KEGG & KGML

Yo-Cheng Chang

張祐誠
XML and RDB

• In a relational database, tables are linked by a pair of a primary key and a foreign key.

• When data is exported as a flat text file from a database, the data fields may have been well described and organized within the database, but that structure gets left behind.

• Use XML to preserve information that identifies and structures data when it is exported from databases for data exchange.
XML in bioinformatics (sequence data)

• The BIOpolymer Markup Language (BIOML)
  – describe experimental information about proteins, genes, and other biopolymers.
• The Protein Sequence Database Markup Language (PSDML)
  – an open-standard markup language used to store protein information in the Protein Information Resource (PIR) database
• The Bioinformatic Sequence Markup Language (BSML)
  – an open-standard protocol for the encoding and display of graphic genomic displays of DNA, RNA, and protein sequence information
• The Genome Annotation Markup Elements (GAME)
  – a markup language used in molecular biology for annotation of a biosequence.
• The Multiple Sequence Alignments Markup Language (MSAML)
  – developed to make manipulation and extraction of multiple sequence alignment information easier by logically defining the parts of an alignment for use in an XML-based application
XML in bioinformatics (microarray data)

• The Gene Expression Markup Language (GEML)
  – an open-standard markup language for DNA microarray and gene expression data, chip patterns, and chip scan profiles.

• The Microarray Markup Language (MAML)
  – a markup language for describing and communicating information about DNA array experiments. Handled by MGED.

• MAGE-ML
  – On 2002-02-08, MAML was replaced by MAGE-ML
XML in bioinformatics (modeling)

- Molecular Dynamics Markup Language (MoDL)
  - a markup language for simulation data from molecular dynamics.
- The Systems Biology Markup Language (SBML)
  - representing and modeling the information in systems biology simulation software, and enabling simulation software to communicate and exchange models across multiple software packages
- KEGG Markup Language (KGML)
  - an XML presentation of the KEGG pathway database
- CellML
  - To store and exchange computer-based biological models of cells. It allows scientists to share models even if they use different model-building software.
Overview of KEGG

- Genomes
  - Information on relations between molecules
- Genes
- Expression
- Pathway
- Orthologs
  - Sequence similarity
- Chemicals and their reactions
KEGG databases

• GENES database
  – genes and proteins
• PATHWAY database
  – enzyme-enzyme relations
  – protein-protein interactions
  – gene expression relations
• LIGAND database
  – cellular processes related data
• EXPRESSION database
  – microarray gene expression profiles
• BRITE database
  – Experiment results of yeast two-hybrid systems
• SSDB database
  – protein coding genes in complete sequenced genomes
  – ortholog groups and group similarities
Relationship of main KEGG databases

- PATHWAY
  - ENZYME
  - REACTION
  - GENES
  - GENOMES
  - COMPOUND
  - GLYCAN
  - LIGAND
The KEGG pathway maps are graphical image maps representing networks of interacting molecules responsible for specific cellular functions.

There are two types of KEGG pathways:

- reference pathways: which are manually drawn
- organism-specific pathways: which are computationally generated based on reference pathways.

The KGML files contain computerized information about graphical objects and their relations in the KEGG pathways.
Reference pathway and organism-specific pathway
The schema of KGML

- The pathway element is a root element, and one pathway element is specified for one pathway map in KGML.

- The entry, relation, and reaction elements specify the graph information, and additional elements are used to specify more detailed information about nodes and edges of the graph.
Main concepts of KGML

• In KGML, the pathway element specifies one graph object with the entry elements as its nodes and the relation and reaction elements as its edges.

• The relation and reaction elements indicate the connection patterns of rectangles (gene products) and circles (chemical compounds), respectively, in the KEGG pathways.
Relation and reaction

• relation

- gene product
  - enzymes or non-enzyme proteins

  relation
  - catalyze
  - regulation
  - modification
  - binding

  ...

- gene product
  - enzymes or non-enzyme proteins

• reaction

- compound of substrate

  - compound of product
The KEGG data representation

Data Representation in KEGG

- Entity: a molecule or a gene
- Binary relation: a relation between two entities
- Network: a graph formed from a set of related entities
- Genome: physical closeness of genes
- Neighbor: similarity of molecules or genes, e.g., sequence similarity
- Hierarchy: hierarchical classification of molecules or genes
- Pathway: biochemical pathway or genetic pathway
KGML data types (from DTD)

- **number.type**  
  Positive integer

- **string.type**  
  Character string

- **id.type**  
  Identification number applied to each entry (positive integer)

- **idref.type**  
  ID numbers reference (positive integer)

- **url.type**  
  String with URL form

- **keggid.type**  
  Character string of KEGGID form [db]:[acc]

- **maporg.type**  
  Alphabet of two or three characters string : organism prefix or "map"

- **mapnumber.type**  
  Five-digit number : map number

- **entry-type.type**  
  It is in any of the following. (enzyme|gene|genes|ortholog|compound|map)

- **reaction-type.type**  
  It is in any of the following. (reversible|irreversible)

- **relation-type.type**  
  It is in any of the following. (ECrel|PPrel|GErel)

- **graphics-type.type**  
  It is in any of the following. (rectangle|circle|roundrectangle)

- **graphics-color.type**  
  The form is a numerical RGB specification. ex) #FFFFFF (this is white.)

- **subtype.type (new in v0.4 !)**
1. The pathway element

- The **pathway element** specifies graph information stored in the KEGG pathway map.
  - attributes:

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>keggid.type</td>
<td>the KEGGID of this pathway map</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>org</td>
<td>maporg.type</td>
<td>map/ot/[org prefix]/mim</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>number</td>
<td>mapnumber.type</td>
<td>the map number of this pathway map</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>title</td>
<td>string.type</td>
<td>the title of this pathway map</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>image</td>
<td>url.type</td>
<td>the resource location of the image file of this pathway map</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>link</td>
<td>url.type</td>
<td>the resource location of the information about this pathway map</td>
<td>IMPLIED</td>
</tr>
</tbody>
</table>
The name attribute (pathway)

- The name attribute contains the KEGG identifier of this pathway map.

- Here ***** represents the pathway map number (five digits) and [org prefix] is a three-letter species code in KEGG.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>path:map*****</td>
<td>the KEGGID of this pathway map</td>
</tr>
<tr>
<td>path:ot*****</td>
<td>ex:</td>
</tr>
<tr>
<td>path:[org prefix]*****</td>
<td>name=&quot;path:map00010&quot;</td>
</tr>
<tr>
<td></td>
<td>name=&quot;path:ot00010&quot;</td>
</tr>
<tr>
<td></td>
<td>name=&quot;path:hsa00010&quot;</td>
</tr>
</tbody>
</table>
The org attribute (pathway)

- The org attribute specifies the classification of this pathway map.

- The distinction of reference pathways and pathways for various organisms is made according to the attribute value.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td>the reference metabolic pathway map represented by EC numbers</td>
</tr>
<tr>
<td>ot</td>
<td>the reference pathway map represented by KO identifiers</td>
</tr>
<tr>
<td>[org prefix]</td>
<td>the organism-specific pathway map for &quot;org&quot;</td>
</tr>
<tr>
<td>mim</td>
<td>the human pathway map with disease genes in OMIM</td>
</tr>
</tbody>
</table>
Difference between map and ot

- map (theoretical reference pathway reconstructed from the LIGAND database)

  <entry id="6" name="ec:2.7.1.24" type="enzyme" reaction="rn:R00130" link="http://www.genome.ad.jp/dbget-bin/www_bget?enzyme+2.7.1.24">

- ot (all orthologs found in all sequenced organisms)

  <entry id="6" name="ko:E2.7.1.24" type="ortholog" reaction="rn:R00130" link="http://www.genome.ad.jp/dbget-bin/www_bget?ko+E2.7.1.24">
Other attributes of the pathway element

• The number attribute (mapnumber.type)
  – specifies the five-digit pathway map number
  – five-digit integer, ex: number="00030"

• The title attribute
  – specifies the title of this pathway map.
  – string, ex: title="Pentose phosphate pathway"

• The image attribute
  – specifies the resource location of the image file for this pathway map in the KEGG Web service.
  – URL string, ex:
    image="http://www.genome.ad.jp/kegg/pathway/map/map00010.gif"

• The link attribute
  – specifies the resource location of information about this pathway map in the KEGG Web service.
  – URL string, ex:
    link="http://www.genome.ad.jp/dbget-bin/show_pathway?map00030"
2. The entry element

- The entry element contains information about a node of the pathway.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>id.type</td>
<td>the ID of this entry in the pathway map</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>name</td>
<td>keggid.type</td>
<td>the KEGGID of this entry</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>type</td>
<td>entry_type.type</td>
<td>the type of this entry</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>link</td>
<td>url.type</td>
<td>the resource location of the information about this entry</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>reaction</td>
<td>keggid.type</td>
<td>the KEGGID of corresponding reaction</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>map</td>
<td>idref.type</td>
<td>the ID of the map entry (this attribute is specified when the node appears in the linked pathway map)</td>
<td>IMPLIED</td>
</tr>
</tbody>
</table>
The name attribute (entry)

- The name attribute contains the KEGG identifier of this entry, which is generally in the form of `db:accession` where db is the database name and accession is the accession number.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
</table>
| path:(accession) | the pathway map  
ex: name="path:map00040" |
| ko:(accession) | the KO (ortholog group)  
ex: name="ko:E3.1.4.11" |
| ec:(accession) | the enzyme  
ex: name="ec:1.1.3.5" |
| cpd:(accession) | the chemical compound  
ex: name="cpd:C01243" |
| [org prefix]: (accession) | the gene product of a given organism  
ex: name="eco:b1207" |
| group:(accession) | the complex of KOs  
ex: name="group:ORC" |
The type attribute (entry)

- The type attribute specifies the type of this entry.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ortholog</td>
<td>the node is a KO (ortholog group)</td>
</tr>
<tr>
<td>enzyme</td>
<td>the node is an enzyme</td>
</tr>
<tr>
<td>gene</td>
<td>the node is a gene product (mostly a protein)</td>
</tr>
<tr>
<td>genes</td>
<td>the node is a complex of gene products (mostly a protein complex)</td>
</tr>
<tr>
<td>compound</td>
<td>the node is a chemical compound (including a glycan)</td>
</tr>
<tr>
<td>map</td>
<td>the node is a linked pathway map</td>
</tr>
</tbody>
</table>
The reaction and map attributes (entry)

• The reaction attribute
  – specifies the KEGGID of the corresponding chemical reaction(s) in the KEGG LIGAND database
  – rn:Rxxxxx, ex: reaction="rn:R02749"

• map attribute
  – When the current pathway map is linked to another pathway map, it is represented by an entry element of "map" type. Some nodes in the linked pathway maps are stored as partners of binary relations in the KGML file, although they are not shown in the actual KEGG pathway map. The map attribute specifies the ID of the linked map entry.
  – ex: map="1"
2.1 The graphics element

- The graphics element is a sub-element of the entry element, specifying drawing information about the graphics object.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>string.type</td>
<td>the label of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>x</td>
<td>number.type</td>
<td>the X axis position of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>y</td>
<td>number.type</td>
<td>the Y axis position of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>type</td>
<td>graphics.type</td>
<td>the shape of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>width</td>
<td>number.type</td>
<td>the width of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>height</td>
<td>number.type</td>
<td>the height of this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>fgcolor</td>
<td>graphics-color.type</td>
<td>the foreground color used by this graphics object</td>
<td>IMPLIED</td>
</tr>
<tr>
<td>bgcolor</td>
<td>graphics-color.type</td>
<td>the background color used by this graphics object</td>
<td>IMPLIED</td>
</tr>
</tbody>
</table>
The type attribute

The type attribute specifies the shape of this object. The default value is "rectangle".

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>rectangle</td>
<td>the shape is a rectangle, which is used to represent a gene product and its complex (including an ortholog group).</td>
</tr>
<tr>
<td>circle</td>
<td>the shape is a circle, which is used to specify any other molecule such as a chemical compound and a glycan.</td>
</tr>
<tr>
<td>roundrectangle</td>
<td>the shape is a round rectangle, which is used to represent a linked pathway.</td>
</tr>
</tbody>
</table>
Other attributes of the graphics element

- The name attribute
  - contains the label that is associated with this graphics object.
  - When two or more name attributes are specified in the same entry element, the first one is taken as the attribute value.
  - When the type attribute value of the entry element is "gene", the gene name is specified for this attribute value.
  - ex: name="1.1.1.43", name="Methane metabolism"

- The x and y attributes
  - specifies the coordinate value of this graphics object in the KEGG pathway map
  - ex: x="190", y="51"

- The width attribute (default 45)
- The height attribute (default 17)

- The fgcolor attributes (default #000000)
- The bgcolor attributes (default #FFFFFF)
  - ex: fgcolor="#000000", bgcolor="#FFFFFF"
2.2 The component element

- The component element is a sub-element of the entry element, and is used when the entry element is a complex node.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>idref.type</td>
<td>the ID of the component which is part of the complex</td>
<td>REQUIRED</td>
</tr>
</tbody>
</table>

- The id attribute
  - specifies the identification number of this component.
  - The entry element of "genes" type is specified by a complete set of component elements.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive integer</td>
<td>the identification number of this component</td>
</tr>
</tbody>
</table>
3. The relation element

- The relation element specifies relationship between two proteins (gene products) or two KOs (ortholog groups), which is indicated by an arrow or a line connecting two rectangles in the KEGG pathways.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>entry1</td>
<td>idref.type</td>
<td>the first (from) entry that defines this relation</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>entry2</td>
<td>idref.type</td>
<td>the second (to) entry that defines this relation</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>type</td>
<td>relation-type.type</td>
<td>the type of this relation</td>
<td>REQUIRED</td>
</tr>
</tbody>
</table>
The type attribute specifies one of the three types of relations, so-called the generalized protein interactions in KEGG.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECrel</td>
<td>enzyme-enzyme relation, indicating two enzymes catalyzing successive reaction steps</td>
</tr>
<tr>
<td>PPrel</td>
<td>protein-protein interaction, such as binding and modification</td>
</tr>
<tr>
<td>GErel</td>
<td>gene expression interaction, indicating relation of transcription factor and target gene product</td>
</tr>
</tbody>
</table>
3.1 The subtype element

- The relation element has a sub-element named the subtype element.
  - When the name attribute value of the subtype element is a value with directionality like "activation", the direction of the interaction is from entry1 to entry2.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>subtype-name.type</td>
<td>Interaction/relation information</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>value</td>
<td>subtype-value.type</td>
<td>Interaction/relation property value</td>
<td>REQUIRED</td>
</tr>
</tbody>
</table>
name and value attributes (subtype)

- The name attribute specifies the subcategory and/or the additional information in each of the three types of the generalized protein interactions.

<table>
<thead>
<tr>
<th>name</th>
<th>value</th>
<th>EC rel</th>
<th>PP rel</th>
<th>GE rel</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>compound</td>
<td>Entry element id attribute value for compound.</td>
<td>*</td>
<td></td>
<td></td>
<td>shared in two successive reactions</td>
</tr>
<tr>
<td>activation</td>
<td>--&gt;</td>
<td>*</td>
<td></td>
<td></td>
<td>positive and negative effects which may be associated with molecular information below</td>
</tr>
<tr>
<td>inhibition</td>
<td>--</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expression</td>
<td>--&gt;</td>
<td>*</td>
<td></td>
<td></td>
<td>indirect effect without molecular details</td>
</tr>
<tr>
<td>repression</td>
<td>--</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>indirect effect</td>
<td>..</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>state change</td>
<td>...</td>
<td>*</td>
<td></td>
<td></td>
<td>association and dissociation</td>
</tr>
<tr>
<td>binding/association</td>
<td>---</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>dissociation</td>
<td>-+-</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>phosphorylation</td>
<td>+p</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>dephosphorylation</td>
<td>-p</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>glycosylation</td>
<td>+g</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>ubiquitination</td>
<td>+u</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
<tr>
<td>methylation</td>
<td>+m</td>
<td>*</td>
<td></td>
<td></td>
<td>molecular events</td>
</tr>
</tbody>
</table>
4. The reaction element

- The reaction element specifies chemical reaction between a substrate and a product indicated by an arrow connecting two circles in the KEGG pathways. The attributes of this element are as follows.

<table>
<thead>
<tr>
<th>attribute name</th>
<th>data type</th>
<th>explanation</th>
<th>REQUIRED/IMPLIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>keggid.type</td>
<td>the KEGGID of this reaction.</td>
<td>REQUIRED</td>
</tr>
<tr>
<td>type</td>
<td>reaction-type.type</td>
<td>the type of this reaction.</td>
<td>REQUIRED</td>
</tr>
</tbody>
</table>

- The type attribute
  - specifies the distinction of reversible and irreversible reactions, which are indicated by bi-directional and uni-directional arrows in the KEGG pathways.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>reversible</td>
<td>reversible reaction</td>
</tr>
<tr>
<td>irreversible</td>
<td>irreversible reaction</td>
</tr>
</tbody>
</table>
4.1 The substrate and product elements

- The reaction element has the **substrate element** and the **product element** as sub-elements.

- The substrate element specifies the substrate node of this reaction.
- The product element specifies the product node of this reaction.

- The name attribute is the only attribute of the substrate element and the product element.
  - The name attribute contains the KEGG identifier of the COMPOUND database or the GLYCAN database.

<table>
<thead>
<tr>
<th>attribute value</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpd:(accession)</td>
<td>ex: cpd:C05378</td>
</tr>
<tr>
<td>gl:(accession)</td>
<td>gl:G00037</td>
</tr>
</tbody>
</table>
The DTD of KGML v0.4 (1)

<!-- DTD for KEGG Markup Language v0.4 -->

<!-- Positive number type --> <!--ENTITY % number.type "NMTOKEN">
<!ENTITY % string.type "CDATA">

<!-- ID type --> <!--ENTITY % id.type "%number.type;">}
<!-- IDREF type --> <!--ENTITY % idref.type "%number.type;">}

<!-- URL type --> <!--ENTITY % url.type "%string.type;">}

<!-- KEGG ID type KEGG ID form : "[db]:[accession]" --> <!--ENTITY % keggid.type "%string.type;">}

<!-- MAPORG type organism prefix or "map" or "enzyme" --> <!--ENTITY % maporg.type "%string.type;">}

<!-- MAPNUMBER type map number : The five-digit number --> <!--ENTITY % mapnumber.type "%string.type;">}

<!-- Type of Entry --> <!--ENTITY % entry-type.type "(enzyme|gene|genes|ortholog|compound|map)">}

<!-- Type of graphics shape --> <!--ENTITY % graphics-type.type "(rectangle|circle|roundrectangle)">}

<!-- Color type of graphics this type is a string that represents the color to be used by the Graphic object. The color is a numerical RGB specification. ex) #FFFFFF --> <!--ENTITY % graphics-color.type "%string.type;">}

<!-- Type of Relation --> <!--ENTITY % relation-type.type "(ECrel|PPrel|GErel)">}

<!-- Name of Subtype (compound|activation|inhibition|expression|repression|indirect effect|state change|binding/association|dissociation|phosphorylation|dephosphorylation|glycosylation|ubiquitination|methylation) --> <!--ENTITY % subtype-name.type "%string.type;">}

<!-- Value of Subtype --> <!--ENTITY % subtype-value.type "%string.type;">}

<!-- Type of Reaction --> <!--ENTITY % reaction-type.type "(reversible|irreversible)">}

flexible !!
The DTD of KGML v0.4 (2)

```xml
<!ELEMENT pathway (entry*, relation*, reaction*)>
<!ATTLIST pathway name %keggid.type; #REQUIRED>
<!ATTLIST pathway number %mapnumber.type; #REQUIRED>
<!ATTLIST pathway org %maporg.type; #REQUIRED>
<!ATTLIST pathway title %string.type; #IMPLIED>
<!ATTLIST pathway image %url.type; #IMPLIED>
<!ATTLIST pathway link %url.type; #IMPLIED>

<!ELEMENT entry (graphics?, component*)>
<!ATTLIST entry id %id.type; #REQUIRED>
<!ATTLIST entry name %keggid.type; #REQUIRED>
<!ATTLIST entry type %entry-type.type; #REQUIRED>
<!ATTLIST entry link %url.type; #IMPLIED>
<!ATTLIST entry reaction %keggid.type; #IMPLIED>
<!ATTLIST entry map %idref.type; #IMPLIED> <!-- If the entry has attribute of map , it is an entry on other pathwaymap. -->

<!ELEMENT graphics EMPTY>
<!ATTLIST graphics name %string.type; #IMPLIED >
<!ATTLIST graphics x %number.type; #IMPLIED >
<!ATTLIST graphics y %number.type; #IMPLIED >
<!ATTLIST graphics type %graphics-type.type; "rectangle"> 
<!ATTLIST graphics width %number.type; "45" >
<!ATTLIST graphics height %number.type; "17" >
<!ATTLIST graphics fgcolor %graphics-color.type; "#000000" >
<!ATTLIST graphics bgcolor %graphics-color.type; "#FFFFFF" >

<!ELEMENT component EMPTY>
<!ATTLIST component id %id.type; #REQUIRED>
```
<ELEMENT relation (subtype*)>
  <!ATTLIST relation entry1 %idref.type; #REQUIRED> <!-- This attribute value indicates attribute of ID defined in the entry. -->
  <!ATTLIST relation entry2 %idref.type; #REQUIRED> <!-- This attribute value indicates attribute of ID defined in the entry. -->
  <!ATTLIST relation compound %idref.type; #IMPLIED> <!-- This attribute value indicates attribute of ID defined in the entry. -->
  <!ATTLIST relation type %relation-type.type; #REQUIRED>
</ELEMENT>

<ELEMENT subtype EMPTY>
  <!ATTLIST subtype name %subtype-name.type; #REQUIRED>
  <!ATTLIST subtype value %subtype-value.type; #REQUIRED>
</ELEMENT>

<ELEMENT reaction (substrate*, product*)>
  <!ATTLIST reaction name %keggid.type; #REQUIRED>
  <!ATTLIST reaction type %reaction-type.type; #IMPLIED>
</ELEMENT>

<ELEMENT substrate EMPTY>
  <!ATTLIST substrate name %keggid.type; #REQUIRED>
</ELEMENT>

<ELEMENT product EMPTY>
  <!ATTLIST product name %keggid.type; #REQUIRED>
</ELEMENT>
Read a KGML file

- Example: map00070, Pantothenate and CoA biosynthesis

  - pathway element (root)

```xml
  <entry ... type="map" ...><graphics ... type="roundrectangle" .../></entry>
  <entry ... type="enzyme" reaction="rn:Rxxxxx" ...><graphics ... type="rectangle" .../></entry>
  <entry ... type="compound" ...><graphics ... type="circle" ... /></entry>
</pathway>
```

- entry elements
The types of entry elements

- entry type → graphics type
  - common attributes of entry: id, name, link
  - common attributes of graphics: name, x, y, fgcolor, bgcolor, weight, height
Differences of links among map, enzyme, and compound

- map:
  - link="http://www.genome.ad.jp/kegg/pathway/map/mapxxxxx.html"

- enzyme:
  - link="http://www.genome.ad.jp/dbget-bin/www_bget?enzyme+x.x.x.x"

- compound:
  - link="http://www.genome.ad.jp/dbget-bin/www_bget?compound+Cxxxxx"
Read a KGML file (cont.)

• relation elements

<relation entry1="xx" entry2="xx" type="ECrel">
  <subtype name="compound" value="xx"/>
</relation>

• reaction elements

<reaction name="rn:Rxxxxx" type="irreversible">
  <substrate name="cpd:Cxxxxx"/>
  <product name="cpd:Cxxxxx"/>
</reaction>

recall the compound id from the entry id
Representation of the relations
Two binary relations between two enzymes with a compound as a link name
Representation of the reactions
KGML reactions

A reaction with two substrates and two products
Cross the defined pathways

<entry id="5" name="path:map00272" type="map" link="http://www.genome.ad.jp/kegg/pathway/map/map00272.html">
...
<entry id="67" name="ec:1.13.11.20" type="enzyme" map="5" link="http://www.genome.ad.jp/dbget-bin/show_pathway?map00272+1.13.11.20"/>
Organism-specific pathways

• How to make this figure?
  – get the figure and the XML file of the reference pathway
  – For each entity, if it appeared in the organism, change the bgcolor of this entity
  – reprint the figure
Compare two or more organism-specific pathways

• How to make this figure?
  – get the figure of the reference pathway
  – get the XML files of the two organisms
  – Compare the entries of these two files, then find the organism-specific entries and conserved entries
  – give them different colors and reprint the figure

• Use the same method to compare more than two organisms
The scoring problem of pathway comparison

- number-based approaches
  - based on number of conserved and organism-specific entities
  - based on number of conserved and organism-specific relations

- graph comparison approaches
  - convert all relations as a connected graph
  - apply well-established graph comparison algorithms

- Which is appropriate for REAL cases?
  - problem of alternative pathway (functional comparison or topological comparison?)
  - still under studying
Thanks for your attention!