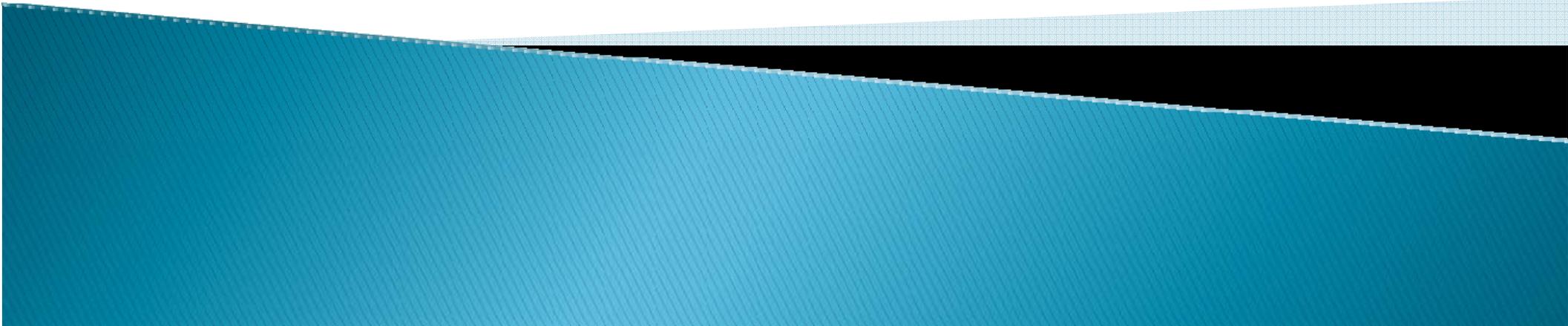


# Learning KEGG – Chemical data –

10 Jun 2008  
Masahiro Hattori



# LIGAND Database



## KEGG LIGAND Database

Molecular building blocks of life in the chemical space

KEGG2 ATLAS PATHWAY BRITE GENES SSDB LIGAND DBGET

### Chemical Substances and Reactions

**KEGG LIGAND** contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

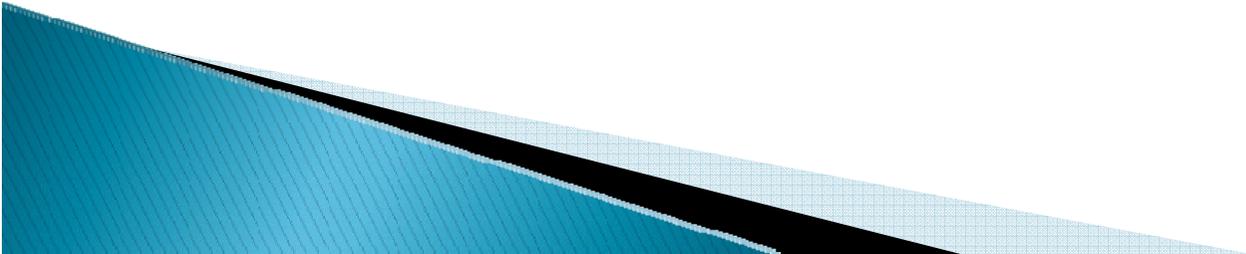
Database	Identifier	Content	Specialized entry point
LIGAND	COMPOUND	C number	Chemical compound structures
	DRUG	D number	Drug structures
	GLYCAN	G number	Glycan structures
	REACTION	R number	Biochemical reactions
	RPAIR	A number	Reactant pair alignments
	ENZYME	EC number	Enzyme nomenclature

Search LIGAND for  Go Clear

<http://www.genome.jp/kegg/ligand.html>

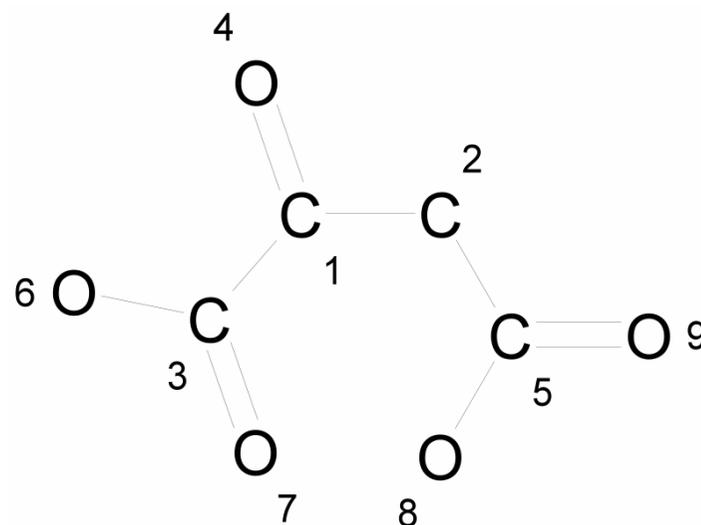
# LIGAND Database

## ▶ Components of LIGAND Database

- **COMPOUND**      Chemical compound structure
  - DRUG              Drug structure
  - GLYCAN           Glycan structure
  - REACTION        Biochemical reaction
  - ENZYME          Enzyme nomenclature
  
  - RPAIR             Reaction pair alignment
- 

# COMPOUND Database

- ▶ **Metabolites** have been collected, including the secondary metabolites etc.
- ▶ Chemical structures are represented and stored as **2D graph** objects,
  - whose vertices and edges are atoms and atomic bonds respectively.
  - Each structure is formatted in **MDL/mol** or **KCF**.



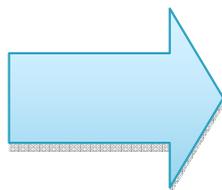


# What is the KCF format?

- ▶ Abbreviation of **KEGG Chemical Function**.
- ▶ KEGG original formats with a new feature.
  - Each atom is described by one of **KEGG Atom Types**.

**KEGG** COMPOUND: C00111 Help

Entry	C00111	Compound
Name	Glycerone phosphate; Dihydroxyacetone phosphate	
Formula	C3H7O6P	
Mass	169.998	
Structure	 C00111 <span>Mol file</span> <span>KCF file</span> <span>DB search</span> <span>Jmol</span> <span>KegDraw</span>	
Reaction	R00842 R00844 R00846 R00848 R00849 R01010 R01011 R01012 R01013 R01014 R01015 R01016 R01068 R01069 R01070 R01785 R01829 R02262 R02263 R02568 R04292 R05378 R05679 R05680	
Pathway	PATH: map00010 Glycolysis / Gluconeogenesis PATH: map00031 Inositol metabolism PATH: map00040 Pentose and glucuronate interconversions PATH: map00051 Fructose and mannose metabolism PATH: map00052 Galactose metabolism PATH: map00561 Glycerolipid metabolism PATH: map00564 Glycerophospholipid metabolism PATH: map00620 Pyruvate metabolism PATH: map00710 Carbon fixation PATH: map00760 Nicotinate and nicotinamide metabolism	
Enzyme	1.1.1.8 1.1.1.94 1.1.1.261 1.1.3.21 1.1.99.4 1.1.99.5 2.3.1.42 2.7.1.29 2.7.1.121 3.1.3.1 4.1.2.2 4.1.2.13 4.1.2.17 4.1.2.19 4.1.2.29 4.1.2.40 4.2.3.3 5.3.1.1	
Other DBS	PubChem: 3411 ChEBI: 16108 PDB-CCD: 13P 3DMET: B00029	
LinkDB	<span>All DBs</span>	
KCF data	<span>Show</span>	



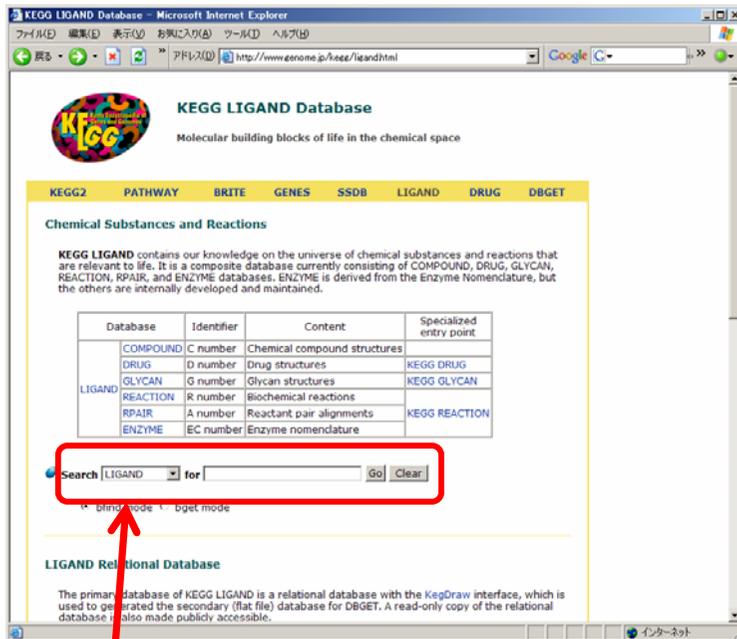
ENTRY	C00111	Compound
ATOM	10	
1	P1b P	22.8555 -17.6695
2	O2b O	21.4279 -17.6628
3	O1c O	22.8992 -19.2106
4	O1c O	22.8950 -16.1215
5	O3b O	24.2890 -17.6497
6	C1b C	20.1911 -16.9523
7	C5a C	20.1911 -15.5247
8	C1b C	20.1911 -14.0971
9	O5a O	21.4787 -15.5247
10	O1a O	21.4279 -13.3801
BOND	9	
1	1 2 1	
2	1 3 1	
3	1 4 1	
4	1 5 2	
5	2 6 1	
6	6 7 1	
7	7 8 1	
8	7 9 2	
9	8 10 1	

# KEGG Atom Types in KCF

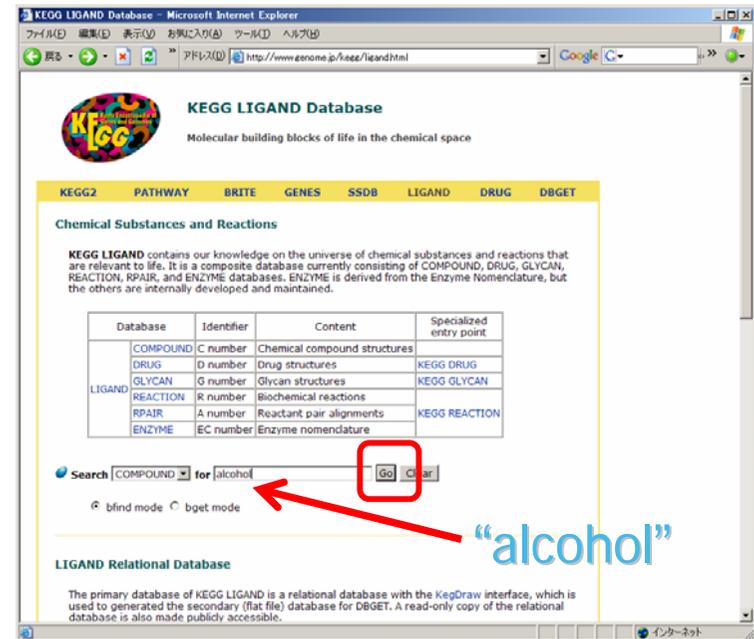
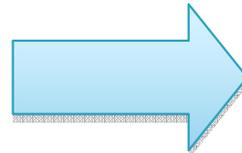
- ▶ Atom typing (atom grouping) based on the classification of **functional groups**.
- ▶ Total **69** atom types are defined by conditions:
  - i) connection type
  - ii) neighboring atom species
  - iii) linear, circular, aromatic
- ▶ Useful for the further analyses like chemical structure comparisons.

KEGG atom types

Atom	Functional group	Atom type	Description	Frequency
C	Alkane	C1a	R-CH3	16473
		C1b	R-CH2-R	20193
		C1c	R-CH(-R)-R	4964
		C1d	R-C(-R)2-R	698
	Cyclic alkane	C1x	ring-CH2-ring	14010
		C1y	ring-CH(-R)-ring	27376
		C1z	ring-CH(-R)2-ring	4463
	Alkene	C2a	R=CH2	634
		C2b	R=CH-R	3965
		C2c	R=C(-R)2	1914
	Cyclic alkene	C2x	ring-CH=ring	2964
		C2y	ring-C(-R)=ring or ring-C(=R)-ring	3722
		C3a	R≡CH	43
		C3b	R≡C-R	282
	Aldehyde	C4a	R-CH=O	350
	Ketone	C5a	R-C(=O)-R	3595
	Cyclic ketone	C5x	ring-C(=O)-ring	2257
	Carboxylic acid	C6a	R-C(=O)-OH	3190
	Carboxylic ester	C7a	R-C(=O)-O-R	1691
C7x		ring-C(=O)-O-ring	869	
Aromatic ring	C8x	ring-CH=ring	19905	
	C8y	ring-C(-R)=ring	20511	
Undefined C	C0		8	
N	Amine	N1a	R-NH2	2440
		N1b	R-NH-R	3003
		N1c	R-N(-R)2	374
		N1d	R-N(-R)3+	105
	Cyclic amine	N1x	ring-NH-ring	806
		N1y	ring-N(-R)-ring	1464
	Imine	N2a	R=N-H	230
		N2b	R=N-R	163
	Cyclic imine	N2x	ring-N=ring	357
		N2y	ring-N(-R)+=ring	14
	Cyan	N3a	R≡N	119
		N4x	ring-NH-ring	785
	Aromatic ring	N4y	ring-N(-R)-ring	840

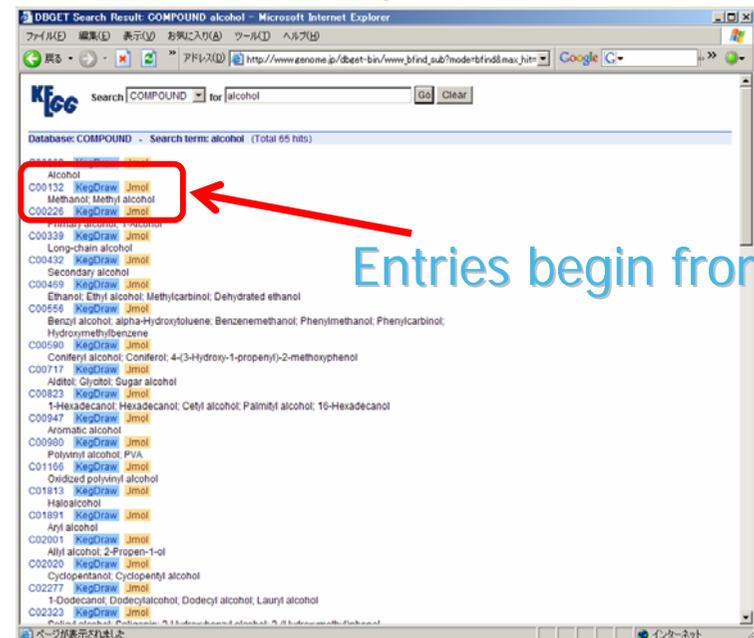
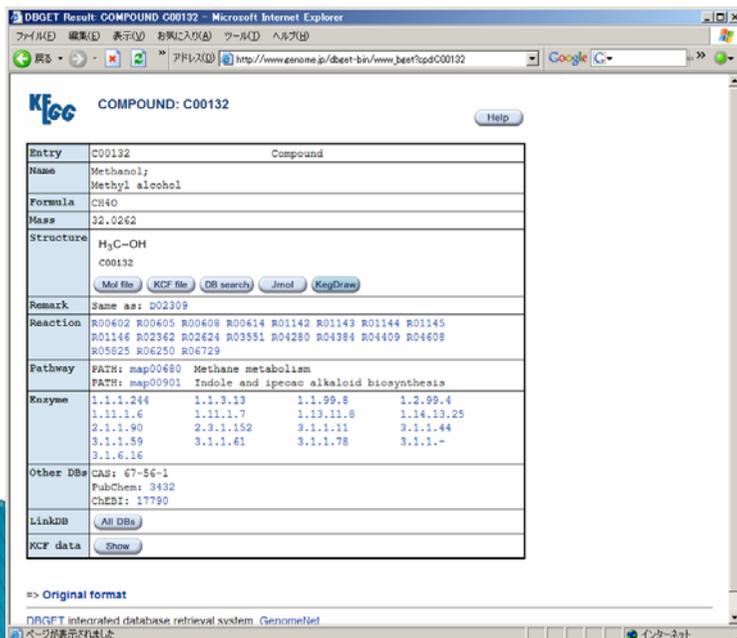
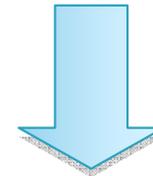


Select "COMPOUND"



"alcohol"

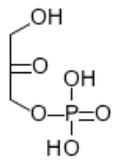
Search example of COMPOUND entry



Entries begin from "C"

# Example of COMPOUND Entry

**KEGG** COMPOUND: C00111 Help

<b>Entry</b>	C00111	Compound																				
<b>Name</b>	Glycerone phosphate; Dihydroxyacetone phosphate																					
<b>Formula</b>	C3H7O6P																					
<b>Mass</b>	169.998																					
<b>Structure</b>	 <p>C00111</p> <p>Mol file KCF file DB search Jmol KegDraw</p>																					
<b>Reaction</b>	<a href="#">R00842</a> <a href="#">R00844</a> <a href="#">R00846</a> <a href="#">R00848</a> <a href="#">R00849</a> <a href="#">R01010</a> <a href="#">R01011</a> <a href="#">R01012</a> <a href="#">R01013</a> <a href="#">R01014</a> <a href="#">R01015</a> <a href="#">R01016</a> <a href="#">R01068</a> <a href="#">R01069</a> <a href="#">R01070</a> <a href="#">R01785</a> <a href="#">R01829</a> <a href="#">R02262</a> <a href="#">R02263</a> <a href="#">R02568</a> <a href="#">R04292</a> <a href="#">R05378</a> <a href="#">R05679</a> <a href="#">R05680</a>																					
<b>Pathway</b>	PATH: <a href="#">map00010</a> Glycolysis / Gluconeogenesis PATH: <a href="#">map00031</a> Inositol metabolism PATH: <a href="#">map00040</a> Pentose and glucuronate interconversions PATH: <a href="#">map00051</a> Fructose and mannose metabolism PATH: <a href="#">map00052</a> Galactose metabolism PATH: <a href="#">map00561</a> Glycerolipid metabolism PATH: <a href="#">map00564</a> Glycerophospholipid metabolism PATH: <a href="#">map00620</a> Pyruvate metabolism PATH: <a href="#">map00710</a> Carbon fixation PATH: <a href="#">map00760</a> Nicotinate and nicotinamide metabolism																					
<b>Enzyme</b>	<table border="0"> <tr> <td><a href="#">1.1.1.8</a></td> <td><a href="#">1.1.1.94</a></td> <td><a href="#">1.1.1.261</a></td> <td><a href="#">1.1.3.21</a></td> </tr> <tr> <td><a href="#">1.1.99.4</a></td> <td><a href="#">1.1.99.5</a></td> <td><a href="#">2.3.1.42</a></td> <td><a href="#">2.7.1.29</a></td> </tr> <tr> <td><a href="#">2.7.1.121</a></td> <td><a href="#">3.1.3.1</a></td> <td><a href="#">4.1.2.2</a></td> <td><a href="#">4.1.2.13</a></td> </tr> <tr> <td><a href="#">4.1.2.17</a></td> <td><a href="#">4.1.2.19</a></td> <td><a href="#">4.1.2.29</a></td> <td><a href="#">4.1.2.40</a></td> </tr> <tr> <td><a href="#">4.2.3.3</a></td> <td><a href="#">5.3.1.1</a></td> <td></td> <td></td> </tr> </table>		<a href="#">1.1.1.8</a>	<a href="#">1.1.1.94</a>	<a href="#">1.1.1.261</a>	<a href="#">1.1.3.21</a>	<a href="#">1.1.99.4</a>	<a href="#">1.1.99.5</a>	<a href="#">2.3.1.42</a>	<a href="#">2.7.1.29</a>	<a href="#">2.7.1.121</a>	<a href="#">3.1.3.1</a>	<a href="#">4.1.2.2</a>	<a href="#">4.1.2.13</a>	<a href="#">4.1.2.17</a>	<a href="#">4.1.2.19</a>	<a href="#">4.1.2.29</a>	<a href="#">4.1.2.40</a>	<a href="#">4.2.3.3</a>	<a href="#">5.3.1.1</a>		
<a href="#">1.1.1.8</a>	<a href="#">1.1.1.94</a>	<a href="#">1.1.1.261</a>	<a href="#">1.1.3.21</a>																			
<a href="#">1.1.99.4</a>	<a href="#">1.1.99.5</a>	<a href="#">2.3.1.42</a>	<a href="#">2.7.1.29</a>																			
<a href="#">2.7.1.121</a>	<a href="#">3.1.3.1</a>	<a href="#">4.1.2.2</a>	<a href="#">4.1.2.13</a>																			
<a href="#">4.1.2.17</a>	<a href="#">4.1.2.19</a>	<a href="#">4.1.2.29</a>	<a href="#">4.1.2.40</a>																			
<a href="#">4.2.3.3</a>	<a href="#">5.3.1.1</a>																					
<b>Other DBs</b>	PubChem: <a href="#">3411</a> ChEBI: <a href="#">16108</a> PDB-CCD: <a href="#">13P</a> 3DMET: <a href="#">B00029</a>																					
<b>LinkDB</b>	<a href="#">All DBs</a>																					
<b>KCF data</b>	<a href="#">Show</a>																					

=> [Original format](#)

DBGET integrated database retrieval system, GenomeNet

Name  
Formula  
Structure files (mol & KCF)

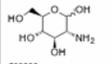
DB search button  
Display structure button  
(Jmol or KegDraw)

Link to other KEGG databases  
(REACTION, PATHWAY, ENZYME)

Link to other databases  
(PubChem, ChEBI, PDB,...)

DBGET Result: COMPOUND C00329 - Microsoft Internet Explorer

COMPONENT: C00329

Entry	C00329	Compound
Name	D-Glucosamine; Chitosamine; 2-Amino-2-deoxy-D-glucose	
Formula	C6H13NO5	
Mass	179.0794	
Structure		
Remark	Same as: R04224	
Reaction	R01200 R01204 R01961 R01962 R01963 R01964 R01965 R01966 R02431 R04225	
Pathway	FATH: map00530 Aminosugars metabolism FATH: map02060 Phosphotransferase system (PTS)	
Enzyme	2.3.1.3 2.7.1.1 2.7.1.8 2.7.1.49 3.2.1.- 3.5.1.33 3.10.1.1	
Other DBs	CAS: 3416-24-8 PubChem: 3623 ChEMBL: 17315	
LinkDB	All DBs	
KCF data	Show	

Buttons: **Jmol** **KegDraw**

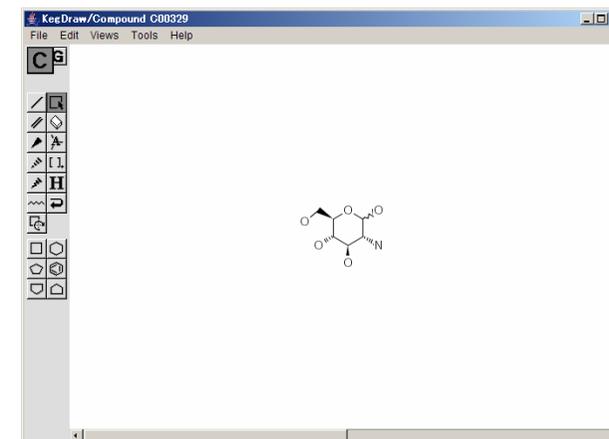
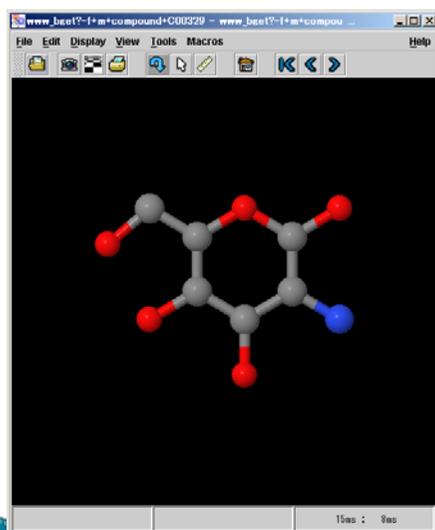
Java Web Start

 DBGET KegDraw  
Kanehisa Laboratories

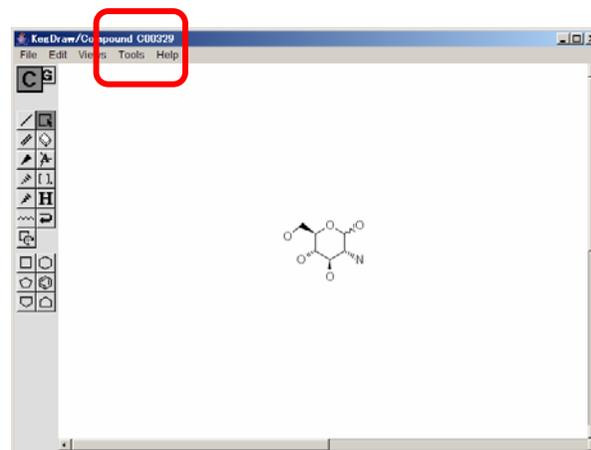
java.sun.com から JRE 1.4 を要求しています

取消し

When a user click the **Jmol** or **KegDraw** button, then the **Java Web Start** launched the **Jmol** or **KegDraw** program.

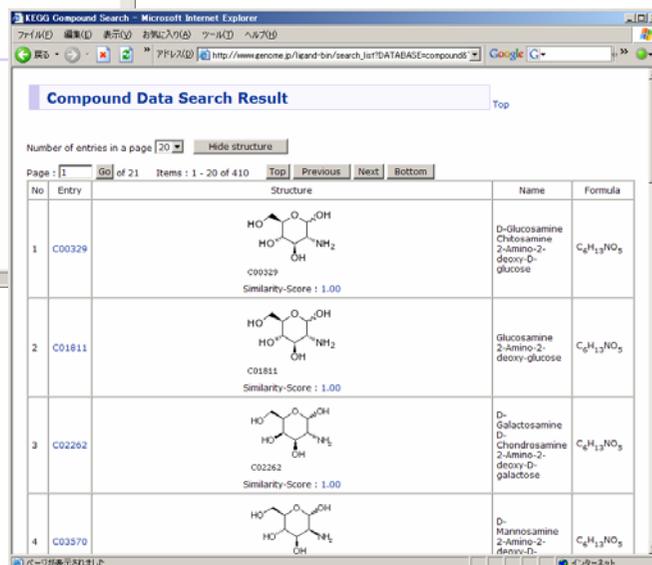
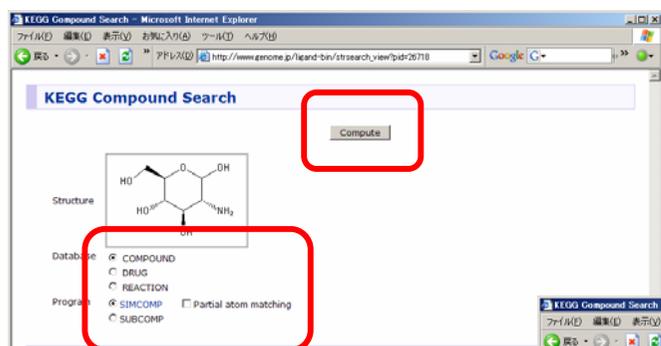


LIGAND (KEGG)

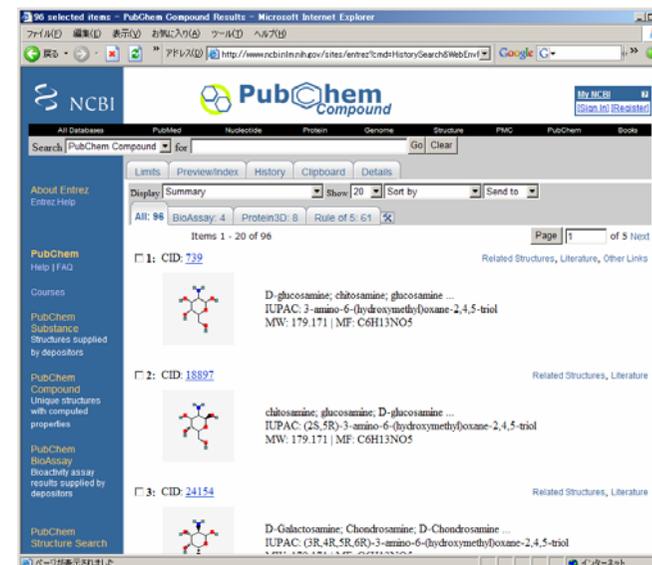


PubChem(NCBI)

A user can search the chemical structure from KegDraw's Tools menu.



No	Entry	Structure	Name	Formula
1	C00329	 C00329 Similarity-Score : 1.00	D-Glucosamine Chitosamine 2-Amino-2-deoxy-D-glucose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
2	C01811	 C01811 Similarity-Score : 1.00	Glucosamine 2-Amino-2-deoxy-glucose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
3	C02262	 C02262 Similarity-Score : 1.00	D-Galactosamine D-Chondrosamine 2-Amino-2-deoxy-D-galactose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
4	C03570	 C03570	D-Mannosamine 2-Amino-2-Deoxy-D-	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>



96 selected items - PubChem Compound Results - Microsoft Internet Explorer

PubChem Compound

Search PubChem Compound for [ ] Go Clear

Display Summary Show 20 Sort by [ ] Send to [ ]

All: 96 BioAssay: 4 Protein3D: 8 Rule of 5: 61

Items 1 - 20 of 96

Page 1 of 5 Next

Related Structures, Literature, Other Links

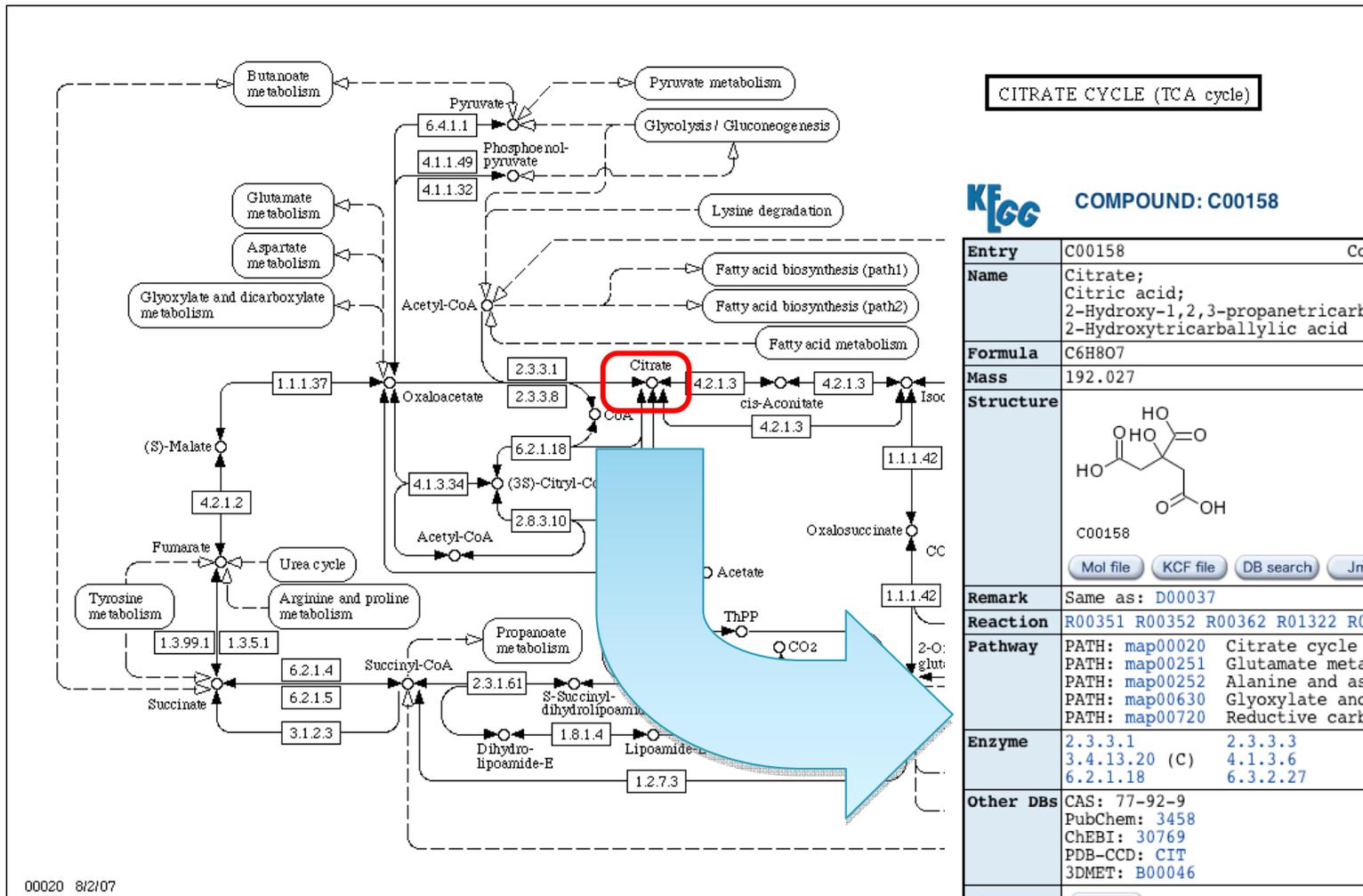
1: CID: 739  
D-glucosamine; chitosamine; glucosamine ...  
IUPAC: 3-amino-6-(hydroxymethyl)oxane-2,4,5-triol  
MW: 179.171 | MF: C6H13NO5

2: CID: 18897  
chitosamine; glucosamine; D-glucosamine ...  
IUPAC: (2S,5R)-3-amino-6-(hydroxymethyl)oxane-2,4,5-triol  
MW: 179.171 | MF: C6H13NO5

3: CID: 24154  
D-Galactosamine; Chondrosamine; D-Chondrosamine ...  
IUPAC: (3R,4R,5R,6R)-3-amino-6-(hydroxymethyl)oxane-2,4,5-triol

<http://www.genome.jp/download/>

# Pathway and Compound

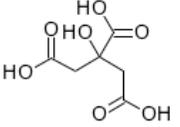


CITRATE CYCLE (TCA cycle)



COMPOUND: C00158

Help

<b>Entry</b>	C00158	Compound
<b>Name</b>	Citrate; Citric acid; 2-Hydroxy-1,2,3-propanetricarboxylic acid; 2-Hydroxytricarballic acid	
<b>Formula</b>	C <sub>6</sub> H <sub>8</sub> O <sub>7</sub>	
<b>Mass</b>	192.027	
<b>Structure</b>	 <p>C00158</p> <p><a href="#">Mol file</a> <a href="#">KCF file</a> <a href="#">DB search</a> <a href="#">Jmol</a> <a href="#">KegDraw</a></p>	
<b>Remark</b>	Same as: D00037	
<b>Reaction</b>	R00351 R00352 R00362 R01322 R01323 R01324 R01325 R04357	
<b>Pathway</b>	PATH: map00020 Citrate cycle (TCA cycle) PATH: map00251 Glutamate metabolism PATH: map00252 Alanine and aspartate metabolism PATH: map00630 Glyoxylate and dicarboxylate metabolism PATH: map00720 Reductive carboxylate cycle (CO <sub>2</sub> fixation)	
<b>Enzyme</b>	2.3.3.1 3.4.13.20 (C) 6.2.1.18	2.3.3.3 4.1.3.6 6.3.2.27
<b>Other DBs</b>	CAS: 77-92-9 PubChem: 3458 ChEBI: 30769 PDB-CCD: CIT 3DMET: B00046	
<b>LinkDB</b>	<a href="#">All DBs</a>	
<b>KCF data</b>	<a href="#">Show</a>	

=> Original format

DBGET integrated database retrieval system, GenomeNet

# LIGAND Database

- ▶ Components of LIGAND Database
  - COMPOUND      Chemical compound structure
  - DRUG            Drug structure
  - GLYCAN        Glycan structure
  - REACTION      Biochemical reaction
  - ENZYME        Enzyme nomenclature
  
  - RPAIR            Reaction pair alignment

# DRUG Database

- ▶ DRUG Database contains the **medicinal compounds**, that is, the compounds whose medicinal effects have been confirmed.
  - Each DRUG entry is also linked to **therapeutic classification number** (therapeutic categories) in KEGG BRITE database.
- ▶ Chemical structure formats are basically the same as the COMPOUND Database.
  - Structures are represented by 2D graph and stored in the MDL/mol and the KCF format.

KEGG LIGAND Database  
Molecular building blocks of life in the chemical space

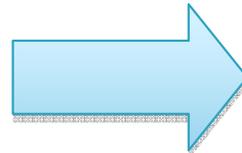
KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

Chemical Substances and Reactions

KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point
LIGAND	COMPOUND	C number	Chemical compound structures
	DRUG	D number	Drug structures
	GLYCAN	G number	Glycan structures
	REACTION	R number	Biochemical reactions
	RPAIR	A number	Reactant pair alignments
	ENZYME	EC number	Enzyme nomenclature

Search **LIGAND** for  Go Clear



KEGG LIGAND Database  
Molecular building blocks of life in the chemical space

KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

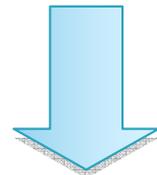
Chemical Substances and Reactions

KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point
LIGAND	COMPOUND	C number	Chemical compound structures
	DRUG	D number	Drug structures
	GLYCAN	G number	Glycan structures
	REACTION	R number	Biochemical reactions
	RPAIR	A number	Reactant pair alignments
	ENZYME	EC number	Enzyme nomenclature

Search **DRUG** for penicillin Go Clear

“penicillin”



Select “DRUG”

Search example of DRUG entry

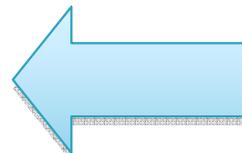
DDGET Result: DRUG D02336

KEGG DRUG: D02336

Entry	D02336	Drug
Name	Benzylpenicillin; Penicillin G	
Formula	C <sub>16</sub> H <sub>18</sub> N <sub>2</sub> O <sub>4</sub> S	
Mass	334.0967	
Structure		
Target	penicillin binding protein (PBP), cell wall synthesis inhibitor [KO:K02345] [PATH:ko00312]	
Remark	Same as: C05551	
Comment	semisynthetic penicillin: narrow spectrum penicillin	
Other DBs	CAS: 61-33-6 PubChem: 7849395	
LinkDB	All DBs	
KCF data	Show	

Original format

DDGET integrated database retrieval system, GenomeNet



DDGET Search Result: DRUG penicillin

Database: DRUG - Search term: penicillin (Total 10 hits)

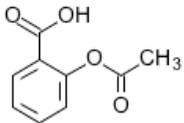
- Phenoxymethylpenicillin potassium (JAN); Penicillin V potassium (USP); Veetids (TN) D01053 [KegDraw](#) [Jmol](#)
- Benzylpenicillin potassium (JAN); Penicillin G potassium (USP); Pfizerpen (TN) D02157 [KegDraw](#) [Jmol](#)
- Phenoxymethylpenicillin hydrate (JP18); Penicillin G benzathine (USP); Bicillin L-A (TN) D02336 [KegDraw](#) [Jmol](#)
- Benzylpenicillin, Penicillin G D02405 [KegDraw](#) [Jmol](#)
- Penicillin V benzathine (USP) D02451 [KegDraw](#) [Jmol](#)
- Penicillin G procaine (USP); Duracillin (TN) D02506 [KegDraw](#) [Jmol](#)
- Phenoxymethylpenicillin benzathine (JAN); Penicillin V benzathine tetrahydrate D05400 [KegDraw](#) [Jmol](#)
- Penicillin G sodium (USP); Penicillin G sodium (TN) D05411 [KegDraw](#) [Jmol](#)
- Penicillin V (USP); V-Cillin (TN) D05413 [KegDraw](#) [Jmol](#)
- Penicillin V hydrabamine (USP)

DDGET integrated database retrieval system, GenomeNet

Entries started from “D”

# Example of DRUG Entry

**KEGG** DRUG: D00109 Help

Entry	D00109 <span style="border: 1px solid red; padding: 2px;">Drug</span>
Name	Aspirin (JP15/USP); Acetylsalicylic acid; Easprin (TN)
Formula	C9H8O4
Mass	180.0423
Structure	 <p>D00109</p> <div style="border: 1px solid red; padding: 2px; display: flex; gap: 5px;"> <span>Mol file</span> <span>KCF file</span> <span>DB search</span> <span>Jmol</span> <span>KegDraw</span> </div>
Target	<span style="border: 1px solid red; padding: 2px;">cyclooxygenase-1 (COX-1) inhibitor [HSA:5742] [EC:1.14.99.1]; cyclooxygenase-2 (COX-2) inhibitor [HSA:5743] [EC:1.14.99.1]</span>
Activity	<span style="border: 1px solid red; padding: 2px;">Analgesic; Antipyretic; Antirheumatic</span>
Remark	<span style="border: 1px solid red; padding: 2px;">Same as: <a href="#">C01405</a> Therapeutic category: <a href="#">1143</a> <a href="#">3399</a> ATC code: <a href="#">A01AD05</a> <a href="#">B01AC06</a> <a href="#">N02BA01</a> <a href="#">BRITE hierarchy</a></span>
Comment	Name previously used: Acetylsalicylic acid Component of Bufferin (TN), Percodan (TN), Darvon compound-65 (TN), E.A.C (TN)
Pathway	<span style="border: 1px solid red; padding: 2px;">PATH: <a href="#">map07110</a> Benzoic acid family PATH: <a href="#">map07219</a> Cyclooxygenase inhibitors</span>
Other DBs	<span style="border: 1px solid red; padding: 2px;">CAS: <a href="#">50-78-2</a> PubChem: <a href="#">7847177</a> ChEBI: <a href="#">15365</a> DrugBank: <a href="#">DB00945</a> PDB-CCD: <a href="#">AIN</a> DailyMed: <a href="#">aspirin</a> LigandBox: <a href="#">D00109</a></span>
LinkDB	<span>All DBs</span>
KCF data	<span>Show</span>

Difference of file format

Name  
Formula  
Structure files (mol & KCF)  
DB search button  
Display structure button  
(Jmol or KegDraw)

Target information  
Therapeutic category  
C entry with the same structure

Link to other PATHWAY  
Link to other databases

=> Original format

DBGET integrated database retrieval system, GenomeNet

[http://www.genome.jp/dbget-bin/www\\_bget?drug+D00109](http://www.genome.jp/dbget-bin/www_bget?drug+D00109)

# Pathway for Drug



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions

KEGG2 ATLAS PATHWAY BRITE GENES SSDB

### Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our molecular interaction and reaction networks for:

- 1. Metabolism**  
Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid  
Glycan PK/NRP Cofactor/vitamin Secondary metabolite Xenobiotics
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

### 6. Drug Development

### Pathway Modules

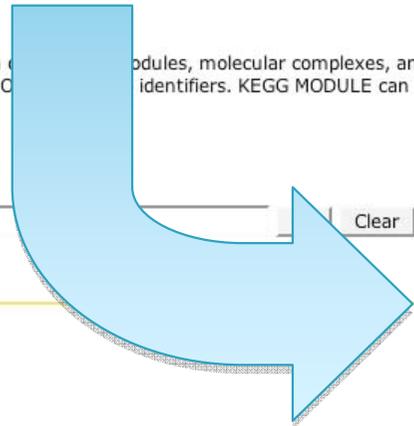
KEGG MODULE is a new collection of modules, molecular complexes, and each represented as a list of KEGG Compound Identifiers. KEGG MODULE can be BRITE hierarchy:

#### KEGG pathway modules

or by the DBGET search.

Search  for

bfind mode  bget mode



## 6. Drug Development

### 6.1 Chronology: Antibiotics

- Penicillins
- Cephalosporins - parenteral agents
- Cephalosporins - oral agents
- Aminoglycosides
- Tetracyclines
- Macrolides and ketolides
- Quinolones
- Rifamycins

### 6.2 Chronology: Antineoplastics

- Antineoplastics - alkylating agents
- Antineoplastics - antimetabolic agents
- Antineoplastics - agents from natural products
- Antineoplastics - hormones
- Antineoplastics - protein kinases inhibitors

### 6.3 Chronology: Nervous System Agents

- Hypnotics
- Anxiolytics
- Anticonvulsants
- Local analgesics
- Opioid analgesics
- Antipsychotics
- Antipsychotics - phenothiazines
- Antipsychotics - butyrophenones
- Antidepressants

### 6.3 Chronology: Other Drugs

- Azoles
- Quinolines
- Statins
- Sulfonamide derivatives - sulfa drugs
- Sulfonamide derivatives - diuretics
- Sulfonamide derivatives - hypoglycemic agents
- Eicosanoids
- Prostaglandins
- Antiarrhythmic drugs
- Antacids
- Antiviral
- Immunosuppressive agents

### 6.5 Target Based Structure Classification

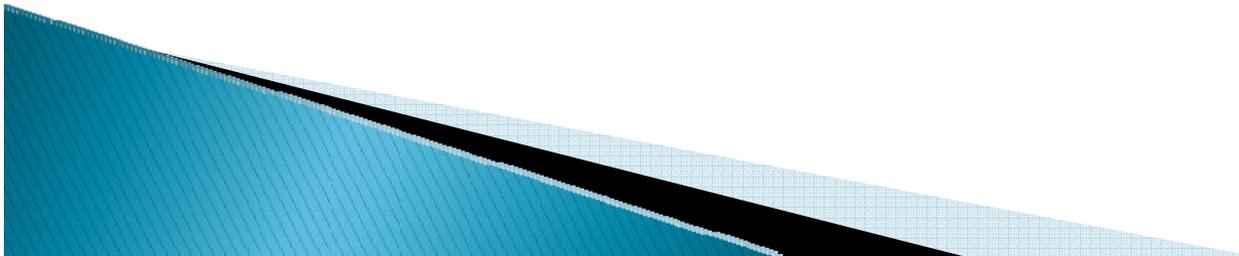
- Serotonin receptor agonists/antagonists
- Dopamine receptor agonists/antagonists
- Histamine receptor antagonists
- beta-Adrenergic receptor agonists/antagonists
- alpha-Adrenergic receptor agonists/antagonists
- Calcium channel blocking agents
- Catecholamine transferase inhibitors
- Angiotensin antagonists

KEGG DRUG

Therapeutic category of drugs  
+ Target  
+ Map number

# Drug Structure Map

- ▶ DRUG structure maps represent our knowledge on **drug development**, that is, the chronology of drug structures.
- ▶ Each map is **graphically** illustrated in a manner similar to KEGG pathway maps.
- ▶ DRUG structure maps are one of categories of the KEGG PATHWAY database.

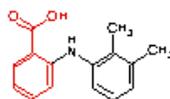


BENZOIC ACID FAMILY

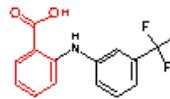
Anthranilic acid derivatives



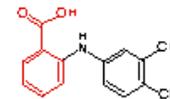
Anthranilic acid



Mefenamic acid

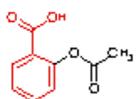


Flufenamic acid

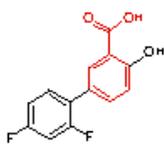


Tolfenamic acid

Salicylic acid derivatives



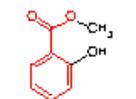
Aspirin



Diflunisal

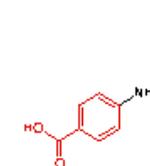


Salicylic acid

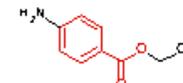


Methyl salicylate

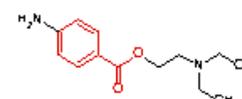
Aminobenzoic acid derivatives



Aminobenzoic acid



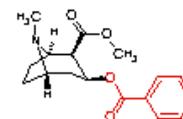
Benzocaine



Procaine



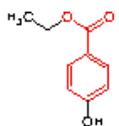
Benzoic acid



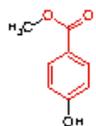
Cocaine

Local analgesic

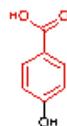
p-Hydroxybenzoic acid derivatives



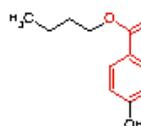
Ethylparaben



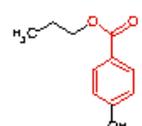
Methylparaben



Hydroxybenzoic acid

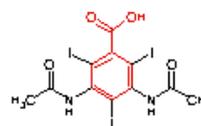


Butylparaben

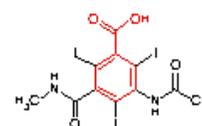


Propylparaben

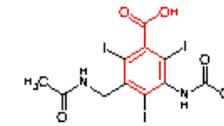
Iodobenzoic acid derivatives



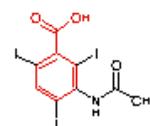
Amidotrizoic acid



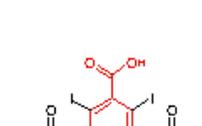
Iotalamic acid



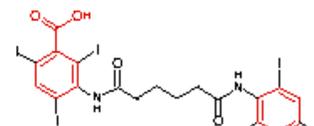
Iodamide



Acetrizoic acid



Metrizoic acid



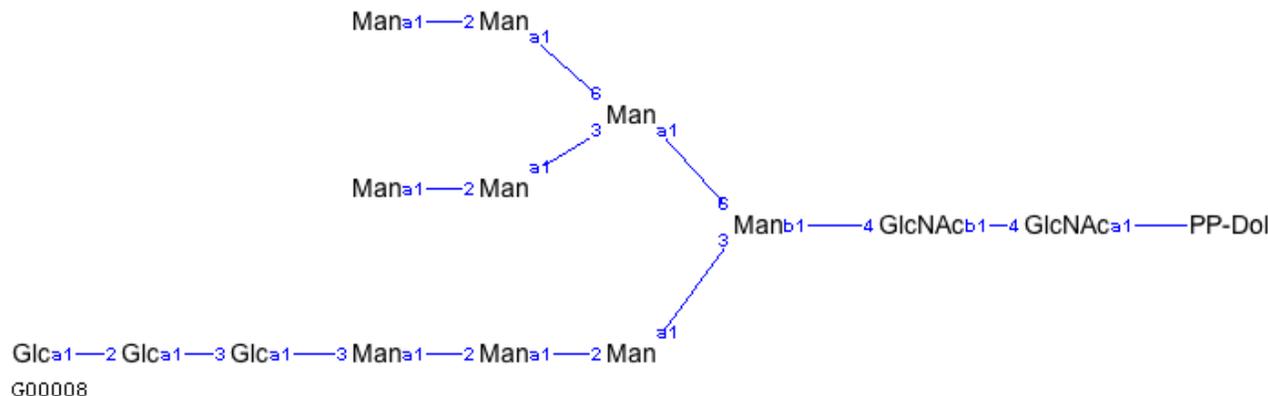
Iodipamide

# LIGAND Database

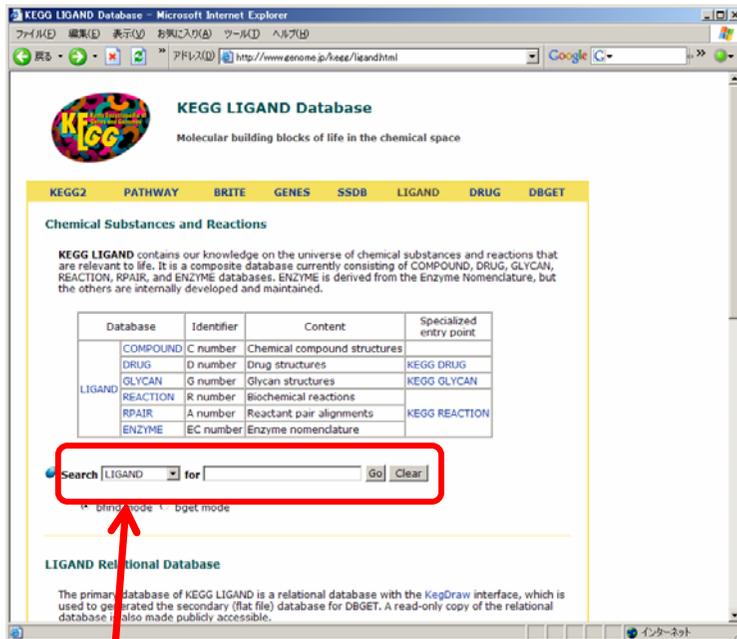
- ▶ Components of LIGAND Database
  - COMPOUND      Chemical compound structure
  - DRUG            Drug structure
  - GLYCAN        Glycan structure
  - REACTION      Biochemical reaction
  - ENZYME        Enzyme nomenclature
  
  - RPAIR            Reaction pair alignment

# GLYCAN Database

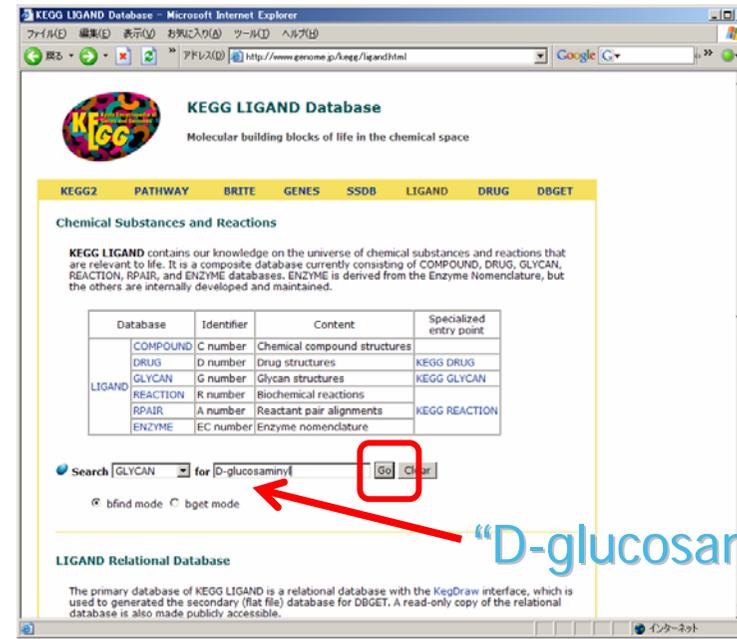
- ▶ Glycan structures have been collected.
- ▶ Each structure is described as **tree**,
  - whose nodes are one of monosaccharides.
  - A tree is a kind of a graph, without ring structures.



- ▶ The structural format is the KCF for Glycan.

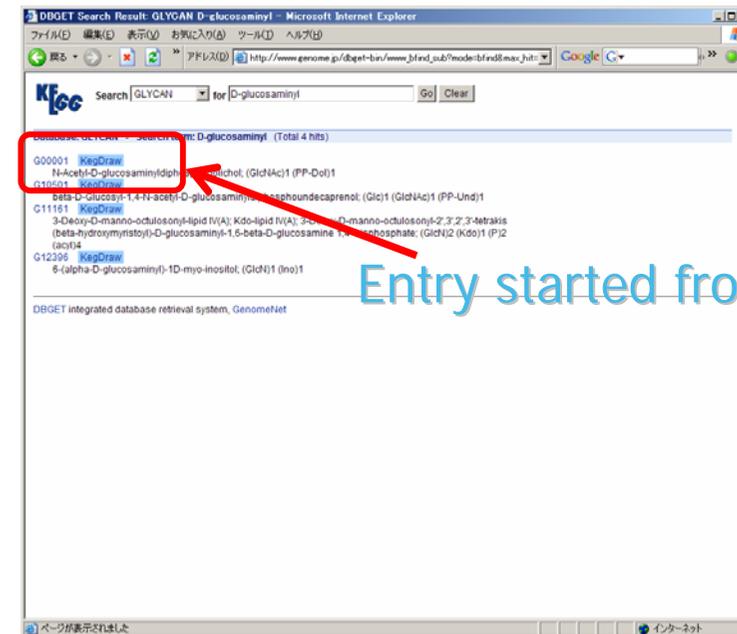
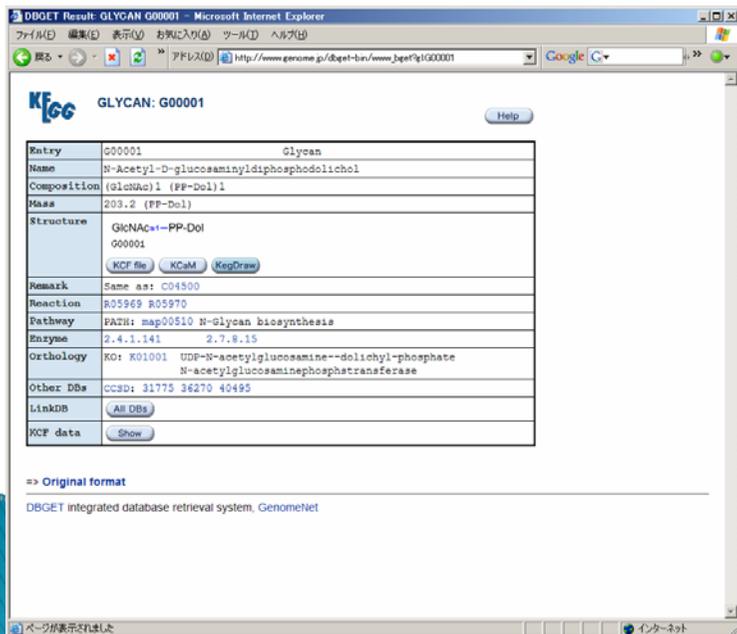


Select "GLYCAN"



"D-glucosaminyl"

Search example of GLYCAN entry



Entry started from "G"

# Example of GLYCAN Entry

**KEGG** GLYCAN: G00008 Help

<b>Entry</b>	G000008 Glycan
<b>Composition</b>	(Glc)3 (GlcNAc)2 (Man)9 (PP-Dol)1
<b>Mass</b>	2352.1 (PP-Dol)
<b>Structure</b>	
<b>Reaction</b>	<a href="#">R05976</a> <a href="#">R06264</a>
<b>Pathway</b>	PATH: <a href="#">map00510</a> N-Glycan biosynthesis
<b>Enzyme</b>	<a href="#">2.4.1.119</a> <a href="#">2.4.1.-</a>
<b>Orthology</b>	KO: <a href="#">K00730</a> dolichyl-diphosphooligosaccharide--protein glycosyltransferase
<b>Other DBs</b>	CCSD: <a href="#">4532</a> <a href="#">6023</a> <a href="#">6993</a> <a href="#">6996</a> <a href="#">6997</a> <a href="#">6999</a> <a href="#">7000</a> <a href="#">7003</a> <a href="#">7005</a> <a href="#">7006</a> <a href="#">7007</a> <a href="#">7012</a> <a href="#">7014</a> <a href="#">7015</a> <a href="#">7018</a> <a href="#">7019</a> <a href="#">7020</a> <a href="#">11900</a> <a href="#">11901</a> <a href="#">27483</a> <a href="#">28887</a> <a href="#">31272</a> <a href="#">31757</a> <a href="#">31774</a> <a href="#">32762</a> <a href="#">32815</a> <a href="#">33845</a> <a href="#">33864</a> <a href="#">37071</a> <a href="#">38410</a> <a href="#">38474</a> <a href="#">47089</a>
<b>LinkDB</b>	<a href="#">All DBs</a>
<b>KCF data</b>	<a href="#">Show</a>

Structure formula (composition)  
 Structure files (KCF)  
 Structure search  
 Display button (KegDraw)

Link to KEGG databases  
 (Reaction, Pathway, Enzyme, KO,  
 C entry with the same structure)

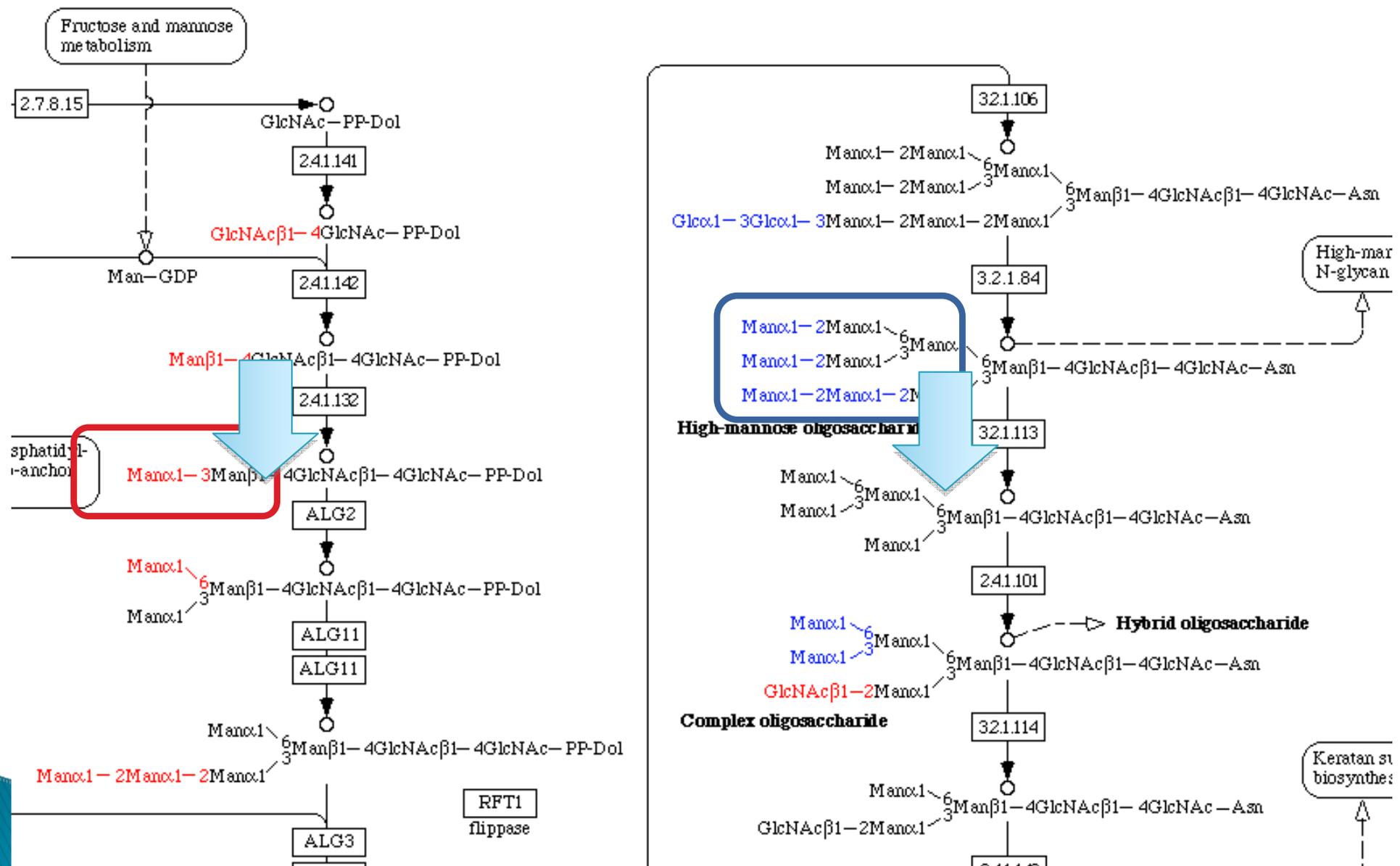
Link to other databases  
 (CCSD)

=> [Original format](#)

DBGET integrated database retrieval system, [GenomeNet](#)



# Pathway for Glycan

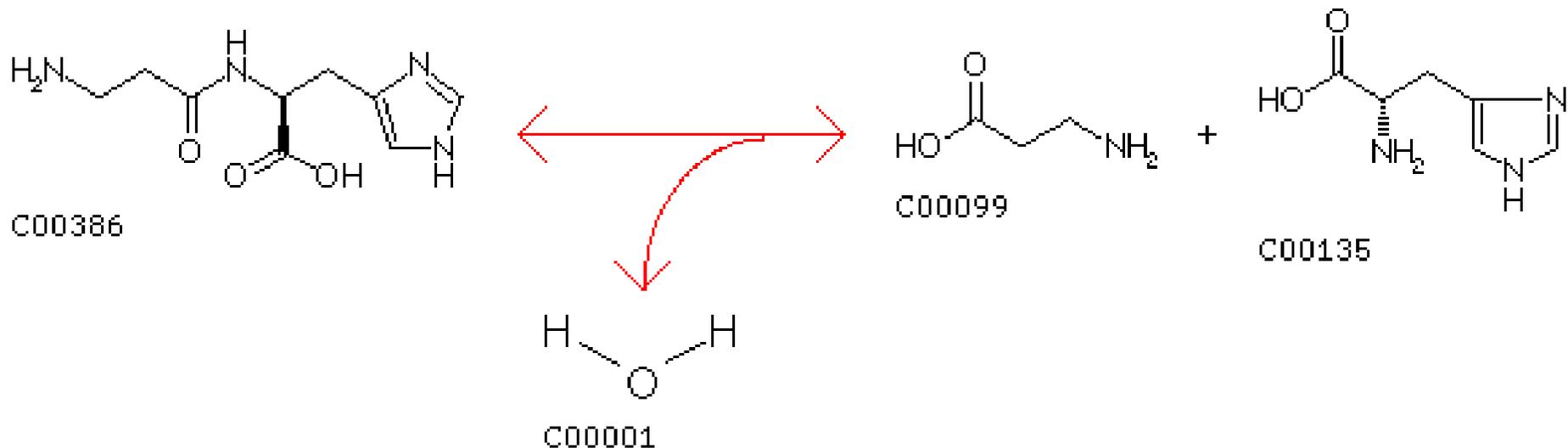


# LIGAND Database

- ▶ Components of LIGAND Database
  - COMPOUND      Chemical compound structure
  - DRUG            Drug structure
  - GLYCAN        Glycan structure
  - REACTION      Biochemical reaction
  - ENZYME        Enzyme nomenclature
  - RPAIR           Reaction pair alignment

# REACTION Database

- ▶ REACTION contains the biochemical reaction information.
  - They are mainly **enzymatic reactions** in metabolism.
- ▶ Information about **binary relations** between reactants are also maintained. (see RPAIR)



KEGG LIGAND Database

Molecular building blocks of life in the chemical space

KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

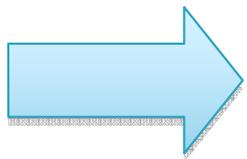
Chemical Substances and Reactions

KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point
LIGAND	COMPOUND	C number	Chemical compound structures
	DRUG	D number	Drug structures
	GLYCAN	G number	Glycan structures
	REACTION	R number	Biochemical reactions
	RPAIR	A number	Reactant pair alignments
	ENZYME	EC number	Enzyme nomenclature

Search LIGAND for [ ] Go Clear

Select "REACTION"



KEGG LIGAND Database

Molecular building blocks of life in the chemical space

KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

Chemical Substances and Reactions

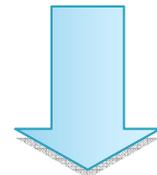
KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point
LIGAND	COMPOUND	C number	Chemical compound structures
	DRUG	D number	Drug structures
	GLYCAN	G number	Glycan structures
	REACTION	R number	Biochemical reactions
	RPAIR	A number	Reactant pair alignments
	ENZYME	EC number	Enzyme nomenclature

Search REACTION for dehydrogenase Go Clear

"dehydrogenase"

# Search example of REACTION entry



DBGET Result: REACTION R05643

REACTION: R05643

Entry	R05643	Reaction
Name	2-carboxybenzaldehyde dehydrogenase	
Definition	2-Carboxybenzaldehyde + NAD+ + H2O <=> Phthalate + NADH + H+	
Equation	C03057 + C00003 + C00001 <=> C01606 + C00004 + C00080	

Comment: PhdR

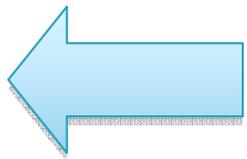
RPair: RP: A00002 C00003 C00004 oofao; RP: A05249 C01606\_C03057 main; RP: A06092 C00001\_C01606 leave

Pathway: PATH: rn00626 Naphthalene and anthracene degradation

Enzyme: 1.2.1.-

Orthology: KO: R00155

LinkDB: [All DBs]



DBGET Search Result: REACTION dehydrogenase

Search [REACTION] for [dehydrogenase] Go Clear

Database: REACTION - Search term: dehydrogenase (Total 21 hits)

R00539 Alkan-1-ol dehydrogenase (acceptor). Primary alcohol + Acceptor <=> Aldehyde + Reduced acceptor

R03449 ATP [pyruvate dehydrogenase (lipoamide)] phosphotransferase: ATP + [Pyruvate dehydrogenase (acetyl-transferring)] <=> ADP + [Pyruvate dehydrogenase (acetyl-transferring)] phosphate

R03450 [Pyruvate dehydrogenase (lipoamide)]-phosphate phosphohydrolase: [Pyruvate dehydrogenase (acetyl-transferring)] phosphate + H2O <=> [Pyruvate dehydrogenase (acetyl-transferring)] + Orthophosphate

R03515 [3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)]-phosphate phosphohydrolase: [3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)] phosphate + H2O <=> [3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)] + Orthophosphate

R03516 ATP [3-methyl-2-oxobutanoate dehydrogenase (lipoamide)] phosphotransferase: ATP + [3-methyl-2-oxobutanoate dehydrogenase (lipoamide)] <=> ADP + [3-Methyl-2-oxobutanoate dehydrogenase (lipoamide)] phosphate

R03732 (2S)-2-[(1-R)-carboxyethyl]amino]pentanoate dehydrogenase (NAD+; L-aminopentanoate-forming): (2S)-2-[(1-R)-Carboxyethyl]amino]pentanoate + NAD+ + H2O <=> L-Norvaline + Phosphate + NADH + H+

R04039 [xanthine dehydrogenase] glutathione-disulfide S-oxidoreductase: [xanthine dehydrogenase] + Glutathione disulfide <=> [xanthine oxidase] + 2 Glutathione

R04401 [isocitrate dehydrogenase (NADP+)] phosphotransferase: ATP + [isocitrate dehydrogenase (NADP+)] <=> ADP + [isocitrate dehydrogenase (NADP+)] phosphate

R05643 2-carboxybenzaldehyde dehydrogenase: 2-Carboxybenzaldehyde + NAD+ + H2O <=> Phthalate + NADH + H+

R05689 NAD-dependent threonine 4-phosphate dehydrogenase: O-Phospho-4-hydroxy-L-threonine + NAD+ <=> 2-Amino-3-oxo-4-phosphonooxybutyrate + NADH + H+

R05689 2-(2-(R)-Hydroxypropylthio)ethanesulfonate dehydrogenase: (R)-2-Hydroxypropyl-CoM + NAD+ <=> 2-Oxopropyl-CoM + NADH + H+

R05690 2-(2-(S)-Hydroxypropylthio)ethanesulfonate dehydrogenase: (S)-2-Hydroxypropyl-CoM + NAD+ <=> 2-Oxopropyl-CoM + NADH + H+

R06401 [4-oxo-3-oxopentanoate] dehydrogenase: [4-oxo-3-oxopentanoate] + NAD+ + H2O <=> [4-oxo-3-oxopentanoate] + NADH + H+

Entries start from "R"

# Example of REACTION Entry

**KEGG** REACTION: R01900 Help

<b>Entry</b>	R01900	Reaction
<b>Name</b>	isocitrate hydro-lyase	
<b>Definition</b>	Isocitrate <=> cis-Aconitate + H2O	
<b>Equation</b>	C00311 <=> C00417 + C00001	
<b>RPair</b>	RP: <a href="#">A01830</a> C00311_C00417 main RP: <a href="#">A06758</a> C00001_C00311 leave	
<b>Pathway</b>	PATH: <a href="#">rn00020</a> Citrate cycle (TCA cycle) PATH: <a href="#">rn00630</a> Glyoxylate and dicarboxylate metabolism PATH: <a href="#">rn00720</a> Reductive carboxylate cycle (CO2 fixation)	
<b>Enzyme</b>	4.2.1.3	
<b>Orthology</b>	KO: <a href="#">K01680</a> aconitate hydratase KO: <a href="#">K01681</a> aconitate hydratase 1 KO: <a href="#">K01682</a> aconitate hydratase 2	
<b>LinkDB</b>	<a href="#">All DBs</a>	

=> [Original format](#)

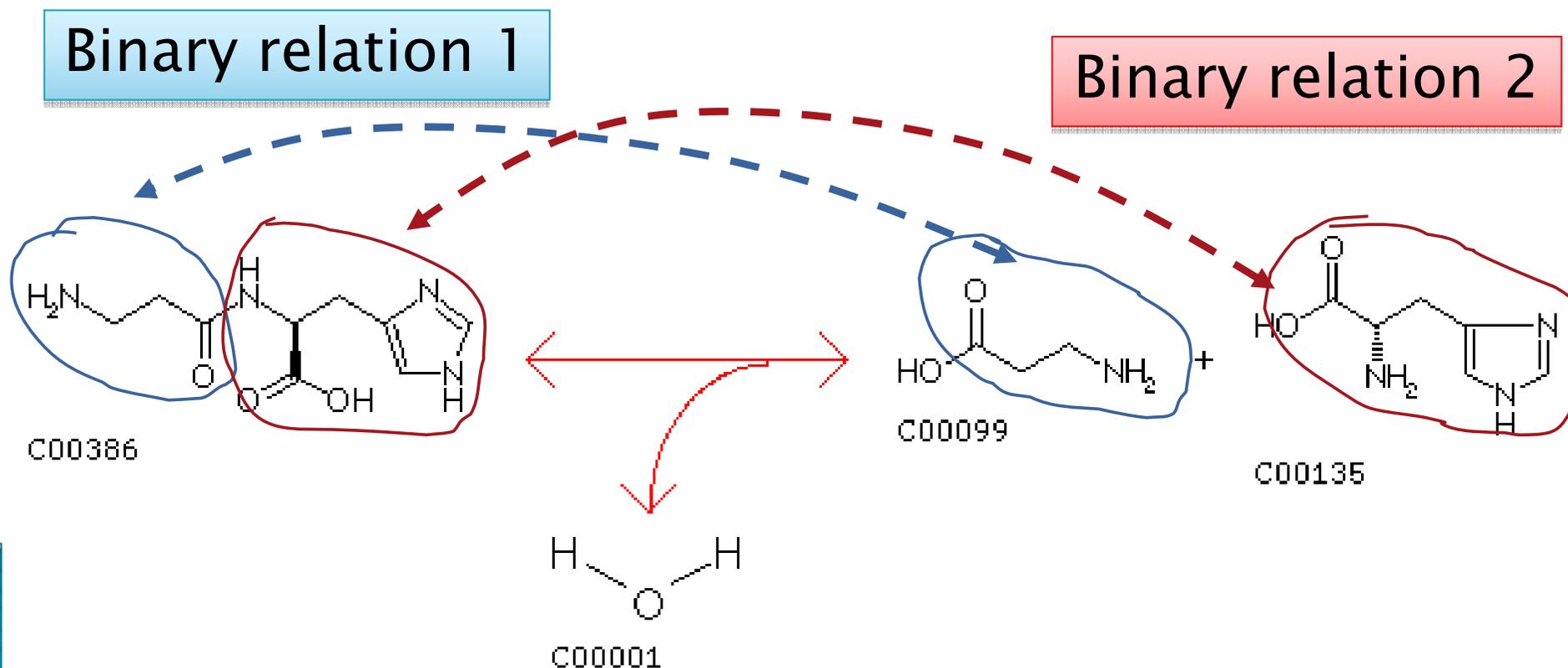
DBGET integrated database retrieval system, [GenomeNet](#)

Reaction formula  
with metabolite names  
with C numbers  
with chemical structures

Link to other KEGG databases  
(RPAIR, PATHWAY, ENZYME,  
KO)

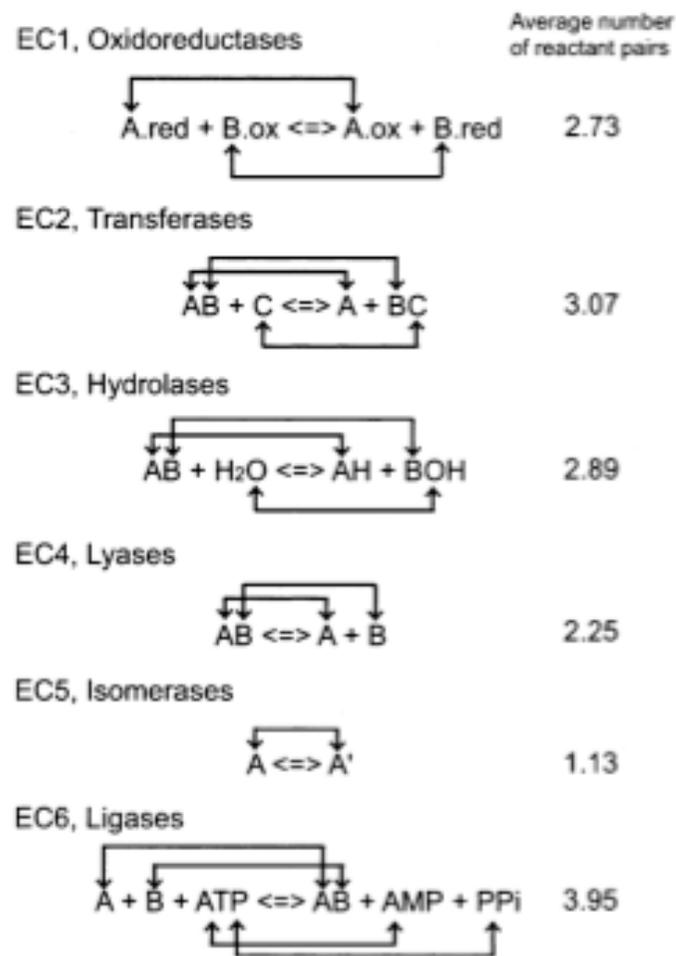
# What is a Binary Relation?

- ▶ The binary relation is defined as a pair of compounds that have atoms or atom groups in common before and after the reaction.



# Typical Relations in Reactions

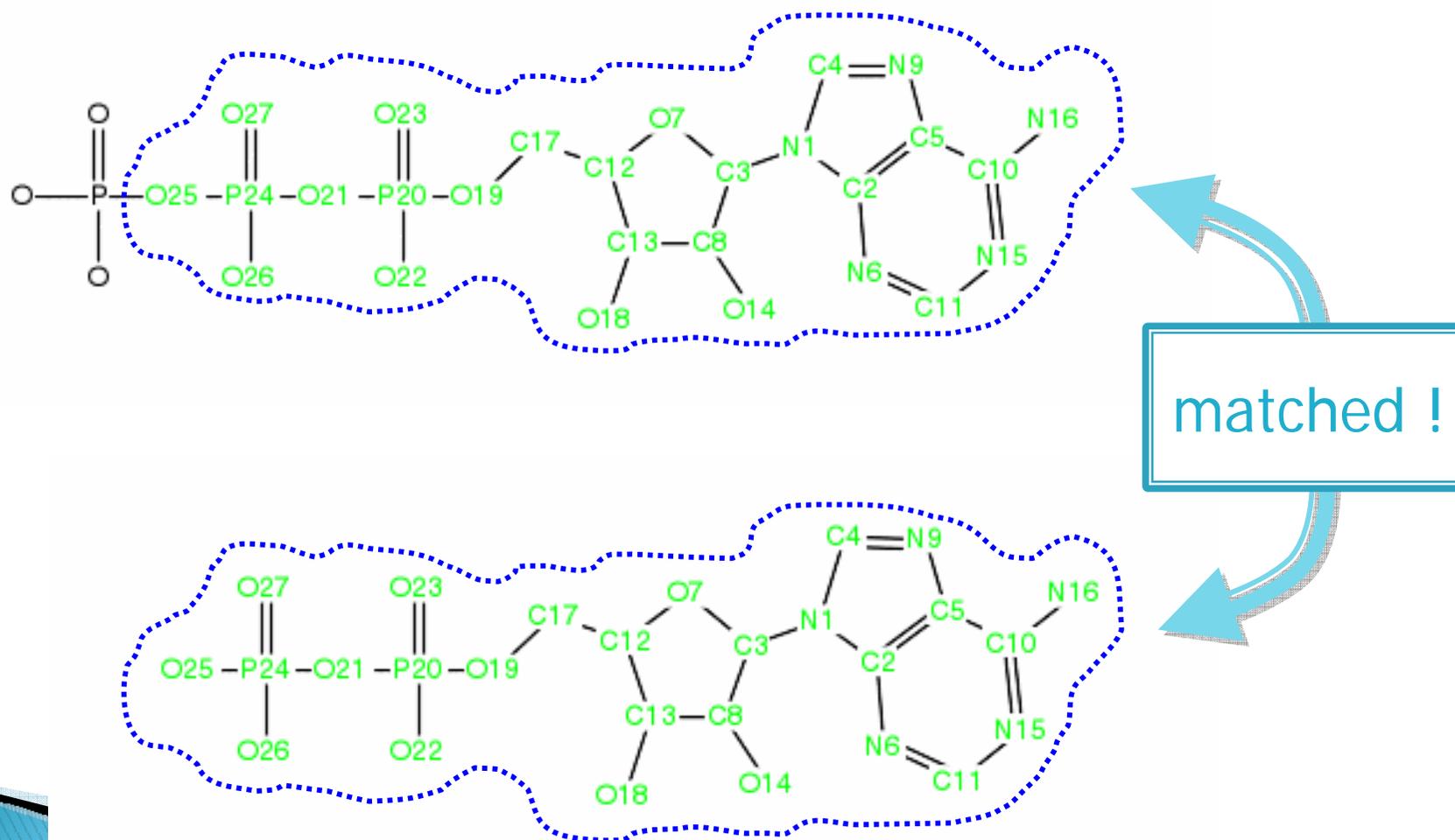
- ▶ Most of enzymatic reactions consist of several and **typical** binary relations.
- ▶ We can assign the relationship of compounds according to such **biochemical** knowledge.



**Figure 1.** Extraction of reactant pairs from reactions of the six EC classes. Reactions in each class show unique topology of the reactant pairs, resulting in the average number of reactant pairs shown on the right.

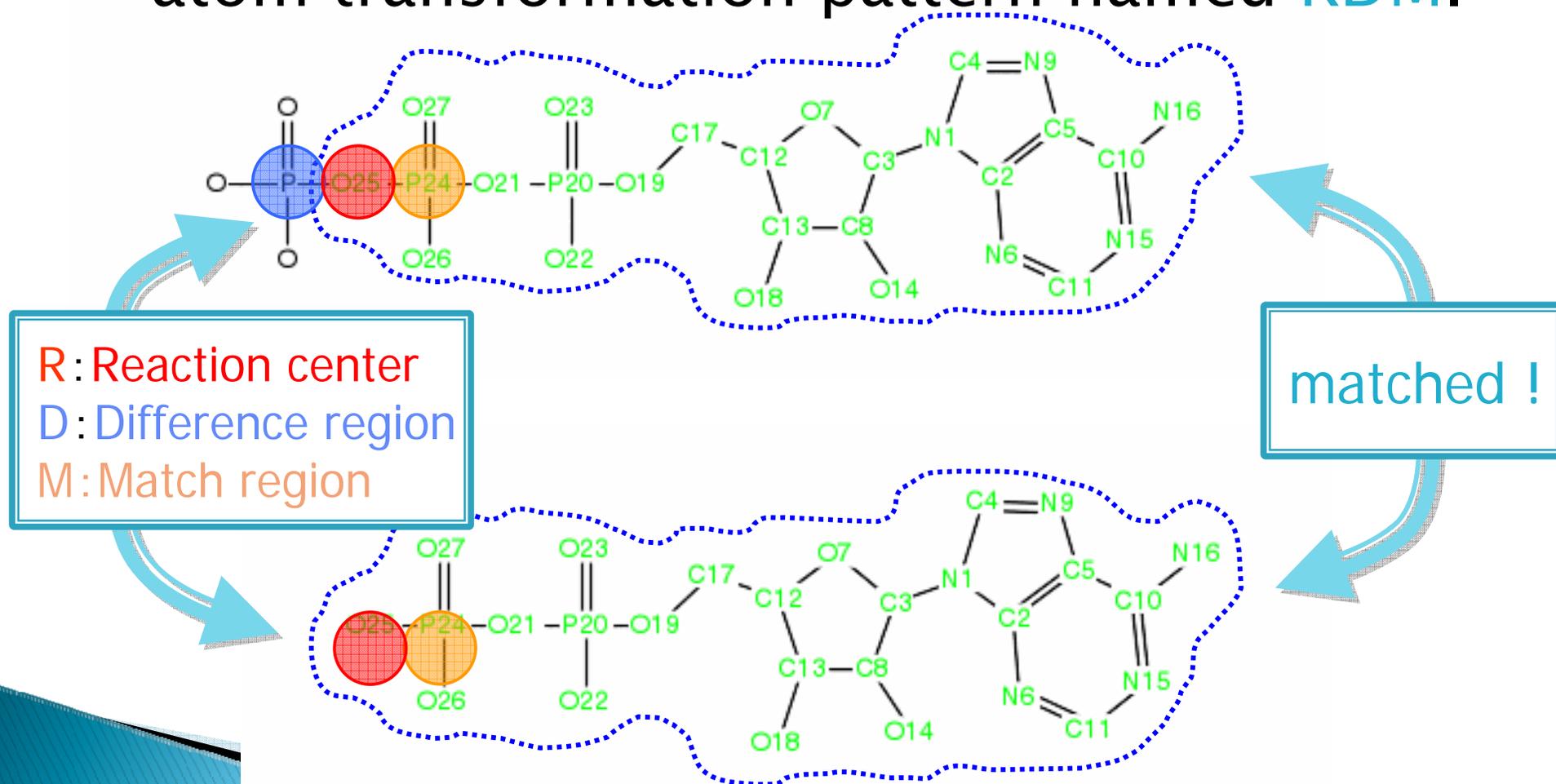
# From Binary Relations,

- ▶ we can compute the **atom alignments**.



# From the Computed Alignment,

- ▶ we can extract the **reaction mechanism** as the atom transformation pattern named **RDM**.



# LIGAND Database

- ▶ Components of LIGAND Database
  - COMPOUND      Chemical compound structure
  - DRUG            Drug structure
  - GLYCAN        Glycan structure
  - REACTION      Biochemical reaction
  - ENZYME        Enzyme nomenclature
- RPAIR            Reaction pair alignment

KEGG LIGAND Database  
Molecular building blocks of life in the chemical space

KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

Chemical Substances and Reactions

KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

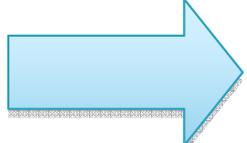
Database	Identifier	Content	Specialized entry point
COMPOUND	C number	Chemical compound structures	
DRUG	D number	Drug structures	KEGG DRUG
GLYCAN	G number	Glycan structures	KEGG GLYCAN
REACTION	R number	Biochemical reactions	
RPAIR	A number	Reactant pair alignments	KEGG REACTION
ENZYME	EC number	Enzyme nomenclature	

Search LIGAND for  Go Clear

blind mode bget mode

LIGAND Relational Database

The primary database of KEGG LIGAND is a relational database with the KegDraw interface, which is used to generate the secondary (flat file) database for DBGET. A read-only copy of the relational database is also made publicly accessible.



KEGG LIGAND Database  
Molecular building blocks of life in the chemical space

KEGG2 PATHWAY BRITE GENES SSDB LIGAND DRUG DBGET

Chemical Substances and Reactions

KEGG LIGAND contains our knowledge on the universe of chemical substances and reactions that are relevant to life. It is a composite database currently consisting of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME databases. ENZYME is derived from the Enzyme Nomenclature, but the others are internally developed and maintained.

Database	Identifier	Content	Specialized entry point
COMPOUND	C number	Chemical compound structures	
DRUG	D number	Drug structures	KEGG DRUG
GLYCAN	G number	Glycan structures	KEGG GLYCAN
REACTION	R number	Biochemical reactions	
RPAIR	A number	Reactant pair alignments	KEGG REACTION
ENZYME	EC number	Enzyme nomenclature	

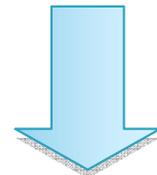
Search RPAIR for A00048 Go Clear

blind mode bget mode

LIGAND Relational Database

The primary database of KEGG LIGAND is a relational database with the KegDraw interface, which is used to generate the secondary (flat file) database for DBGET. A read-only copy of the relational database is also made publicly accessible.

"A00048"



Select "RPAIR"

Search example of RPAIR entry

DBGET Result: RPAIR A00048

ReactantPair

Entry	A00048
Name	C00088_C00244

Compound: C00088 Nitrite  
C00244 Nitrate

Type: main leave

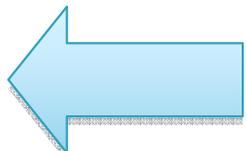
RDM: 1  
1 N2b-N2b:--O3a:O1b+O3a-O1b+O3a

Reaction: R00791 R00792 R00794 R00796 R00798 R00800 R01106 R02165 R03071

KEG data Show

=> Original format

DBGET integrated database retrieval system, GenomeNet



DBGET Search Result: RPAIR A00048

Search RPAIR for A00048 Go Clear

Database RPAIR Search term: A00048 (Total 1 hit)

A00048  
C00088\_C00244

DBGET integrated database retrieval system, GenomeNet

Entry started from "A"

# Example of REACTION Entry

KfGC RPAIR: A04458 Help

Entry	A04458	ReactantPair
Name	C00025_C00624	
Compound	C00025 L-Glutamate C00624 N-Acetyl-L-glutamate	
Type	main trans	
RDM	1 1 N1a-N1b:*-C5a:C1c-C1c	
Related pair	A00245 A00258 A00269 A00279 A00335 A00423 A00425 A00428 A00431 A00508 A00611 A00824 A00863 A00864 A00881 A00975 A00991 A01038 A01042 A01057 A01071 A01130 A01172 A01186 A01300 A01330 A01331 A01338 A01391 A01607 A01668 A01931 A02027 A02036 A02037 A02039 A02060 A02199 A02248 A02249 A02291 A02319 A02446 A02448 A02494 A02496 A02498 A02500 A02714 A02828 A03482 A03520 A03533 A03556 A03587 A03738 A03793 A03874 A03875 A03947 A03994 A04010 A04217 A04219 A04240 A04262 A04266 A04322 A04409 A04457 A04483 A04484 A04485 A04539 A04559 A04641 A04795 A04802 A05279 A05285 A05371 A07758 A08213 A09150 A09214	
Reaction	R00259 R02282	
KCF data	<input type="button" value="Show"/>	

Alignment result

Character string notation  
of RDM

Link to other databases  
(COMPOUND, REACTION)

Alignment result (KCF)

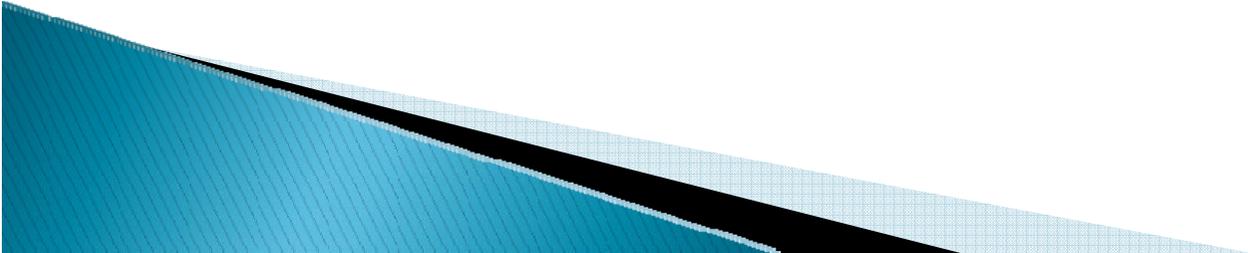
=> [Original format](#)

DBGET integrated database retrieval system, GenomeNet

# LIGAND Database

- ▶ Components of LIGAND Database
  - COMPOUND      Chemical compound structure
  - DRUG            Drug structure
  - GLYCAN        Glycan structure
  - REACTION      Biochemical reaction
  - ENZYME        Enzyme nomenclature
  - RPAIR            Reaction pair alignment

# ENZYME Database

- ▶ Traditional enzyme database by [IUBMB](#).
    - LIGAND ENZYME are compiled from the original enzyme database and linked to lots of databases (KEGG databases and other databases).
  - ▶ EC number is necessary for connecting the enzyme proteins (gene encoded information) to metabolic compounds or metabolic reactions (chemical information).
- 

# Chemical Computational Tools

## Computational Tools

### ● Search similar compound structures

**SIMCOMP**: maximal common subgraph search -- a portion of the query compound is optimally matched to a portion of the database compound [[reference](#)]

**SUBCOMP**: isomorphic subgraph search -- the query compound is fully matched to a portion of the database compound

### ● Search similar glycan structures

**KCaM**: local or global search for matching tree structures [[reference](#)]

### ● Predict reactions for given reactant pairs

**e-zyme**: automatic assignment of EC numbers [[reference](#)]

### ● Generate possible reaction paths

**PathComp**: generate possible reaction paths from an enzyme list

*Last updated: March 21, 2008*

[Feedback](#)

[KEGG](#)

[GenomeNet](#)

Kanehisa Laboratories

<http://www.genome.jp/kegg/ligand.html>

# Provided Tools in KEGG

- ▶ Computation of chemical compound graphs

- SIMCOMP to get the maximal common subgraph
- SUBCOMP to solve the subgraph isomorphism

- ▶ Computation of glycan structures

- KCaM Developed by Kiyoko Aoki-Kinoshita.

- ▶ Desktop application for compounds and glycans

- KegDraw Java based software to input structures

- ▶ Automatic assignment of EC number

- e-zyme to predict EC numbers for a set of pairs

# About SIMCOMP

- ▶ SIMCOMP can detect the **maximal common subgraph** between two chemical compounds.
- ▶ The algorithm is based on finding the **maximal clique** within the docking graph (association graph) defined by two original graphs.
- ▶ This problem is NP-hard and very time consuming.

# SIMCOMP: Start page



**SIMCOMP**  
for chemical structure search

Home Help

**About SIMCOMP**

SIMCOMP (SIMilar COMpound) search is a graph-based method for comparing chemical structures. It has been implemented in the KEGG system for searching against chemical structure databases.

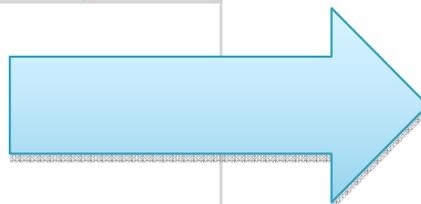
**Input form**

To run the query against KEGG use the following form:

**SIMCOMP search**

*Last updated: April 1, 2008*

KEGG GenomeNet Kyoto University Bioinformatics Center



## KEGG Compound Search

Enter query compound: (in one of the four forms)

Compound ID  (Example) C00022

MOL File Name

MOL File Text

SMILES

**Select target database:**

COMPOUND  DRUG  REACTION

**Select program and option:**

SIMCOMP  Partial atom matching  
 SUBCOMP

[ [LIGAND](#) | [KEGG](#) ]

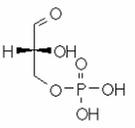
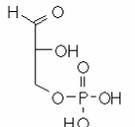
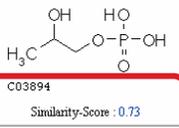
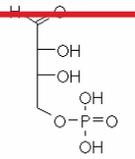
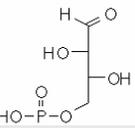
# SIMCOMP: Result example

KEGG Compound Search - Microsoft Internet Explorer

Compound Data Search Result [Top](#)

Number of entries in a page: 20

Page: 1 of 6 Items: 1 - 20 of 116

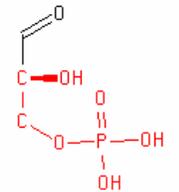
No	Entry	Structure	Name	Formula
1	C00118	 C00118 Similarity-Score : 1.00	(2R)-2-Hydroxy-3-(phosphonoxy)propanal D-Glyceraldehyde 3-phosphate	C <sub>3</sub> H <sub>7</sub> O <sub>6</sub> P
2	C00661	 C00661 Similarity-Score : 1.00	Glyceraldehyde 3-phosphate	C <sub>3</sub> H <sub>7</sub> O <sub>6</sub> P
3	C03894	 C03894 Similarity-Score : 0.73	Propane-1,2-diol 1-phosphate	C <sub>3</sub> H <sub>7</sub> O <sub>5</sub> P
4	C00279	 C00279 Similarity-Score : 0.69	D-Erythrose 4-phosphate	C <sub>4</sub> H <sub>9</sub> O <sub>7</sub> P
5	C03109	 C03109 Similarity-Score : 0.69	D-Threose 4-phosphate	C <sub>4</sub> H <sub>9</sub> O <sub>7</sub> P

ページが表示されました インターネット

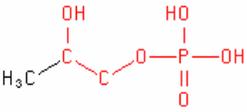
KEGG Compound Search - Microsoft Internet Explorer

Similarity-Score : 0.73

Query :



Entry : C03894

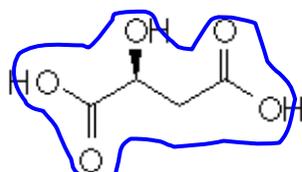


ページが表示されました インターネット

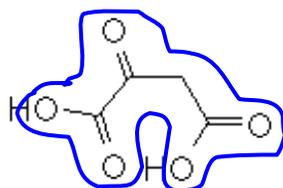
# Similarity Score in SIMCOMP

$$JC(G_1, G_2) \equiv \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} = \frac{|MCS(G_1, G_2)|}{|G_1 + G_2 - MCS(G_1, G_2)|} = \frac{|MCS(G_1, G_2)|}{|G_1| + |G_2| - |MCS(G_1, G_2)|}$$

(s)-malate

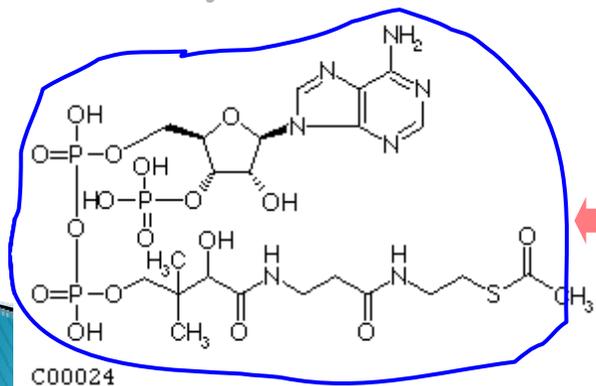


oxaloacetate

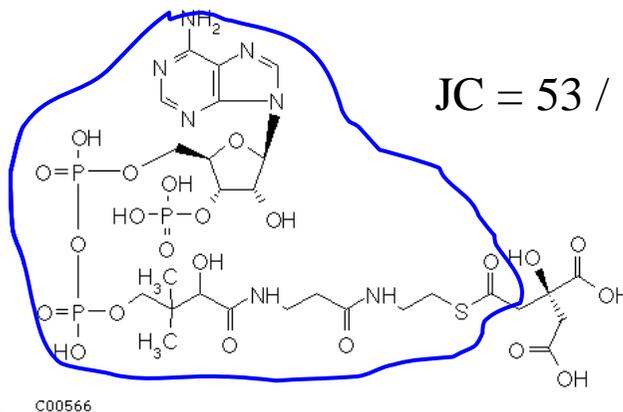


$$JC = 9 / (9 + 9 - 9) = 1.000$$

Acetyl-CoA



(3S)-citryl-CoA



$$JC = 53 / (53 + 62 - 53) = 0.854$$

# About the Option in SIMCOMP

KEGG Compound Search

Compute Return

Query C00118

Structure

Database

- COMPOUND
- REACTION
- SIMCOMI
- SUBCOMI

Program

- SIMCOMI
- SUBCOMI

Partial atom matching

[ LIGAND | KEGG ]

ページが表示されました

インターネット

When the “**Partial atom match**” is turned on, the KEGG Atom Types will not be considered.



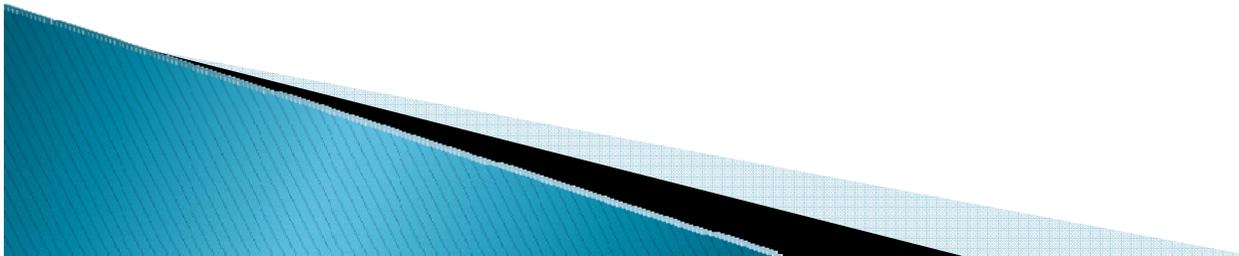
The resulting alignment will become **larger** and thus the score will also become **larger**.

# Provided Tools in KEGG

- ▶ Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- ▶ Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- ▶ Desktop application for compounds and glycans
  - KegDraw Java based software to input structures
- ▶ Automatic assignment of EC number
  - e-zyme to predict EC numbers for a set of pairs

# About SUBCOMP

- ▶ SUBCOMP tests if the query structure can be found within each entry of database.
- ▶ This type of problem has been known as **the subgraph isomorphism** in the graph theory.
- ▶ A user can invoke the SUBCOMP computation via the same interface (the same URL) with SIMCOMP.



KEGG Compound Search - Microsoft Internet Explorer

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

アドレス(D) http://www.genome.jp/ligand-bin/se Google

## KEGG Compound Search

View structure Clear

Enter query compound: (in one of the three forms)

Compound ID  (Example)

MOL File Name

MOL File Text

Select target database:

COMPOUND  REACTION

Select program and option:

SIMCOMP  Partial atom match

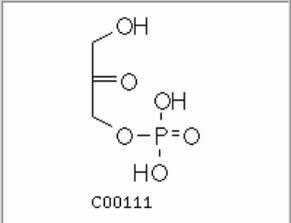
SUBCOMP

KEGG Compound Search

Compute Return

Query C00111

Structure



C00111

Database  COMPOUND  REACTION

Program  SIMCOMP  Partial atom matching  SUBCOMP

[ LIGAND | KEGG ]

KEGG Compound Search - Microsoft Internet Explorer

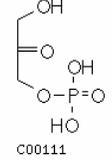
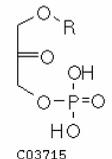
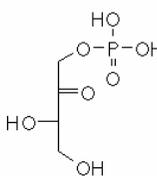
ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)

アドレス(D) http://www.genome.jp/ligand-bin/search\_list?DATABASE=compound Google

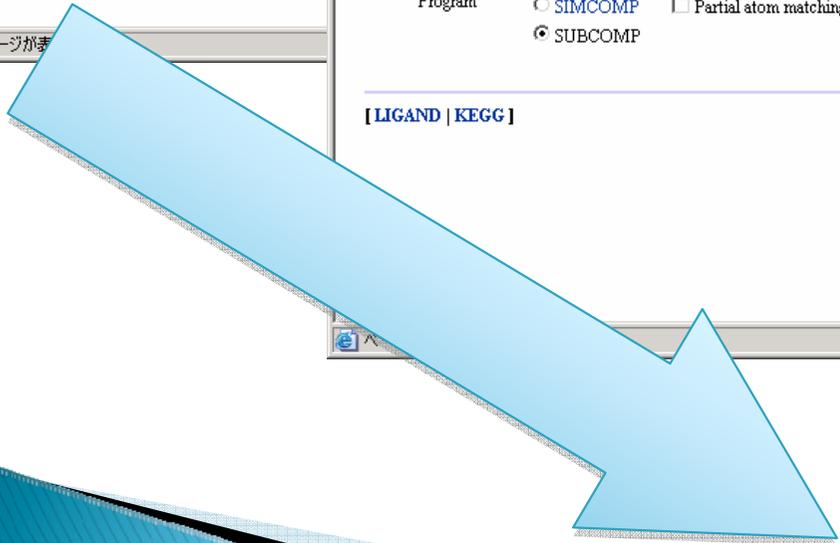
## Compound Data Search Result

Number of entries in a page 20 Hide structure

Items: 1 - 20 of 20

No	Entry	Structure	Name	Formula
1	<a href="#">C00111</a>	 <p>C00111 Similarity-Score : 1.00</p>	Glycerone phosphate Dihydroxyacetone phosphate	C <sub>3</sub> H <sub>7</sub> O <sub>6</sub> P
2	<a href="#">C03715</a>	 <p>C03715 Similarity-Score : 0.91</p>	O-Alkylglycerone phosphate Alkyl-glycerone 3-phosphate Dihydroxyacetone phosphate alkyl ether	C <sub>3</sub> H <sub>6</sub> O <sub>6</sub> PR
3	<a href="#">C03394</a>	 <p>C03394 Similarity-Score : 0.83</p>	Erythrulose 1-phosphate	C <sub>4</sub> H <sub>7</sub> O <sub>7</sub> P

ページが表示されました インターネット



# Provided Tools in KEGG

- ▶ Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- ▶ Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- ▶ Desktop application for compounds and glycans
  - KegDraw Java based software to input structures
- ▶ Automatic assignment of EC number
  - e-zyme to predict EC numbers for a set of pairs

# Provided Tools in KEGG

- ▶ Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- ▶ Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- ▶ Desktop application for compounds and glycans
  - **KegDraw** Java based software to input structures
- ▶ Automatic assignment of EC number
  - e-zyme to predict EC numbers for a set of pairs

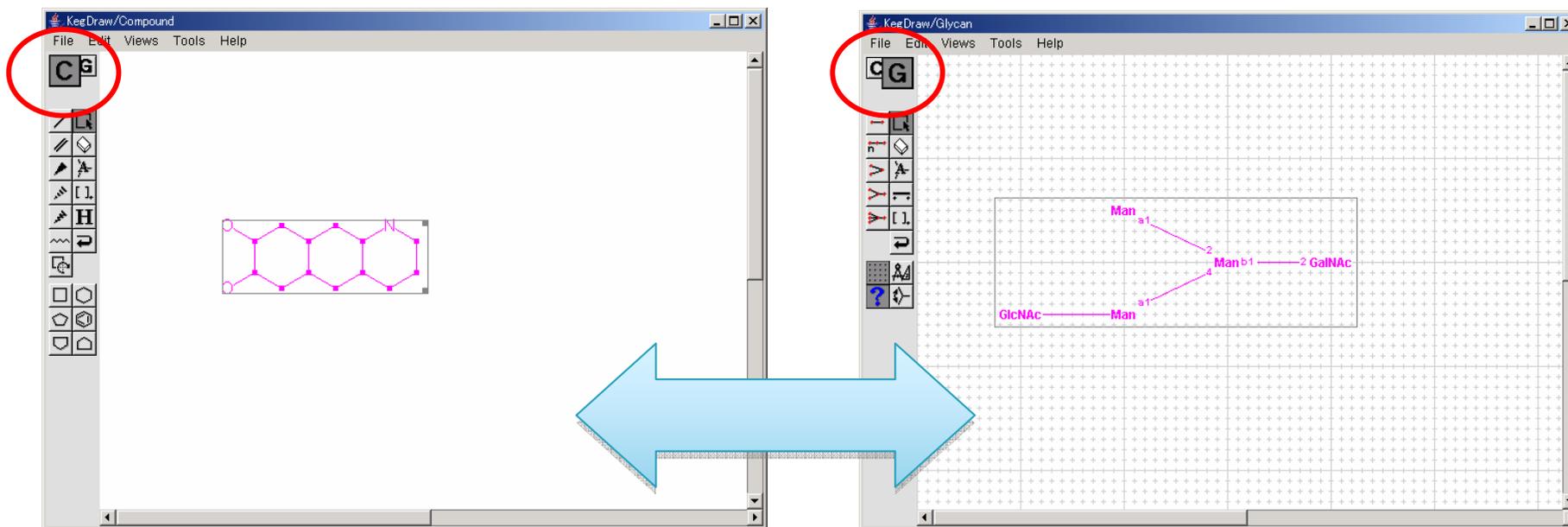
# About KegDraw

- ▶ Desktop **application** to edit chemical compound structures, like the ChemDraw or ISIS/Draw.
- ▶ KegDraw can also edit the glycan structures.
- ▶ A user can perform the structural search against KEGG or PubChem database via the Tools menu of KegDraw.

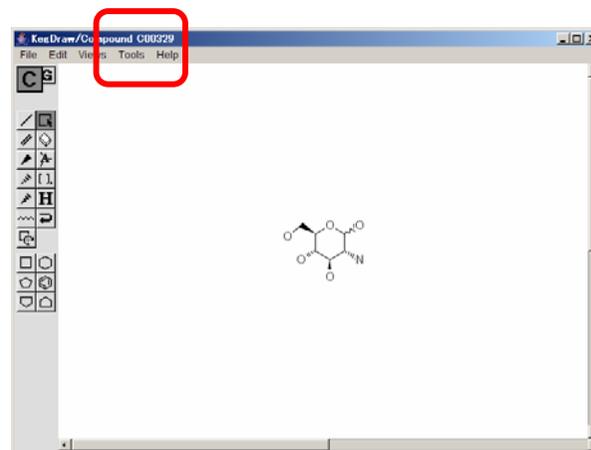


# Draw mode in KegDraw

- ▶ Two drawing modes for **compound** or for **glycan** structures are selectable by clicking the upper left icon labeled as **C** or **G**.

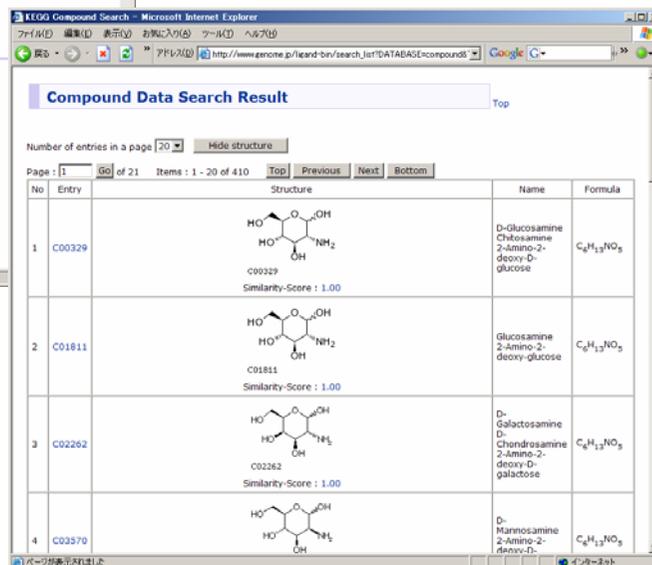
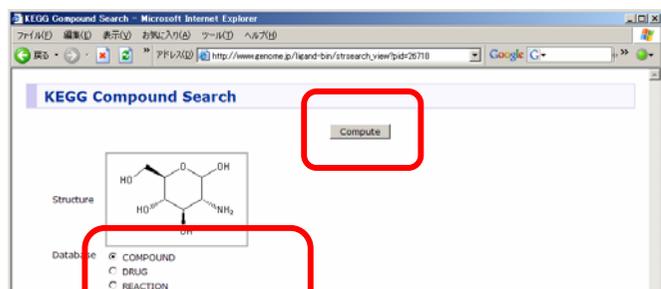


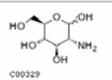
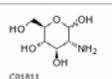
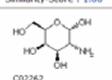
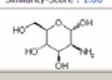
LIGAND (KEGG)

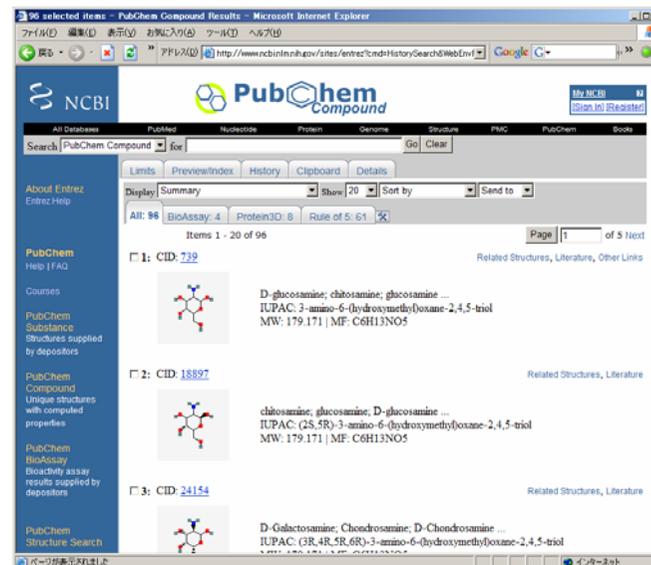


PubChem(NCBI)

A user can search the chemical structure from KegDraw's Tools menu.



No	Entry	Structure	Name	Formula
1	C00329		D-Glucosamine Chitosamine 2-Amino-2-deoxy-D-glucose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
2	C01811		Glucosamine 2-Amino-2-deoxy-glucose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
3	C02262		D-Galactosamine D-Chondrosamine 2-Amino-2-deoxy-D-galactose	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>
4	C03570		Mannosamine 2-Amino-2-Deoxy-D-	C <sub>6</sub> H <sub>13</sub> NO <sub>5</sub>



96 selected items - PubChem Compound Results - Microsoft Internet Explorer

PubChem Compound

Search PubChem Compound for [ ] Go Clear

Display Summary Show 20 Sort by Send to

All: 96 BioAssay: 4 Protein3D: 8 Rule of 5: 61

Items 1 - 20 of 96

Page 1 of 5 Next

Related Structures, Literature, Other Links

1: CID: 739  
D-glucosamine; chitosamine; glucosamine ...  
IUPAC: 3-amino-6-(hydroxymethyl)oxane-2,4,5-triol  
MW: 179.171 | MF: C6H13NO5

2: CID: 18897  
chitosamine; glucosamine; D-glucosamine ...  
IUPAC: (2S,5R)-3-amino-6-(hydroxymethyl)oxane-2,4,5-triol  
MW: 179.171 | MF: C6H13NO5

3: CID: 24154  
D-Galactosamine; Chondrosamine; D-Chondrosamine ...  
IUPAC: (3R,4R,5R,6R)-3-amino-6-(hydroxymethyl)oxane-2,4,5-triol

# Provided Tools in KEGG

- ▶ Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- ▶ Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- ▶ Desktop application for compounds and glycans
  - KegDraw Java based software to input structures
- ▶ Automatic assignment of EC number
  - e-zyme to predict EC numbers for a set of pairs

# About e-zyme

- ▶ Once the **atom transformation patterns** between query compounds are calculated by SIMCOMP, then the resulting patterns are verified **against the RPAIR** database, that is, the pre-computed pattern database (= high quality knowledge of enzymatic reaction mechanisms).
- ▶ The prediction accuracy is **remarkably high** if the **only top 3 digits of EC** are considered.
  - EC number consists of 4 digits and the top 3 of them may be **relevant** for the reaction mechanism. The last digit is thought to be just the sequential number.

# e-zyme: Start page



**e-zyme**

for prediction of enzymatic reactions

[Home](#)

[Help](#)

## About e-zyme

The e-zyme software has been developed in the KEGG project as a means to better computerize the information about enzymatic reactions. It currently performs an automatic assignment of EC numbers based on the RDM system for reaction classification. In the future this software will be expanded to predict reaction pathways.

## Input form

To run the query, use the following form:

[e-zyme prediction](#)

Last updated: April 1, 2008

[KEGG](#)

[GenomeNet](#)

[Kyoto University Bioinformatics Center](#)

## KEGG Reaction Prediction

Query Mode

### Enter Pair 1 (in one of the three forms)

Compound ID	<input type="text"/>	<input type="text"/>
MOL File Name	<input type="text"/> 選択...	<input type="text"/> 選択...
MOL File Text	<input type="text"/>	<input type="text"/>

### Enter Pair 2 (in one of the three forms)

Compound ID	<input type="text"/>	<input type="text"/>
MOL File Name	<input type="text"/> 選択...	<input type="text"/> 選択...
MOL File Text	<input type="text"/>	<input type="text"/>

### Enter Pair 3 (in one of the three forms)

Compound ID	<input type="text"/>	<input type="text"/>
MOL File Name	<input type="text"/> 選択...	<input type="text"/> 選択...
MOL File Text	<input type="text"/>	<input type="text"/>

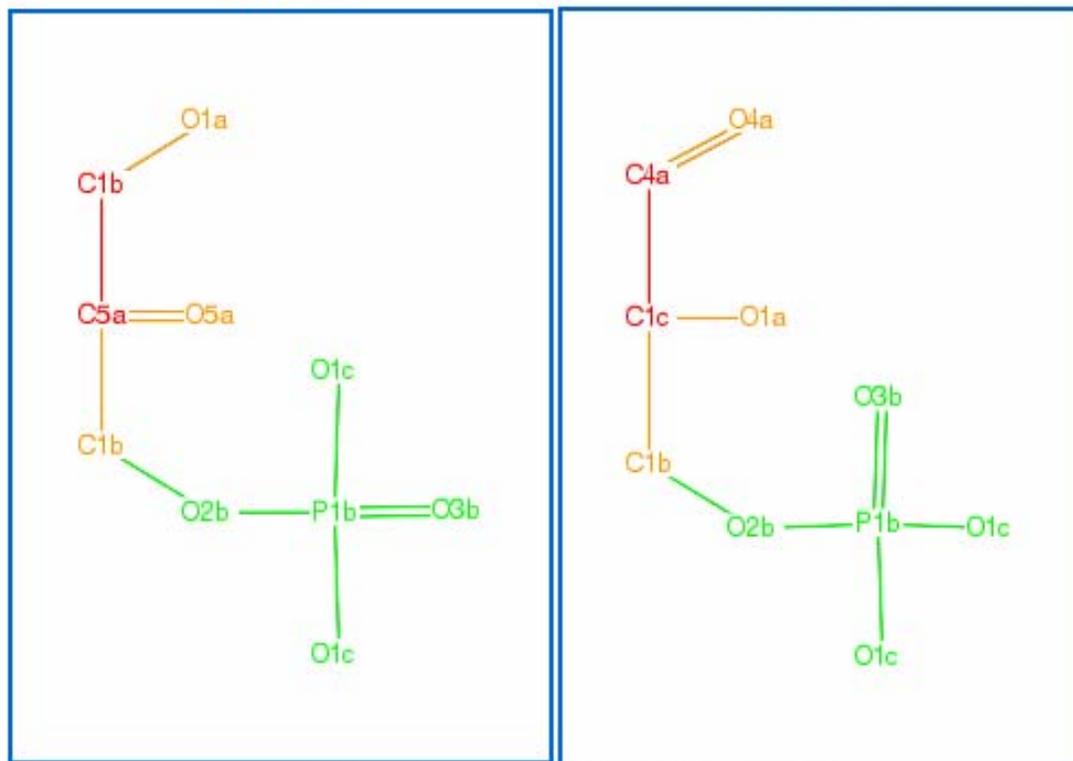
# e-zyme result

Pair 1

Edit

C00111

C00118



C5a-C1c:\*-\*:C1b+C1b+O5a-C1b+C4a+O1a  
C1b-C4a:\*-\*:C5a+O1a-C1c+O4a

Alignment result

Character string notation  
of RDM

Predicted EC numbers  
with Weighted scores  
with Observed freq.

## EC assignments

EC number Weighted score Observed freq.

5.3.1	25.8	1.0
3.5.4	0.9	
6.-.-	0.9	

# e-zyme: Weighted scoring

- ▶ Key idea: use of the correlation between template RDM patterns

RDM pattern profile: a set of profiles describing the frequency of related EC numbers

EC (e.g., 1.1.1, 1.1.2, ..., 2.4.1, ..., 4.3.1, ..., 6.6.1)

	EC1	EC2	EC3	EC4	EC5	...	EC p	
RDM pattern 1	(3,	1,	0,	0,	0,	...	1)	} correlation coefficient : $corr(RDM_i, RDM_j)$ $(i, j = 1, 2, \dots, n)$
RDM pattern 2	(1,	1,	1,	0,	2,	...	0)	
RDM pattern 3	(0,	1,	0,	0,	2,	...	0)	
...								
RDM pattern n	(1,	0,	4,	0,	0,	...	1)	

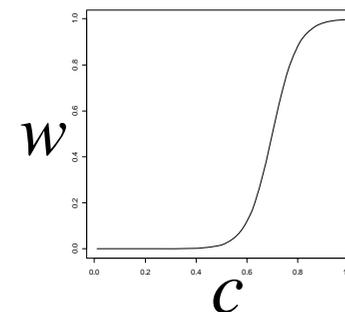
# e-zyme: Weighted scoring

- ▶ Assume that the RDM pattern of the query reaction is computed (denoted by  $RDM_{query}$ )
- The old method used in Kotera's paper: select  $EC_k$  with the highest score by

$$S(EC_k) = f_k, \text{ where } f_k : \text{frequency of } EC_k \text{ in } RDM_{query}$$

- New method: select  $EC_k$  with the highest score by

$$S(EC_k) = \sum_{i=1}^{|RDM|} w_i f_{ik}, \text{ where } \begin{cases} f_{ik} : \text{frequency of } EC_k \text{ in } RDM_i \\ c_i = \text{corr}(RDM_{query}, RDM_i) \\ w_i = \text{sigmoid}(c_i) \end{cases}$$



# e-zyme: Multiple mode



REACTION: R01811

Help

Entry	R01811	Reaction
Name	N-Acetylneuraminate pyruvate-lyase	
Definition	N-Acetylneuraminate $\rightleftharpoons$ N-Acetyl-D-mannosamine + Pyruvate	
Equation	C00270 $\rightleftharpoons$ C00645 + C00022	
	<p>The diagram illustrates the reversible reaction of N-Acetylneuraminate (C00270) to N-Acetyl-D-mannosamine (C00645) and Pyruvate (C00022). N-Acetylneuraminate is a complex aminosugar with multiple hydroxyl groups and an acetyl group. N-Acetyl-D-mannosamine is a simpler aminosugar with an acetyl group. Pyruvate is a three-carbon carboxylic acid. The reaction is reversible, as shown by the double-headed red arrow.</p>	
RPair	RP: <a href="#">A00421</a> C00270_C00645 main RP: <a href="#">A01773</a> C00022_C00270 main	
Pathway	PATH: <a href="#">rn00530</a> Aminosugars metabolism	
Enzyme	<a href="#">4.1.3.3</a>	
Orthology	KO: <a href="#">K01639</a> N-acetylneuraminate lyase	
LinkDB	<a href="#">All DBs</a>	

Entry	A00421	ReactantPair
Name	C00270_C00645	
Compound	C00270 N-Acetylneuraminic acid C00645 N-Acetyl-D-mannosamine	
Type	main	
RDM	3 1 O2x-O1a:C1z-*:C1y-C1y 2 C1y-C1y:C1x+*-+O2x:C1y+O1a-C1y+O1a 3 O1a-O2x:*-C1y:C1c-C1y	
Related pair	A00295 A00429 A00430 A00548 A00945 A01132 A01262 A01269 A01284 A01544 A01654 A01758 A01778 A01888 A02088 A02144	

This is the real solution  
for pair 1.

This is pair 2.

Entry	A01773	ReactantPair
Name	C00022_C00270	
Compound	C00022 Pyruvate C00270 N-Acetylneuraminic acid	
Type	main	
RDM	2 1 C1a-C1x:*-C1y:C5a-C1z 2 C5a-C1z:*-O2x:C1a+C6a+O5a-C1x+C6a+O1a	
Reaction	R01811	
KCF data	<a href="#">Show</a>	

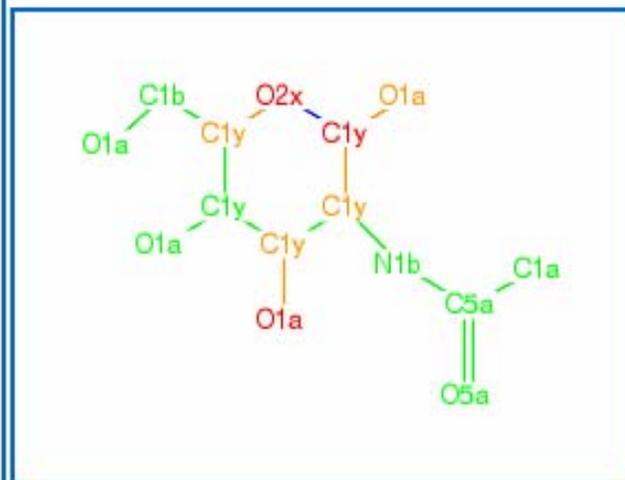
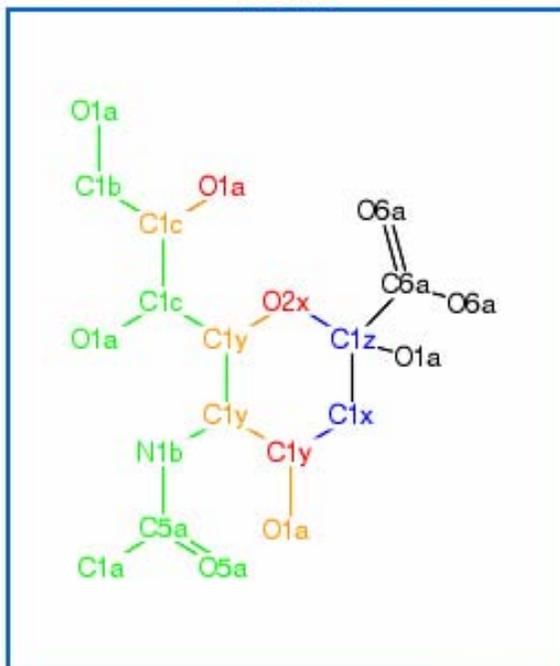
# e-zyme result

Pair 1

Edit

C00270

C00645



O2x-O1a:C1z-\*:C1y-C1y  
C1y-C1y:C1x+\*-\*+O2x:C1y+O1a-C1y+O1a  
O1a-O2x:\*-C1y:C1c-C1y

When invoking only the pair 1 with the single mode, the predicted ECs are not good.

## EC assignments

EC number	Weighted score	Observed freq.
2.5.1	71.6	1.0
4.1.3	15.7	1.0
3.1.1	2.0	
4.1.99	1.3	
4.4.1	0.9	
1.5.99	0.9	
3.3.1	0.5	

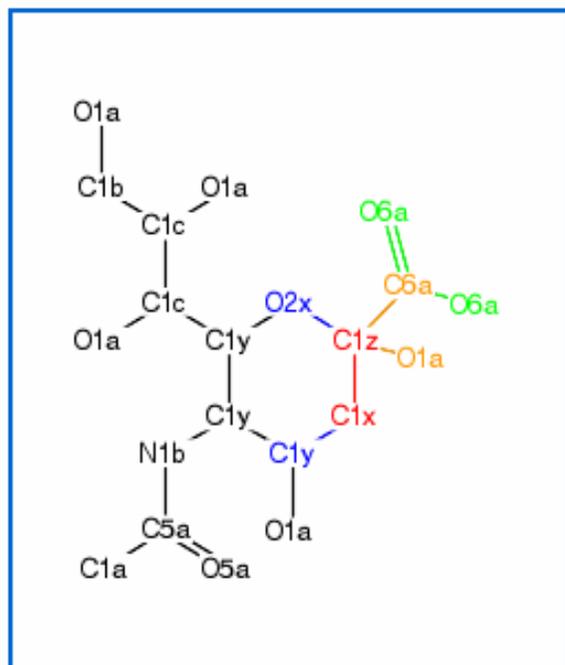
C1a O5a

O2x-O1a:C1z-\*:C1y-C1y  
 C1y-C1y:C1x+\*-\*:O2x:C1y+O1a-C1y+O1a  
 O1a-O2x:\*-C1y:C1c-C1y

Pair 2

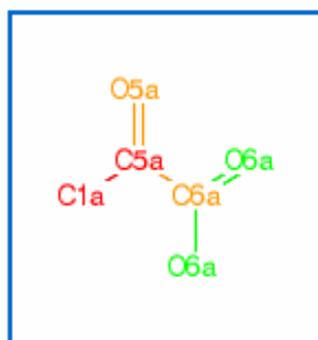
Edit

C00270



C1z-C5a:O2x-\*:C1x+C6a+O1a-C1a+C6a+O5a  
 C1x-C1a:C1y-\*:C1z-C5a

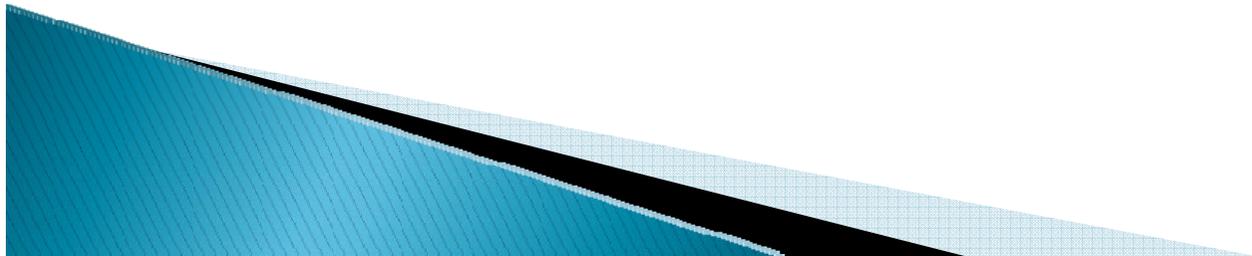
C00022



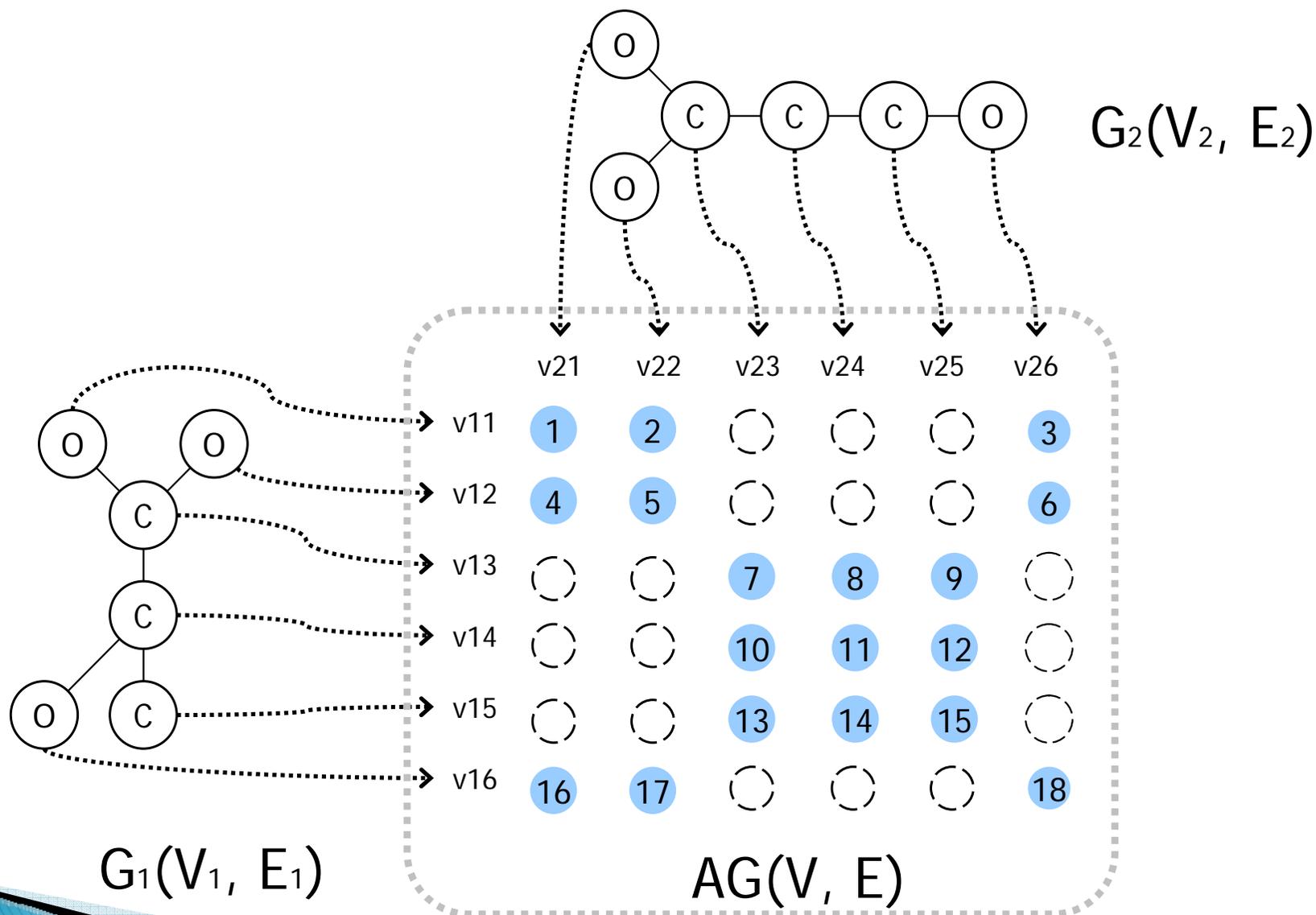
When invoking both pair 1 and 2 with the multiple mode, the resulting ECs were **improved**.

## EC assignments

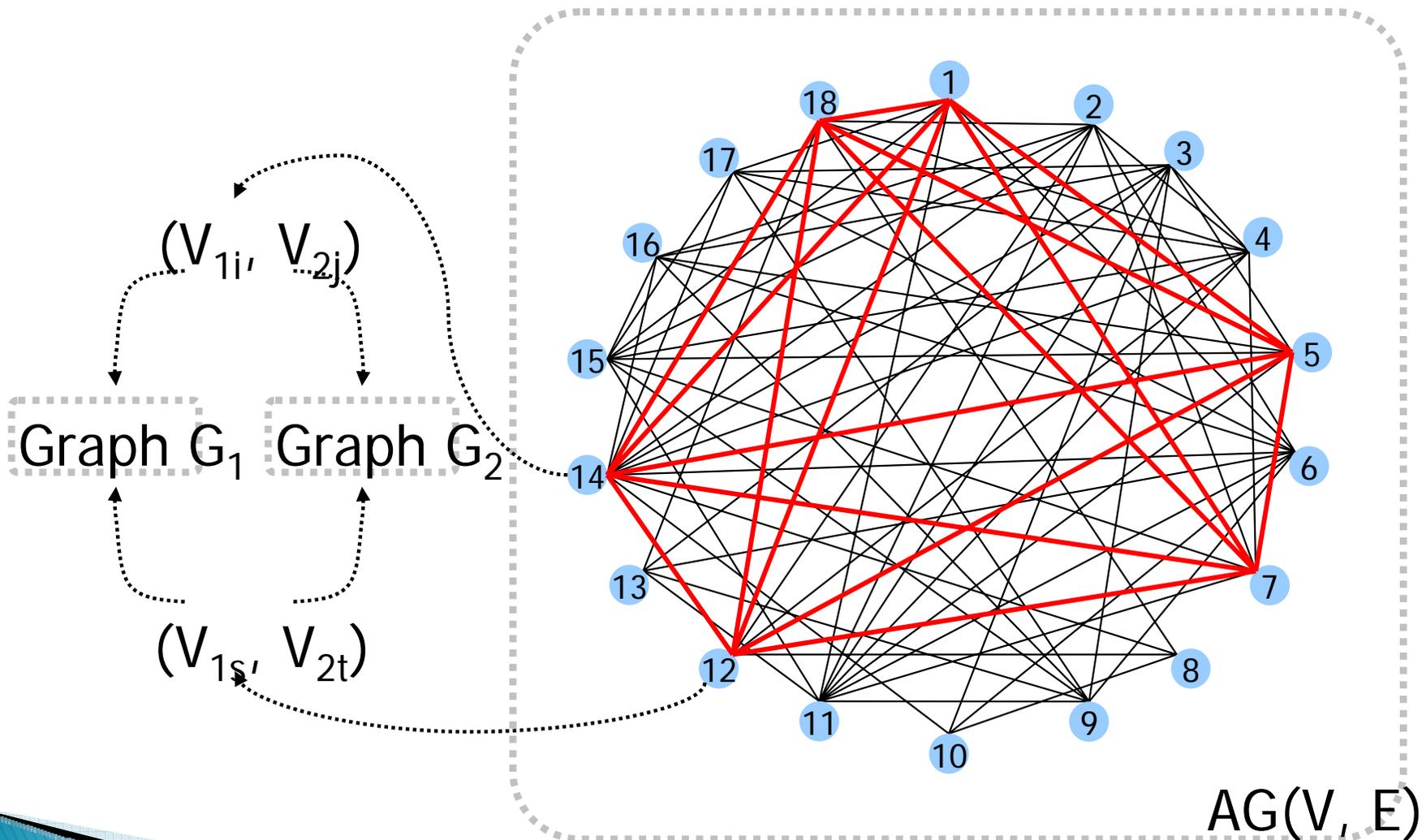
EC number	Weighted score	Observed freq.
4.1.3	22.0	1.0
4.1.2	1.5	
4.1.1	1.0	
4.1.99	0.5	



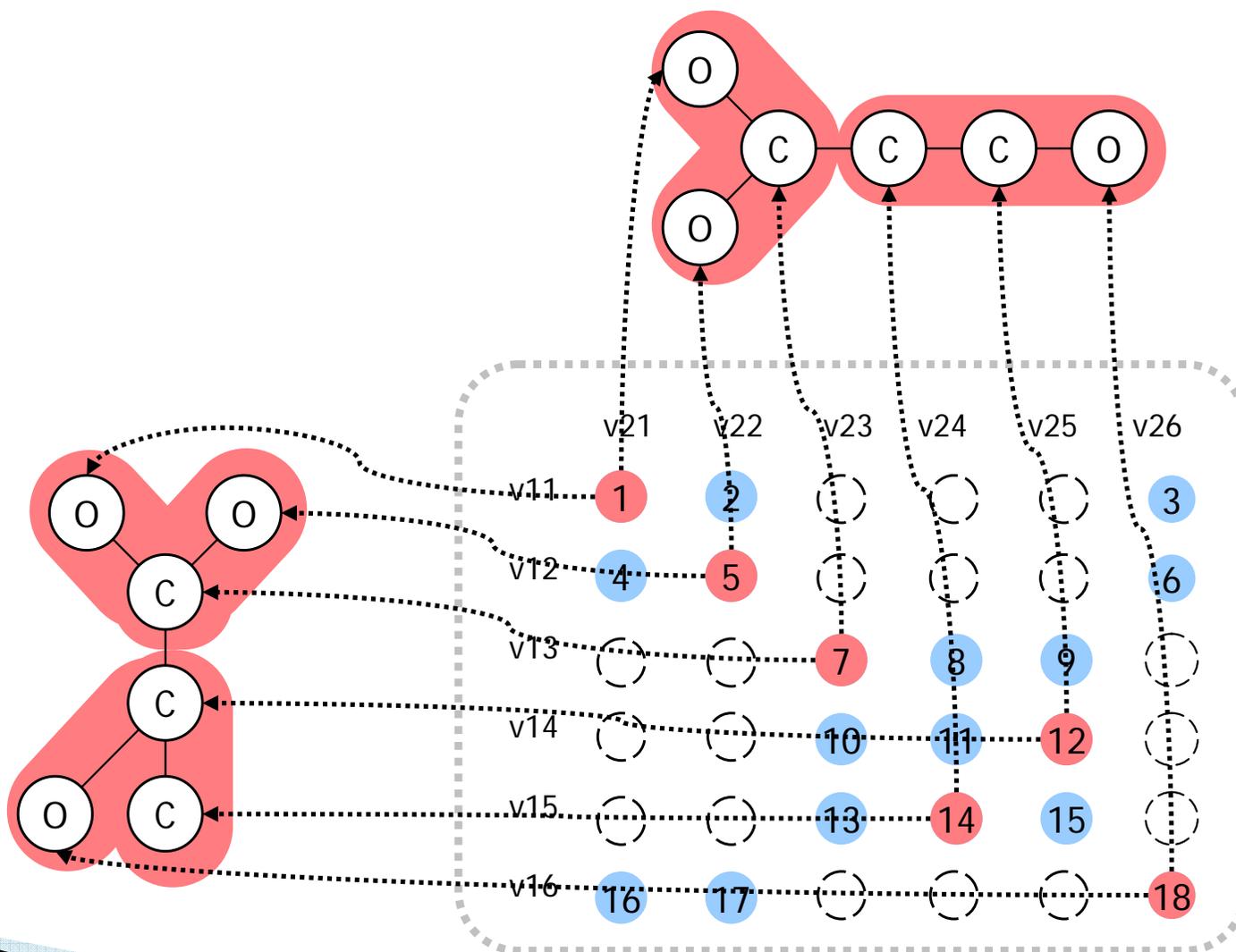
# SIMCMOP: Appendix



# SIMCMOP: Appendix



# SIMCMOP: Appendix



# e-zyme: Appendix

RDM matching was followed:

1. Compute the RDM pattern for a query reaction
2. Search template reactions sharing the same RDM pattern in the database
3. Collect all the EC numbers assigned to the matched template reactions
4. Select the most frequent EC number by majority vote approach

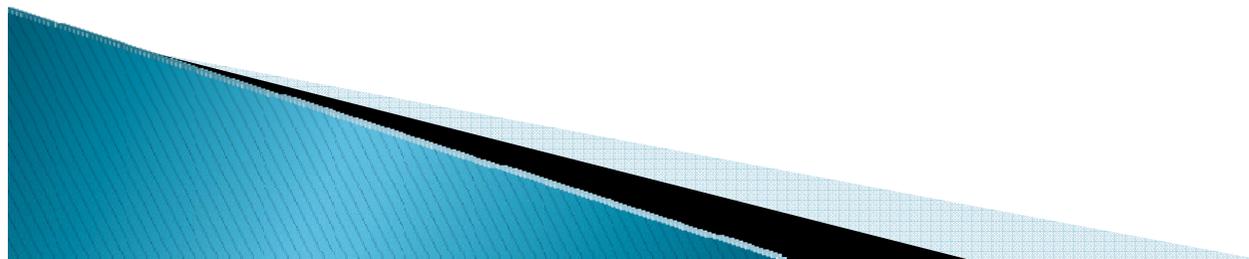
(Kotera et al., *J. Am. Chem. Soc.*, 2004)



# e-zyme: Appendix

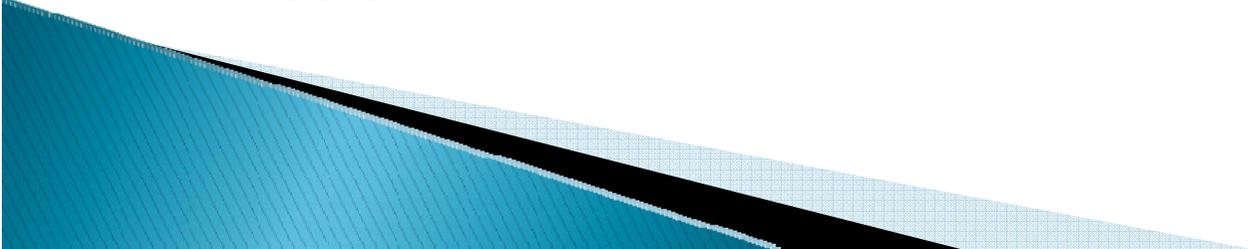
## Loosing the restriction of exact RDM matching

- ▶ Problem: Kotera's method cannot work when the query RDM is not matched with any template RDMs in the database
- ▶ Proposed procedure:
  1. Use the full RDM for matched reactions
  2. Use RD only when the query RDM is a singleton (no hits in the database)



# e-zyme: Appendix

Validation of prediction power by jackknife tests

1. Take one reaction from 5008 template reactions as a query, and compute the RDM pattern
  2. Predict the EC number of the query, based on the information of the remaining 5007 reactions
  3. Evaluate the prediction result:
    - True positive: the predicted EC is the real EC of the query
    - False positive: the predicted EC is NOT the real EC of the query
  4. Repeat the above procedure to all the 5008 reactions
- 

# e-zyme: Appendix

	coverage	accuracy (main)	accuracy (sub)	accuracy (subsub)
Kotera method	66.6	92.1	88.5	83.2
New 1: Loosing	81.3	92.2	88.3	82.3
New 2: Loosing+ Weighted	81.3	93.5	90.6	85.0

- Coverage (%): the rate of predictable queries (that have hits in the database)
- Accuracy (%): the rate of correctly predicted reactions