A Graph Query Algebra for Biological Pathway Exploration

Keqin Wu, Member, IEEE, Liang Sun, Carl Schmidt, Jian Chen, Member, IEEE

Abstract—We introduce an example-based pathway graph query algebra (PGQA) for inner- and between-pathway relationship exploration. Studying pathways of intricate relationships in context of rich information is challenging. Similar to query-by-example language, PGQA interprets relationship query given by selected examples to find a match for, extract the identical parts between, or tracing a path from any interested pathway components. By allowing selecting one or more objects on screen as query input and using the query output as next query input, PGQA provides intuitiveness, concurrence, and dynamics for biologists to explore pathway graphs.

Index Terms—Biological pathway, graph visualization, visual analysis, relational query

1 INTRODUCTION
Pathways partition the complicated biological network into smaller pieces relevant to individual biological processes. One important use of pathway graphs is to help biologists investigate the complex interactions and understand important molecular mechanisms, which can guide them towards the most valuable experiments producing biological findings. To achieve this, biologists conduct various activities which mostly involve sorts of queries, such as finding interested proteins and paths. However, it is challenging to effectively query the rich information from pathway graphs with intricately interrelated components.

Pathway graph query algebra (PGQA) is a tool developed to relieve biologists’ work load when exploring pathways of complex relationships. It is partly inspired by a database query language — Query-by-Example (QBE) [1] which allows users to make queries by creating examples on the screen. Concurrency is inherent in PGQA since multiple examples can be used in a single query which can be made across multiple pathway graphs at the same time.

PGQA is possibly the first tool of its kind designed for pathway graph exploration. It is also based on the participatory design and a characterization of the questions asked in pathway analysis. PGQA is currently integrated in a suite of visualization techniques available in a bubble-based visualization framework [2] where query is made across a set of grouped pathways. However, it can also be used with other tools in need of an intuitive and interactive graph exploration.

2 BACKGROUND AND RELATED WORK
2.1 Pathway Relationship
Biological pathways have complicated relationships within and between pathways. The within-pathway relationship is usually expressed as a compound graph depicting both interactions between biomolecules and containing relationships between pathway components of three levels: pathway (PW), compartment (CP), and biomolecule (BM). The between-pathway relationship is formed by all sorts of cross-talks between different pathways and often expressed by a link between shared proteins.

2.1.1 Pathway graph
Pathway graph consists of nodes representing biomolecules including protein, complex, small molecule, etc. and directional edges indicating the input, output, or catalyst of the reactions. Biomolecules are located inside certain compartments which exist in certain pathways.

An example pathway graph is shown in Fig. 1. We obtain pathway data sets from a popular public pathway database — Reactome [3]. Our graph layout integrates both free-directed algorithm (to reduce edge crossing and edge lengths) [4], the hierarchical layout (to emphasizes compartment information) [5], and tapered edges (to emphasize direction) [6]. We also slightly edit the node shapes, edge shapes, and colors used by Reactome for a better information presentation.

Fig. 1. (a) The graph of pathway "Translation". (b) Symbols used in pathway graphs.

2.1.2 Cross-talking between pathways
Two pathways are considered related when they share nodes, have edges linking one pathway to the other, and/or one is contained or referenced in the other [7]. Cross-talking proteins are the most common forms of interconnection between pathways in which one pathway affects the others and are critical in analysis scenarios such as judging effects of drugs [7].

2.2 Relationship Exploration Study
2.2.1 Pathway relationship exploration
Tools used for pathway relationship exploration include many general purpose graph tools. Several pathway study tools focus on designing graph layouts or multi-view interactions to facilitate the relationship exploration. For example, interconnection between pathways could be visualized in an overall network as used in KEGG Ailla [8] or be highlighted by Focus + Context methods such as Entourage [7] where one pathway map is the focus linked to subsets of related pathways. A work similar to ours is enRoute [9] which interactively explores experimental data along the paths selected by biologists. Another one is ConTour [10] which explores multi-relational datasets for drug discovery through brushing and linking. PGQA aims at providing a quick relationship discovery through interactive query without asking biologists to browse through the large and complicated pathway graphs.
2.2.2 Query-by-Example
Different from SQL, Query-by-Example (QBE) [1] allows users to make queries by creating examples on the screen. For example, Spatial-Query-by-Sketch [11] allows a user to formulate a spatial query by drawing the desired configuration and translates it into a symbolic representation used to query a geo-graphics database. This paper is similar to QBE in terms of using examples created on the screen, other than SQL statements.

3 Pathway Query Algebra

3.1 Analytical Tasks
Our design started with observing biologists’ activities under two topics: gene expression analysis which involves exploring individual pathway graphs and cross-talks between pathways. Except for viewing data content, there are mainly analytical tasks of three types: (1) the search of interesting pathway components, (2) tracing paths, and (3) compare (identifying shared parts) between compartments and pathways. For example, biologists may be interested in certain proteins or a path of a protein destined for secretion. Biologists’ query spans different levels (pathways, compartments, and nodes in pathway graphs). They also wish to perform queries based on previous query results or query multiple objects at the same time.

We list six questions that biologists often ask during pathway exploration in Table 1 and categorize them according to the relationship types that they involve and levels of pathway components that they cover. Given the size and the complexity of pathway graphs, biologists cannot answer those questions easily by visual browsing or using text search box.

To address these questions, we advocate a pathway query algebra, PQGA, based on a query-by-example paradigm [1]. A biologist makes a query by simply clicking example objects and selecting query type and level via menu, the system automatically interprets the example objects and highlights objects found in pathway graphs. Simplicity is achieved from biologists’ perspective by implicit object relationship presentation, without having to use query languages such as SQL.

3.2 Basic Concepts
Users interact with pathway graph through PQGA by simply selecting an example configuration (objects and expected output level). PQGA parses this input and translates it into a graph query.

(1) Query level. A user selects a query level so that the query will output objects of the given level. It could be biomolecule (BM), compartment (CP), or pathway (PW) and not necessarily the level of example objects selected. Since biologists are mainly interested in proteins, by default, selecting biomolecule level will output proteins.

(2) Semantic object. A semantic object is defined as a set of objects with a containing relationship. In a query, when a user selects a set of items with containing relationships, it implies that the query is to seek items with same containing relationships. If more than one object is contained by another object, or verse visa, more than one semantic object will be abstracted. Fig. 2 shows all the possible semantic objects composed of items of one to three levels. Fig. 2h shows an example of a user input where more than one semantic object will be extracted.

For example, if a user wants to find all proteins named SMAD3 contained by compartment Nucleoplasm, the user select a SMAD3 and a Nucleoplasm containing it, and set the search level as BM. This outputs all SMAD3s contained by Nucleoplasm in the grouped pathway graphs. If the user select the same input, but set the search level as CP, the output will be all Nucleoplasm containing SMAD3 in the group instead.

3.3 Query Algebra
A typical query consists of three steps: (1) select example objects, (2) set query level, and (3) query (match, compare, or trace path of) each selected object in grouped pathways. All the query cases are enumerated in Table 2.

(1) Match. The outputs are the items matched the names and relationships implied by the input sample semantic object. Table 2(1) gives all the input and output cases of the search for the items matched with a single semantic object. If a user selects more than one semantic object, the output will be the union of the results of individual input objects. For example, Match(Obj1, Obj2, ..., Objn, Level) = Match(Obj1, Level) ∪ Match(Obj2, Level) ... Union(Objn, Level)...

(2) Compare. Comparison is made between two compartments or pathways to find out all shared items of them. Different comparison cases are shown in Table 2(2) and example objects must be in the same level and be one of CP or PW. If more than one semantic object is selected, the output will be the intersec of the results of comparing every two input objects. For example, Compare(Obj1, Obj2, ..., Objn, Level) = Compare(Obj1, Obj2, Level) ∩ ... Compare(Objn, Objn, Level) ... Union(Objn, Objn, Level) ... Compare(Objn, Objn, Level).

(3) Path tracing. There are two path tracing types: one is finding paths from the selected objects while another is finding paths connecting selected objects. The two cases are shown in Table 2 (3) and (4). If more than one semantic object is selected, the output will be, for first case, union of the results of individual input objects, for second case, union of the paths connecting every input objects.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Levels</th>
<th>Relationships</th>
<th>Analytical tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Is a protein in other compartments?</td>
<td>X</td>
<td>X</td>
<td>Within-pathway</td>
</tr>
<tr>
<td>Q2 Is a protein in other pathways?</td>
<td>X</td>
<td>X</td>
<td>Between-pathway</td>
</tr>
<tr>
<td>Q3 Are there any shared proteins in given compartments?</td>
<td>X</td>
<td>X</td>
<td>Within-pathway</td>
</tr>
<tr>
<td>Q4 Are there any shared proteins in given pathways?</td>
<td>X</td>
<td>X</td>
<td>Between-pathway</td>
</tr>
<tr>
<td>Q5 What are the gene products linked to a given protein?</td>
<td>X</td>
<td>Within-pathway</td>
<td>Path Tracing</td>
</tr>
<tr>
<td>Q6 What are the paths connecting two given proteins?</td>
<td>X</td>
<td>Within-pathway</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2 Sematic objects defined by containing relationships. (a) <BM1, CP1, PW1> | BM2 ∈ CP2 and CP1 ∈ PW1; (b) <CP1, PW1> | CP1 ∈ PW1; (c) <BM1, CP1> | BM2 ∈ CP1; (d) <BM1, CP1> | BM2 ∈ CP1; (e) PW1; (f) CP1; (g) BM1; (h) Four sematic objects selected: Obj1=<<BM1, CP1, PW1> | BM2 ∈ CP1 and CP1 ∈ PW1>, Obj2=<<BM2, CP1, PW1> | BM2 ∈ CP1 and CP1 ∈ PW1>, Obj3=<<CP1, PW1> | CP2 ∈ PW1>, and Obj4=<<BM1, PW2> | BM2 ∈ PW2>.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Levels</th>
<th>Relationships</th>
<th>Analytical tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Is a protein in other compartments?</td>
<td>X</td>
<td>X</td>
<td>Within-pathway</td>
</tr>
<tr>
<td>Q2 Is a protein in other pathways?</td>
<td>X</td>
<td>X</td>
<td>Between-pathway</td>
</tr>
<tr>
<td>Q3 Are there any shared proteins in given compartments?</td>
<td>X</td>
<td>X</td>
<td>Within-pathway</td>
</tr>
<tr>
<td>Q4 Are there any shared proteins in given pathways?</td>
<td>X</td>
<td>X</td>
<td>Between-pathway</td>
</tr>
<tr>
<td>Q5 What are the gene products linked to a given protein?</td>
<td>X</td>
<td>Within-pathway</td>
<td>Path Tracing</td>
</tr>
<tr>
<td>Q6 What are the paths connecting two given proteins?</td>
<td>X</td>
<td>Within-pathway</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Questions for the analysis of pathway relationships.
Fig. 3. Find items of interest. (a) Search protein SMAD7 that is located in compartment Nucleoplasm. (b) Search shared proteins by two pathways. The input sample objects are highlighted with solid orange boundaries and objects found are with orange halos. Pairs of matched items are linked with blue lines.

Fig. 4. Find paths of interest. The input objects are highlighted with solid orange boundaries and objects found are with orange halos. (a) Search all paths in biomolecule level linked to complex pre-TGFB1 complex in three reaction steps. (b) Search all paths in compartment level linked to compartment Plasma Membrane. (c) Search paths linking proteins TGFBR2 and SMURF2. (d) Search paths from proteins TGFBR2 and a complex to SMURF2. The resulted two paths are highlighted in two colors respectively.
As shown in Table 2, PGQA can actually answer more questions than the 6 questions listed in Table 1, providing a potential for biologists to explore larger problem domains in different scenarios.

### 3.4 Examples

#### 3.4.1 Query Nodes of Interest

Biologists input example objects by direct click and find matched or shared items of the selected objects across pathways through menu operations before a query.

Fig. 3 shows two typical query results. The items selected by biologists as input are highlighted in solid orange boundaries, output items are highlighted with orange halos and each pair of matched items are linked with blue lines. Fig. 3a shows the results for finding all proteins named SMAD7 contained by compartment Nucleoplasm. Fig. 3b shows all the proteins shared by the two pathways.

#### 3.4.2 Query Paths of Interest

PGQA provides two path tracing operations. (1) Search “all paths” from a given starting node: highlights all the edges and nodes connected with the node in certain reaction steps. Applying this operation successively traces through all paths reachable from the selected node set. (2) Search “linking paths” between selected nodes.

Some resulted paths are shown in Fig. 4. Fig. 4a shows the result for finding all paths connected to a complex in three reaction steps. Fig. 4b shows the result for finding all paths from a compartment. Fig. 4c shows the path linking two proteins TGFBR2 and SMURF2. Fig. 4d shows the paths found after a path searching action made for TGFBR2, SMURF2, and a complex.

### 4 Case Study and Discussion

#### 4.1 Study Downstream Effect and Cross-talking between Pathways

This case study demonstrates the usefulness of PGQA in identifying downstream effects from a differentially expressed gene based on expression data and quickly finding cross-talk cases.

![Fig. 5. Study downstream effect of expressed proteins and cross-talking in two TGF pathways.](image)

Biologists obtain gene transcriptome data by wet experiment. The large amount of differentially expressed genes presents biologists a challenge to make biological sense of them. After loading the expression data and a pathway “TGFBR2 MSI Frameshift Mutants in Cancer”, gene SMAD4 and SMAD7 are found up-expressed (marked in red as shown in Fig. 5). Interested in the effect of knocking down gene SMAD4, we use path tracing operations to highlight all the genes connecting with SMAD4. Quickly browsing through the highlighted downstream paths, we find one path leading to protein complex p-2S-SMAD2/3-SMAD4 which will either directly or indirectly regulate the expression of gene MYC, JUNB, etc. inside nucleus (nucleoplasm). To further inspect if the downstream effect impact other TGF-beta signalling pathways, we then open one related pathway, e. g. TGF-beta receptor signalling activates SMADs, and use “match” option to find all the matched proteins in the new pathway for those in the highlighted paths. We find SMAD7 among the shared proteins. Eventually, through PGQA, we learn that that knocking down gene SMAD4 will affect the final expression of downstream genes in cell nucleus and knocking down gene SMAD7 would also affect pathway TGF-beta receptor signalling activates SMADs. The downstream effect of knocking down gene SMAD7 in this cross-talking pathway can also be traced via our method. The visualizations built during this case study are shown in Fig. 5.

#### 4.2 Discussion

Our biologist collaborators who are also the co-authors valued the exploratory capabilities of PGQA. They found that query becomes easy and fast by using PGQA. We believe that PGQA can be useful for other graph applications or be extended to address larger problem domains. For example, current PGQA added with species level may be used to study inference of pathways of different species.

### 5 Conclusion

This paper presents PGQA (pathway graph query algebra), a novel query method that supports biologists for interactive pivoting of multiplex biological pathways of intricate relationships. We base its query mechanisms on a simple example-based query that relieve biologists’ work load. As demonstrated, PGQA enables a dynamic and concurrent query on different pathway components, providing biologists new capabilities and experiences that other tools may not.

### Acknowledgments

This work was supported in part by NSF, Award No. 1147216.

### References


