Towards an Integrated Software System for Biological Pathways

Z. M. Ozsoyoglu[1,2], J. Nadeau[1,3], G. Ozsoyoglu[1,2], M. Tasan[1,2]

[1]Center for Computational Genomics
[2]Department of Electrical Engineering and Computer Science
[3]Department of Genetics

Case Western Reserve University, Cleveland, Ohio
(meral, joseph.nadeau, tekin, murat.tasan)@case.edu

Abstract

In this paper, we describe the Pathways Database System, being developed at Case Western Reserve University. Pathways Database System is an integrated software system for storing, managing, analyzing, and querying biological pathways at different levels of genetic, molecular, biochemical and organismal detail. We summarize the data model and the architecture of the Pathways Database System, and then briefly discuss two recently developed Pathways Database System tools, namely, the Pathway Explorer (a tree-structured query interface), and the Pathway Editor (to create and update pathways). We present a list of database research problems associated with pathways, and conclude with a summary of future work on the Pathways Database System.

1 Introduction

Biological pathways are used for presentation, modelling genomic data, studying genome sequences, and in functional studies of attributing functions to genes. Pathways are the sequential and cumulative action of genetically distinct, but functionally related molecules. Each reaction in a pathway is a biochemical step from specific substrates (input molecules) to products (output molecules) that are chemically modified substrates. Each step may also use combinations of molecules as cofactors, activators, inhibitors, and regulators, and usually involves at least one genetically unique gene product that catalyzes the reaction step.

The three general classes of biological pathways are metabolic and biochemical pathways, transcription and protein synthesis pathways, and signaling pathways. Metabolic pathways are the most known pathways. They involve the chemical reactions that provide basic biological functions such as energy metabolism, fatty acid synthesis, glycolysis, citric acid cycle, and many others. Transcription and protein synthesis pathways refer to synthesis of proteins from DNA. Signaling pathways are responsible for coordinating metabolic processes and other molecular and cellular activities, and transcription and protein synthesis processes.

Pathways, in general, illustrate the functional relations between molecules. Functional annotations include, for example, the identity of the substrate(s), product(s), cofactors, activators, inhibitors, enzymes or

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

other processing molecules, RNA and protein expression patterns, reaction kinetics, and associated phenotypic variation and diseases. Ultimately, many other kinds of information (or knowledge) can be incorporated. Such information forms a rich research resource that integrates genomic and biological information which can be managed, analyzed, queried and displayed in dynamic ways at various levels of biological and genetic detail to provide insight into diverse biological processes in health and disease.

Pathways Database System[K+a03] is an integrated software system for storing, managing, analyzing, and querying biological pathways at different levels of genetic, molecular, biochemical and organismal detail. At the computational level, the system allows users to visualize pathways in multiple abstraction levels, and to pose predetermined as well as ad hoc queries using a graphical user interface. Pathways are represented as graphs, and implemented on top of a relational database. Pathways Database System has multiple levels, with multiple tools at each level. In this paper, we summarize Pathways Database System's data model and architecture in general. We then present in more detail Pathway Explorer (a Tree-structured Query Interface. We conclude with what we view as the database research problems about pathways systems, and our future work.

2 Data Model

The design of the underlying database in the Pathways Database System is the primary influence on how data is to be acquired, presented, edited, and entered. While different tools can provide slightly different approaches, the logical steps for handling the data must be the same, and ultimately this is reflected in an overall similarity in all tools that access and work with the database. In what follows, we describe, in very basic ways, the underlying biological terminology as well as the corresponding entities and relationships captured in our data model. For more details on the data model, see [T04].

Molecular Entities. All biochemical pathways can be simply described as molecular entity interactions throughout the organism. Basically, molecules that are floating around in (and out of) the cell sometimes get together, resulting often in changes to the molecules. These changed molecules then go off and interact with other molecules, continuing the procedure, and thus giving birth to cascading pathways. So, at the very foundation of the notion of a pathway is the single molecule, which is referred to here as a molecular entity. Lumping all molecular entities together though into a single group is not a good modeling practice as different classes of entities (e.g., RNAs, amino acids) possess unique attributes. Therefore, the molecular entities are broken up into different classes, e.g., genes, RNAs, proteins, amino acids, and basic molecules.

Processes. The process refers to an interaction between molecular entities. A process is also sometimes seen as a single step (or reaction) in a metabolic pathway. Most of the chemical reactions that are included in metabolic pathways are catalyzed by some particular enzyme. There are also cases where no catalyzing enzyme (or catalyzing RNA) is present, and the reactants engage in a process entirely on their own. In either sort of process, there are a series of roles that different molecular entities can play. For each process in the Pathways Database, all of the known reactants are stored (with associated quantity attributes, to allow for correct stoichiometry), as well as some other process-specific attributes (such as whether or not the process is reversible, cellular location, etc.).

Each molecular entity involved in a process falls into some category or role-players for that process. The major categories of roles are those entities that are consumed (or destroyed), those that are created (or produced), and those that are not modified. Catalyzing enzyme allows molecules that would otherwise not interact to be brought into contact with each other in such a way that a reaction occurs. Once the process is complete, the catalyzing enzyme releases any reactants and moves on, completely unmodified (hence the term catalyzing enzyme). The next two roles are the substrates and the products, which are the primary reactants in the process. If a molecular entity is described as a substrate in a process, it will be consumed during the course of the reaction, and will not be present (at least in the same quantity) after the process has completed. Similarly, a product is not present at the start of the process, but will be at the end. It is through the substrates and the products that most pathways are defined. Cofactors are secondary substrates (cofactor-in) and secondary products (cofactor-out). Cofactors are also consumed or produced during the course of a process. There are also other molecular entities that can be involved in the reaction that are not modified besides the catalyzing protein. They are often called regulators when it is not entirely clear how they affect the process. If the reaction is dependent on a regulator, or if the regulator speeds up the rate of the reaction, they are usually called activators. On the other hand, inhibitors are regulators that play a negative role in the completion (or initiation) of the process.

Some processes are reversible, and are thus governed by concentration equilibria usually. Since we must store the roles of the process in at least one manner, for those reactions that are reversible it is clear that the choice is somewhat arbitrary. When a process is reversed, the entire set of substrates can become the set of products, and vice versa. The same dichotomy exists for cofactors (in becomes out, out becomes in), as they are functionally often the same as products and substrates. For regulators, it is not always clear if an activator will become an inhibitor and vice versa, and indeed since it is not fully understood biochemically, it can only be modeled if the understanding of its behavior is available.

Next we describe the data modeling of two pathways classes, namely, metabolic and signaling pathways.

Metabolic Pathways. The metabolic pathway is where we can see the processes in action; a pathway is simply a collection of processes. Typically, interrelated processes to describe a particular aspect of life function are grouped into a pathway, and named accordingly, e.g., fatty acid synthesis pathway.

Logically, a metabolic pathway becomes a graph of processes, with edges between them if there exists a substrate/product relationship between the processes (i.e. a product of one process is used as a substrate of another). A disconnected graph can be a flag for a possible error, e.g. maybe the curator should consider making each component a separate pathway itself. From the database perspective, a pathway is stored very easily with a few attributes (name, type, references, etc.) and a collection of the processes it consists of.

Signaling Pathways. In all multicellular organisms, an elaborate intercellular communication network coordinates the growth, differentiation, and metabolism of cells in diverse tissues and organs. Cells communicate in terms of extra *cellular signaling molecules*, which are synthesized and released by *signaling cells* and produce a specific response only in *target cells* that have *receptors* or *transporters* for the signaling molecules. A large variety of molecules, including proteins, peptides, etc. are used in cell-to-cell communication, called *signal transduction* (or *signaling*) *pathway*, that is, the process of converting extra cellular signals into cellular responses. Signaling pathways regulate many different aspects of cellular metabolism, function, and development. Specific signals promote changes in gene expression, cell morphology, and cell movements.

Communication by extra cellular signaling usually involves five parts (captured in our data model): synthesis (release of a signaling molecule by signaling cell), transport (of signal to target cell), signal detection (by a specific receptor protein), a change in cellular metabolism/function, and removal of signal to terminate further cellular response. Based on the operation distance of signaling molecules, there are three types of signaling pathways: *endocrine signaling pathways* (hormones carried over blood into target cells), *paracrine signaling pathways* (signaling molecules having an effect on cells in close proximity), and *autocrine signaling pathways* (cells responding to molecules released by themselves).

Signaling pathways may be viewed as consisting of multiple *signaling steps* where one or more *signaling molecules* signal one or more *signaled molecules*. It is possible that a signaling step may involve the substrate and the product, or enzyme of a metabolic pathway process. Molecule locations are captured and involve tissue (loosely defined as a collection of cells performing a specified function) and organelle information (i.e., the membrane, cytoplasm, nucleus membrane, or nucleus). Locations of signaling/signaled molecules may be different. Thus a signaling step can be viewed as a (usually single) directed edge from signaling molecules to signaled molecules, enhanced by location information, where location consists of organelle and tissue information.

A large number of information about signaling molecules can be captured in a data model. While our data model continually evolves, presently we capture location, role (e.g., enzyme, hormone, receptor, transcription factor, etc.), molecule families (super/subfamily), and component/subcomponent information, among others.

In comparison with metabolic pathways, there is much less agreement as to what signaling steps together constitute a specific signaling pathway with a certain biological function. Our present solution to this problem is to name and represent all proposed variations in the database.

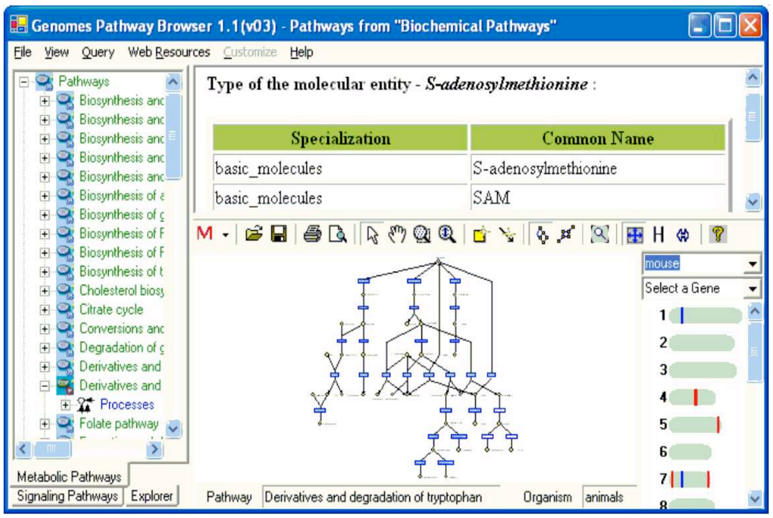

Figure 1: Snapshot of the Pathway System main GUI. The leftmost panel is the Pathway Browser (metabolic pathway or signaling pathway). The upper-right pane displays query results. The lower-right pane is the Pathway Viewer within the main GUI (it can also be a separate window).

3 Architecture

Pathway Database System is developed using the .Net platform (a Java version is also being developed [R03, PSRJ03]). The system architecture has three layers. At the top layer, we have a set of GUI tools that are continually improving and evolving, and increasing in number. At the middle layer, we have two service subsystems, Querying Services and Data Extraction Services. At the lowest level, the Pathways Database, currently a relational database, resides. Next we discuss only the top-level tools. For a discussion of the other tools, please see [K+a03] and the technical reports of Pathway System Release #2 [PSR04].

The user interface tools at the top level are all presented through a single interface, with the only exception that users can visualize multiple pathways in multiple separate windows. However, all the tools are developed independently as plug-ins to the main interface, and communicate with each other at the user command level asynchronously. At the present time, we have developed four tools, which are described next.

Pathway Browser provides a tree-structured view of the metabolic pathways, and signaling pathways available in the database. Metabolic pathways, processes and molecular entities can be browsed in different orders (Figure 1, leftmost pane), and a set of useful queries [X04] can be specified about pathways, processes and molecular entities, from predefined simple queries to parameterized path selection and neighborhood queries. Similarly, for signaling pathways, signaling steps, and signaling and signaled molecules and their components can be browsed, and a set of useful queries are presently being developed. For more details, see [K03].

Pathway Viewer visualizes a single pathway (Figure 1, lower middle pane) or multiple pathways in an organism-specific or organism-independent manner, provides a large number of collapsing-zooming, hiding-expanding, etc. functionalities (Figure 1, top of middle pane), allows XML versions of pathways in the database to be exported, saved, and loaded back to the viewer and re-visualized. It also provides a seamless transition from metabolic pathways to signaling pathways, and from signaling steps that belong to metabolic pathway processes into the relevant pathways. Pathway viewer for signaling pathways is currently under construction. For more details, see [M04].

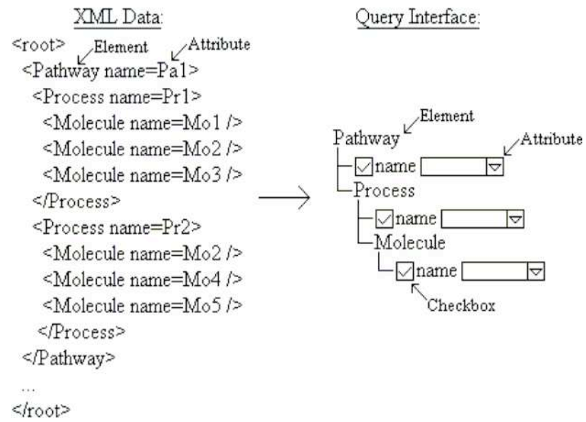

Figure 2: From Schema Elements to the Tree-Structured Interface

Pathway Explorer provides a generic user interface for posing a large class of queries about pathways using tree-structured views; see section 4 below.

Pathway Editor allows view-based database updates; that is, users can insert/update/edit pathways via the high-level user views of the pathway viewer. The pathway editor, while enforcing the integrity of the existing database (sometimes interactively with the user), maps them as updates to the underlying relational database. See section 5 below.

4 Pathways Explorer: A Tree Structured Query Interface

As it is clear from the discussion on the data model, entities involved in the pathways database can be better viewed by generalization/specialization hierarchies, e.g., classes of pathways, super pathways, functional classifications in arbitrary granularities, etc.. In addition, pathway components, i.e., processes in a pathway, and molecular entities involved in each process can naturally be viewed in the form of a tree-structured (hierarchical) view. Motivated by this observation, Pathways Explorer is a query interface which allows users to query the database using tree-structured views. As an example, consider the tree-structured (hierarchical) schema Pathways-Process-Molecular entity, where a pathway consists of several processes, and each process is associated with several molecular entities. This may also be visualized as a logical schema of a nested relation (Pathways (Process (Molecular_Entity)*)*), or an XML view. Thus, Pathways Explorer can also be viewed as a query interface for querying tree-structured views for a relational database, as shown in Figure 2.

Given the schema of an XML view, or a nested relation view of the underlying relational database, query interface (template) structure follows that of the view schema. As shown in Figure 2, the schema declares that pathway elements contain process elements that contain molecule elements, so the tree in the user-interface would contain only these nodes, along with the node attributes as children. Each attribute's possible values may then be seen individually by opening its drop-down box. Choosing a value for an attribute effectively creates a query constraint declaring that only nodes with the chosen attribute value are of interest. As the user chooses more constraints, the number of choices in each drop-down box decreases because the source XML view is being dynamically trimmed in the background as the constraints narrow its content (in reality, there is no actual trimming as this would be inefficient; but this paradigm makes the behavior easy to grasp).

In general, the tree in the interface need not have a one-to-one correspondence to that of the schema. The user may construct the tree (and effectively the output as well) using any of the elements in any arrangement. In the above example, the structure of the interface may instead be a molecule associated with a process which is in turn associated with a pathway (i.e., molecule is the root; pathway is the leaf). It could also be a process in a

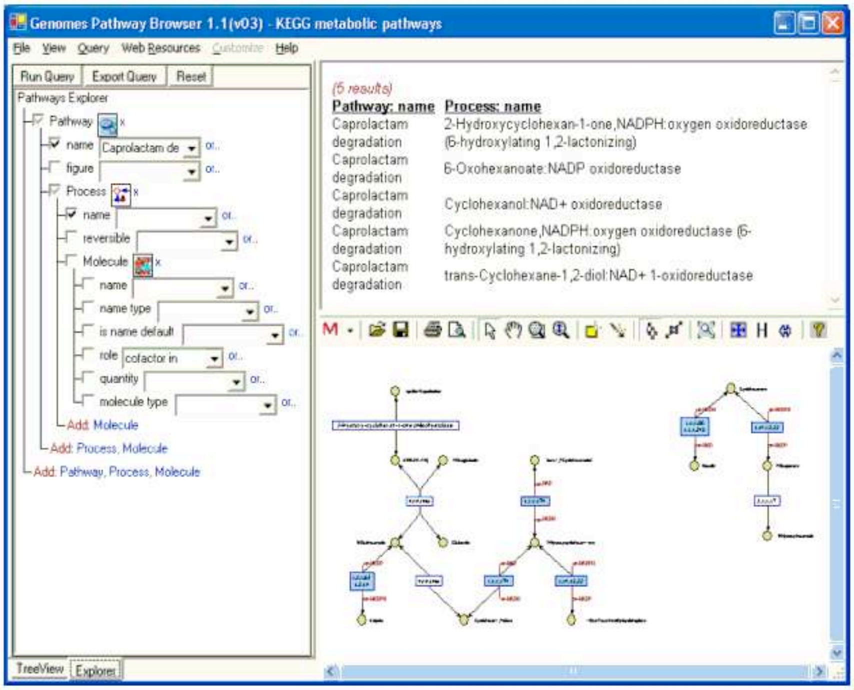

Figure 3: From Schema Elements to the Tree-Structured Interface

pathway and containing a molecule (i.e., process is the root; and pathway and molecule are its two children). As long as there is an ancestor-descendent relationship between two elements in the schema, these elements may be in a parent-child or child-parent relationship in the tree.

The Pathway Explorer interface has been implemented [NO04]. Query capabilities include selection, projection, equijoins, and self-joins. Figure 3 illustrates a query using Pathway Explorer, where the left pane shows the tree-structured query template, top pane displays the query results, and the bottom right pane shows graphs of the user-requested biological pathways or processes. For more details, see [NO04].

5 Pathway Editor Tool

Pathway Editor allows users to create or update pathways in the pathways database. The basic issues with and the properties of the Pathway Editor are listed below.

- *Transaction-oriented operations*. Operations such as creating or editing a pathway or a process are multi-level, multi-step operations with each step at each level involving database updates. It is possible that, at an intermediate step, the overall operation needs to be cancelled and all previous database updates are rolled back. Thus, all pathway editor updates are implemented as atomic transactions.

- *Interactive Communication*. The pathways database contains a myriad of biologically-inspired integrity constraints, such as the cofactor-in ATP in a process is accompanied by the cofactor-out ADP, or pathways are usually connected, or the requested update for a certain organism is not permissible, etc. As users perform a step in an update, the associated integrity constraint can be satisfied in multiple ways. This necessitates an interactive communication between the Pathway Editor and the users, with the Pathway Editor deciding on which route to take and the users ultimately controlling the update route and the enforcement of the integrity constraints.

- *Extensibility and Flexibility*. The Pathway System is continually evolving with new pathway types and functionality, and new extensions to the underlying pathway data model. To allow for these extensions, the Pathway Editor is designed to be extensible and flexible.

- *Familiar User Interface*. The Pathway Editor has a user interface similar to the other tools in the system. consistent user interface reduces the study curve for users.

- *Minimum-effort pathway editing*. In most biological databases, data gathering and data editing are among the most difficult tasks, time-consuming and error-prone. Pathway Editor is continually improved to ease the user workload by providing pop-up and drop-down menus of the existing objects in the database, prompting users with reminder-type questions and steps, and implementing a step-by-step enforcement of data integrity-related issues.

- *Complexity Hiding*. As much as possible, Pathway Editor is designed to hide from users the complexity of the operations it performs associated with pathway updates.

6 Other Pathway Systems

For metabolic pathways, there are a large number of systems with varying capabilities, which include PathDB system for pathways (http://www.ncgr.org/pathdb/architecture.html), Kyoto Encyclopedia of Genes and Genomes (www.genome.ad.jp/kegg),Metabolic Pathways of Biochemistry (www.gwu.edu/ mpb), Encyclopedia of E. Coli Genes and metabolism (http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html), Biocarta (www.biocarta.com) with statically available metabolic and signaling pathways, Enzymes and Metabolic Pathways (emp.mcs.anl.gov), AceDB (www.acedb.org), Soybean metabolic pathways (http://cgsc.biology.yale.edu/metab.html), Nicholson minimaps (www.tcd.ie/Biochemistry/IUBMB-Nicholson), ExPASy Biochemical pathways (www.expasy.ch/cgi-bin/search-biochem-index), and others. These systems range from simply displaying picture images of metabolic pathways to database-backed advanced systems. Among the systems listed above, in terms of capabilities and tools, PathDB is closest to our Pathways Database System with the QueryTool (similar to our Pathway Browser), the Pathway Viewer, and the Discovery Tool (that provides data analysis approaches). PathDB however only deals with metabolic pathways.

On the signaling pathways side, the existing systems include CSNDB (Cell Signaling Networks database at http://geo.nihs.go.jp/csndb), TRANSPATH (http://transpath.gbf.de), Roche Apoptosis Pathway (www.grt.kyushu-u.ac.jp/spad), Database of Quantitative Cellular Signaling (DQCS, at http://doqcs.ncbs.res.in), Signaling Pathway database (SPAD, at www.grt.kyushu-u.ac.jp/spad/index.html). All of these systems model a large number of biological and pathways information; however, they are specifically for signaling pathways.

7 Database Research Problems

We have found early on that building biological data content and data curation (which is better done by a life scientist who is familiar with computer science), data provenance and data integrity enforcement are among the most difficult and important issues, if not the most important one, with pathway systems. Pathways data in biological resources and on the web are incomplete, and curation requires expert knowledge and data extraction. Moreover, as with any biological information, all pathways-related rules have various exceptions and variations-making the task of data integrity enforcement a challenging one.

An integrated pathways system potentially incorporates a very large number of different types of biological information, e.g., about genes and proteins, such information include synonyms, gene and protein sequence information, genetic map location, polymorphisms, transcription factors, expression patterns (RNA, proteins), biochemistry, consequences of pharmacological perturbations, phenotypes, evolution, etc. Building all of these information in a pathway system is impossibly complex and unnecessary; a better alternative is to extract this information offline or online from the web-accessible resources. Ultimately, for difficult to extract data, one may envision the use of data extraction systems specialized for the tasks at hand. Finally, there is always the seamless data integration issue of the data available in the local database with the data in remote resources, on different components and in different levels.

8 Future Work

Our Pathways Database System is a typical computer science application for structured databases, involving querying and updating using views, XML views, data integration, query optimization. And, there are several graph traversals, path queries, for which special query evaluation techniques have potential to improve efficiency and need to be investigated.

An important goal is the integration of the three classes of pathways so that the information on different classes of pathways can be queried and visualized over the web in an integrated manner. Our long-range goal is to use the metabolic/genomic Pathways Database System for a number of (high-level) problems, including the relation between genetic mutations and functional consequences (such as metabolic dysfunction), and predictions of pathway composition and function.

Acknowledgments. Since its conception four years ago, a number of people contributed and helped develop Pathways Database System, including F. Akgul, A. Cakmak, M. Kirac, T. Kitami, L. Krishnamurthy, S. Newman, Y. Mei, C. Parker, N.Ratprasartporn, G. Schaeffer, W. Xu, B. Zhang. We also acknowledge Yoh-Han Pao for his continued support. This research has been funded by a gift from the Charles Wang Foundation and the NSF research grant DBI-0218061.

References

[K+a03] Krishnamurty, L., et al, *Pathways Database System: an Integrated System for Biological Pathways* Bioinformatics, Vol. 19, No. 8, 2003.

[K03] Krishnamurty, L., *Pathways Browser* M.S.Thesis, CWRU, 2003.

[M04] Mei, Y., *Pathway Viewer Tool* Version 1.0, Tech. Rep., CWRU, July 2004.

[NO04] Newman, S., Ozsoyoglu, M., *A Tree-Structured Query Interface for Querying Semi-Structured Data*, SSDBM Conf., June 2004.

[PSR04] Pathway System Release #2, available at http://nashua.cwru.edu/Pathways, 2004.

[PSRJ03] Java-based Pathway System Release #1, available at http://nashua.cwru.edu/Pathways, 2004.

[R03] Ratprasartporn, N., *J-Pathway Browser and Viewer Tool* Version 1.0, CWRU, July 2003.

[T04] Tasan, M., *Pathways Database User Guide*, Version 2.0, Tech. Rep., CWRU, July 2004

[X04] Xu, W., *Pathway Querying Services for Genome Pathways Database*, CWRU, July 2004.