

5 May 2005

MetNet DB xml files and formats description

introduction

The MetNet DB file consists of a group of two or more xml-format text files zipped into a single archive. A single compressed archive is used to save disk space and make file interaction easier for the user by reducing the number of files that must be opened.

All MetNet DB zipped archives must contain one contents.xml file and one topology.xml file. The archive may also contain an index.xml file, and/or pathways.xml file. These files are optional but may only occur once each if present.

All MetNet DB xml files should be saved as plain ascii text with the name of the file type as the name of the file (for example, all topology files should be called topology.xml).

topology.xml file format

The topology file lists the nodes and edges of the graph. All of the other files contain additional details about the biological significance of each node; the topology file contains all of the information required to draw the graph.

The topology file is made up of node and edge listings, the description of which follows. The order in which the nodes and edges are listed does not matter.

- each node entry consists of:
 - attribute `id` - string (unique ID) required
 - attribute `molID` - string (molecule ID) required
 - attribute `nodeName` - string (default node name) required
 - element `nodeType` - string (type of molecule represented by the node) 1 required
 - element `location` - string (cellular location of the node) 1 required
- each edge entry consists of:
 - attribute `id` - string (unique ID) required
 - attribute `tail` - string (unique ID of node edge is from) required
 - attribute `head` - string (unique ID of node edge points to) required
 - attribute `directed` - boolean (if edge is directed) required
 - attribute `strength` - decimal (reaction strength represented) optional
 - element `edgeType` - string (type of reaction represented by the edge) zero or one
 - element `certainty` - string (certainty that the edge is correct) zero or one

Every node has two ID codes, a *unique ID* and a *molecule ID*. The unique ID is the internal name for the node, and can contain both letters and numbers. There must not be any repeated unique IDs in the entire file. The molecule ID identifies the molecule that the node represents. Many nodes may share the same molecule ID. For example, the compound glucose is present in many reactions and many places in a cell. As a result, there are multiple glucose nodes, each of which will have a different unique ID but the same molecule ID. All reaction nodes have the same molecule ID (generally "reaction"). The molecule ID can be made of letters or numbers or both.

contents.xml file format

contents.xml is the shortest and simplest xml file of the data set. The contents file contains background information about the data included in the zipped archive and boolean values to indicate if the optional files (pathways, index, or extended) are present in the archive.

contents.xml consists of a single fileSet entry made up of:

- attribute `topologyPresent` – boolean (if file is present) required
- attribute `indexPresent` – boolean (if file is present) required
- attribute `pathwaysPresent` – boolean (if file is present) required
- attribute `extendedPresent` – boolean (if file is present) required
- element `createdBy` – string (person(s) that created the archive) zero or more
- element `dateCreated` – date (date archive created) zero or one
- element `projectName` – string (name of project that created the archive) zero or one
- element `institution` – string (institutions(s) that created the archive) zero or more
- element `description` – string (project/data description; may be multiple paragraphs) zero or one
- element `organism` – string (organism(s) the data comes from) zero or more
- element `dataSource` – string (source(s) of the data [MetNet database, etc.]) zero or more

index.xml file format

index.xml contains information on alternative names for the molecules represented by the nodes in the graph and summarizes some of the information from topology.xml. The molecule ID from every node in the topology file (except reaction nodes) will have a single entry in index.xml.

Nested within each molecule ID's entry will be a listing of synonyms and abbreviations for the molecule (other than the default `nodeName`) and a list of all the locations in which the molecule occurs in the graph.

index.xml consists of `molecule` listings, one for each unique molecule ID (except reaction nodes) in the associated topology.xml. Each `molecule` entry is composed of:

- attribute `molID` – string (molecule ID of the molecule) required
- attribute `nodeName` – string (default node name, same as in topology.xml) required
- element `abbrev` - string (abbreviation for the molecule) zero or more
- element `synonym` - string (synonyms of the molecule's name) zero or more
- element `location` - string (all locations where the molecule is found) zero or more

pathways.xml file format

pathways.xml contains information on the named biologically-significant pathways present in the data set. Each pathway is made of nodes and edges from topology.xml, all of which are listed as `member` entries. Multiple names (`synonyms`) may be included for each pathway in this file; all other information about the pathway (references, etc.) is contained in extended.xml.

pathways.xml consists of `pathway` listings, each of which are composed of:

- attribute `id` - string (pathway unique ID) required
- attribute `pathwayName` – string (default name for the pathway) required
- element `synonym` - string (other name(s) for the pathway) zero or more
- element `member` - string (unique ID of a node or edge in the pathway) one or more