CODATA 2002

A Proposition of XML Format for Proteomics Database

Ken'ichi KAMIJO, Toshimasa YAMAZAKI, and Akira TSUGITA

Proteomics Research Center,

Fundamental Research Labs., NEC Corp.

Data Format Standardization

- Download entries from public DBs as a flat-file
  - easy for a person to read
  - different formats for every DB
  - sometimes needs special access methods and special applications for each format

- Needs machine-readable formats for software tools

- To boost studies by exchanging data among researchers

Activates standardization

XML format

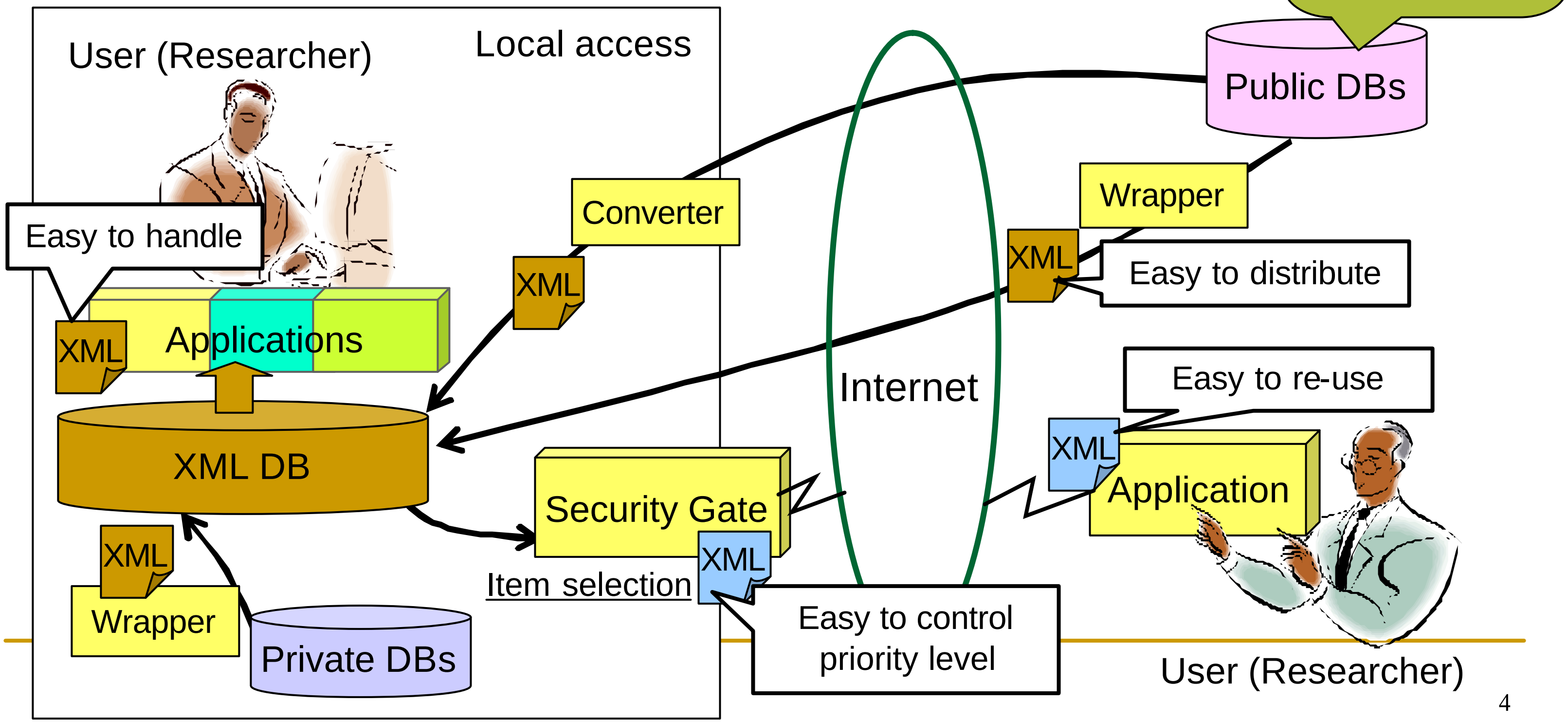
- XML (eXtensible Markup Language)
  - Highly readable for machine and person
  - Can represent information hierarchy and relationships
  - Details can be added right away
- Convenient for exchanging data
  - Easy to translate to other formats
  - Logical-check by a Document Type Definition (DTD)

```
<tag_source  element_growth=“8 weeks”>
  rice leaf
</tag_source>
```

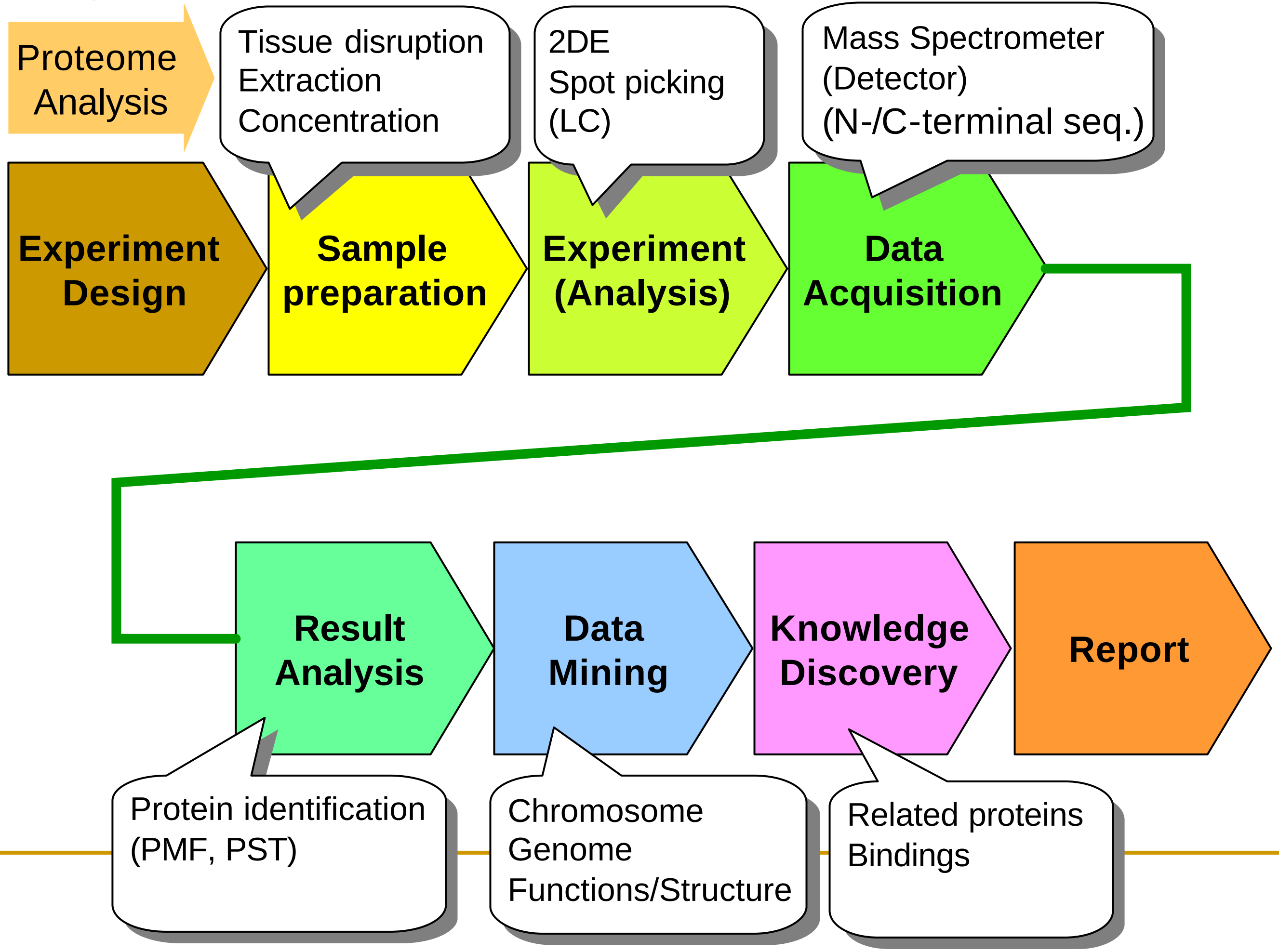
Example

XML in Bioinformatics

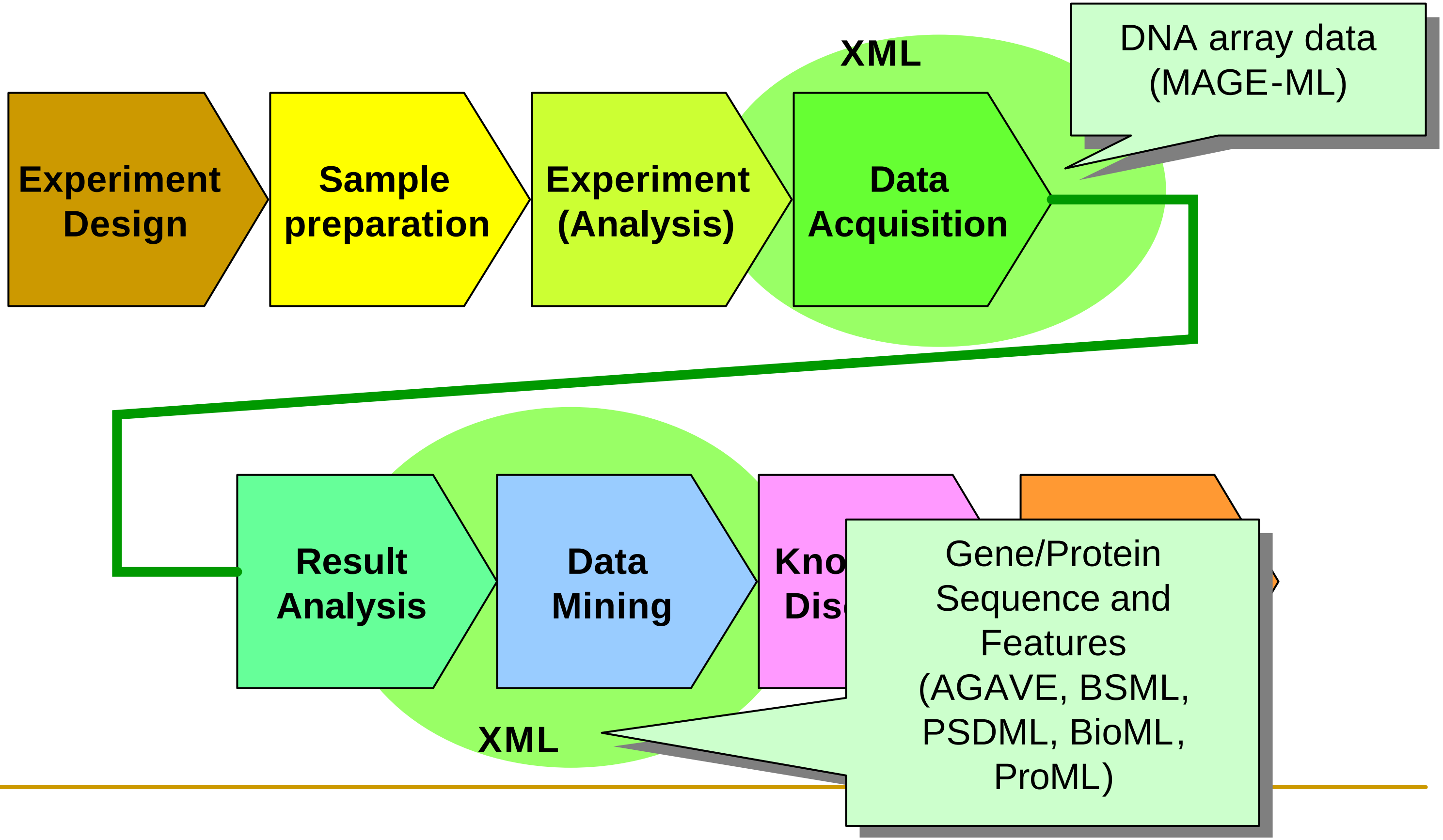
"The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web." -- W3C XML Web site, 2000-07-06.



Analysis flow in Life Science



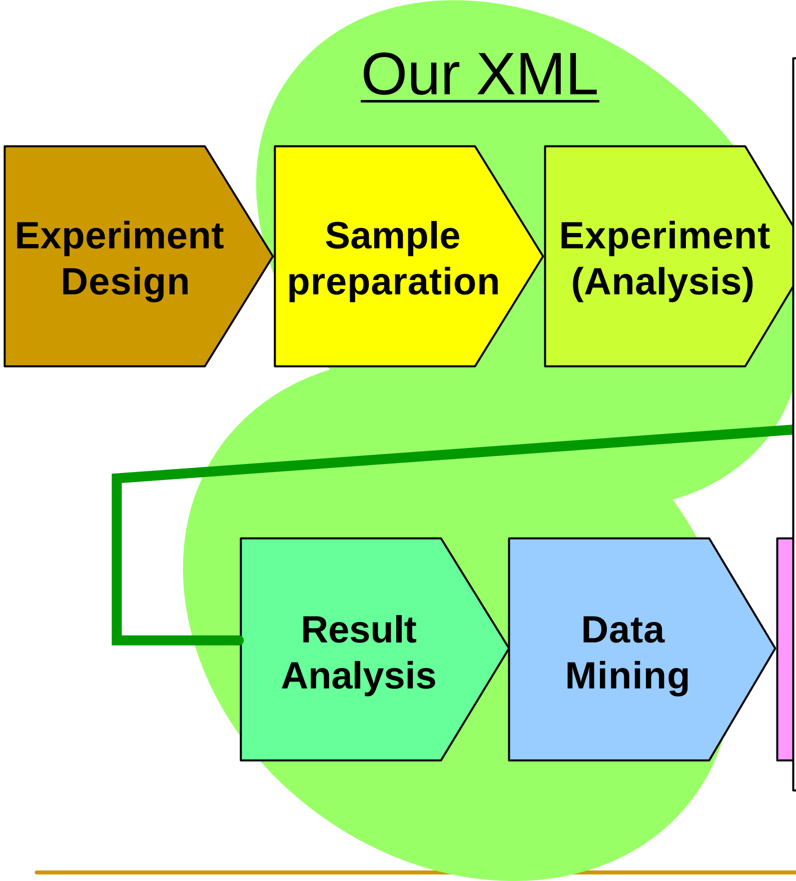
Conventional XMLs in Life Science



DNA array data (MAGE-ML)

Gene/Protein Sequence and Features (AGAVE, BSML, PSDML, BioML, ProML)

Our XML-based data model



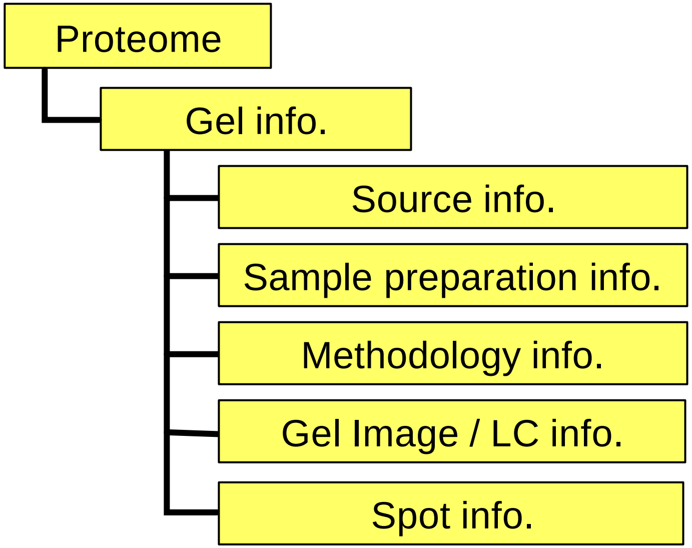
- Proteome-analysis oriented
- Describes
  - Sample preparation
  - Methodology
  - 2D gel image / LC results
  - Spot information
  - Sequence and feature
  - 3D structure
- Includes other open XMLs used in life science

Now Available : HUP-ML (Human Proteome Markup Language) DTD and Editor

http://www.jhupo.org/

XML for Proteomics

- Information Structure:



Proteome

Gel info.

Source info.

Sample preparation info.

Methodology info.

Gel Image / LC info.

Spot info.

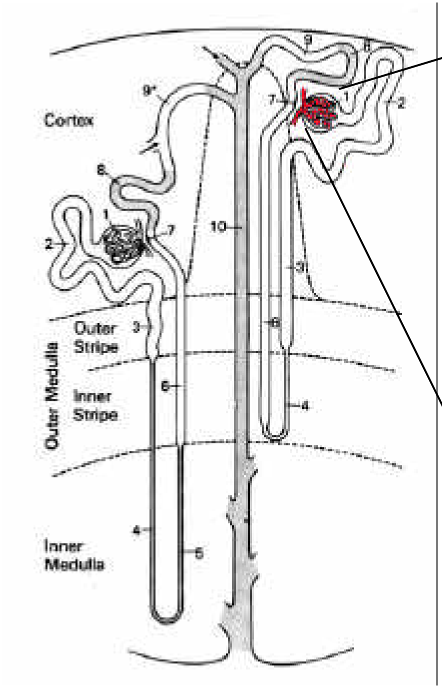
```
<proteome>
  <gel   id="1">
      <source_info>
      <gel_img >
      <sample_preparation>
      <gel_conditions>
      <marker>
      <detection>
      <gel_image>
      <spot   id="1">
          ....
          ....
      <spot   id="2">

      ....
  <gel   id="2">
```

CODATA 2002

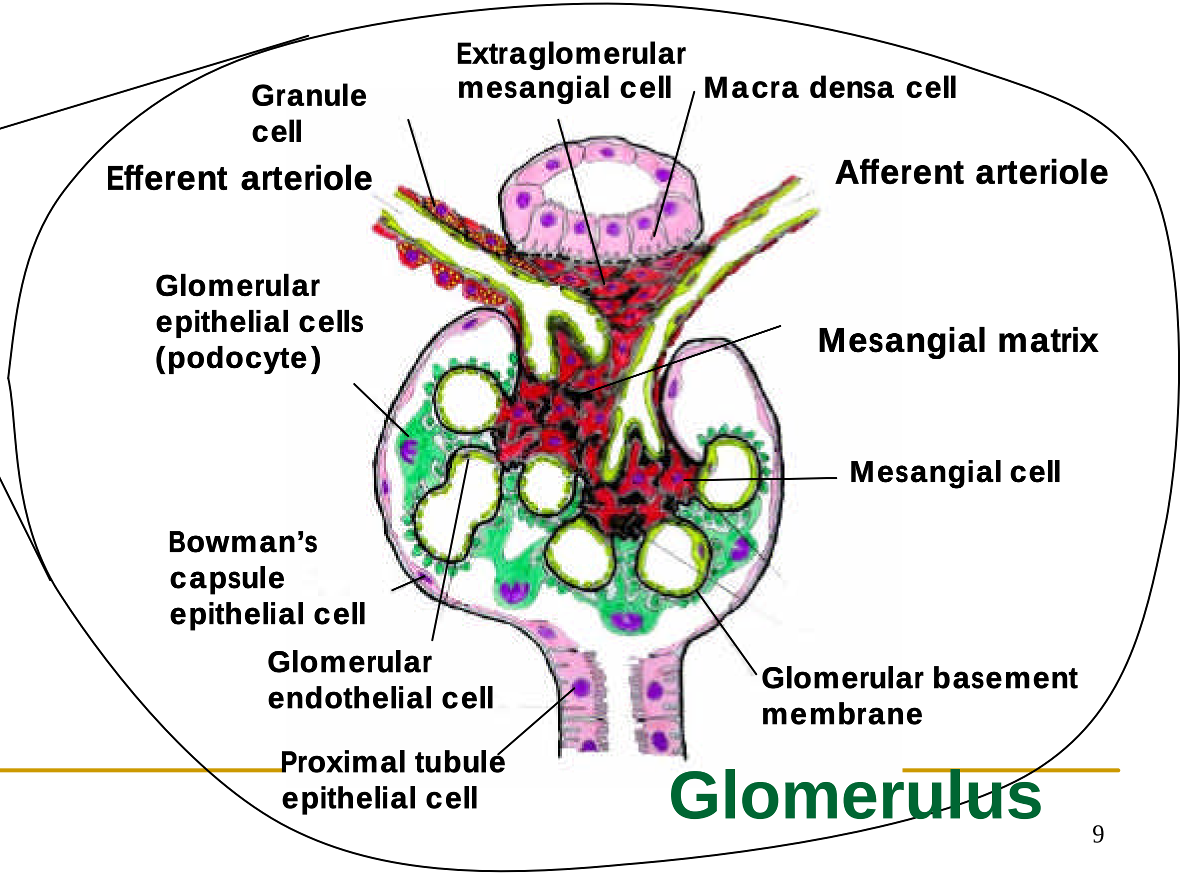
Example:

By A. Tsugita et al.(2002)

Human Kidney Glomerulus Proteome



Nephron



Extraglomerular mesangial cell

Macra densa cell

Granule cell

Efferent arteriole

Afferent arteriole

Glomerular epithelial cells (podocyte)

Mesangial matrix

Mesangial cell

Bowman's capsule epithelial cell

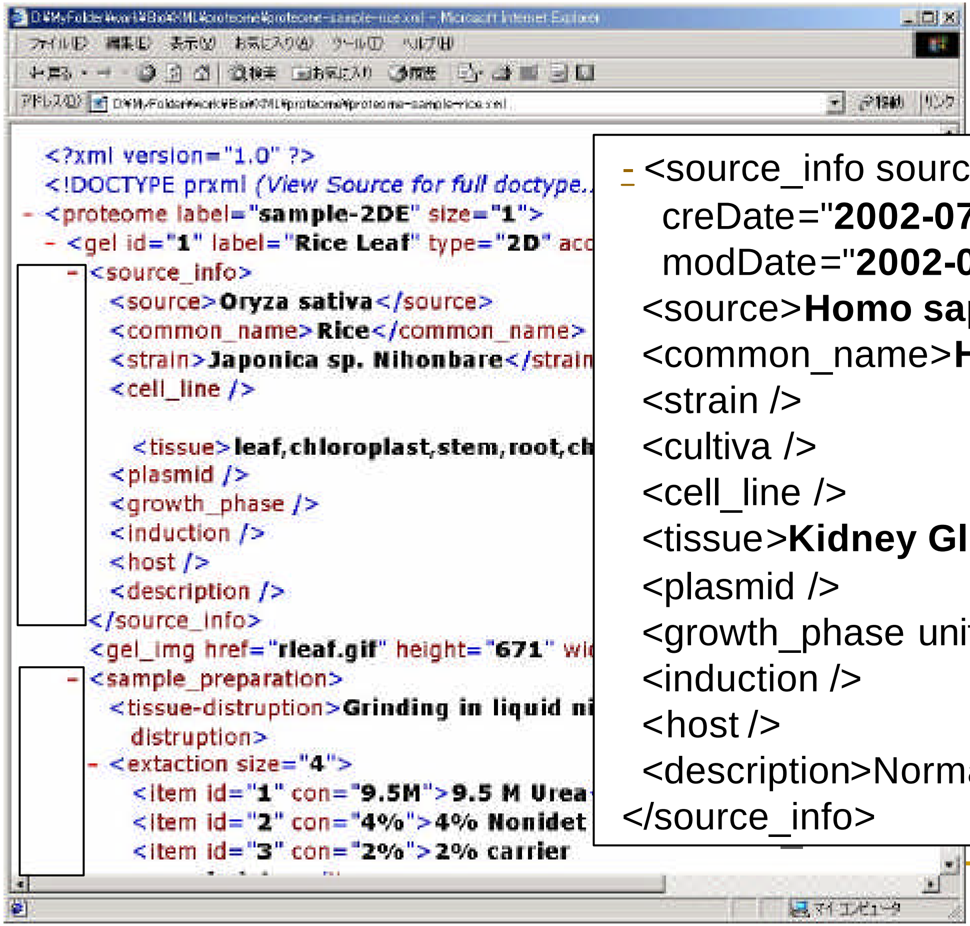
Glomerular endothelial cell

Glomerular basement membrane

Proximal tubule epithelial cell

Glomerulus

Sample of ProteomeXML (1)


D:¥MyFolder¥work¥BioXML¥proteome¥proteome-sample-rice.xml - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り 履歴
アドレス(D) D:¥MyFolder¥work¥BioXML¥proteome¥proteome-sample-rice.xml
移動 リンク
マイ コンピュータ

```
  <?xml version="1.0" ?>
  <!DOCTYPE prxml (View Source for full doctype..
- <proteome label="sample-2DE" size="1">
  - <gel id="1" label="Rice Leaf" type="2D" acc
    - <source_info>
        <source>Oryza sativa</source>
        <common_name>Rice</common_name>
        <strain>Japonica sp. Nihonbare</strain
        <cell_line />

          <tissue>leaf,chloroplast,stem,root,ch
        <plasmid />
        <growth_phase />
        <induction />
        <host />
        <description />
      </source_info>
      <gel_img href="rleaf.gif" height="671" wi
    - <sample_preparation>
        <tissue-distruption>Grinding in liquid ni
          distruption>
      - <extaction size="4">
          <item id="1" con="9.5M">9.5 M Urea
          <item id="2" con="4%">4% Nonidet
          <item id="3" con="2%">2% carrier
```

Source information

```
- <source_info source_info_ID="HKG-1"
    creDate="2002-07-20T12:00:00"
    modDate="2002-08-10T17:20:00">
  <source>Homo sapiens</source>
  <common_name>Human</common_name>
  <strain />
  <cultiva />
  <cell_line />
  <tissue>Kidney Glomerulus</tissue>
  <plasmid />
  <growth_phase unit="year">48</growth_phase>
  <induction />
  <host />
  <description>Normal</description>
</source_info>
```

Sample of ProteomeXML (2)

Sample preparation

```
- <sample_preparation>
  <tissue-disruption>Standard sieving technique using
four stainless sieves. The glomeruli on the 150 micro
m sieves were collected ice cold phosphate-buffered
saline (PBS).</tissue-disruption>
- <extraction>
 - <procedure>
      <process seq="1" action="spin-down"
                 sample="collection" />
       <process seq="2" action="homogenize"
                 sample="precipitate" >
                 <add_solution solution_ID="sol-A"/>
       </process>
       <process seq="3" action="stand"
                  time="60" time_unit="min"
                  temp="37" temp_unit="degree in C" />
       <process seq="4" action="centrifuge"
                 sample="suspension"
                 time="20" time_unit="min">
                <times_g>12000</times_g>
        </process>
```

```
        <process seq="5" action="store"
                      sample="supernatant"
                      te
                      tir
  </procedure>
  <comment_ext
  </extraction>

- <solution solution_ID="sol-A" label="2-DE lysis solution">
  <item_solution con="9.8" unit="M" name="Urea" />
  <item_solution con="2" unit="% w/v" name="NP-40" />
  <item_solution con="2" unit="% v/v" name="Pharmalyte(pH3-10)"
/>
  <item_solution con="10" unit="mM" name="DDT" />
  <item_solution con="0.5" unit="micro g/mL" name="E-64" />
  <item_solution con="0.5" unit="mM" name="PMSF" />
  <item_solution con="40" unit="micro g/mL" name="TLCK" />
  <item_solution con="1" unit="micro g/mL" name="aprotinin" />
  <item_solution con="10" unit="micro g/mL" name="chymostain"
/>
  <item_solution c
  <item_solution c
  <comment_solu
  </solution>
```

Procedure :

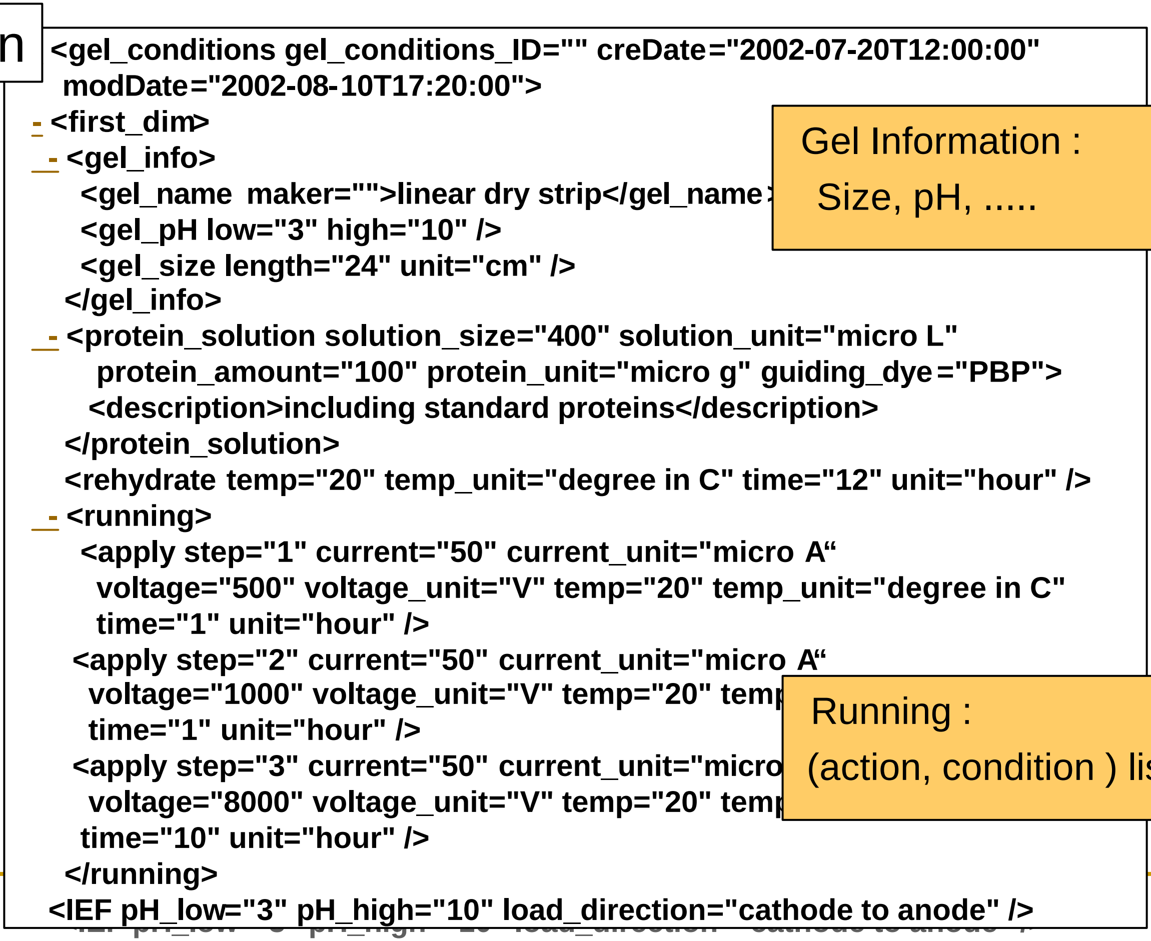
(action, target, condition) lists

Solution list :

solution item information

Sample of ProteomeXML (3)

Gel condition



```
<gel_conditions gel_conditions_ID="" creDate="2002-07-20T12:00:00"
 modDate="2002-08-10T17:20:00">
- <first_dim>
  - <gel_info>
      <gel_name  maker="">linear dry strip</gel_name
      <gel_pH low="3" high="10" />
      <gel_size length="24" unit="cm" />
    </gel_info>
  - <protein_solution solution_size="400" solution_unit="micro L"
        protein_amount="100" protein_unit="micro g" guiding_dye="PBP">
       <description>including standard proteins</description>
    </protein_solution>
    <rehydrate temp="20" temp_unit="degree in C" time="12" unit="hour" />
  - <running>
      <apply step="1" current="50" current_unit="micro  A“
        voltage="500" voltage_unit="V" temp="20" temp_unit="degree in C"
        time="1" unit="hour" />
     <apply step="2" current="50" current_unit="micro  A“
       voltage="1000" voltage_unit="V" temp="20" temp
       time="1" unit="hour" />
     <apply step="3" current="50" current_unit="micro
       voltage="8000" voltage_unit="V" temp="20" temp
      time="10" unit="hour" />
    </running>
  <IEF pH_low="3" pH_high="10" load_direction="cathode to anode" />
```

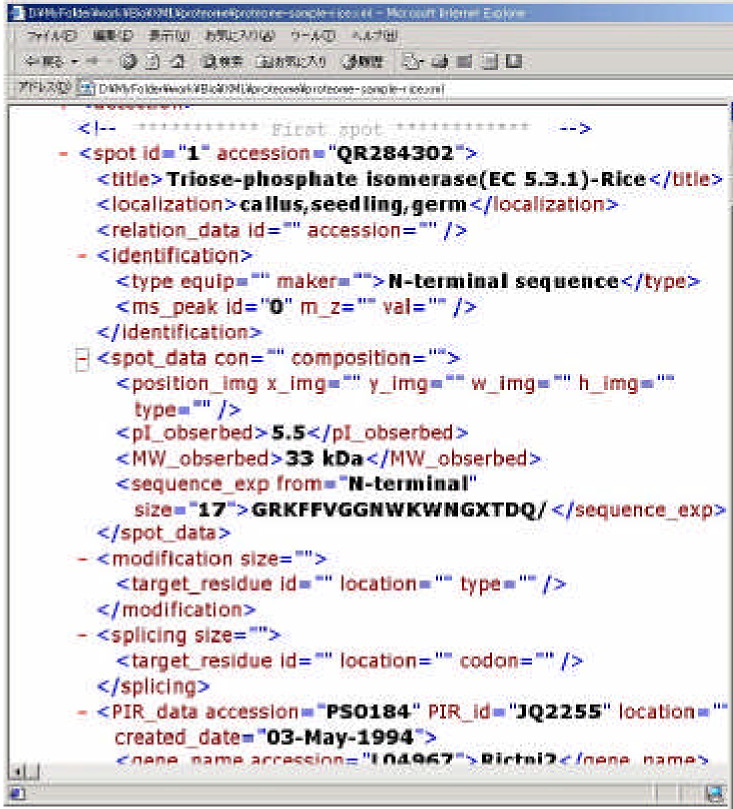
Gel Information :

Size, pH, ...

Running :

(action, condition) lists

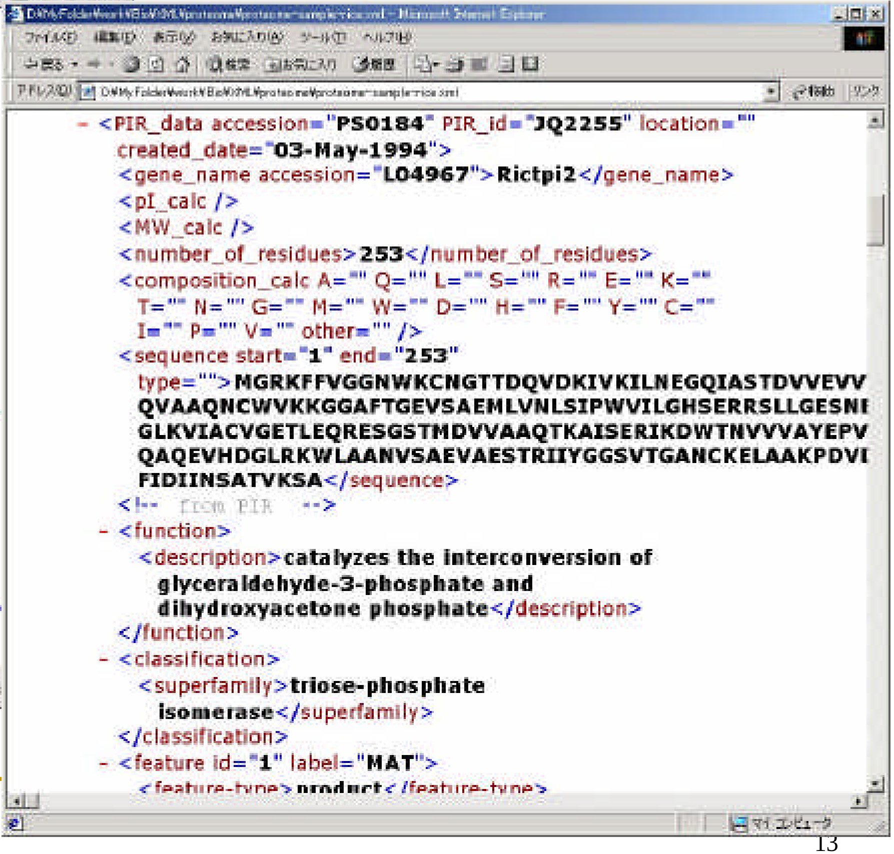
Sample of ProteomeXML (4)


ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
D:¥My Folder¥work¥BioXML¥proteome¥proteome-sample-rice.xml

```
  <!-- ************* First spot ************* -->
- <spot id="1" accession="QR284302">
    <title>Triose-phosphate isomerase(EC 5.3.1)-Rice</title>
    <localization>callus,seedling,germ</localization>
    <relation_data id="" accession="" />
  - <identification>
      <type equip="" maker="">N-terminal sequence</type>
      <ms_peak id="0" m_z="" val="" />
    </identification>
  - <spot_data con="" composition="">
      <position_img x_img="" y_img="" w_img="" h_img=""
        type="" />
      <pI_obserbed>5.5</pI_obserbed>
      <MW_obserbed>33 kDa</MW_obserbed>
      <sequence_exp from="N-terminal"
        size="17">GRKFFVGGNWKWNGXTDQ/</sequence_exp>
    </spot_data>
  - <modification size="">
      <target_residue id="" location="" type="" />
    </modification>
  - <splicing size="">
      <target_residue id="" location="" codon="" />
    </splicing>
  - <PIR_data accession="PS0184" PIR_id="JQ2255" location=""
      created_date="03-May-1994">
      <gene_name accession="L04967">Rictpi2</gene_name>
```

Spot information area

PIR data area


ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
D:¥My Folder¥work¥BioXML¥proteome¥proteome-sample-rice.xml
マイ コンピュータ

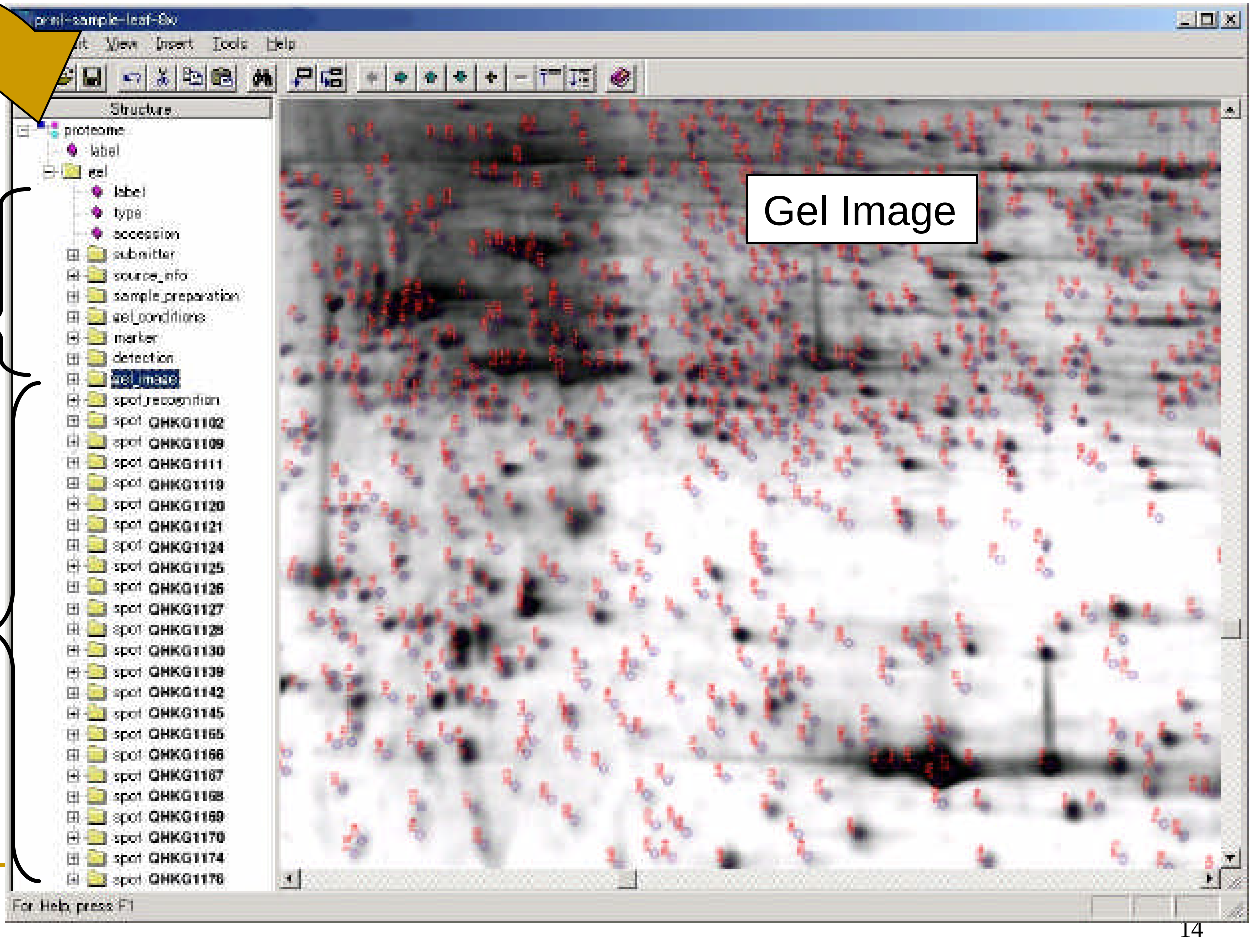
```
- <PIR_data accession="PS0184" PIR_id="JQ2255" location=""
    created_date="03-May-1994">
    <gene_name accession="L04967">Rictpi2</gene_name>
    <pI_calc />
    <MW_calc />
    <number_of_residues>253</number_of_residues>
    <composition_calc A="" Q="" L="" S="" R="" E="" K=""
      T="" N="" G="" M="" W="" D="" H="" F="" Y="" C=""
      I="" P="" V="" other="" />
    <sequence start="1" end="253"
      type="">MGRKFFVGGNWKCNGTTDQVDKIVKILNEGQIASTDVVEVV
      QVAAQNCWVKKGGAFTGEVSAEMLVNLSIPWVILGHSERRSLLGESNI
      GLKVIACVGETLEQRESGSTMDVVAAQTKAISERIKDWTNVVVAYEPV
      QAQEVHDGLRKWLAANVSAEVAESTRIIYGGSVTGANCKELAAKPDVI
      FIDIINSATVKSA</sequence>
    <!-- from PIR -->
  - <function>
      <description>catalyzes the interconversion of
        glyceraldehyde-3-phosphate and
        dihydroxyacetone phosphate</description>
    </function>
  - <classification>
      <superfamily>triose-phosphate
        isomerase</superfamily>
    </classification>
  - <feature id="1" label="MAT">
      <feature-type>product</feature-type>
```

XML Editor for Proteomics Information

Our XML Document

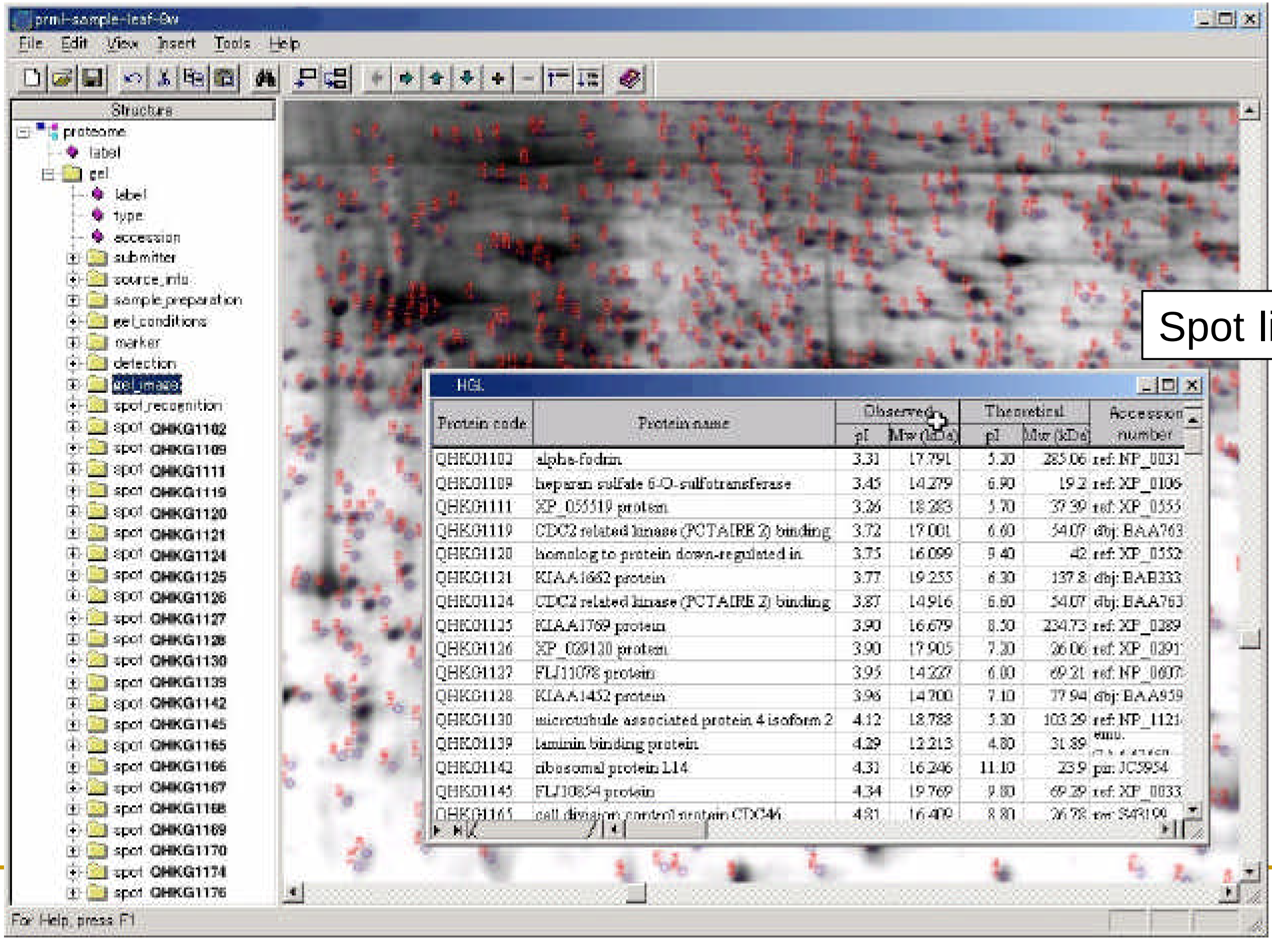
Gel Info.

Spot Info.



Gel Image

XML Editor (Example)

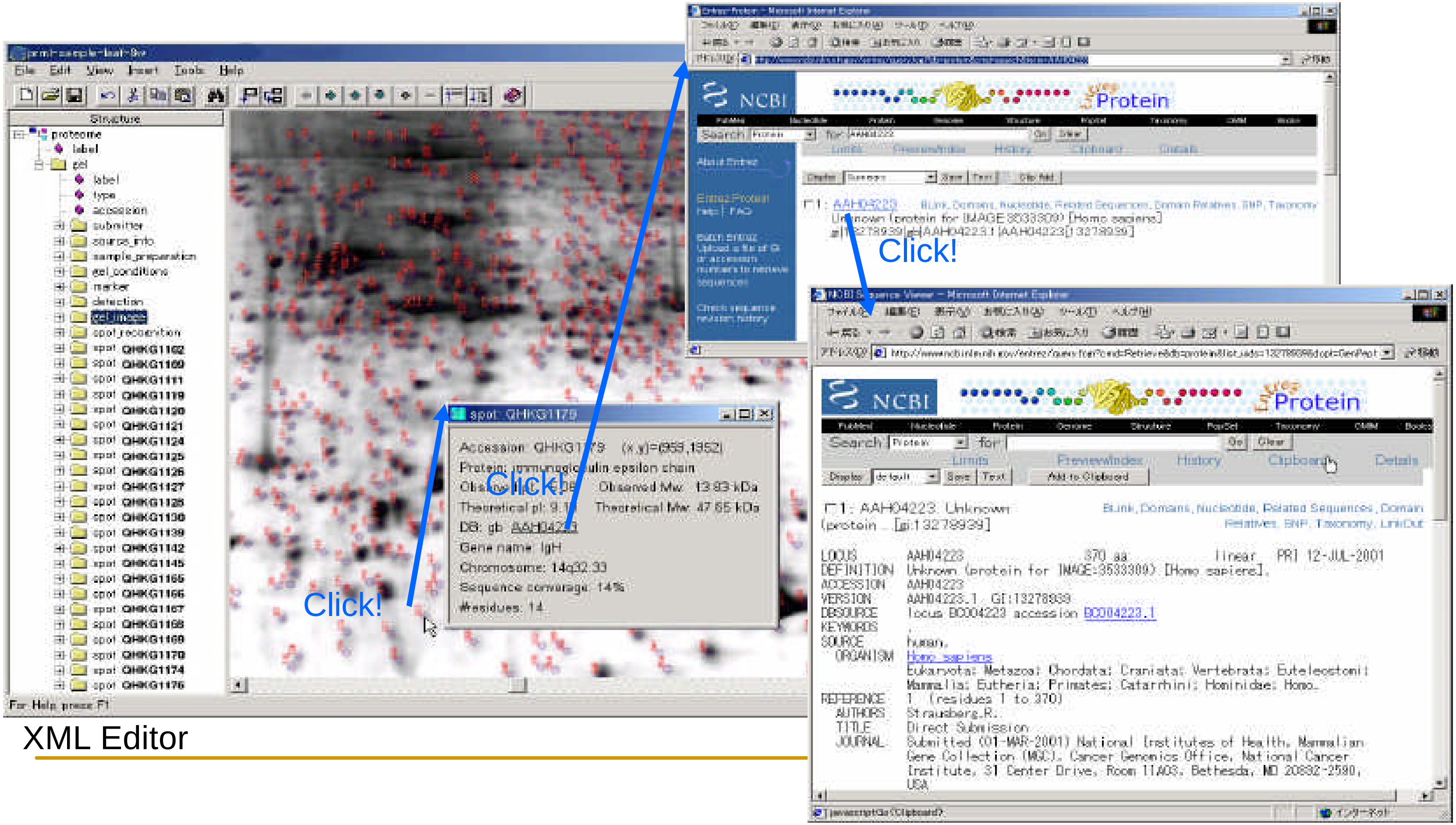


Protein code	Protein name	Observed		Theoretical		Accession number
		pI	Mw (kDa)	pI	Mw (kDa)	
QHKG1102	alpha-fodrin	3.31	17.791	5.20	285.06	ref: NP_0031
QHKG1109	heparan sulfate 6-O-sulfotransferase	3.45	14.279	6.90	19.2	ref: XP_0106
QHKG1111	XP_055519 protein	3.26	18.283	5.70	37.39	ref: XP_0555
QHKG1119	CDC2 related kinase (PCTAIRE 2) binding	3.72	17.001	6.60	54.07	dbj: BAA763
QHKG1120	homolog to protein down-regulated in	3.75	16.099	9.40	42	ref: XP_0552
QHKG1121	KIAA1662 protein	3.77	19.255	6.30	137.8	dbj: BAB333
QHKG1124	CDC2 related kinase (PCTAIRE 2) binding	3.87	14.916	6.60	54.07	dbj: BAA763
QHKG1125	KIAA1769 protein	3.90	16.679	8.50	234.73	ref: XP_0289
QHKG1126	XP_029130 protein	3.90	17.905	7.20	26.06	ref: XP_0291
QHKG1127	FLJ11078 protein	3.95	14.227	6.00	69.21	ref: NP_0607
QHKG1128	KIAA1452 protein	3.96	14.700	7.10	77.94	dbj: BAA959
QHKG1130	microtubule associated protein 4 isoform 2	4.12	18.788	5.30	103.29	ref: NP_1121
QHKG1139	laminin binding protein	4.29	12.213	4.80	31.89	emb:
QHKG1142	ribosomal protein L14	4.31	16.246	11.10	23.9	pir: JC5954
QHKG1145	FLJ10854 protein	4.34	19.769	9.80	69.29	ref: XP_0033
QHKG1165	cell division control protein CDC46	4.81	16.409	8.80	26.78	sw: S43190

Spot list

XML Editor (Browsing)

CODATA 2002



spot: QHKG1179

Accession: QHKG1179 (x,y)=(959,1952)
Protein: immunoglobulin epsilon chain
Observed pI: 5.05 Observed Mw: 13.93 kDa
Theoretical pI: 9.18 Theoretical Mw: 47.65 kDa
DB: gb AAH04223
Gene name: IgH
Chromosome: 14q32.33
Sequence coverage: 14%
#residues: 14

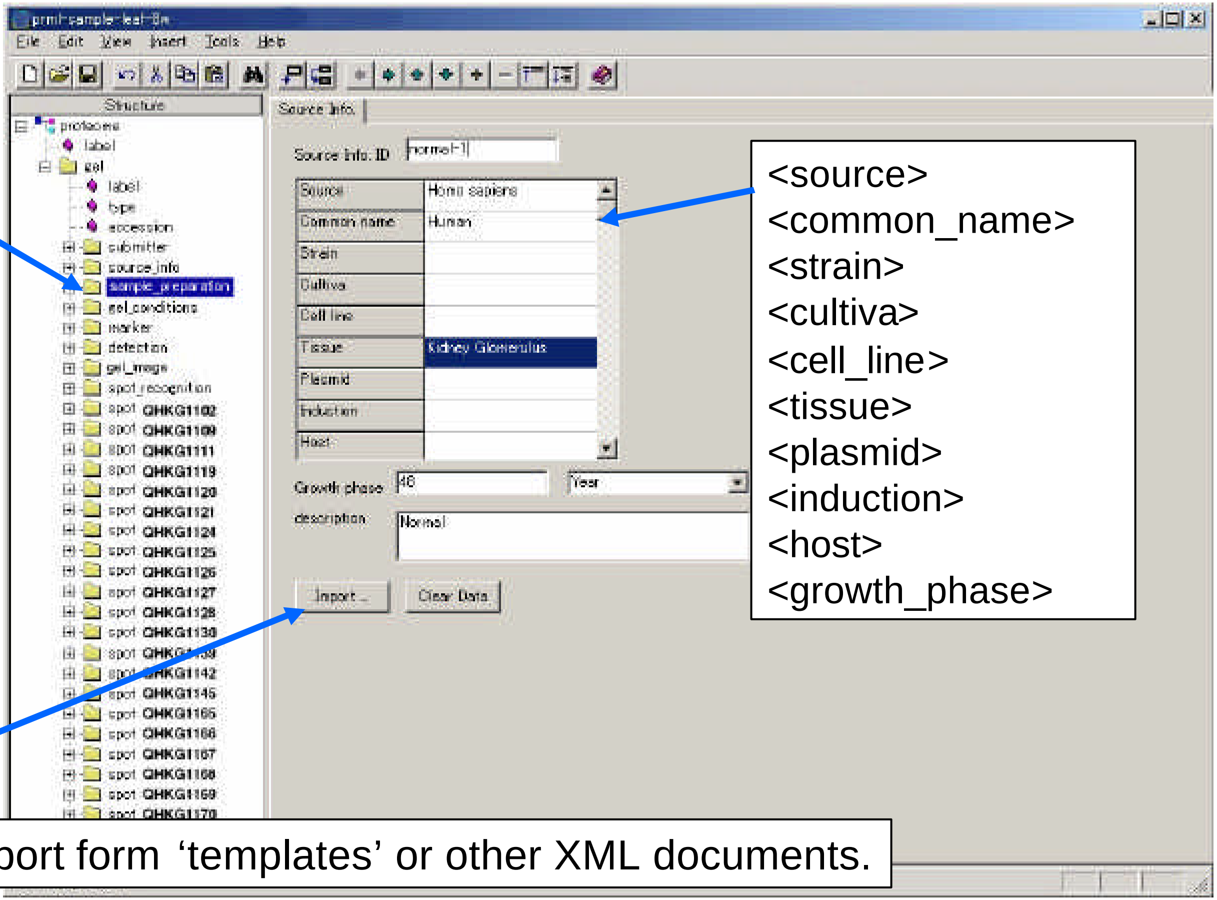
1: AAH04223 BLink, Domains, Nucleotide, Related Sequences, Domain Relatives, SNP, Taxonomy
Unknown (protein for IMAGE:3533309) [Homo sapiens]
gi|13278939|gb|AAH04223.1|AAH04223[13278939]

1: AAH04223. Unknown (protein ... [gi:13278939]

```
LOCUS       AAH04223                 370 aa            linear   PRI 12-JUL-2001
DEFINITION  Unknown (protein for IMAGE:3533309) [Homo sapiens].
ACCESSION   AAH04223
VERSION     AAH04223.1  GI:13278939
DBSOURCE    locus BC004223 accession BC004223.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 370)
  AUTHORS   Strausberg,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (01-MAR-2001) National Institutes of Health, Mammalian
            Gene Collection (MGC), Cancer Genomics Office, National Cancer
            Institute, 31 Center Drive, Room 11A03, Bethesda, MD 20892-2580,
            USA
```

XML Editor

XML Editor (Source Information)



- proteome
  - label
  - gel
    - label
    - type
    - accession
    - submitter
    - source_info
    - sample_preparation
    - gel_conditions
    - marker
    - detection
    - gel_image
    - spot_recognition
    - spot QHKG1102
    - spot QHKG1109
    - spot QHKG1111
    - spot QHKG1119
    - spot QHKG1120
    - spot QHKG1121
    - spot QHKG1124
    - spot QHKG1125
    - spot QHKG1126
    - spot QHKG1127
    - spot QHKG1128
    - spot QHKG1130
    - spot QHKG1138
    - spot QHKG1142
    - spot QHKG1145
    - spot QHKG1165
    - spot QHKG1166
    - spot QHKG1167
    - spot QHKG1168
    - spot QHKG1169
    - spot QHKG1170

Source Info. ID: normal-1

Source	Homo sapiens
Common name	Human
Strain	
Cultiva	
Cell line	
Tissue	Kidney Glomerulus
Plasmid	
Induction	
Host	

Growth phase: 48 Year

description: Normal

Source Information

<source>
<common_name>
<strain>
<cultiva>
<cell_line>
<tissue>
<plasmid>
<induction>
<host>
<growth_phase>

It is possible to import form ‘templates’ or other XML documents.

Features of our data model

Our proteomics XML:

- describes sample preparations
  - Improves reliability of analysis results
- can distribute experimental information
  - share know-how
  - improves skills
- handle both gel-image and analysis results
- describes analysis information
  - image recognition

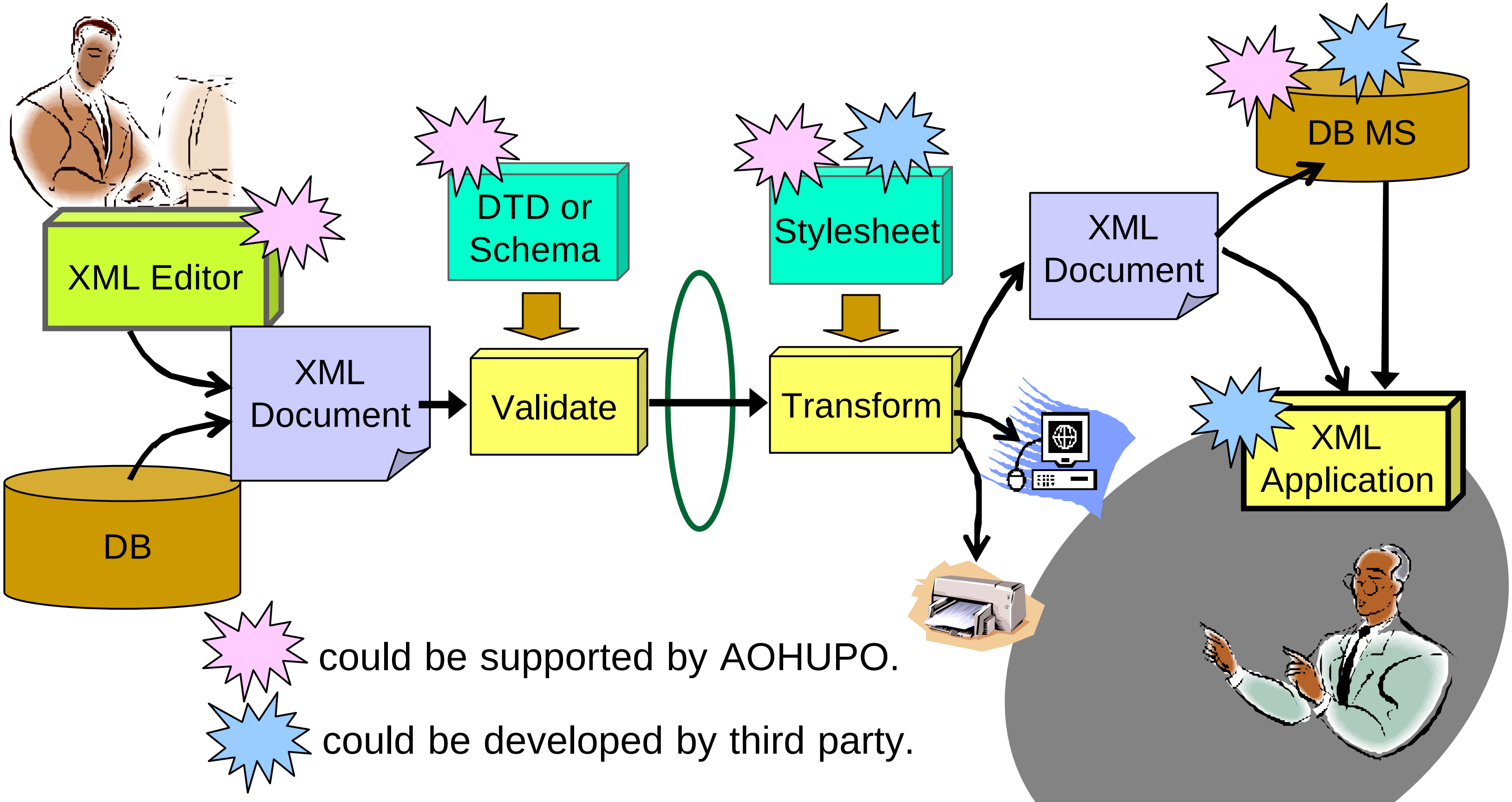
Now Available : HUP-ML (Human Proteome Markup Language) DTD and Editor
http://www.jhupo.org/

Future works

- Open DTD and/or XML Schema
  - Collaboration with AOHUPO
- Develop XML viewer for free distribution
- Prototype WWW-based management system
  - for registration, viewing, and retrieval of entries
- Convert from other XML formats
- Relation to other analysis tools
  - image-analysis software
  - homology-analysis tools, etc.

AOHUPO: Asia Oceania Human Proteome Organiazaion

Our XML Workflows



✸ could be supported by AOHUPO.

✸ could be developed by third party.

Now Available : HUP-ML (Human Proteome Markup Language) DTD and Editor
http://www.jhupo.org/