CONDITIONAL PATHWAY INTEGRATION

Alexandra Skolozub[†], Ofer Sarig[§] and Ron Y. Pinter[†§ *]

[†]*Department of Computer Science and* [§]*the Rappaport School of Medicine*
Technion – Israel Institute of Technology, Haifa 32000, Israel
Email: pinter@cs.technion.ac.il

Many biological pathways that describe complex cellular processes are available in public and commercial databases as well as in the literature. However, each item focuses on a particular cellular function. Moreover, pathways differ in the way they are described in different sources, emphasizing complementary aspects of the biological system under study. Considering related pathways in a unified framework is essential for understanding their behavior and for elucidating and refining open issues involving such systems. To address these challenges we have developed a conditional pathway algebra, in which pathways are enriched with both new node types as well as additional edge types providing significantly more expressive power for the description of existing biological phenomena. During conditional pathway integration, some interactions are made dependent upon a specific predicate (the presence/absence of a protein, extracellular factors, etc.). Moreover, such integration enables distinguishing between different data sources and points out problematic interactions in the given pathways. We provide a formal definition of the algebra and prove some properties of its operations, such as closure, commutativity, and the lack of associativity. Some of these operations are essential when applied to several pathways to form an entire (sub)system. Our algebra is embodied in the Pathway Integration Environment (PIE) as a plugin for Cytoscape. To demonstrate the utility and effectiveness of our method, we have applied it to three well characterized yeast signaling pathways: (i) Pheromone response, (ii) Filamentous growth, and (iii) High osmolarity glycerol pathways. Most of our computational observations are confirmed in the literature.
Availability: upon request.

1. INTRODUCTION

Many biological networks and pathways that describe complex cellular processes are being reconstructed and are ready for analysis. This information pertains to a variety of functional aspects and different levels of abstraction, *e.g.* metabolism, signal transduction, and regulation, each providing a partial perspective on the biological issues at hand. Currently, the descriptions of individual pathways is stored in several – both public as well as commercial – databases (such as KEGG [1-3], MetaCyc [4, 5], iPath [6], Reactome [7], and STKE [44]; see [8] for a review) using a variety of representations and presentation mechanisms; a wealth of information can also be found in the literature (in *e.g.* PubMed [45], where typically each paper describes the properties of a specific pathway or subsystem. Consequently, it is nearly impossible for a life science researcher to glean a comprehensive and integrative view of several pathways under one framework, which is a clear need for the systematic study of biological processes beyond encapsulated, specific functions. For example, when trying to elucidate drug response, it is critical to capture the impact of a substance (be it constitutive or administered) in its entirety.

Examination of several pathways at the same time can enhance our understanding of the biological system under study. The pertinent pathways can be different, or be separate instances of the same pathway coming from various data sources, each emphasizing distinct aspects of the biological process. By combining these resources, we can recognize *e.g.* regulating factors in pathways and crosstalk between them. Such deep understanding of the biological systems can benefit in the investigation of *e.g.* positive response, resistance to treatment, and adverse drug reaction in pharmacogenetic studies [9-11].

Thus we propose a method – and a tool that is based on it – for the integration of pathways reflecting inconsistencies among the data sources and producing a coherent subsystem, which – subsequently – can be analyzed by both static as well as dynamic methods and tools. For example, the latter type of analysis may consist of simulations of the resulting conditional pathway that can be now treated as one subsystem, assuming – of course – that the simulation engine can accommodate the richness of the conditional

* Corresponding author.

description, as is the case with several advanced analysis tools that have been devised recently, such as[12, 13].

The challenge of pathway integration is compounded by both biological as well as technical issues. It is well known that cells of different types and from distinct tissues under various environmental conditions have different gene expression profiles, which are the results of activating slightly different pathways [14-18]. Activating factors also play an important role in protein[b] modes and localization within a cell. Data inconsistency between different data sources, *e.g.* the same reaction being described differently, should not be forgotten either. Moreover, studied pathways are often divided into several sub-pathways due to technical limitations. Therefore, inappropriate cutoffs should also be taken into consideration.

Hence, during pathway integration all of the abovementioned stumbling blocks should be accommodated. The commonly used abstraction of a biological pathway as a labeled directed graph is not expressive enough to reflect minor but potentially critical differences between them. Many of those differences, such as tissue, cell type, and cell condition, are expressed verbally by biologists but cannot be included in current representations. Thus we developed a *conditional pathway algebra,* in which a "simple" pathway graph is enriched both with new node types as well as additional edge types accompanied with various attributes. We provide a formal definition of the algebra and prove some of its properties, such as closure and commutativity, and demonstrate its lack of associativity. Of these the most important is closure, *i.e.* ensuring that the resulting pathway is always contained in the set of pathways to which further operations are applicable.

The algebra is embodied in the *Pathway Integration Environment (PIE)* as a plugin for Cytoscape, a general-purpose, open-source environment for the large scale integration of molecular interaction network data and its visualization [19]. We applied PIE to all possible two-way as well as the three-way integration of three well characterized yeast signaling pathways: (i) Pheromone; (ii) Filamentous growth and (iii) High osmolarity glycerol pathways; the obtained results are confirmed in the literature.

[b] We refer to proteins, but this holds for other biological entities as well.

2. METHODS

To perform integration of pathways they must be represented as mathematical objects to which we can apply well-defined operations, thereby forming an algebra. We extend the common abstraction by which a pathway is represented as a graph, where proteins, genes, and small molecules are represented as labeled nodes, and interactions are represented as edges[19-22]: We add both node and edge types to reflect conditioning of activation. In this section we first define our conditional pathway algebra which includes various entities and operations for pathway integration. Then we describe the algorithms supporting the algebra's operations and prove a variety of properties that they preserve. Then we provide details concerning the implementation of a software tool embodying the algebra. We conclude this section with a description of the biological model system we used for validating our method. Please note that readers who are less mathematically inclined can go directly to Section 2.4.

2.1. Model

We model a biological pathway as a labeled, directed graph $G(V, E)$, where the nodes represent biological entities (proteins, genes, small molecules, enzymes, mRNA, etc.) and the edges represent interactions and other relations between them. The set of nodes V comprises three types: *regular* biological entities (V_b), *external* (boundary) conditions (V_e), and *putative* conditions (V_p), which support the pathway integration process. Each node is labeled with a name, as detailed below. The set of edges E contains both *common* interactions between pairs of nodes in V_b ($E_{comm} \subseteq (V_b \times V_b)$), as well as *conditional* edges (E_{cond}, see formal definition below) where each starts at a node (of any type) and points to a common edge [13]. The latter are used to model the case in which the activity of the common edge that is pointed at depends on the state of the start node, as explained below. Both types of edges can reflect either a positive or a negative effect: common edges denote *e.g.* activation or repression in regulation and signaling pathways, and conditional edges can represent either a positive or a negative dependency.

Names of regular nodes are taken from a common namespace of genes, proteins, etc.; we assume that the same name is being used to denote a certain biological

entity in all pathways under study (or that they are normalized by a pre-processing step) so as to allow their identification during the integration process. Additional attributes can be assigned both to whole pathways as well as to individual nodes: the former include cell type, tissue and physiological conditions, and the latter are *e.g.* cellular localization and activation state.

When several common edges go into a single node their joint effect is open to interpretation and there are various dynamic models that can be used. One prevalent model, used for both signaling and regulation pathways, is the sigmoid model (*e.g.*[23]). This issue is beyond the scope of this paper, but we found it useful – for purposes of the integration process – to qualify a node as having the AND attribute if all its incoming edges should be active in the system for this node to be activate; for example, let X be a complex; then all its components should be present in the system for its formation. Further interpretation of the combined effect of the incoming edges is left to the ensuing analysis tools.

Finally, we are ready to define conditional edges. For a given pathway *P*, let $cond_w(e)$ be a conditional edge between node *w* ($w \in V$) and an edge $e=(u_b,v_b)$, denoted $w \rightarrow e$ or $w \rightarrow (u_b,v_b)$. $cond_w(e)$ expresses the fact that the activity of the pointed-at edge *e* is conditioned upon the status of the node *w*, *e.g.* whether it is active or not. There are 3 options for *w*:

1. It is a common node [$w \in V_b$]: Note that ub in addition to being the source node of the edge e can also function as w, *i.e.* $w=u_b$.
2. It is an external conditions node [$w \in V_e$]: $w \rightarrow (u_b,v_b)$
3. It is a putative conditions (PC) node [$w \in V_p$]: $w \rightarrow (u_b,v_b)$

Notice, that the biological meaning of each type of conditional edges is different. Specifically, the last kind expresses a possible explanation for an edge *e* which is beyond a pathway's scope. One interpretation of a source PC node may be the presence of an additional protein or mRNA missing in the original pathway that has a role in regulating an edge it points to. By specifying pathways in which an edge *e* is present, the PC node can help in emphasizing that this edge is missing in other pathways.

Furthermore, note that a conditional edge is a triple (rather than a pair) of nodes, *i.e.* $E_{cond} \subseteq ((V_b \cup V_e \cup V_p) \times V_b \times V_b)$; thus – formally – it can be regarded as a hyperedge of size 3. Hence pathways in our representation are in fact directed hypergraphs where the maximum size of a hyperedge is 3. Still, for sake of presentation, unless it is necessary we refer to conditional hyperedges as edges; similarly, we disregard the signs of edges (positive or negative). To summarize, a pathway is represented as *P(V,E)*, where: $V = V_b \cup V_e \cup V_p, E = E_{comm} \cup E_{cond}$ with labels associated with nodes, and signs – with edges.

Next we define four merge operations between two pathways, P_1 and P_2:

1. *Graph Union* (denoted $P_1 + P_2$), performed as the set union of both nodes and edges and is based on their name equality (Fig. 1a).
2. *Positive Conditional Merge* ($P_1 \oplus P_2$, Fig. 1b). The result is a conditional pathway which is a refinement of the results obtained by graph union (Operation 1 above) where conditional edges are added that point to differences between P1 and P2 thereby giving possible explanations to these discrepancies. The refinement is performed using conditional edges of the positive type that point to "edges that must be explained" (as the result of the Conditional Merge Algorithm, defined in Section 2.2 below).
3. *Negative Conditional Merge* ($P_1 \ominus P_2$, Fig. 1c). As in Operation 2, but the conditional edges that are added are of the negative type.
4. *Mixed Conditional Merge* ($P_1 \oplus\!\!\ominus P_2$, Fig. 1d). The result is a conditional graph composed of the result of operation (1) and both positive and negative conditional edges pointing to "edges that have to be explained".

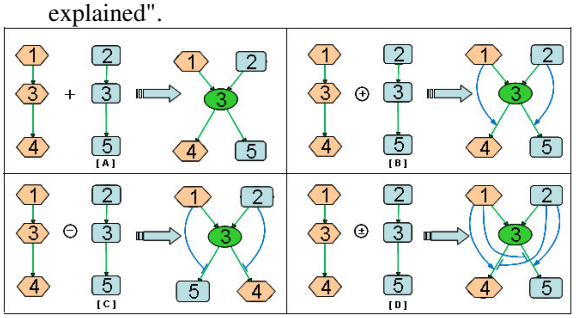

Fig. 1. The conditional pathway algebra operations. (A) *Graph Union*; (B) *Positive Conditional Merge*; (C) *Negative Conditional Merge*; and (D) *Mixed Conditional Merge.*

2.2. The Conditional Merge Algorithm

The operations defined above are realized by the *Conditional Merge* algorithm. We first perform the union of both graphs that will serve as the basis for further refinement. The union process detects the nodes that are present in both pathways, but some (or all) of whose outgoing edges are present in only one of the pathways; we call these outgoing edges "edges that must be explained" (denoted diffEdgeList). Then, according to the parameter mergeType, the algorithm finds appropriate explanations for these differences. There are five possible explanations, denoted: AND attribute, local, upstream, external (boundary) conditions, and putative conditions (see *2.2.1*). The order in which we search for an explanation is defined by the user (see *2.2.2*) using the parameter mergeMode. In case an explanation for a specific edge was found, the search for an explanation for this edge is stopped.

The pseudo-code for this algorithm is as follows; some of the support sub-functions are described briefly below.

Procedure ConditionalMerge (*A*, *B*, mergeType, mergeMode)

Input:
Graphs $A(V_A, E_A)$, $B(V_B, E_B)$
mergeType: one of {
Graph Union, Positive Conditional Merge, Negative Conditional Merge, Mixed Conditional Merge}
mergeMode: one of {
"Upstream", "Local, Upstream, External Conditions", "External Conditions, Local, Upstream"}
Output: Graph *G(V,E)* // a merge result of *A* and *B*

G: $A + B$;
if mergeType!= Graph Union **then**
for each v in $V_A \cap V_B$
let v_A: the node v in A, v_B: the node v in B;
denote the outgoing edges from node v in graph G by out(v_G);
// common edges leaving v in A but not in B
diffEdgeListA:= out(v_A) / out(v_B);
// common edges leaving v in B but not in A
diffEdgeListB:= out(v_B) / out(v_A);
removeExplainedEdges(A,B,diffEdgeListA);
removeExplainedEdges(B,A,diffEdgeListB);
if diffEdgeListA $\neq \emptyset$
then findExplanation(diffEdgeListA,v_A, mergeMode);
if diffEdgeListB $\neq \emptyset$
then findExplanation(diffEdgeListB,v_B, mergeMode);
return(G);

The procedure removeExplainedEdges removes the edges that have an explanation from its 3rd parameter. The procedure findExplanation finds the explanations (as per its last parameter) for the edges in its 1st parameter as explained in Section 2.2.1 below. In addition to the resulting merged graph, the algorithm also produces a list of inconsistent conditional edges.

2.2.1. *Kinds of explanation*

As mentioned above, there are five possibilities for an edge $e=(u,v)$ that must be explained (without loss of generality, we assume $e \in V_A$). The first option is to ascribe an AND attribute to the node v which might be a complex; in this case that one of v's compounds is missing in Pathway B, the complex v will not be created and – therefore – edges from other compounds to it will be missing. The second possible explanation is *local*, in which a difference in node u's attributes explains e. A possible biological scenario: in cytosol u activates protein v which is not present in the nucleus, therefore activation is impossible. The third one is *upstream* in which we are looking for some protein w upstream of u that is present only in pathway A and which possibly stimulates v's activation by u, as described by a conditional edge. In case we do not find such a w, a PC node and a corresponding conditional edge are added; this provides the fourth possible explanation. The fifth kind of an explanation is an *external condition* in which we compare pathway attributes and add an external condition node and a conditional edge from it to an edge e. Such an explanation can be used in case that some processes are activated in heat shock but depressed in cold shock.

2.2.2. *Merge mode*

For a given edge list, diffEdgeList, that contains edges to be explained, the order in which potential explanations are checked is defined by the *merge mode* we are in. This order, that is provided as a parameter (that can assume one of several possibilities) to the algorithm, can lead to significantly different merge results. Each of the possible modes reflects a slightly different biological approach: The "Upstream" mode, in which we look only for upstream differences in the graphs, is preferable in cases where we want to suppress system condition effects and are interested in investigating the effect of the pathway's structure. When

we are interested in the detection of the differences starting from the node state and localization, *i.e.* the differences that are the closest to the node we are dealing with, the "Local, Upstream, External Conditions" mode should be chosen. External condition attributes are the last to look at since those signal effects come from outside the cell. The last mode, "External Conditions, Local, Upstream", is pretty similar to the previous one besides the assumption that the external condition attributes should be examined first. The intuition is that it is meaningless to merge pathways that occur in different external conditions.

2.2.3. *Inconsistency between edges*

For a given pathway A, let $cond_w(e)$ be a conditional edge between a node $w \in V_A$ and an edge $e=(u,v)$, (where $e \in E_A$). During a conditional merge, we evaluate whether each of the previously added conditional edges is consistent with the merged pathways. This evaluation is based on the following truth-table:

Case	x: Regulation type in A	y: $e \in E_B$	z: $w \in V_B$	XOR(x, y,z)	Whether consistent
1	+	+	+	1	Yes
2	+	+	-	0	No
3	+	-	+	0	No
4	+	-	-	1	Yes
5	-	+	+	0	No
6	-	+	-	1	Yes
7	-	-	+	1	Yes
8	-	-	-	0	No

Legend: In column X (regulation type) "+" means positive, "-" means negative; in columns Y, Z "+" means that the membership predicate holds and "-" means that it does not hold.

Intuitively, when the regulation type is positive (denoted by +) the meaning is that w stimulates a reaction from u to v, *i.e.* in case w is inactive the edge from u to v will have no effect. The opposite is valid for a negative (denoted by -) type of regulation. Therefore, in Case 1, since we have positive regulation and w is present in Pathway B, it is reasonable that the edge e will be present as well. However, in Case 5, w that represses e is present, therefore e should not be present, contrary to what we have in the pathway and therefore it is a case of inconsistency. All the other cases can be analyzed in a similar fashion. Moreover, it is evident from the table that this logic can be expressed mathematically by the XOR operator.

2.3. Properties of the Conditional Pathway Algebra

Our algebra satisfies two important properties: *closure* and *commutativity*. Recall that a set is closed under an operation if when applied to any members of the set the operation returns a value that is a member of the same set; similarly, it is closed under a collection of operations if it is closed under each of the operations individually. In our case, the system under consideration comprises graphs that describe pathways and the operations defined above. For lack of space, we omit the proof of this property, but we note that it allows us to apply the various operations on the results of previously performed operations without any restrictions.

Commutativity means that for any two pathway graphs, P_1 and P_2, that are members of the set of conditional graphs, the order of the operands does not matter, *i.e. e.g.* $P_1 \textcircled{+} P_2 = P_2 \textcircled{+} P_1$ (and likewise for the other operations). Again, we omit the proofs for lack of space.

Unfortunately, *associativity* does not hold for our algebra (except for the simple Graph Union operation). In other words, for three given graphs P_1, P_2, and P_3 it is not necessarily the case that $(P_1 \otimes P_2) \otimes P_3 = P_1 \otimes (P_2 \otimes P_3)$ where $\otimes$ is one of the other three operations. When our merge algorithm encounters inconsistency in the explanations, it allows for the user's intervention or leaves the previous explanation. Therefore, the result of the second operation depends on condition edges that were previously added as well as user intervention.

2.4. The Pathway Integration Environment (PIE)

PIE is a software tool that realizes our conditional pathway algebra. It was implemented in Java as a plug-in for Cytoscape, a free bioinformatics software platform for visualizing molecular interaction pathways and integrating these interactions with gene expression profiles and other static data [19]. Therefore, we benefit from existing Cytoscape features: import/export of different file formats, pathway drawing and editing, searching and filtering of nodes and edges according to various attributes, and pathway visualization showing a variety of data attributes using visual means.

We can divide PIE's functionality into two, as follows:

1. Performing the Union and Merge operations; these operations were defined earlier in this section and are implemented in PIE.
2. Enriching pathways with known data:

- **Attributes' data entry:** Different conditions are stored in pathways, nodes, and edges as attributes. Whereas cell type, tissue, and physiological conditions (such as starvation and heat shock) that can be additionally assigned to a specified sub-pathway are more appropriate to be referred to as pathway attributes, protein localization and activation state (phosphorylation, ubiquitination, glycozilation, etc.) are naturally represented as node attributes. Since PIE is implemented as a plug-in to Cytoscape, we could use an existing attributes platform. However, in the current version of Cytoscape an attribute value is assigned to all the proteins with the same id in all opened pathways. Sometimes we are interested in assigning different attribute values to the same protein in different pathways. Since it is critical, we implement this feature differently allowing a shorter and more intuitive way to assign attributes to nodes.
- **Adding conditional edges:** Known dependencies that are expressed by conditional edges can be manually entered using the Cytoscape editor (which was enhanced by the HyperEdgeEditor plug-in to support this feature). Those conditional edges can utter conditional regulation on protein state, and localization within the cell by adding an activation or a repressing edge from the node itself to one of its outgoing edges. Similarly, by adding edges from another node to an edge that represents regulation we want to impose a condition on it. In addition, there is a possibility to add additional activating or repressing mediators to an already conditioned edge.

  Note that a user can add another conditional edge in which a source node points to an existing conditional edge. Our algorithm will not generate such a situation; it can, however, work with such a pathway as an input.
- **Adding an AND attribute to a node:** Not less important is the ability to emphasize that for the creation of some proteins, complexes, mRNAs, etc. there is a need in simultaneous regulation of several proteins, enzymes, etc. This property is implemented by PIE as an AND attribute of the node that is conditioned on others.

2.5. Pathways Under Study

To validate our method we applied it to three well characterized *Saccharomyces cerevisiae* (yeast) pathways: the Filamentous Growth (FGP), Pheromone Signaling (PSP), and High Osmolarity Glycerol (HOGP) pathways as they are described in the Science magazine cell signaling database STKE [44]. Yeast is an important model system for eukaryotic organisms, and the selected pathways represent well-studioned and highly curated biological functions, allowing us to evaluate our *in silico* predictions: the pathways (as described in the rest of this section) each comprise a small and simple subsystem, but they are strongly interrelated so it is possible to check the predictions that were obtained when integrating them. Still, each STKE pathway entry had to be corrected to reflect later updates in the literature and some curation errors; these changes are specified herein for each of the cases under *network correction*.

• *Filamentous Growth Pathway (FGP)* [48]

In response to nitrogen starvation and other signals, diploid a/α yeast cells undergo a developmental change and switch to a filamentous form of growth called pseudohyphal development. This transition includes cell elongation, a switch to a unipolar budding pattern, maintenance of the attachment between mother and daughter cells, and the consequent ability to invade semisolid media. This morphological change is likely to cause a foraging response that allows cells to scavenge for nutrients. In haploid cells this switch is termed haploid invasive growth [24, 25].

Network correction. The database entry includes the Fus3 repressing Tec1 edge which is not functionally related to this network; it exists only when the pheromone pathway is active as one way to represent cross-talk between the FGP and PSP [26, 27]. The authors probably added this edge to the network to show a broader view of regulation for this pathway. We, however, thought that moving this edge to the PSP (Fig. 2a) will provide a more correct view of the networks.

Another correction was made with the Ste11 and Ste50 nodes. All three pathways under study include these nodes; in two of the networks the edge between these nodes is described as activation, whereas in the third network it appears as neutral. Base on networks

description and literature [28-31], we decided to accept the neutral form of the edge (Fig. 2b).
Overall this network included 32 nodes and 40 edges.

• *Pheromone Signaling Pathway (PSP)* [49]
Yeast cells can exist as either haploid or diploid cells. Haploid cells of the opposite mating type (a or α) can mate, *i.e.* fuse and form a diploid. Cellular responses to mating include: arrest in the G1 phase of the cell-cycle, oriented growth towards the mating partner, and – ultimately – fusion of the plasma membranes of the mating partners, followed shortly thereafter by the fusion of their nuclei [24, 32].
Network correction. We added the Tec1 node and the Fus3 repressing Tec1 edge from the FGP (see text above for explanation). We also changed the edge between the nodes Ste11 and Ste50 to be neutral as in the FGP (Fig. 2b, text above). Furthermore, based on the networks explanation and literature [24, 25, 32], we decided to change Fus3 to Dig1/2 edges from activation to repression (Fig. 2c).
Overall this network included 32 nodes and 46 edges.

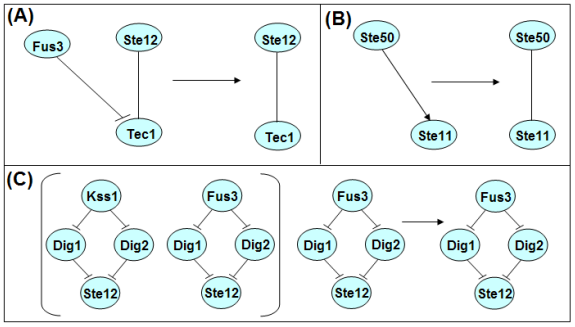

Fig. 2. Network corrections for the pathways under study.

• *High Osmolarity Glycerol Pathway (HOGP)* [50]
The internal osmolarity of a growing yeast cell is maintained to be higher than the external osmolarity. Increasing external osmolarity is a commonly encountered stress for a yeast cell in various natural environments such as a split-open grape drying under the sun, a Petri dish left open in the incubator, or the start of fermentation when sugar is added. The high osmolarity glycerol (HOG) MAPK pathway is activated by an increased environmental osmolarity and results in a rise of the cellular glycerol concentration so as to adapt the intracellular osmotic pressure [24, 33].
Network correction: No changes were necessary for this network.
Overall this network included 29 nodes 31 edges.

3. RESULTS AND DISCUSSION

We defined a conditional pathway algebra that safely extends the traditional graph theoretical-based pathway description model to include *e.g.* protein localization and external conditions in which pathways are activated, and take these factors into consideration during pathway integration. Not only is it the case that no information is lost during the integration process, but rather new information regarding either of the pathways and about crosstalk between them – as well as possible effects of some proteins on specific regulations – can be generated.

PIE is a bioinformatic tool that implements the formal algebra and the integration algorithm. It was applied successfully to several cases, as described in this section. Moreover, it proved to be an excellent research tool: each of the conditional edges that were generated by the conditional merge algorithm contains important information about the relationships between the involved pathways, the reliability of the edges in their original description, or lack of pertinent information in it. Each conditional edge can then be validated by performing wet experiments or a literature search. In the integration experiments that we performed and report herein, no contradictions were detected between the literature and conditional edges that were added to the graphs. Moreover, in one of the cases, when a contradiction seemed to appear, we found a more recent paper that corrected it.

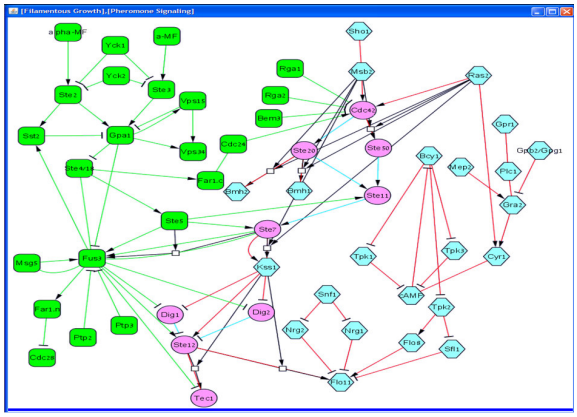

Fig. 3. *Positive Conditional Merge* of the pheromone signaling (nodes shown as rounded rectangles) and filamentous growth (hexagonal nodes) pathways. Nodes in the intersection of the pathways appear as ellipses.

- *Pheromone Signaling pathway* ⊙ *Filamentous Growth pathway*

The resulting network includes 62 nodes, 9 of which are common. Out of 105 edges, only 9 edges have to be explained and this is done by adding merely 12 conditional edges. Still, each and every one of these edges reflects either some biological observation or the lack of important data in the original description, as explained next.

The FGP is composed of at least three signaling modalities that control the switch from budding to filamentous growth in *S. cerevisiae* [34]. The core of the main pathway is a three-tiered mitogen-activated protein kinase (MAPK) cascade [34]. This cascade shares multiple components with the PSP that uses almost the same MAPK cascade, while the other two modalities do not show any overlap with the PSP ([24, 35] and Fig. 3). In fact, these two pathways are an excellent example for the case in which two pathways that are quite different in their original composition converge to one unit centered around an almost identical subcircuit that in itself is highly conserved in evolution. The main difference between the two pathways is that they use two different MAPKs as the target of the phosphorylation cascade: Kss1 in the FGP and Fus3 in the PSP [34, 36]. In the pheromone signaling pathway, Fus3 – together with the MAPKK Ste7 and the MAPKKK Ste11 – is bound to Ste5, a scaffold protein that is PSP specific [34, 36]. In addition, Fus3 – in response to a mating signal – is specifically down regulating Tec1, a FGP specific transcription factor [27]. In the end of the MAPK FGP, Tec1 forms a heterodimer with another transcription factor, Ste12, to mediate various gene expression responses [24, 25]. In the PSP, Tec1 is down regulated and two Ste12 molecules join together to form a homodimer that induces or represses genes that are required for successful mating [24, 32].

The conditional merge algorithm succeeded in pinpointing the right players that are specific to one of the pathways, meaning: Ste5 and Fus3 in the PSP and Kss1 and Tec1 in the FGP. We also expected that the algorithm would identify other pathway specific components that are more upstream to the MAPK core of the pathways, and indeed the Ste20 to Bmh1/2 edges are examples for such components (Fig. 4a): Bmh1 and Bmh2 are two genes in yeast that show strong similarity to the 14-3-3 proteins (acidic dimeric molecules that likely play a role in signal transduction). Bmh1 and Bmh2 – when associated with Ste20 – are required for FGP but not for the PSP [37]. Ras2 and Msb2 are the two Ste20 upstream components of the FGP that are specific to this pathway ([34] and Fig. 3). The algorithm correctly identified this relationship.

The conditional merge algorithm also identified the proteins that are specific to only one of the pathways and that will be used as explanations for the edges that must be explained. The edge from Ste7 to Kss1 (Fig. 4b) is a good example for this: as explained previously, one of the main differences between the FGP and the PSP is the use of two different MAPKs, Kss1 and Fus3. As described in Section 2.5, Kss1 is used only in the FGP; furthermore, Ras2 and Msb2 are the inputs for the FGP ([34] and Fig. 3). The algorithm recognized that Kss1 is unique to the FGP and inferred that Ras2 and Msb2 are responsible for this edge. Moreover, the algorithm correctly connected between the FGP Kss1 MAPK and this pathway's unique output Tec1 and Flo11 (the edges from Ste12 to Tec1 and from Ste12 to Flo11; see Fig. 4c).

An additional feature of our algorithm is to identify edges with low confidence or incomplete data. The Cdc42 and Ste50 nodes appear in all of the three pathways, but this edge exists only in the FGP (Fig. 4b). The algorithm found this conflict and marked it for further questioning. Although this edge exists in the FGP, it appears there with low confidence. The three networks we used for our analysis were last update in 2005. Interestingly, a more recent paper reports that this edge exists in the HOGP as well [38]. Apparently the data for this edge is not complete; the algorithm recognized this fact and used the available data to call Ras2 and Msb2 as the best explanations for this edge using information that is captured in those pathways.

Finally, an important property of our algorithm is the ability to use different merge operations (activation, repression, or both repression and activation). The usage of these operations may provide varying perspectives of the networks and highlight interesting nodes. When choosing conditional activation/repression, we can obtain different graphs (Fig. 5). In case of the FGP and PSP, on the edge Ste7 to MAPK (Fig. 4d, 4e) we get two different MAPKs and explanations, which are both reasonable. The algorithm explains the edge by the nearest explanation. After examination of the results of different merge operations for the networks we worked

with, we concluded that the most appropriate operation was indeed "conditional merge activation".

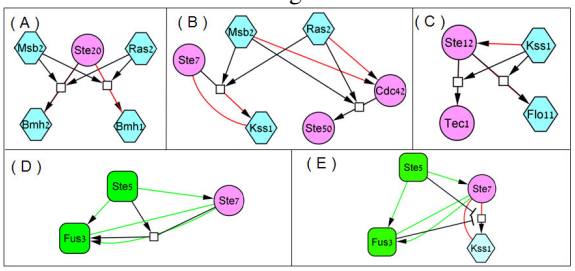

Fig. 4. Details of the *Positive Conditional Merge* of the pheromone signaling and filamentous growth pathways (excerpts from Fig. 3).

• *Pheromone Signaling pathway* ⊕ *HOG pathway*

The PSP and the HOGP share some components and regulation, but this commonality is less than what we found between the pheromone signaling and FGP ([24, 33, 39] and Fig. 6). Thus, in their integration we can see more examples of "Putative Conditions" (*e.g.* PC:Pheromone Signaling and PC:HOG).

The obtained network includes 63 nodes, 6 of which are common. Out of 105 edges, only 6 edges have to be explained and it is done by 7 conditional edges.

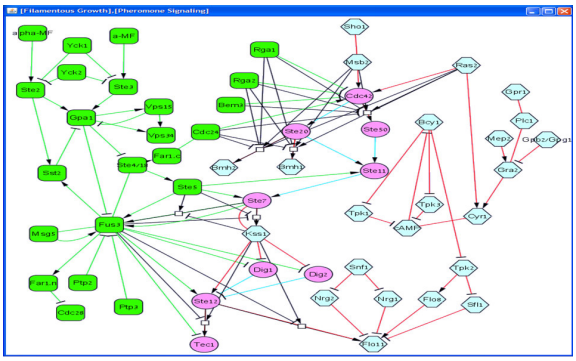

Fig. 5. *Mixed Conditional Merge* of the pheromone signaling and filamentous growth pathways. The shape coding is as in Fig. 3.

Notably, here the algorithm identified pathway specific components that are more upstream, for example in the case of the Ste11 to Pbs2 edge (Fig. 7a) the algorithm detected that the Ste11 node is present in both pathways but the Pbs2 node is present only in the HOGP (Fig. 6). As explained in Section 2.2.1, the algorithm looks upstream to find an appropriate explanation to this edge, and it found that the unique elements of the HOG pathway that can be responsible for this edge are Msb2 and Sho1. When we looked in the literature we found that indeed these proteins were associated with Pbs2 activation through Ste11 [40, 41].

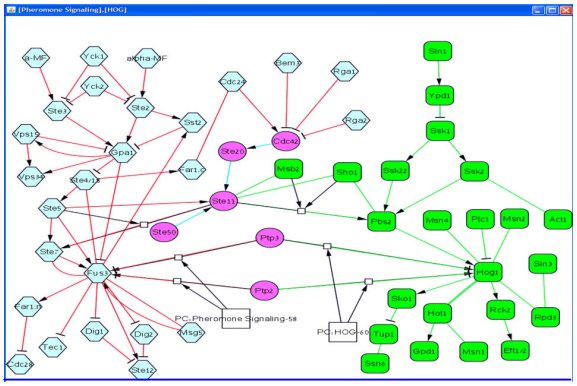

Fig. 6. *Positive Conditional Merge* of the pheromone signaling (hexagons) and HOG (rounded recrtangles) pathways. Nodes in the intersection of the pathways appear as ellipses, and Putative Condition (PC) nodes are rectangles.

A similar situation (one node present in both pathways and another that is unique to only one pathway) is manifested in two edges: Ptp2/3 to Fus3 and Ptp2/3 to Hog1. There is, however, one important difference compared to the previous situation: whereas then the algorithm had upstream components to resolve the conflict with, now there are no upstream nodes to explain the conflict (Fig. 6). In this situation the algorithm adds a PC node and an edge from it to the pathway's original edge (Fig. 7b); this unknown node is a sign to the user that he or she needs to find some other explanation to the conflict.

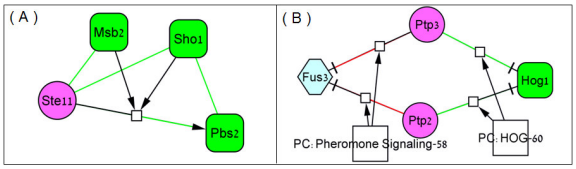

Fig. 7. Details of the *Positive Conditional Merge* of the pheromone signaling and HOG pathways (excerpts from Fig. 6).

• *Filamentous Growth pathway* ⊕ *Pheromone Signaling pathway* ⊕ *HOG pathway*

When merging three pathways the situation is more complicated since inconsistencies between some of the edges can appear. Furthermore, the outcome may depend on the order of the merging, due to the lack of associativity. On the other hand, we may obtain a more accurate explanation after adding the 3rd pathway for an edge for which we had a poor explanation when merging only two of the three pathways. For example, if we merge the three pathways in this order: (FGP ⊕ PSP) ⊕ HOGP, we get the conflict shown in Fig. 8a. If, however, we perform the merging in another order: FGP ⊕ (PSP ⊕ HOGP), we get a new edge with one

explanation (Fig. 8b). This, incidentally, supports Ras2 as being the better explanation.

Let us consider the results of the first ordering, namely (FGP ⊙ PSP) ⊙ HOGP. All new regulations that were added to the (FGP ⊙ PSP) pathway after merging it with the HOGP pathway were the same as those observed and explained for the (PSP ⊙ HOGP) merger (described above). This result is not surprising since the FGP and the HOGP hardly share any proteins. No additional regulations were received also in the case of the other merging order, namely FGP ⊙ (PSP ⊙ HOGP), for the same reason as above.

Note, however, that as the number of merged pathways increases (more than two), the results could be more cluttered and the user may get new inconsistent explanations (as exemplified in Figs. 8). In case the merge order is FGP ⊙ (PSP ⊙ HOGP), we obtain 2 new inconsistencies that were not present in the previous merge order. In this situation, the user must use the algorithm with care and to decide – based on other sources of information – what the best explanation would be.

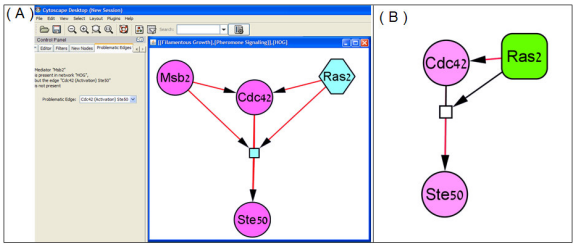

Fig. 8. Inconsistent edges during the *Positive Conditional Merge* of the filamentous growth, pheromone signaling, and HOG pathways

4. SUMMARY AND FUTURE WORK

PIE is a powerful tool for combining data from different sources that describe pathways based on experiments in a variety of conditions. It is based on the conditional pathway algebra that we have defined, enabling its users to enrich biological pathway representation with knowledge that comes from the experimental conditions that were used and from previous studies. It also allows systems biologists to add to the pathways' representations important information that has up to now been described informally (in words), such as "this regulation occurs only if a certain protein is phosphorilated and is located in the cytoplasma or is conditioned on the presence/absence of some other protein in the system". In other cases, we are also able now to express the notion that for activating some protein X, there is a need in simultaneous co-regulation of three other proteins, X1, X2, and X3. Moreover, PIE can point out interactions that are conditioned by specific regulators (presence/absence of proteins, co-regulation, extracellular factor etc.); using simple graph union we would miss these issues, with no way for reconstruction. Finally, PIE can be used not only for safe information integration; it can also be leveraged as a research tool. During pathway integration, the user – using different modes of merging – can see both "core" differences between pathways as well as edges with low confidence or without enough support information. Focusing on and further investigation of these differences can enhance our understanding of the biological systems under study.

A natural next step would be the integration of pathways of different types, *e.g.* signaling, regulation, and metabolic pathways that all pertain to the same biological function, into one framework. This would lead us to deeper understanding of biological systems as a whole.

Finally there is the technical challenge of dealing with pathways that are represented in different, divergent file formats that are being used for pathway retrieval. These are both XML-based representations, such as BioPAX [46], SBML [42], KGML [47], and XGMML [43], as well as text formats, such as SIF. Platforms like Cytoscape are making progress towards the convergence of this issue.

Funding

We are grateful to the Wolfson Family Charitable Trust, the Center for Complexity Science (Horowitz Foundation), and the Galil Center for Medical Informatics, Telemedicine, and Personalized Medicine for their generous support.

Acknowledgments

We would like to thank Allan Kuchinsky (Agilent Labs) for his help with Cytoscape matters, and Sivan Bercovici, Michael Shmoish, and Noa Tzunz for their feedback on the PIE system. We would also like to thank Nili Avidan, Ariel Miller, and Tami Paperna for helpful discussions.

References

1. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. *Nucl. Acids Res.* 2008; **36**: D480-484.
2. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* 2000; **28**: 27-30.
3. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res.* 2006; **34**: D354-357.
4. Karp PD, Arnaud M, Collado-Vides J, Ingraham J, Paulsen IT, Milton H, Saier J. The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database. *ASM News* 2004; **70**: 25-30.
5. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* 2008; **36**: 623-631.
6. Letunic I, Yamada T, Kanehisa M, Bork P. iPath: interactive exploration of biochemical pathways and networks. *Trends in Biochemical Sciences* 2008; **33**: 101-103.
7. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005; **33**: D428–432.
8. Vihinen M. Signal transduction-related bioinformatics services. *Brief Bioinform.* 2003; **4**: 325-331.
9. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nature Reviews Genetics* 2003; **4**: 937-947.
10. Meyer JM, Ginsburg GS. The path to personalized medicine. *Current Opinion in Chemical Biology* 2002; **6**: 434-438.
11. Roses AD. Pharmacogenetics. *Human Molecular Genetics* 2001; **10**: 2261-2267.
12. Elkon R, Vesterman R, Amit N, Ulitsky I, Zohar I, Weisz M, Mass G, Orlev N, Sternberg G, Blekhman R, Assa J, Shiloh Y, Shamir R. SPIKE – a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics* 2008; **9**: 110
13. Rubinstein A, Gurevich V, Kasulin-Boneh Z, Pnueli L, Kassir Y, Pinter RY. Faithful modeling of transient expression and its application to elucidating negative feedback regulation. *Proceedings of the National Academy of Sciences* 2007; **104**: 6241-6246.
14. Chi JT, Wang Z, Nuyten DSA, Rodriguez EH, Schaner ME, Salim A, Wang Y, Kristensen GB, Helland As, Borresen-Dale AL, Giaccia A, Longaker MT, Hastie T, Yang GP, van de Vijver MJ, Brown PO. Gene Expression Programs in Response to Hypoxia: Cell Type Specificity and Prognostic Significance in Human Cancers. *PLoS Medicine* 2006; **3**: e47.
15. Staudt LM, Brown PO. Genomic Views of the Immune System. *Annual Review of Immunology* 2000; **18**: 829-859.
16. Liu Y, Krueger JG, Bowcock AM. Psoriasis: genetic associations and immune system changes. *Genes Immun.* 2006; **8**: 1-12.
17. Bowcock AM. Krueger JG. Getting under the skin: the immunogenetics of psoriasis. *Nature Reviews Immunology* 2005; **5**: 699-711.
18. Lindberg RL, Kappos L. Transcriptional profiling of multiple sclerosis: towards improved diagnosis and treatment. *Expert Rev Mol Diagn* 2006; **6**: 843-55.
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 2003; **13**: 2498-2504.
20. Deville Y, Gilbert D, van Helden J, Wodak SJ. An overview of data models for the analysis of biochemical pathways. *Brief Bioinform* 2003; **4**: 246-259.
21. Fukuda K, Takagi T. Knowledge representation of signal transduction pathways. *Bioinformatics* 2001; **17**: 829-837.
22. Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucl. Acids Res.* 2000; **28**: 4021-4028.
23. Li F, Long T, Lu Y, Ouyang Q, Tang C. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America* 2004; **101**: 4781-4786.
24. Gustin MC, Albertyn J, Alexander M, Davenport K. MAP Kinase Pathways in the Yeast Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* 1998; **62**: 1264-1300.

25. Stanhill A, Schick N, Engelberg D. The Yeast Ras/Cyclic AMP Pathway Induces Invasive Growth by Suppressing the Cellular Stress Response. *Mol. Cell. Biol.* 1999; **19**: 7529-7538.
26. Bao MZ, Schwartz MA, Cantin GT, Yates 3rd JR, Madhani HD. Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signaling specificity in yeast. *Cell* 2004; **119**: 991-1000.
27. Chou S, Huang L, Liu H. Fus3-regulated Tec1 degradation through SCFCdc4 determines MAPK signaling specificity during mating in yeast. *Cell* 2004; **119**: 981-90.
28. Grimshaw SJ, Mott HR, Stott M, Nielsen R, Evetts KA, Hopkins LJ, Nietlispach D, Owen D. Structure of the Sterile {alpha} Motif (SAM) Domain of the Saccharomyces cerevisiae Mitogen-activated Protein Kinase Pathway-modulating Protein STE50 and Analysis of Its Interaction with the STE11 SAM. *J. Biol. Chem.* 2004; **265**: 1071-87.
29. Jansen G, Buhring F, Hollenberg CP, Ramezani Rad M. Mutations in the SAM domain of STE50 differentially influence the MAPK-mediated pathways for mating, filamentous growth and osmotolerance in Saccharomyces cerevisiae. *Mol Genet Genomics* 2001; **265**: 102-17.
30. Posas F, Witten EA, Saito H. Requirement of STE50 for Osmostress-Induced Activation of the STE11 Mitogen-Activated Protein Kinase Kinase Kinase in the High-Osmolarity Glycerol Response Pathway. *Mol. Cell. Biol.* 1998; **18**: 5788-5796.
31. Ramezani-Rad M. The role of adaptor protein Ste50-dependent regulation of the MAPKKK Ste11 in multiple signalling pathways of yeast. *Curr Genet* 2003; **43**: 161-70.
32. Bardwell L. A walk-through of the yeast mating pheromone response pathway. *Peptides* 2004; **25**: 1465-76.
33. Dihazi H, Kessler R, Eschrich K. High Osmolarity Glycerol (HOG) Pathway-induced Phosphorylation and Activation of 6-Phosphofructo-2-kinase Are Essential for Glycerol Accumulation and Yeast Cell Proliferation under Hyperosmotic Stress. *J. Biol. Chem.* 2004; **279**: 23961-23968.
34. Truckses DM, Garrenton LS, Thorner J. Jekyll and Hyde in the Microbial World. *Science* 2004; **306**: 1509-1511.
35. Sabbagh W Jr., Flatauer LJ, Bardwell AJ, Bardwell L. Specificity of MAP kinase signaling in yeast differentiation involves transient versus sustained MAPK activation. *Mol Cell* 2001; **8**: 683-91.
36. Schwartz MA, Madhani HD. Principles of MAP kinase signaling specificity in Saccharomyces cerevisiae. *Annu Rev Genet* 2004; **38**: 725-48.
37. Roberts RL, Mosch UH, Fink GR. 14-3-3 Proteins Are Essential for RAS/MAPK Cascade Signaling during Pseudohyphal Development in S. cerevisiae. *Cell* 1997; **89**: 1055-1065.
38. Tatebayashi K, Yamamoto K, Tanaka K, Tomida T, Maruoka T, Kasukawa E, Saito H. Adaptor functions of Cdc42, Ste50, and Sho1 in the yeast osmoregulatory HOG MAPK pathway. *Embo J* 2006; **25**: 3033-44.
39. Dohlman HG, Thorner J. Regulation of G protein–initiated signal transduction in yeast: Paradigms and Principles. *Annual Review of Biochemistry* 2001; **70**: 703-754.
40. Cullen PJ, Sabbagh W Jr., Graham E, Irick MM, van Olden EK, Neal C, Delrow J, Bardwell L, Sprague GF Jr. A signaling mucin at the head of the Cdc42- and MAPK-dependent filamentous growth pathway in yeast. *Genes & Development* 2004; **18**: 1695-1708.
41. Posas F, Saito H. Osmotic Activation of the HOG MAPK Pathway via Ste11p MAPKKK: Scaffold Role of Pbs2p MAPKK. *Mol. Cell. Biol.* 1997; **18**: 1702-1705.
42. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003; **19**: 524-531.
43. Punin J, Krishnamoorthy M. Extensible Graph Markup and Modeling Language (XGMML) Specification. *http://www.cs.rpi.edu/#puninj/XGMML/draftxgmml.html* 1999.

Web Site References

44. STKE (Signal Transduction Knowledge Environment): http://stke.sciencemag.org/
45. PubMed: http://www.ncbi.nlm.nih.gov/pubmed/
46. BioPAX (Biological Pathways Exchange): http://www.biopax.org/
47. KGML (KEGG Makeup Language): http://www.genome.jp/kegg/xml/
48. STKE - Filamentous growth pathways: http://stke.sciencemag.org/cgi/cm/stkecm;CMP_14554
49. STKE – Pheromone signaling pathway: http://stke.sciencemag.org/cgi/cm/stkecm;CMP_13999
50. STKE – High osmolarity grycerol pathway: http://stke.sciencemag.org/cgi/cm/stkecm;CMP_14620