Graph Data Representation in Oracle Database 10*g*: Case Studies in Life Sciences

Susie M. Stephens
Oracle Corporation
10 Van de Graaff Drive
Burlington
MA 01803
USA
susie.stephens@oracle.com

Johan Rung
McGill University and
Genome Québec Innovation Centre
740 Doctor Penfield Ave.
Montreal, QC, H3A 1A4
Canada
johan.rung@mail.mcgill.ca

Xavier Lopez
Oracle Corporation
One Oracle Drive
Nashua
NH 03062
USA
xavier.lopez@oracle.com

Abstract

*New technologies have been developed in the life sciences that allow researchers to study biological systems in rich detail. These advances have resulted in an abundance of data that describes the relations between the fundamental components of biological systems, such as genes, proteins, and metabolites. The network of relations between the components holds insights as to how biological systems function, and consequently can help researchers understand the mechanisms behind disease. Biological networks are commonly managed and analyzed in a graph representation. Oracle Database 10*g *has the functionality to model data as a graph, and thereby has the potential to greatly facilitate research. In this paper we describe the Oracle implementation and provide case studies from the life sciences.*

1 Introduction

Graph theory is an established field where users represent data as a series of nodes and edges [Pearl88]. A node represents an object of interest and an edge indicates a relationship between two nodes. Edges can have a direction indicating an ordering between the nodes, or be undirected. A graph typically assumes very little about the data it describes, enabling the nodes and edges of a graph to form a web of interconnectivity.

The behavior of systems is influenced by the topology of the graph, which represents the nature of the connections within the graph. Much early work in graph theory dealt with the properties of random graphs, where it is assumed that connections between any two nodes in a network are equally probable [ER60]. Many natural and man-made systems exhibit a scale-free topology where most nodes have just a few connections and only a few nodes form highly connected hubs [BA99]. Scale-free networks are considered to be robust as the risk of a random error to occur in a highly connected node is low. Such graphs also display a tendency to cluster, have shorter average path lengths between nodes, and are better able to demonstrate sustained growth [AB04, Sear03]. Graphs have been extensively investigated, for example in the social [Mil67], man-made [Al99] and biological domains [BO04, Alon].

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

For advances to be made in the understanding of biological systems and disease, it is critical for researchers to be able to gain insights into the role of different biological entities and their interactions. These interactions guide the life of a cell, allowing it to function and to respond to environmental stimuli. The size and complexity of biological systems makes analyzing their interactions challenging. Most cellular processes are a result of a cascade of events mediated by proteins. For example, a gene may express a transcription factor that regulates the expression of a different set of genes, a gene may express an enzyme that activates a set of proteins, two proteins may bind to each other to form a functional complex, or a gene may express an enzyme which catalyzes the production of a metabolic compound which in turn inhibits another enzyme.

Analysis algorithms such as nearest neighbor, minimum cost spanning tree, and within distance are commonly used with graphs to help researchers identify new insights in the data. These analytical techniques are especially powerful when combined with methods that enable cost or weight values to be assigned to edges, as this allows researchers to quickly limit a search to the desired targets. Cost values are typically used to represent interaction strengths, or to label an edge with a statistical value. The restriction of a graph to a sub-network of interest is critical as it enables graph visualization to become more manageable, and it reduces the computational complexity of subsequent analysis.

Managing life sciences data as a graph can facilitate the modeling of complex interactions because the emphasis is placed on the relationship between the objects. Graphs can enable complex networks to be visualized in a straightforward manner that captures the essential information about the structure of the system. Further, graphs that support hierarchies of information, where a set of nodes and edges can be contained within a parent node, are well suited to modeling the different organizational levels of biological. As a consequence of the suitability of graph modeling to life sciences data, its popularity as a method for data representation has increased over recent years. The range of applications of graph theory have also grown from primarily being used for metabolic pathway representation [Jeong], to areas as diverse as protein-protein interaction networks [Gag04, Reiss, Betel, Uetz], analyzing cell images [Gunduz], gene function determination [Sch03] and text mining [Je01].

Graphs have also been used in combination with fixed vocabularies and ontologies to aid scientists in answering questions of a more open-ended nature. For example, if protein data is represented as a graph and is mapped to an ontology that contains a concept such as INTERACTS-WITH, it becomes possible to identify all protein-protein interactions. A graph data representation can enable scientists to discover new insights by using pre-determined rules and ontologies to traverse across data sets.

The increasing number of data sources that contain biological pathway and interaction data, such as KEGG [Kan97, Kan00], BIND [BIND03], IntAct [IntAct04], DIP [DIP00], MINT [MINT02] and Ecocyc [Karp00], demonstrate the importance of systems data to the life sciences. As these data sources become more prevalent, and increase in size and complexity, it becomes more important that these data sources can be managed as a graph representation in a relational database management system (RDBMS). A RDBMS offers users the ability to store data in a secure, highly available and scalable environment. To date, it appears Oracle is the only relational database vendor to offer graph functionality in the database. This paper focuses on the implementation of a graph data model in Oracle Database 10*g*, and provides examples of its use within the field of bioinformatics.

2 Implementation

Oracle Database 10*g* includes a Network Data Model (NDM) as part of Oracle Spatial that enables users to model and analyze data as a graph [OraNDM]. NDM stores objects of interest as nodes, the relationship between nodes as links, and an ordered list of links that contain no repeating links or nodes as paths. NDM can be used to represent directed, undirected, random, scale free and hierarchical graphs. In addition, NDM has the ability to support logical graphs, and graphs that relate to spatial information. NDM is a model that represents networks in object-relational form in the database and as Java objects in the client or application tier.

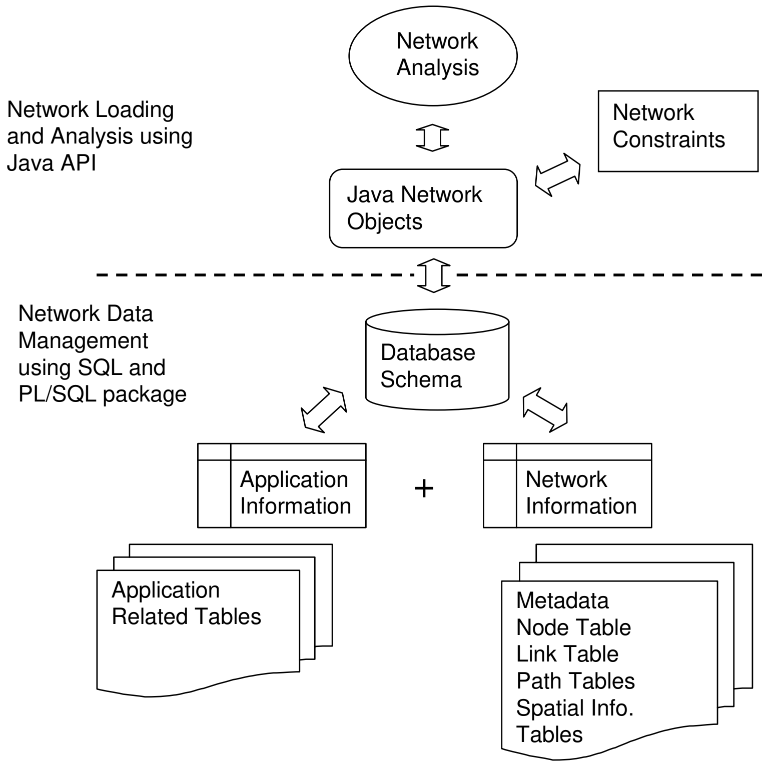

Figure 1: The Architecture of Oracle Spatial Network Data Model.

The NDM schema contains tables for storing the node, link and path data, and the network metadata. The NDM schema also includes a PL/SQL package for querying network data, maintaining referential integrity, and updating link, node and path tables. Standard SQL queries and PL/SQL functions and procedures can be performed on the data. Only generic connectivity information is captured in the data model, thereby providing a clean separation between the data and application information.

The Java API enables network loading, network representation and network analysis. NDM provides analytical capabilities for tracing, shortest path, minimum cost spanning tree, within cost and nearest neighbors. NDM also supports network constraints that are useful filters for network analysis, including minimum bounding rectangle, path cost, path depth, and must avoid nodes and links.

3 Network Data Model Case Studies

In the life sciences there are many possible use cases for a graph data model. In this section two possible use cases will be highlighted: metabolic pathway analysis and the integration of gene networks with gene expression data.

3.1 Metabolic Pathway Analysis

A graph data model is well suited to storing and analyzing metabolic pathways. Data from the KEGG metabolic pathway database from the University of Kyoto was loaded into NDM as a directed network. The chemical compounds were stored as nodes (table 1) and the enzymes were stored as links (table 2). Once the graph was

Table 1: Pathway Data Definition: Node table

NODE_ID	NODE_NAME	ACT	COSTS	SAMPLE_ID	ENTRY_ID
1	C00022	Y	1	Pyruvate	49
2	C00122	Y	1	Fumarate	50
3	C00036	Y	1	Oxaloacetate	51
4	C05379	Y	1	Oxalosuccinate	52
5	C00074	Y	1	Phosphoenolpyruvate	53
6	C00024	Y	1	Acetyl-CoA	54
7	C00149	Y	1	(S)-Malate	55
8	C00311	Y	1	Isocitrate	56
9	C00417	Y	1	cis-Aconitate	57
10	C00042	Y	1	Succinate	58

Table 2: Pathway Data Definition Link table

LINK_ID	LINK_NAME	START_NODE_ID	END_NODE_ID	ACT	COST	SAMPLE_ID
1	1.1.1.42 (rn:R00268)	4	19	Y	1	isocitrate dehydrogenase (NADP)
2	1.1.1.42 (rn:R00268)	19	4	Y	1	isocitrate dehydrogenase (NADP)
3	1.1.1.42 (rn:R00268)	4	20	Y	1	isocitrate dehydrogenase (NADP)
4	1.1.1.42 (rn:R00268)	20	4	Y	1	isocitrate dehydrogenase (NADP)
5	4.1.1.49 (rn:R00341)	3	19	Y	1	phosphoenolpyruvate carboxykinase (ATP)
6	4.1.1.49 (rn:R00341)	3	5	Y	1	phosphoenolpyruvate carboxykinase (ATP)
7	1.1.1.37 (rn:R00342)	3	7	Y	1	malate dehydrogenase
8	6.4.1.1 (rn:R00344)	1	3	Y	1	pyruvate carboxylase

stored in NDM, a Java visualizer was used to view the metabolic pathways and to interrogate the data using analysis including shortest path and all paths. Metabolic pathway analysis can guide researchers as to which drug discovery targets are likely to yield the most beneficial results.

3.2 Gene Network Analysis

Gene networks describe relations within a group of genes. Common approaches to gene network analysis include using the edge to represent correlations between gene expression profiles [Sz99], high mutual information [Butte], or the impact of a gene mutation on gene expression [Rung]. More complex examples include using a mathematical framework of probabilistic graphical models, such as Bayesian networks, to learn how the structure and parameters of a network fit observed data [Fr04, Hec95]. Gene networks can also be used to describe interactions between transcription factors and genes, where an edge typically represents a transcription factor physically binding to a regulatory site in the vicinity of a gene [Lee02].

This case study describes how gene expression data can be mapped to a signaling pathway, thereby providing additional insights to the perturbed component of the biological system. Gene expression profiles are of significant interest in life sciences as they can provide valuable insights into the underlying mechanisms of disease [Ra02]. In this example, signaling pathways are stored in NDM with genes represented as the nodes and interactions between the genes as the edges. Gene expression data from biological samples is stored in a separate relational table. Queries can be performed that span both the signaling pathway and the gene expression data, enabling researchers to identify subsets of the signaling pathway that are likely to be impacted by the changes in the gene expression profile (fig. 2). With NDM it becomes simple to retrieve connectivity information about genes.

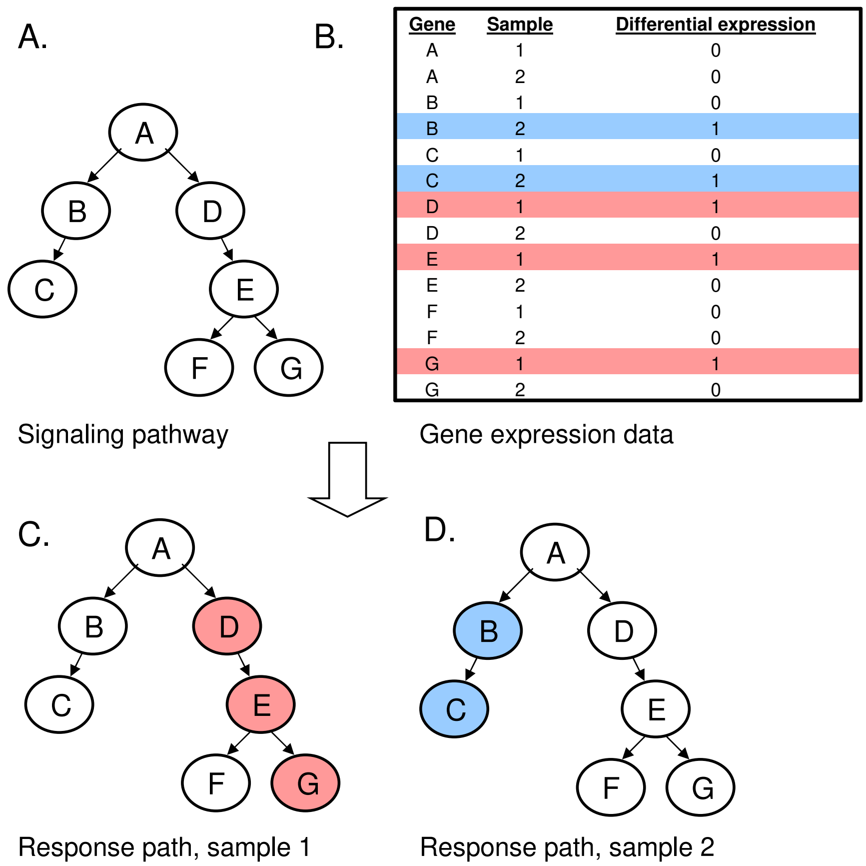

Figure 2: Case study for gene network analysis. A depicts a signaling pathway stored in NDM. B shows gene expression data stored in a relational table. C and D highlights the differential pathway expression between sample 1 (red) and sample 2 (blue) that was determined by performing a query that spans both data sets.

4 Conclusions

NDM enables scientists to manage and analyze data as a graph representation in Oracle Database 10*g*. NDM has been implemented by Oracle as an open and generic model that separates application logic from data management. This approach enables industries ranging from telecommunications, geographic information systems, electronics and life sciences to take advantage of the functionality.

In this paper, case studies were provided as to how NDM can be used for storage and analysis of metabolic pathways and gene networks. As increasing volumes of data in the life sciences are represented as a graph, it becomes important that a secure, reliable and scalable environment is provided for the data. As relational databases already manage much data in the life sciences, they should provide a strong platform for graph data management in drug discovery.

5 Acknowledgements

Data for table 1 and 2 used with the kind permission of Prof. Susumu Goto.

References

[Al99] R. Albert *et al.* Diameter of the World-Wide Web. *Nature*, 401: 130-131, 1999.

[AB04] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74: 47-97, 2002.

[Alon] U. Alon. Biological Networks: The Tinkerer as an Engineer. *Science*, 301: 1866-1867, 2003.

[BIND03] G.D. Bader *et al.* BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31: 248-50, 2003.

[BA99] A.-L. Barabási and R. Albert. Emergency of Scaling in Random Networks, *Science*, 286: 509-512, 1999.

[BO04] A.-L. Barabási and Z.N. Oltvai. Network Biology: Understanding the Cells Functional Organization, *Nature Rev. Gen.*, 5: 101-113, 2004.

[Betel] D. Betel *et al.* Analysis of domain correlations in yeast protein complexes. *Bioinformatics*, 20 Suppl. 1: 55-62, 2004.

[Butte] A.J. Butte *et al.* Discovering Functional Relationships Between RNA Expression and Chemotherapeutic Susceptibility using Relevance Networks, *Proc. Natl. Acad. Sci. USA*, 97: 12182-12186, 2000.

[ER60] P. Erdös and A. Rényi. On the evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5: 17-61, 1960.

[Fr04] N. Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models, *Science*, 303: 799-805, 2004.

[Gag04] J. Gagneur *et al.* Modular decomposition of protein-protein interaction networks. *Genome Biology*, 5:R57, 2004.

[Gunduz] C. Gunduz *et al.* The cell graphs of cancer. *Bioinformatics*, 20 Suppl. 1: 145-151, 2004.

[Hec95] D. Heckerman *et al.*. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, *Machine Learning*, 20: 197-243, 1995.

[IntAct04] H. Hermjakob *et al.* IntAct - an open source molecular interaction database. *Nucleic Acids Res.*, 32: D452-D455, 2004.

[Je01] T.-K. Jenssen *et al.* A Literature Network of Human Genes for High-throughput Analysis of Gene Expression. *Nature Gen.*, 28: 21-28, 2001.

[Jeong] H. Jeong *et al.* The large-scale organization of metabolic networks. *Nature*, 407: 651-654, 2000.

[Kan97] M. Kanehisa. A database for post-genome analysis. *Trends Genet.*, 13: 375-376, 1997.

[Kan00] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28: 27-30, 2000.

[Karp00] P.D. Karp *et al.* The ecocyc and metacyc databases. *Nucleic Acids Res.*, 28: 56-59, 2000.

[Lee02] T.O. Lee *et al.* Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science*, 298: 799-804, 2002.

[Mil67] S. Milgram. The small-world problem. *Psychology Today*, 1: 60-67, 1967.

[OraNDM] http://www.oracle.com/technology/products/spatial/pdf/10g_network_model_twp.pdf

[Pearl88] J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, (Morgan Kaufmann Publishers), 1988.

[Ra02] S. Ramaswamy and T.R. Golub. DNA Microarrays in Clinical Oncology, *J. Clin. Oncology*, 20: 1932-1941, 2002.

[Reiss] D.J. Reiss and B. Schwikowski. Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics*, 20 Suppl. 1: 274-282, 2004.

[Rung] J. Rung *et al.* Building and Analyzing Genome-wide Disruption Networks. *Bioinformatics,* 18 Supp.2: 202-210, 2002.

[Sch03] T. Schlitt *et al.* From Gene Networks to Gene Function. *Genome Research*, 13: 2568-2576, 2003.

[Sear03] D. Searls. Data integration - connecting the dots. *Nature Biotech.*, 21: 844-845, 2003.

[Sz99] Z. Szallasi. Genetic Network Analysis in light of massively parallel biological data acquisition. *Pacific Symposium on Biocomputing*, 4: 5-16, 1999.

[Uetz] P. Uetz *et al.* A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces cerevisiae. *Nature*, 406: 623-627, 2000.

[DIP00] I. Xenarios *et al.* DIP: The Database of Interacting Proteins. *Nucleic Acids Res.*, 28: 289-91, 2000.

[MINT02] A. Zanzoni *et al.* MINT: a Molecular INTeraction database. *FEBS Letters*, 513: 135-140, 2002.