsis and yeast which differ from the other 4 species most drastically show low amount of overlap.

Fig. 8 shows 3 examples of redescriptions found to be common for various species for a Jaccards thresh-

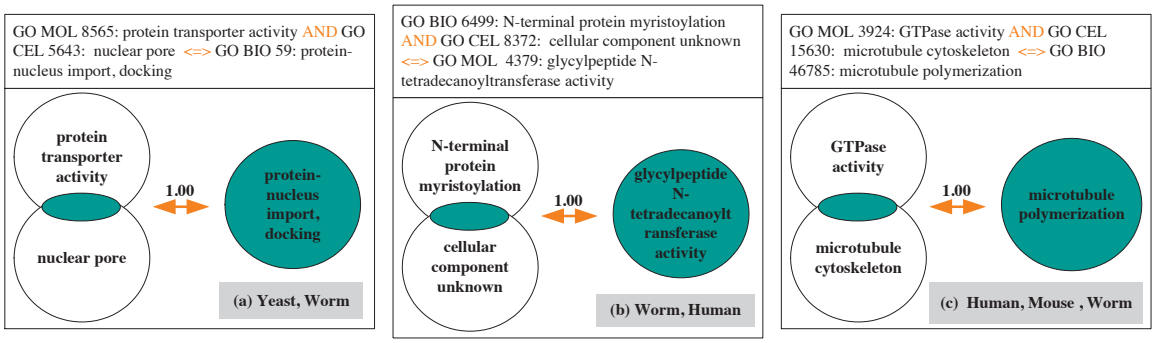

Figure 8: Examples of redescriptions that hold with Jaccard's coefficient for more that 1 species: (a) A redescription common to yeast and worm genomes, (b) A redescription common to the worm and human genomes, (c) A redescription common to the human, mouse and worm genomes.

Table 5: GO categories for which redescriptions were found common to all 6 species.

GO category	Type	Description
GO:0003735	MOL	structural constituent of ribosome
GO:0004672	MOL	protein kinase activity
GO:0004674	MOL	protein serine/threonine kinase activity
GO:0004812	MOL	tRNA ligase activity
GO:0005840	CEL	ribosome
GO:0006413	BIO	translational initiation
GO:0006418	BIO	tRNA aminoacylation for protein translation
GO:0006468	BIO	protein amino acid phosphorylation
GO:0008452	MOL	RNA ligase activity
GO:0016310	BIO	phosphorylation
GO:0016875	MOL	ligase activity, forming carbon-oxygen bonds
GO:0016876	MOL	ligase activity, forming aminoacyl-tRNA and related compounds
GO:0016886	MOL	ligase activity, forming phosphoric ester bonds
GO:0043038	BIO	amino acid activation
GO:0043039	BIO	tRNA aminoacylation

old of 1. Fig. 8 (a) shows a redescription commomn to the yeast and worm genome. This redescription involves an intersection between a GO MOL and a GO CEL category related to a GO BIO category. It involves 12 genes in the yeast genome and 4 genes in the worm genome. Fig. 8 (b) shows a redescription common to the worm and human genome. This redescription again involves an intersection with the GO CEL category "cellular category unknown" that is conserved across the two species. It involves 4 genes in the human genome and 3 genes in the worm genome. Fig. 8 (c) shows a redescription common to the human, mouse and worm genome. It involves the intersection between a GO MOL and GO CEL category related to a GO BIO category. This redescription involves 45 genes for human, 30 genes for mouse and 16 genes for the worm genome. Out of the redescription counts shown in Table 4, 920 redescriptions involving 15 categories were found to be constant across all 6 species. These 15 categories are listed in Table 5. All these categories lie quite high in the GO hierarchy and involve a lot of genes. Thus, there is no example of a redescription involving a very specific and precise functional category that could be found to be conserved across the 6 species.

5 Discussion

We hope to have shown here that redescription mining provides a domain-neutral way to cast complex data mining scenarios in terms of simpler primitives. This work makes possible to formulate and solve entirely new classes of research problems that are vital to biological knowledge discovery. The key to success in our approach is the use of domain-scientist-defined object sets (i.e., descriptors) as the starting point of analysis, ensuring relevance of mined results. As scientists are empowered to create their own vocabularies and descriptors and reason with them, there will be greater understanding of scientific datasets. Redescription mining promises to be an important tool in this endeavor.

Acknowledgements

This work was supported in part by NSF grants CNS-0615181, ITR-0428344, EMT-0829835, and CNS-0103708, and NIH Grant 1R01EB0080161-01A1. A significant amount of this work is the result of collaboration with many colleagues, including Deept Kumar, Richard F. Helm, Malcolm Potts, and Lizhuang Zhao. Deept Kumar helped gather the results presented in Section 4.4.

References

[1] R. Agrawal, T. Imielinski, and A.N. Swami. Mining Association Rules between Sets of Items in Large Databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*, pages 207–216, May 1993.

[2] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB'94)*, pages 487–499, Sep 1994.

[3] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of Gene Regulatory Networks by Strategic Gene Disruptions and Gene Overexpressions. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 695–702, 1998.

[4] C.A. Ball, I.A. Awad, J. Demeter, J. Gollub, J.M. Hebert, T. Hernandez-Boussard, H. Jin, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, P.O. Brown, and G. Sherlock. The Stanford Microarray Database Accomodates Additional Microarray Platforms and Data Formats. *Nucleic Acids Research*, Vol. 1(33):pages D580–D582, Jan 2005.

[5] B.-J. Breitkreutz, C. Stark, and M. Tyers. The GRID: The General Repository for Interaction Datasets. *Genome Biology*, Vol. 3(12), Nov 2002.

[6] N.H. Bshouty. Exact Learning Boolean Functions via the Monotone Theory. *Information and Computation*, Vol. 123(1):pages 146–153, 1995.

[7] A.E. Carpenter and D.M. Sabatini. Systematic Genome-Wide Screens of Gene Function. *Nature Reviews Genetics*, Vol. 5(1):pages 11–22, Jan 2004.

[8] S.Y. Chan and D.R. Appling. Regulation of S-Adenosylmethionine Levels in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, Vol. 278(44):pages 43051–43059, Oct 2003.

[9] U. de Lichtenberg, L.J. Jensen, S. Brunak, and P. Bork. Dynamic Complex Formation During the Yeast Cell Cycle. *Science*, Vol. 307(5710):pages 724–727, Feb 2005.

[10] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. *PNAS*, Vol. 95(25):pages 14863–14868, Dec 1998.

[11] D.H. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning*, Vol. 2(2):pages 139–172, 1987.

[12] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, and P.O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, Vol. 11:pages 4241–4257, 2000.

[13] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, T. Raida, M. annd Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and Superti-Furga G. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature*, Vol. 415(6868):pages 141–147, Jan 2002.

[14] K.C. Gunsalus and F. Piano. RNAi as a Tool to Study Cell Biology: Building the Genome-Phenome Bridge. *Current Opinion in Cell Biology*, Vol. 17(1):pages 3–8, Jan 2005.

[15] M.A. Matzke and J.A. Birchler. RNAi-Mediated Pathways in the Nucleus. *Nature Reviews Genetics*, Vol. 6(1):pages 24–35, Jan 2005.

[16] M.A. Matzke and A.J.M. Matzke. Planting the Seeds of a New Paradigm. *PLoS Biology*, Vol. 2(5):pages 0582–0586, May 2004.

[17] R.S. Michalski. Knowledge Acquisition through Conceptual Clustering: A Theoretical Framework and Algorithm for Partitioning Data into Conjunctive Concepts. *International Journal of Policy Analysis and Information Systems*, Vol. 4:pages 219–243, 1980.

[18] M. Mizunuma, K. Miyamura, D. Hirata, H. Yokoyama, and T. Miyakawa. Involvement of S-adenosylmethionine in G1 Cell Cycle Regulation in *Saccharomyces cerevisiae*. *PNAS*, Vol. 101(16):pages 6086–6091, Apr 2004.

[19] L. Parida and N. Ramakrishnan. Redescription Mining: Structure Theory and Algorithms. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, pages 837–844, July 2005.

[20] Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying Regulatory Networks by Combinatorial Analysis of Promoter Elements. *Nature Genetics*, Vol. 29(2):pages 153–159, Sep 2001.

[21] K.Q. Pu and A.O. Mendelzon. Concise Descriptions of Subsets of Structured Sets. *ACM Transactions on Database Systems*, Vol. 30(1):pages 211–248, Mar 2005.

[22] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, Vol. 1(1):pages 81–106, 1986.

[23] N. Ramakrishnan, M. Antoniotti, and B. Mishra. Reconstructing Formal Temporal Models of Cellular Events using the GO Process Ontology. In *Proceedings of the Eighth Annual Bio-Ontologies Meeting (ISMB'05 Satellite Workshop)*, June 2005.

[24] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R.F. Helm. Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 266–275, Aug 2004.

[25] E. Segal, N. Friedman, D. Koller, and A. Regev. A Module Map showing Conditional Activity of Expression Modules in Cancer. *Nature Genetics*, Vol. 36(10):pages 1090–1098, Oct 2004.

[26] E Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module Networks: Identifying Regulatory Modules and their Condition-Specific Regulators from Gene Expression Data. *Nature Genetics*, Vol. 34(2):pages 166–176, June 2003.

[27] J. Singh, D. Kumar, N. Ramakrishnan, V. Singhal, J. Jervis, A. Desantis, J. Garst, S. Slaughter, M. Potts, and R.F. Helm. Transcriptional Response of *Saccharomyces cerevisiae* to Desiccation and Rehydration. *Applied and Environmental Microbiology*, Vol. 71(12):pages 8752–8763, Dec 2005.

[28] A. Sturn, J. Quackenbush, and Z. Trajanoski. Genesis: Cluster Analysis of Microarray Data. *Bioinformatics*, Vol. 18(1):pages 207–208, 2002.

[29] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *PNAS*, Vol. 102(43):pages 15545–15550, Oct 2005.

[30] N. Sudarsan, J.E. Barrick, and R.R. Breaker. Metabolite-binding RNA Domains are present in the Genes of Eukaryotes. *RNA*, Vol. 9:pages 644–647, 2003.

[31] W.C. Winkler, A. Nahvi, N. Sudarsan, J.E. Barrick, and R.R. Breaker. An mRNA Structure that Controls Gene Expression by Binding S-adenosylmethionine. *Nature Structural Biology*, Vol. 10:pages 701–707, 2003.

[32] J.J. Wyrick, F.C. Holstege, E.G. Jennings, H.C. Causton, D. Shore, M. Grunstein, E.S. Lander, and R.A. Young. Chromosomal Landscape of Nucleosome-Dependent Gene Expression and Silencing in Yeast. *Nature*, Vol. 402:pages 418–421, 1999.

[33] M. Zaki and N. Ramakrishnan. Reasoning about Sets using Redescription Mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'05)*, pages 364–373, Aug 2005.

[34] L. Zhao, M. Zaki, and N. Ramakrishnan. BLOSOM: A Framework for Mining Arbitrary Boolean Expressions over Attribute Sets. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2006)*, pages 827–832, Aug 2006.