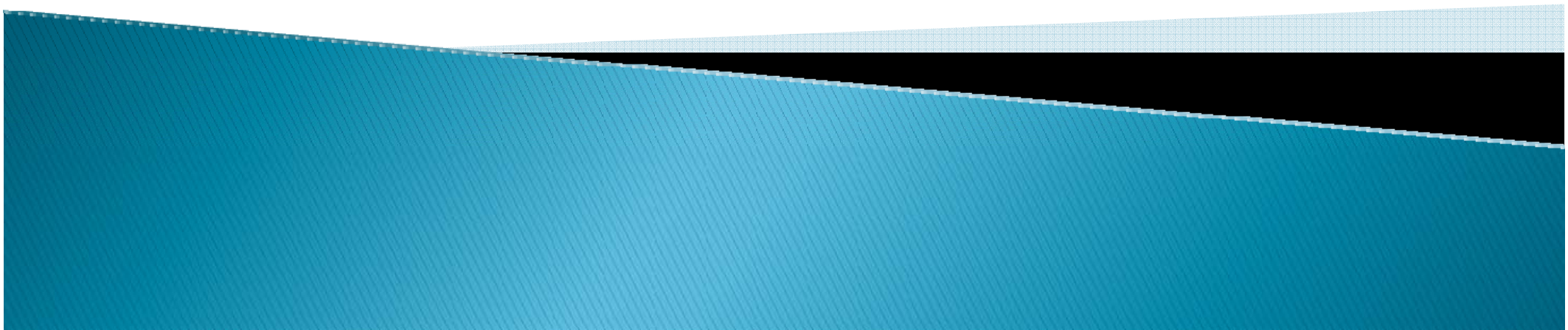
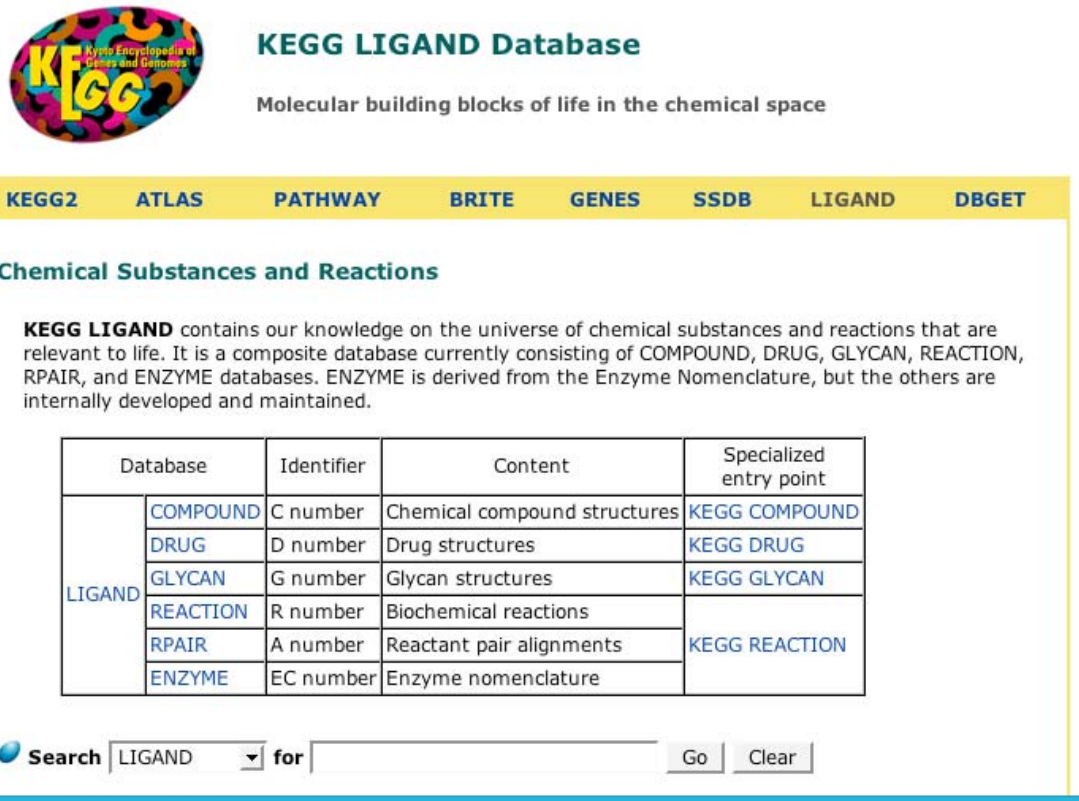
Learning KEGG
– Chemical data –

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	KEGG COMPOUND
	DRUG	D number	Drug structures	KEGG DRUG
	GLYCAN	G number	Glycan structures	KEGG GLYCAN
	REACTION	R number	Biochemical reactions	KEGG REACTION
	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

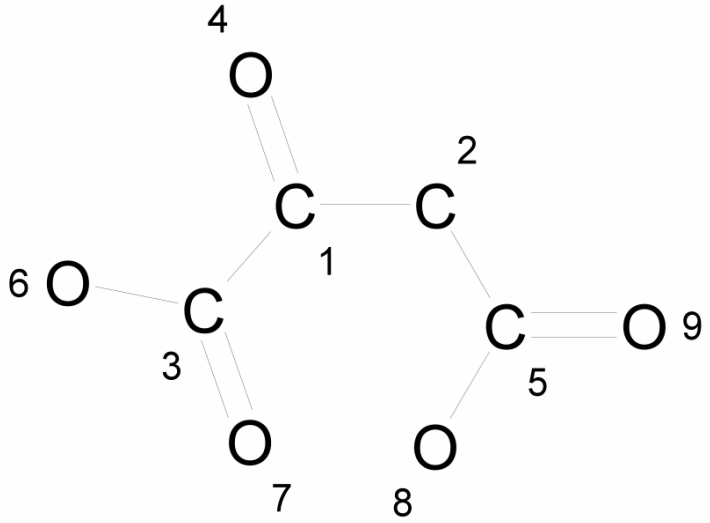
- Components of LIGAND Database
  - COMPOUND Chemical compound structure
  - DRUG Drug structure
  - GLYCAN Glycan structure
  - REACTION Biochemical reaction
  - ENZYME Enzyme nomenclature
  - RPAIR Reaction pair alignment

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.

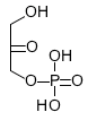
  - Each structure is formatted in MDL/mol or KCF.

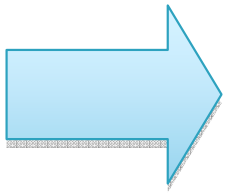


What is the MDL/mol format?

- Developed by MDL (now Elsevier MDL).
- De-facto standard of molecule formats available in most of applications.

KEGG **COMPOUND: C00111** Help

Entry	C00111 Compound
Name	Glycerone phosphate; Dihydroxyacetone phosphate
Formula	C3H7O6P
Mass	169.998
Structure	 C00111 Mol file KCF file DB search Jmol KegDraw
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	All DBs
KCF data	Show



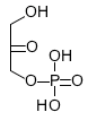
```
 10  9  0  0  0  0  0  0  0  0999 V2000
   22.8555  -17.6695    0.0000 P   0  0  3  0  0  0  0  0  0  0  0  0
   21.4279  -17.6628    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   22.8992  -19.2106    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   22.8950  -16.1215    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   24.2890  -17.6497    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   20.1911  -16.9523    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   20.1911  -15.5247    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   20.1911  -14.0971    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   21.4787  -15.5247    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   21.4279  -13.3801    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0      0  0
  1  3  1  0      0  0
  1  4  1  0      0  0
  1  5  2  0      0  0
  2  6  1  0      0  0
  6  7  1  0      0  0
  7  8  1  0      0  0
  7  9  2  0      0  0
  8 10  1  0      0  0
M  END
```

http://www.mdl.com/downloads/public/ctfile/ctfile.pdf

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 Mol file KCF file DB search Jmol KegDraw
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	All DBs
KCF data	Show

```
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
```

Hattori et al., J Am Chem Soc 125:11853-65 (2003)

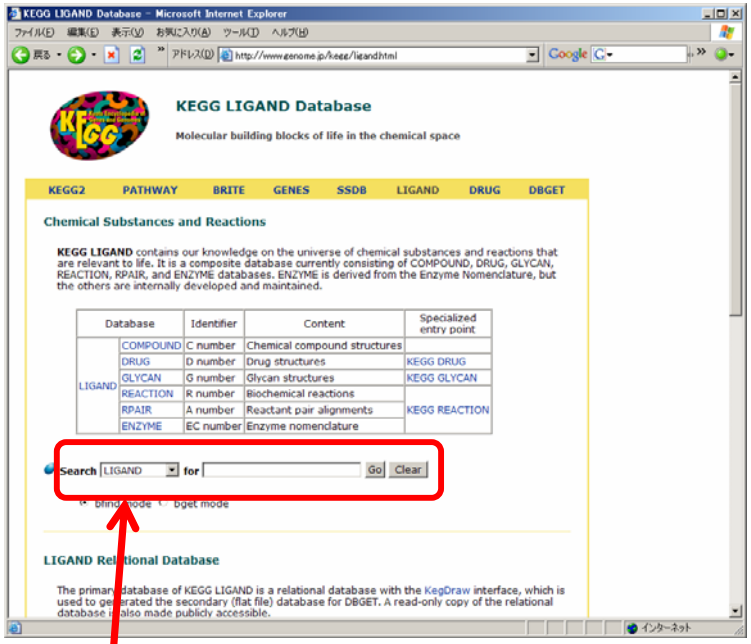
KEGG Atom Types in KCF

- 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.

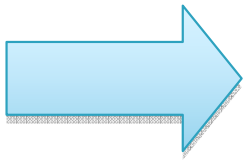
KEGG atom types

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
	Alkyne	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
	Aromatic ring	N4x	ring-NH-ring	785
		N4y	ring-N(-R)-ring	840

http://www.genome.jp/kegg/reaction/KCF.html

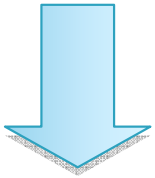


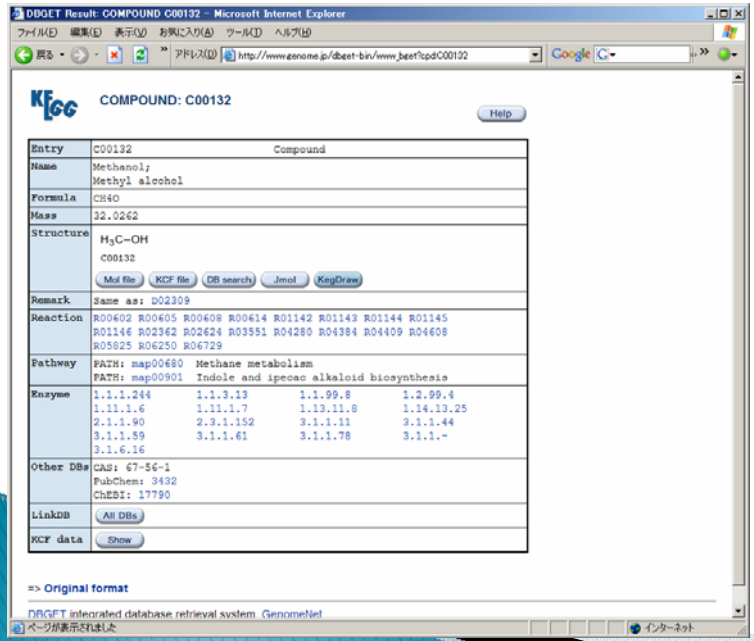
Select “COMPOUND”

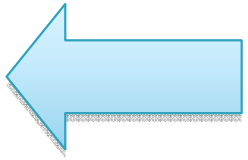


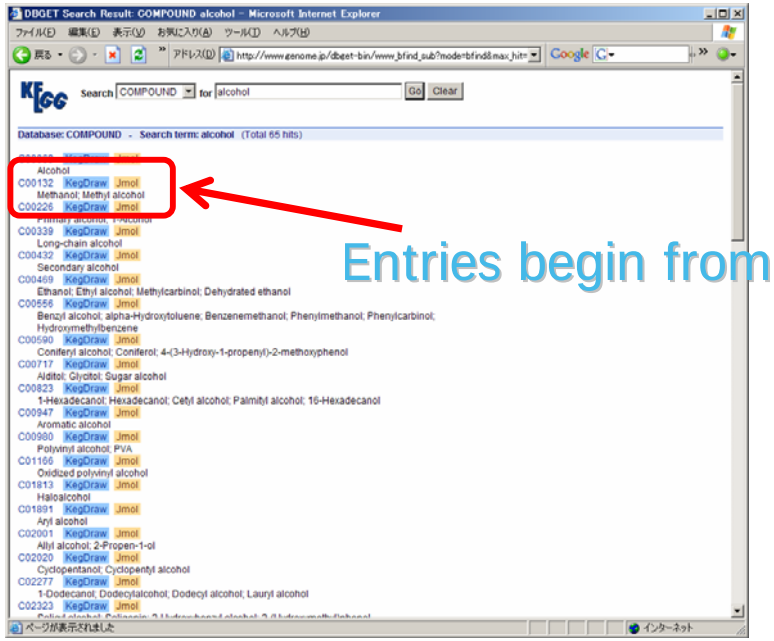
“alcohol”

Search example of COMPOUND entry





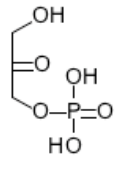




Entries begin from “C”

Example of COMPOUND Entry

KEGG **COMPOUND: C00111** Help

Entry	C00111 Compound
Name	Glycerone phosphate; Dihydroxyacetone phosphate
Formula	C3H7O6P
Mass	169.998
Structure	 C00111 Mol file KCF file DB search Jmol KegDraw
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	All DBs
KCF data	Show

=> Original format

DBGET integrated database retrieval system, GenomeNet

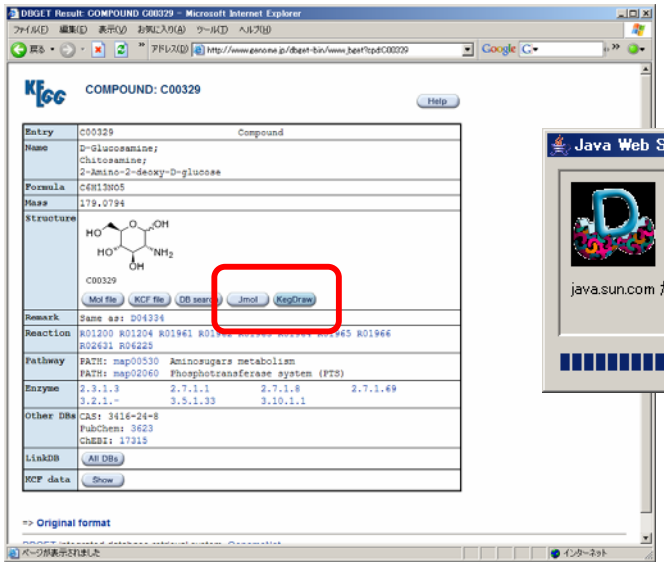
Name
Formula
Structure files（mol & KCF）

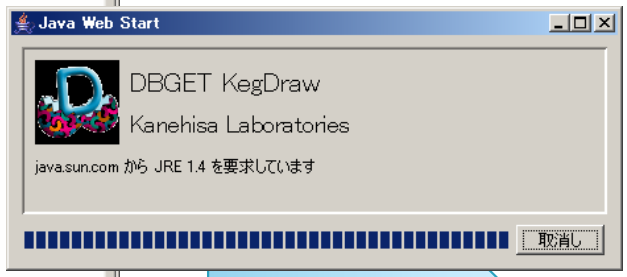
DB search button
Display structure button
(Jmol or KegDraw)

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

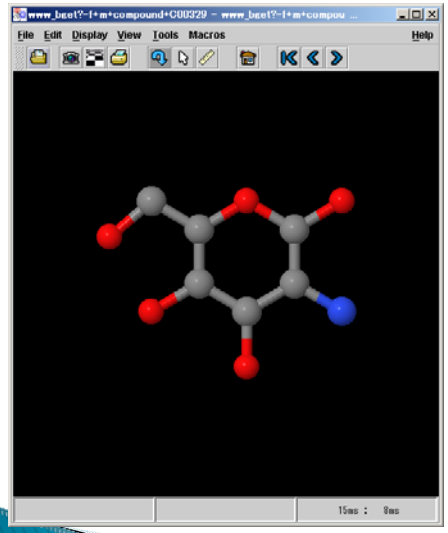
Link to other databases
(PubChem, ChEBI, PDB,,,)

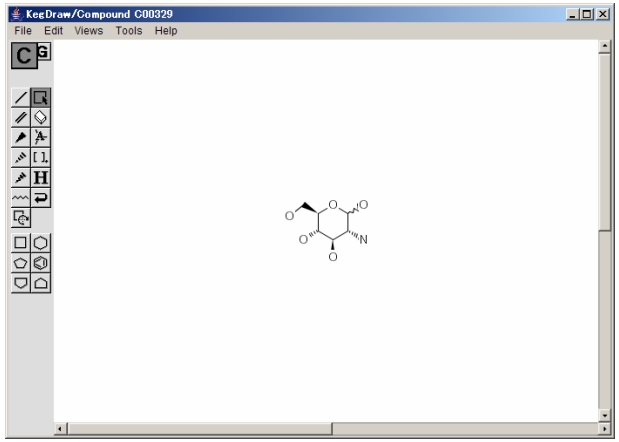
http://www.genome.jp/dbget-bin/www_bget?compound+C00111





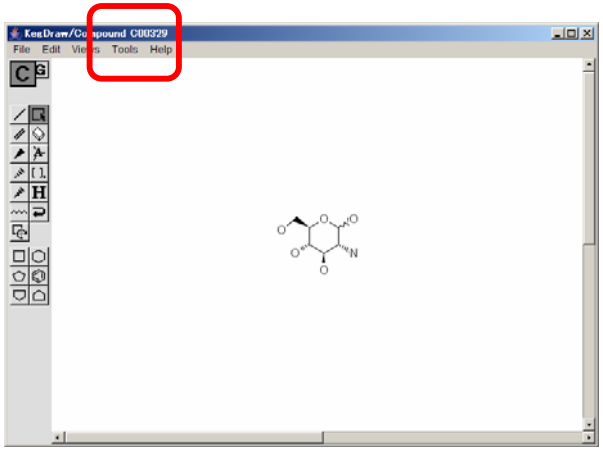
When a user click the Jmol or KegDraw button, then the Java Web Start launched the Jmol or KegDraw program.





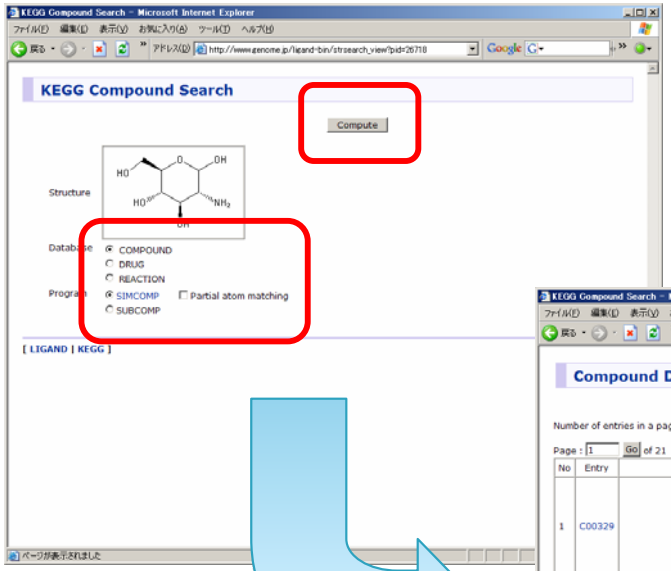
http://www.genome.jp/dbget-bin/www_bget?compound+C00329

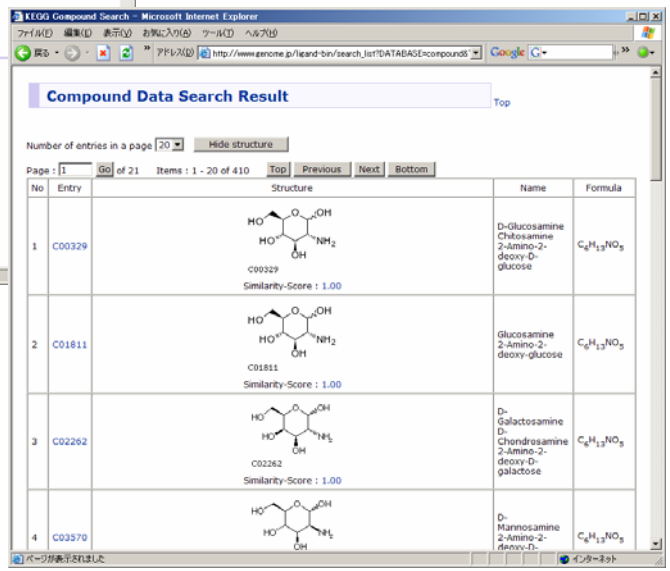
LIGAND (KEGG)

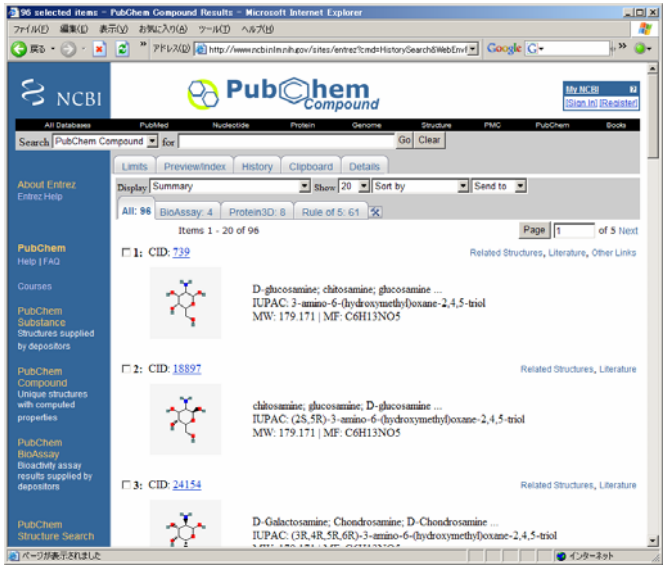


PubChem(NCBI)

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

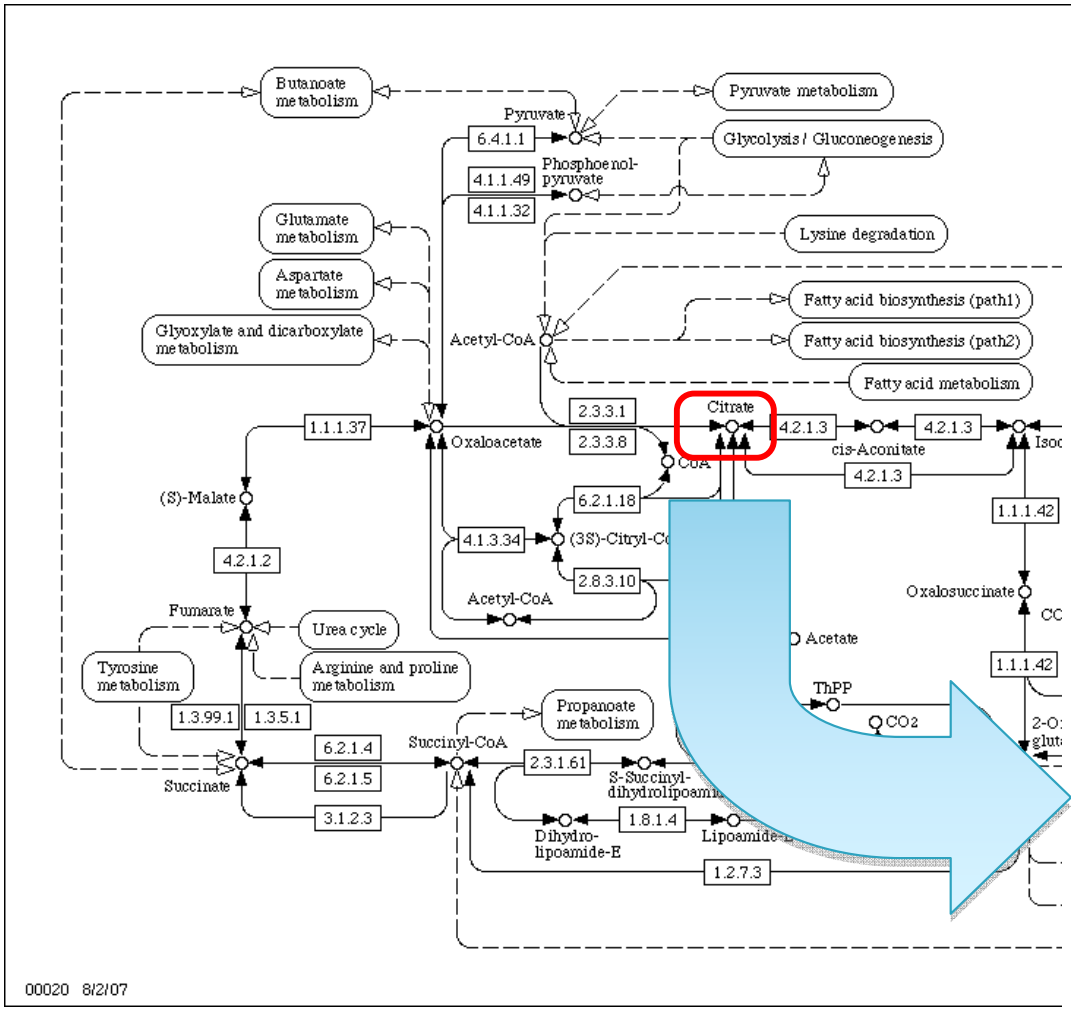






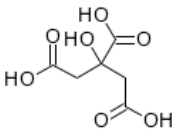
http://www.genome.jp/download/

Pathway and Compound



KEGG **COMPOUND: C00158**

Help

Entry	C00158 Compound
Name	Citrate; Citric acid; 2-Hydroxy-1,2,3-propanetricarboxylic acid; 2-Hydroxytricarballylic acid
Formula	C6H8O7
Mass	192.027
Structure	 C00158 Mol file KCF file DB search Jmol KegDraw
Remark	Same as: D00037
Reaction	R00351 R00352 R00362 R01322 R01323 R01324 R01325 R04357
Pathway	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 (CO2 fixation)
Enzyme	2.3.3.1 2.3.3.3 2.3.3.8 2.8.3.10 3.4.13.20 (C) 4.1.3.6 4.2.1.3 4.2.1.4 6.2.1.18 6.3.2.27
Other DBs	CAS: 77-92-9 PubChem: 3458 ChEBI: 30769 PDB-CCD: CIT 3DMET: B00046
LinkDB	All DBs
KCF data	Show

=> **Original format**

DBGET integrated database retrieval system, GenomeNet

http://www.genome.jp/kegg/pathway/map/map00020.html

LIGAND Database

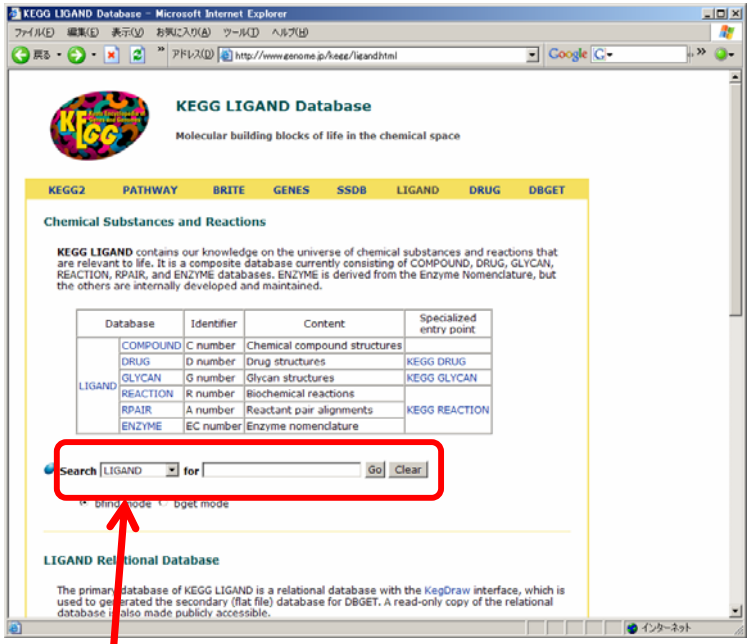
- Components of LIGAND Database
  - COMPOUND Chemical compound structure
  - DRUG Drug structure
  - GLYCAN Glycan structure
  - REACTION Biochemical reaction
  - ENZYME Enzyme nomenclature
  - RPAIR Reaction pair alignment

DRUG Database

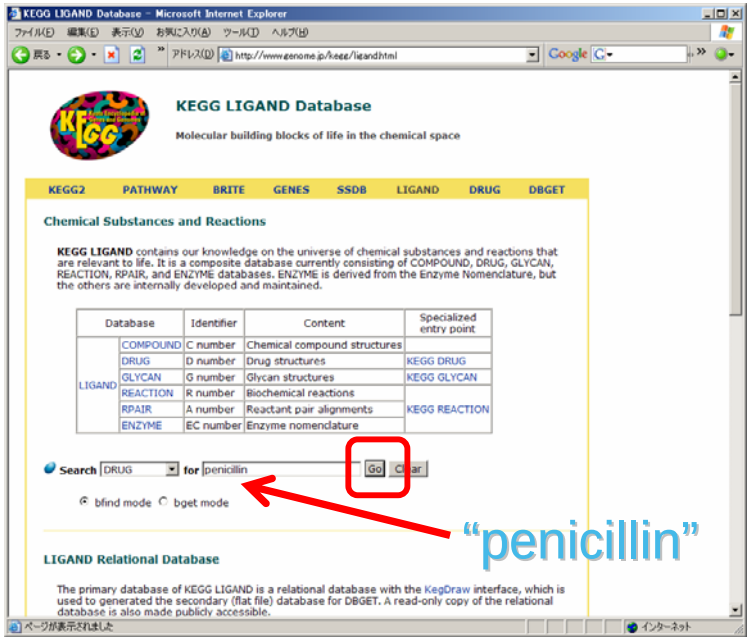
- DRUG Database contains the medicinal compounds, that is, the compounds whose medicinal effects have been confirmed.
  - Each DRUG entry is also linked to therapeutic classification number (therapeutic categories) in KEGG BRITE database.

- 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.

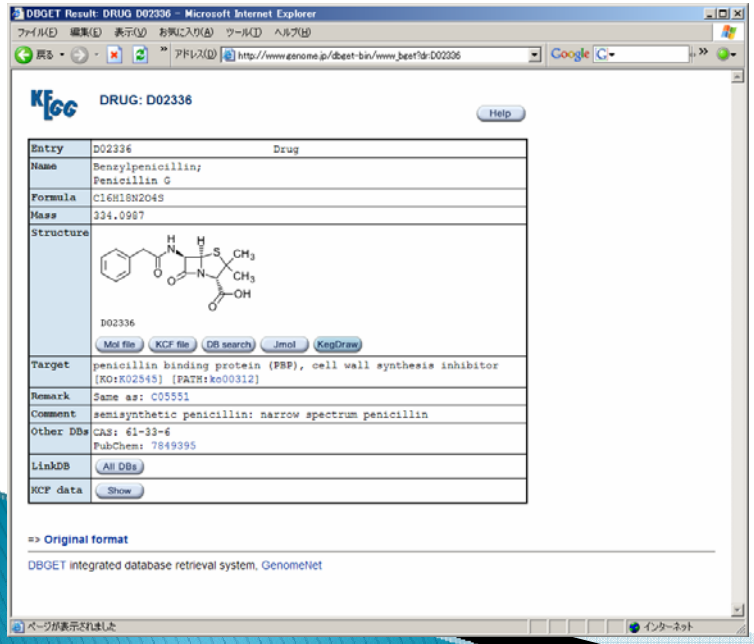
Search example of DRUG entry

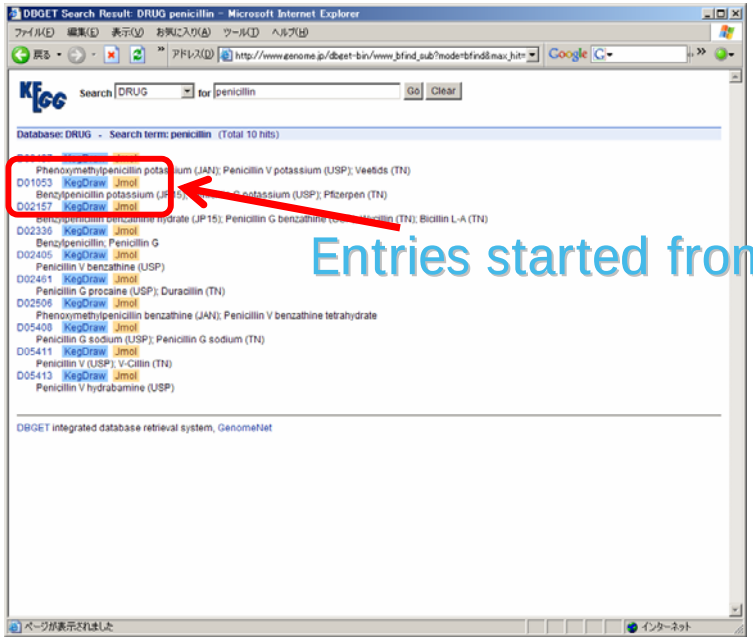


Select “DRUG”



“penicillin”





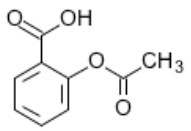
Entries started from “D”

Example of DRUG Entry

KEGG **DRUG: D00109**

Difference of file format

Help

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

=> **Original format**

DBGET integrated database retrieval system, GenomeNet

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

http://www.genome.jp/dbget-bin/www_bget?drug+D00109

Pathway for Drug

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

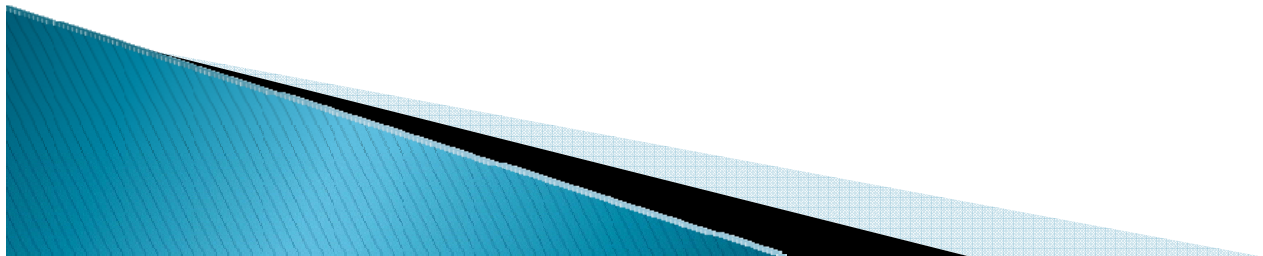
6.5 Target Based Structure Classification
- Serotonin receptor agonists/antagonists
- Dopamine receptor agonists/antagonists
- Histamine receptor antagonists
- beta-Adrenergic receptor agonists/antagonists
- alpha-Adrenergic receptor agonists/antagonists
- Calcium channel blocking agents
- Catecholamine transferase inhibitors
- Angiotensin antagonists

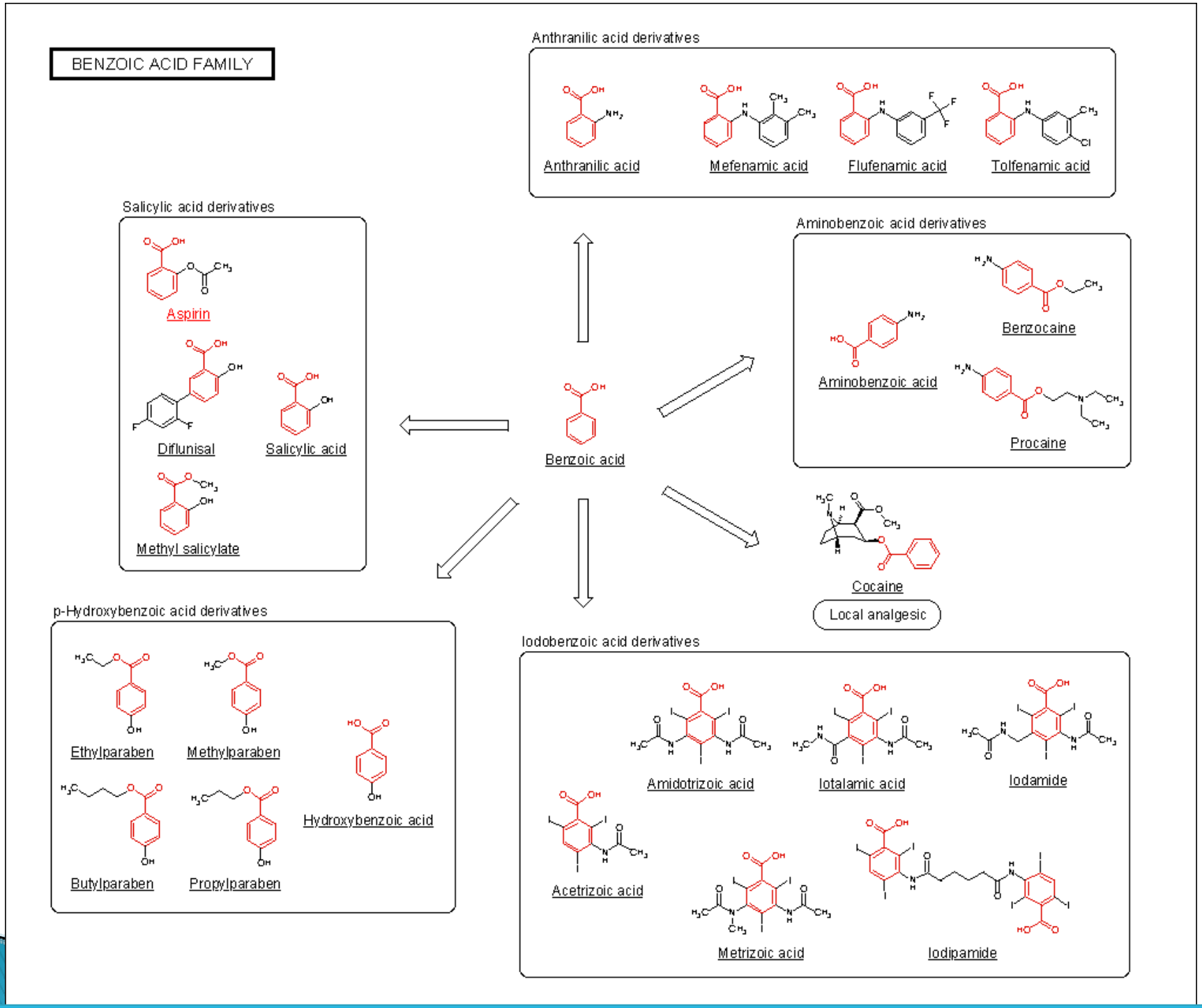
KEGG DRUG

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.

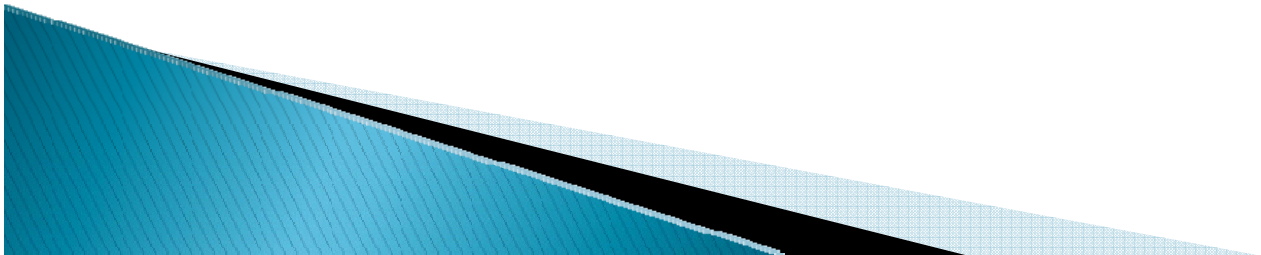




http://www.genome.jp/kegg/pathway/map/map07110.html

LIGAND Database

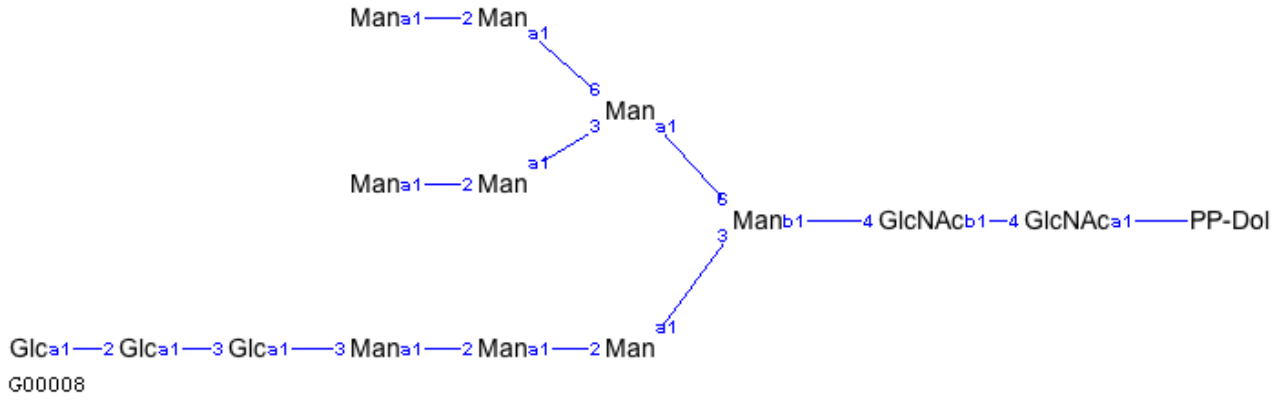
- Components of LIGAND Database
  - COMPOUND Chemical compound structure
  - DRUG Drug structure
  - GLYCAN Glycan structure
  - REACTION Biochemical reaction
  - ENZYME Enzyme nomenclature
  - RPAIR Reaction pair alignment



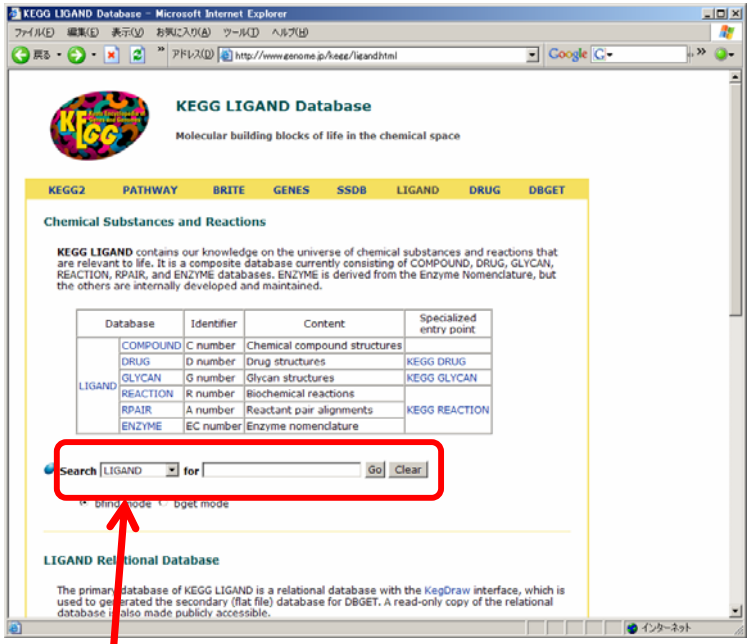
GLYCAN Database

- Glycan structures have been collected.

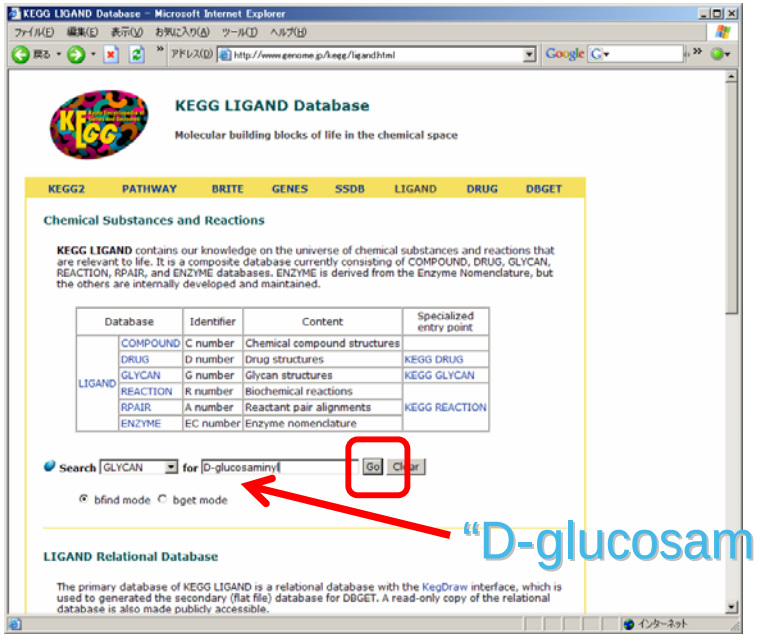
- Each structure is described as tree,
  - whose nodes are one of monosaccharides.
  - A tree is a kind of a graph, without ring structures.



- The structural format is the KCF for Glycan.

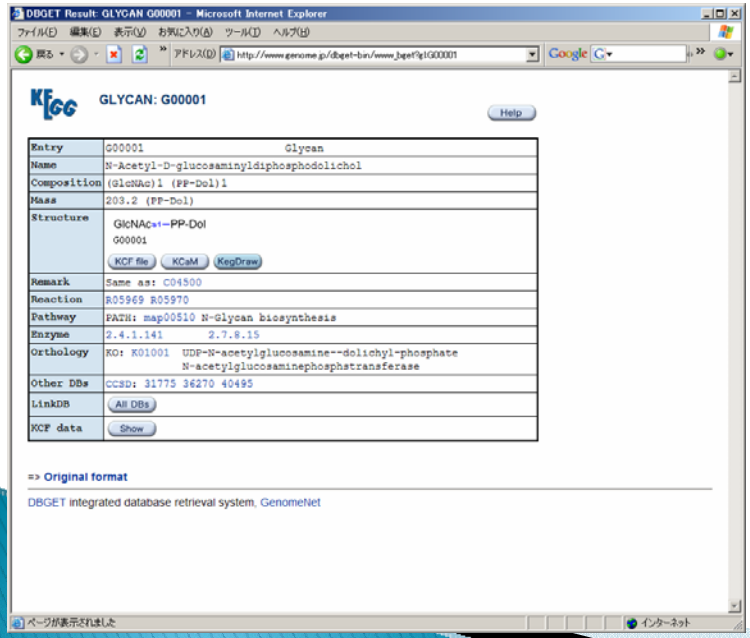


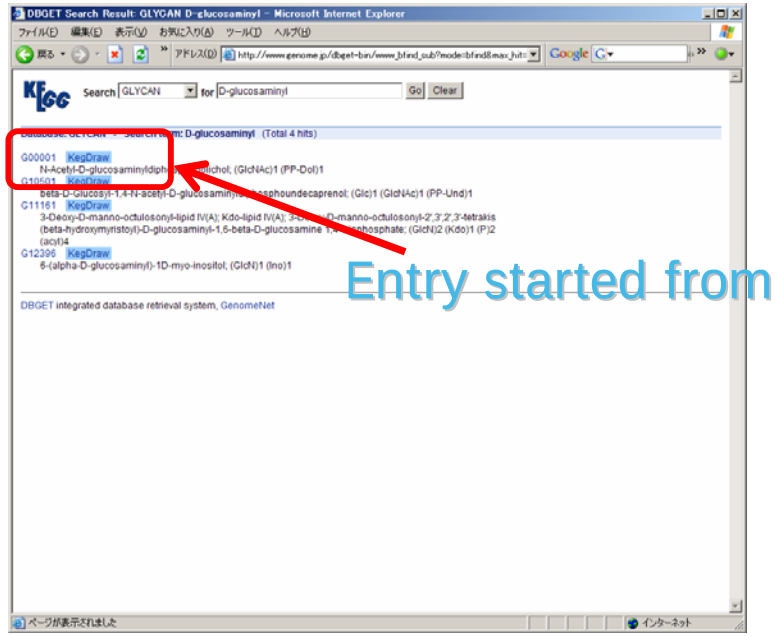
Select “GLYCAN”



“D-glucosaminyl”

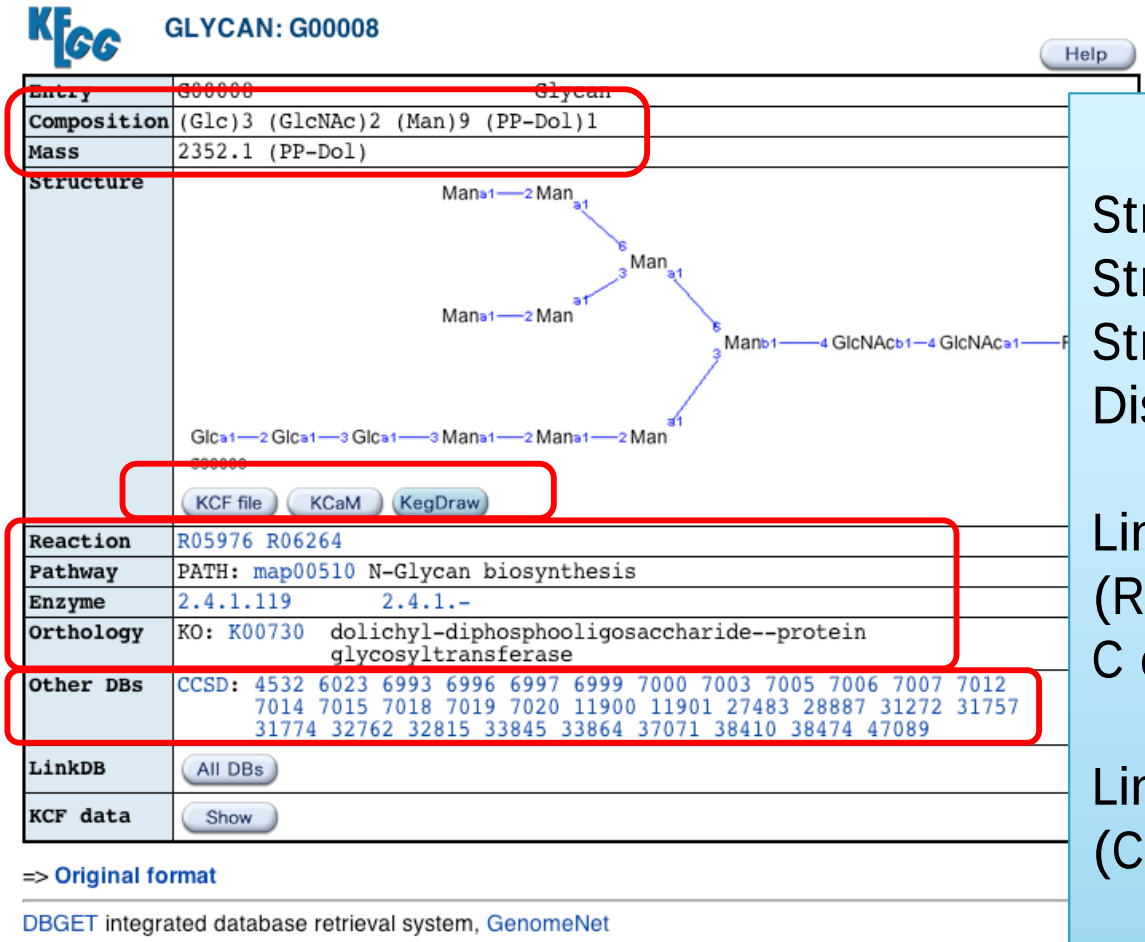
Search example of GLYCAN entry





Entry started from “G”

Example of GLYCAN Entry



KEGG GLYCAN: G00008

Help

Entry	G00008 Glycan
Composition	(Glc)3 (GlcNAc)2 (Man)9 (PP-Dol)1
Mass	2352.1 (PP-Dol)
Structure	Man α1—2 Man α1 / Man α1—2 Man α1 / 6 3 / Man α1 / Man β1—4 GlcNAc β1—4 GlcNAc α1—R / 6 3 / Man α1 / Glc α1—2 Glc α1—3 Glc α1—3 Man α1—2 Man α1—2 Man α1 / G00008
	KCF file KCaM KegDraw
Reaction	R05976 R06264
Pathway	PATH: map00510 N-Glycan biosynthesis
Enzyme	2.4.1.119 2.4.1.-
Orthology	KO: K00730 dolichyl-diphosphooligosaccharide--protein glycosyltransferase
Other DBs	CCSD: 4532 6023 6993 6996 6997 6999 7000 7003 7005 7006 7007 7012 7014 7015 7018 7019 7020 11900 11901 27483 28887 31272 31757 31774 32762 32815 33845 33864 37071 38410 38474 47089
LinkDB	All DBs
KCF data	Show

=> Original format

DBGET integrated database retrieval system, GenomeNet

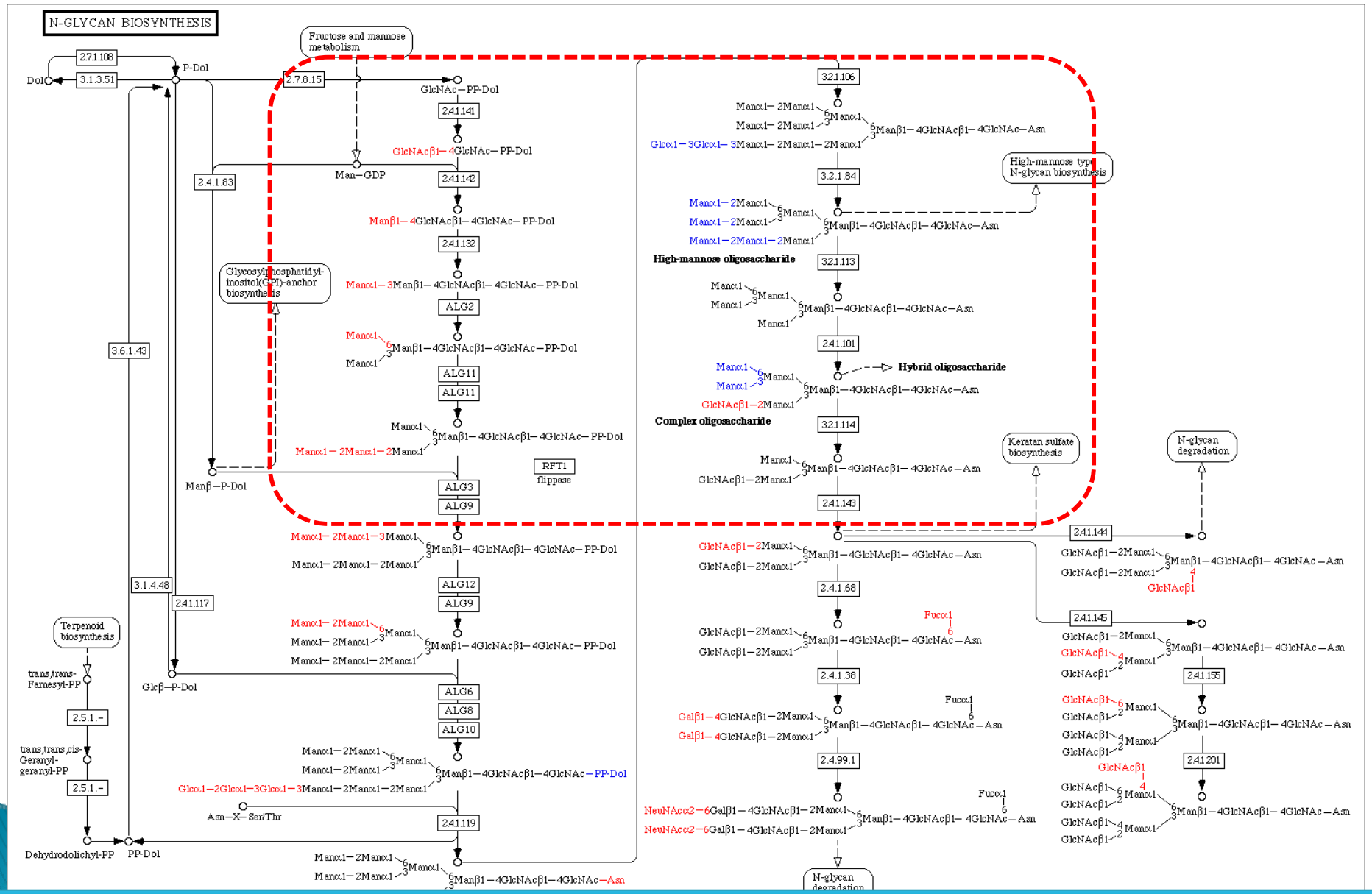
Structure formula (composition)
Structure files (KCF)
Structure search
Display button (KegDraw)

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

Link to other databases
(CCSD)

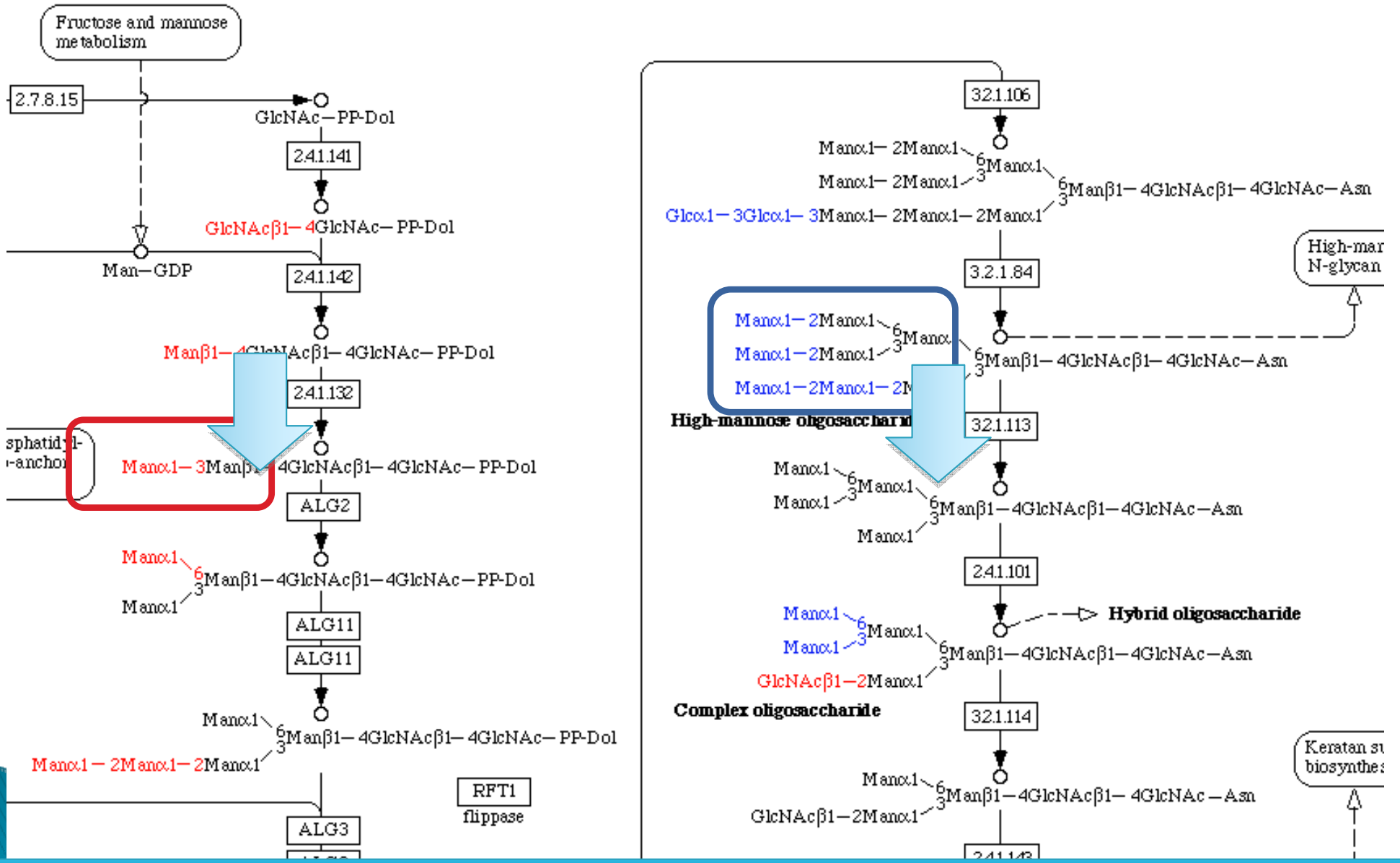
http://www.genome.jp/dbget-bin/www_bget?glycan+G00008

Pathway for Glycan



http://www.genome.jp/kegg/pathway/map/map00510.html

Pathway for Glycan



http://www.genome.jp/kegg/pathway/map/map00510.html

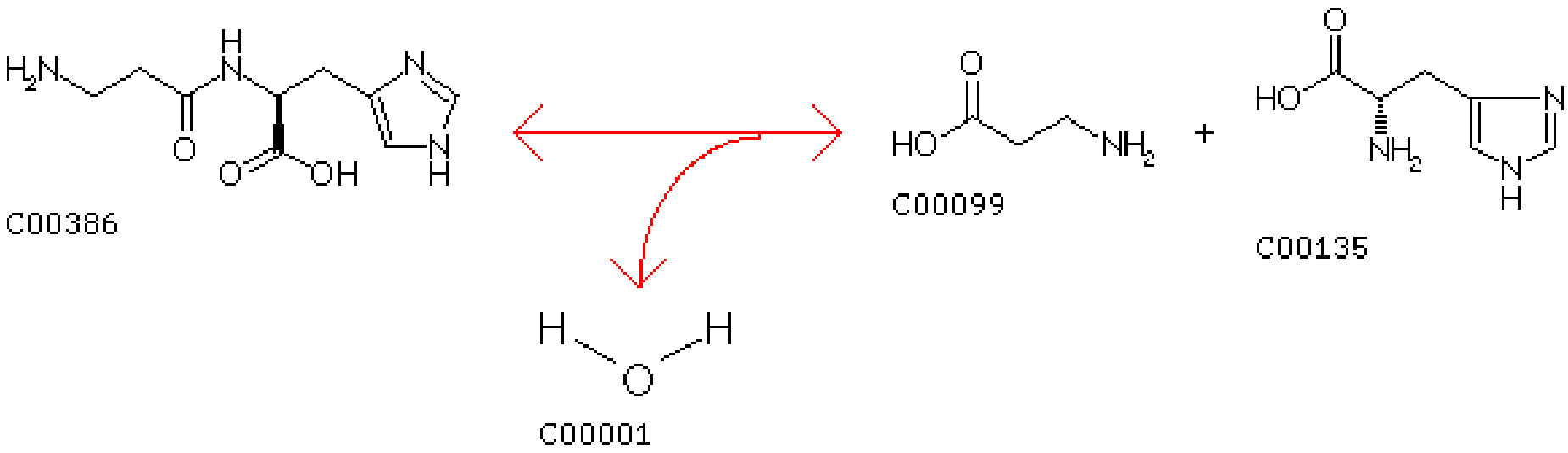
LIGAND Database

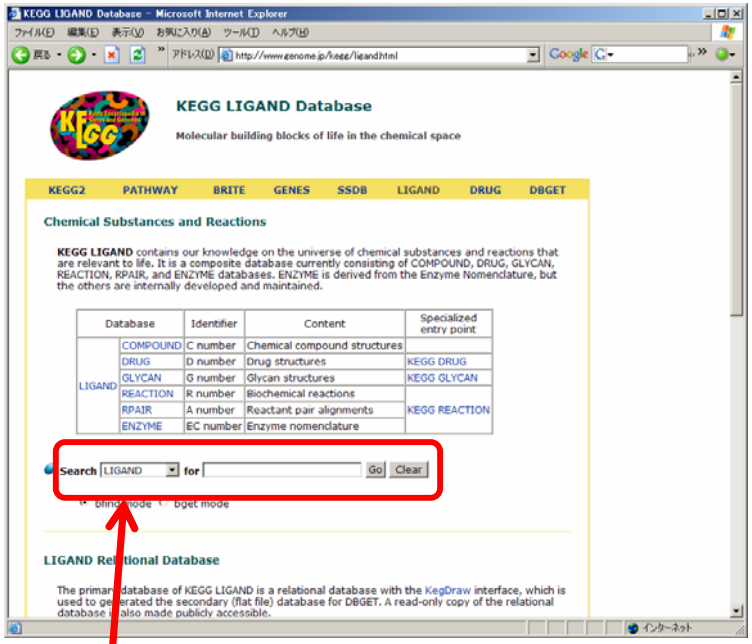
- Components of LIGAND Database
  - COMPOUND Chemical compound structure
  - DRUG Drug structure
  - GLYCAN Glycan structure
  - REACTION Biochemical reaction
  - ENZYME Enzyme nomenclature
  - RPAIR Reaction pair alignment

REACTION Database

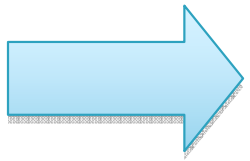
- REACTION contains the biochemical reaction information.
  - They are mainly enzymatic reactions in metabolism.

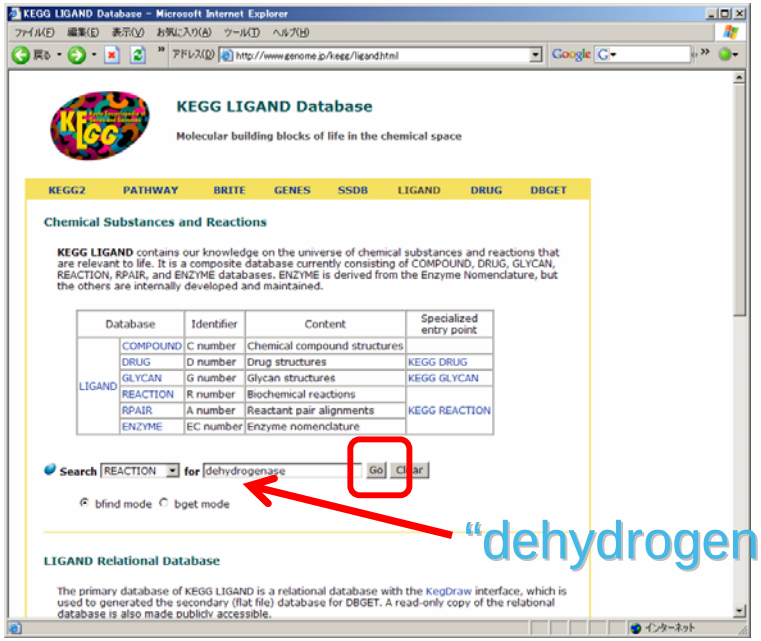
- Information about binary relations between reactants are also maintained. (see RPAIR)





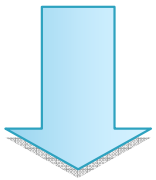
Select “REACTION”

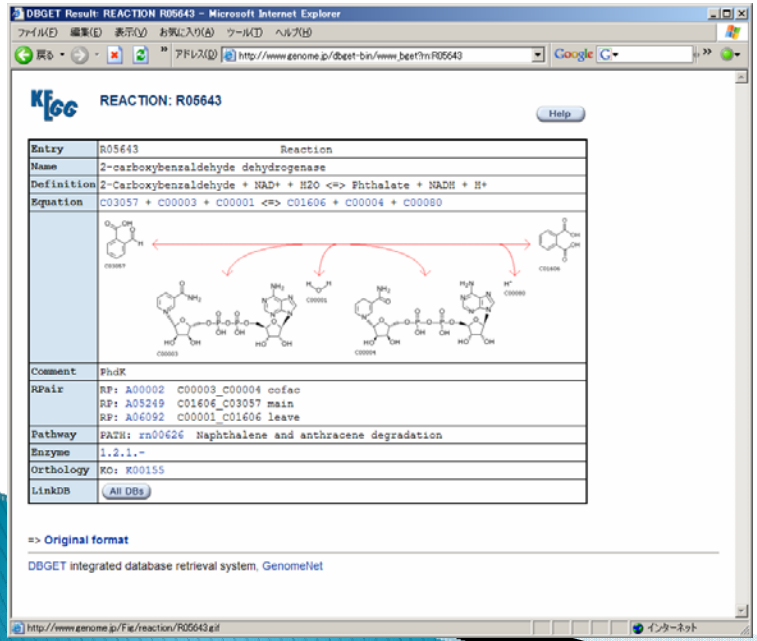


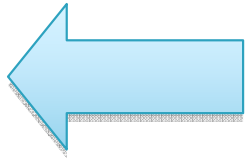


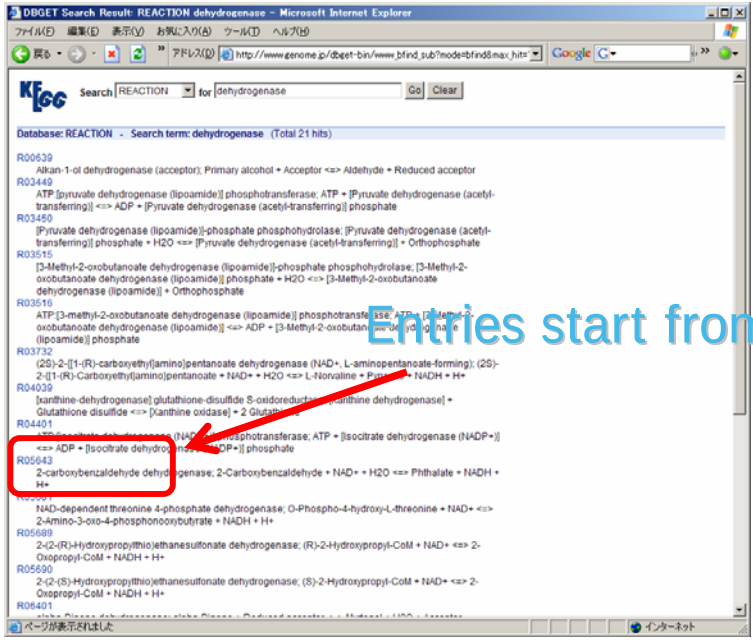
“dehydrogenase”

Search example of REACTION entry







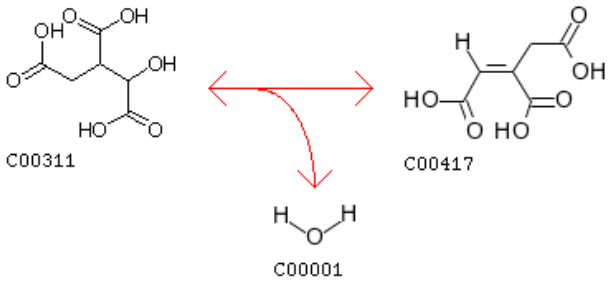


Entries start from “R”

Example of REACTION Entry

KEGG **REACTION: R01900**

Help

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

=> Original format

DBGET integrated database retrieval system, GenomeNet

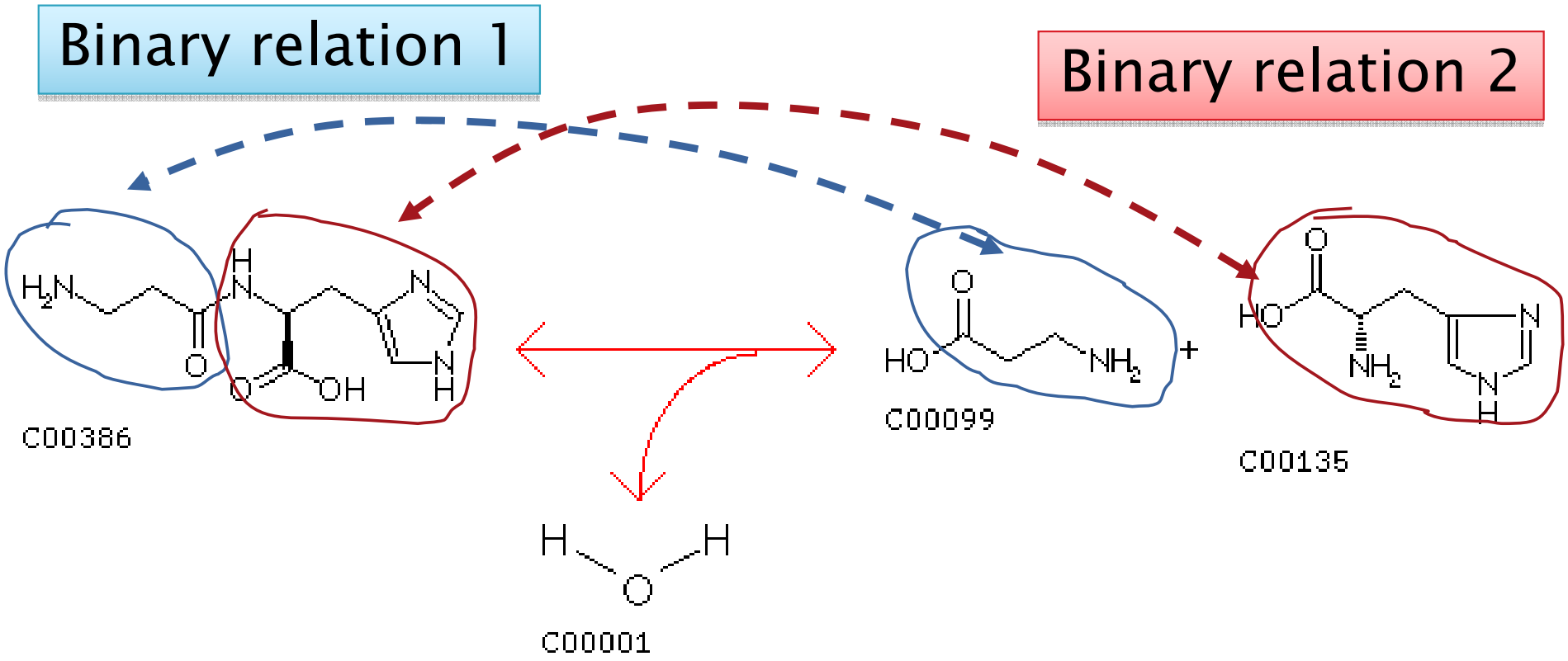
Reaction formula
- with metabolite names
- with C numbers
- with chemical structures

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

http://www.genome.jp/dbget-bin/www_bget?reaction+R01900

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.

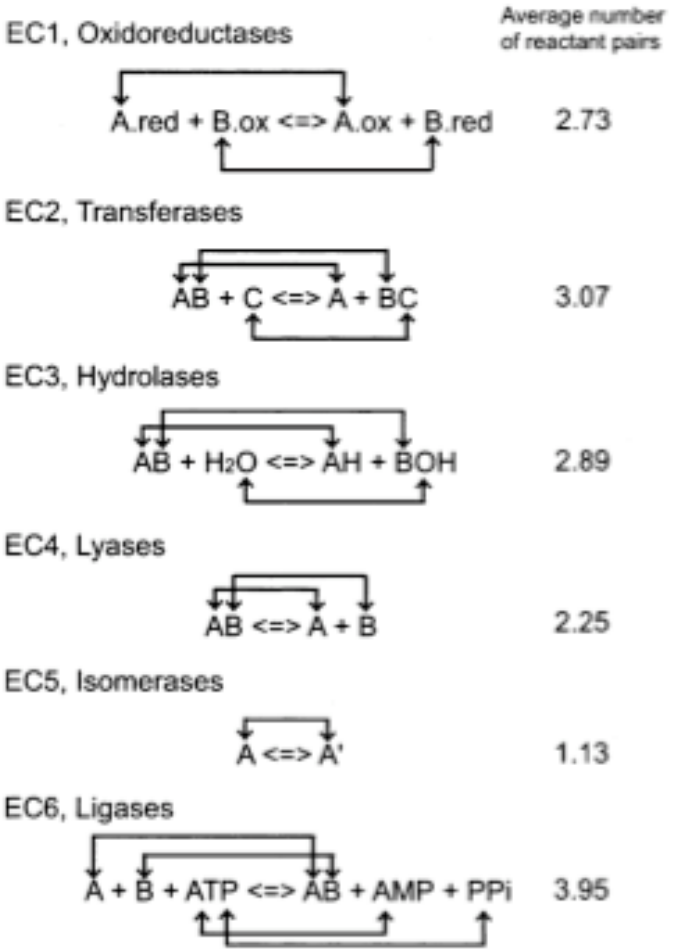
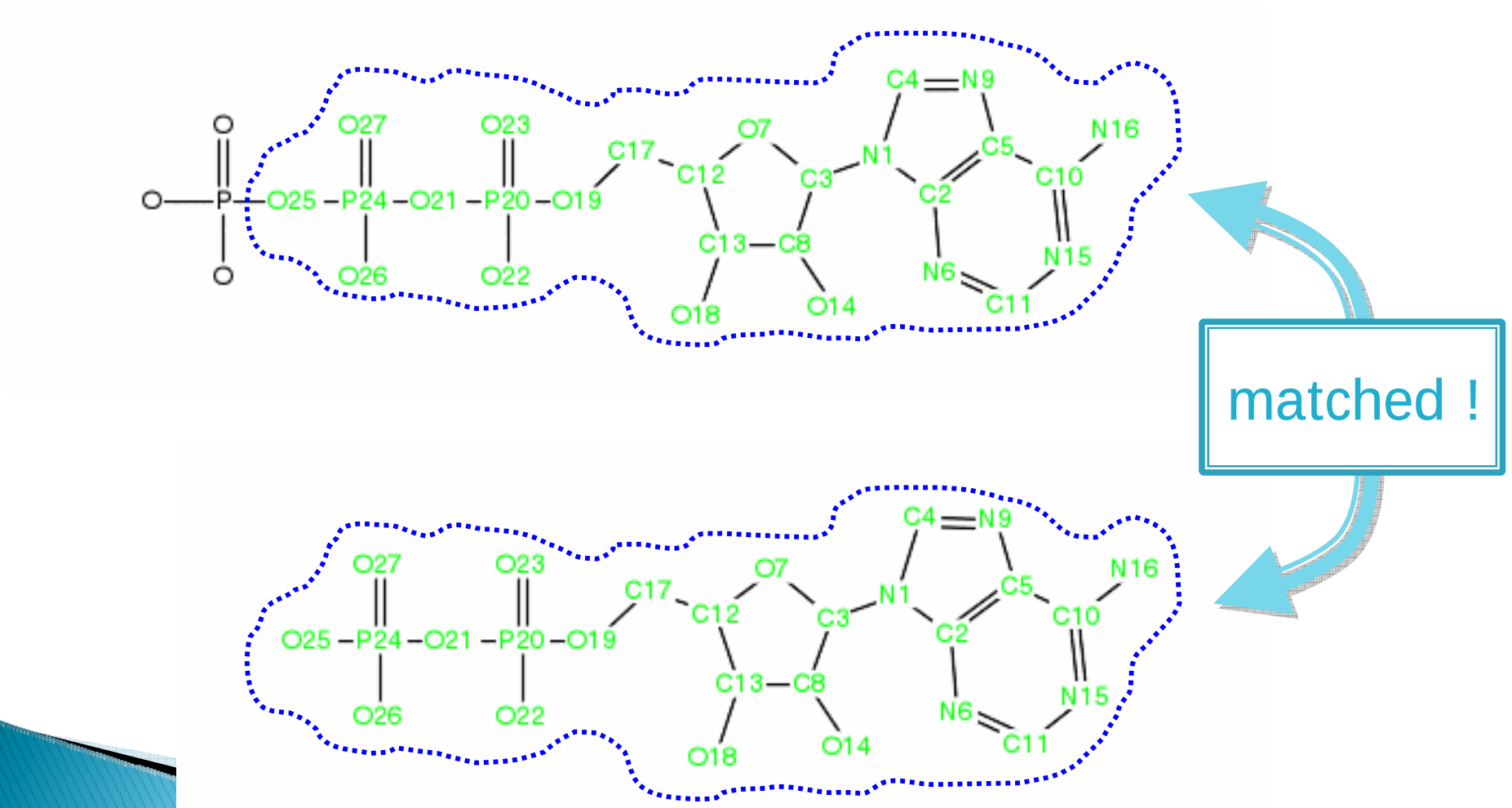


Figure 1. Extraction of reactant pairs from reactions of the six EC classes. Reactions in each class show unique topology of the reactant pairs, resulting in the average number of reactant pairs shown on the right.

Kotera et al., J Am Chem Soc 126:16487-16498 (2004)

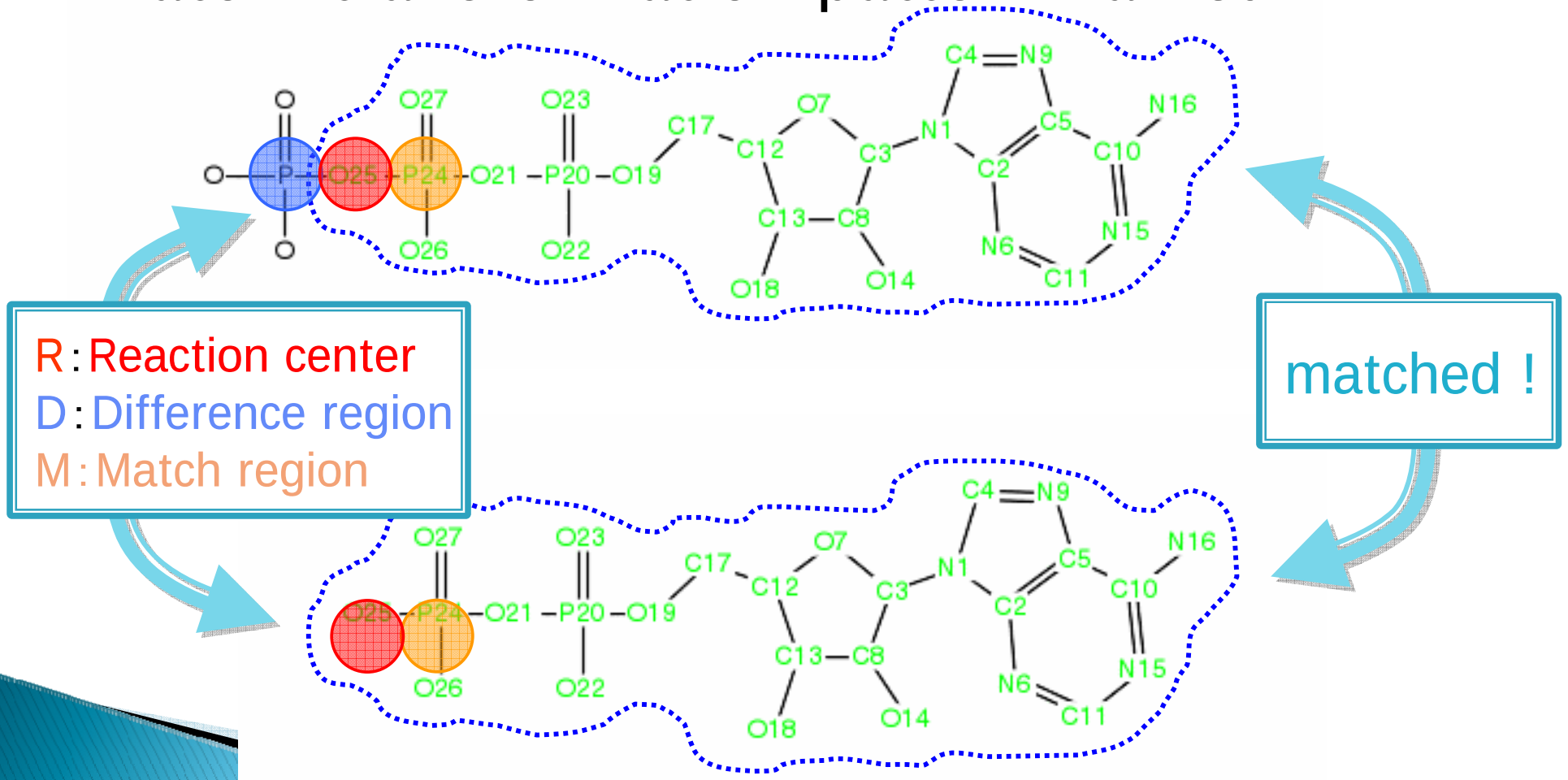
From Binary Relations,

- we can compute the atom alignments.



From the Computed Alignment,

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

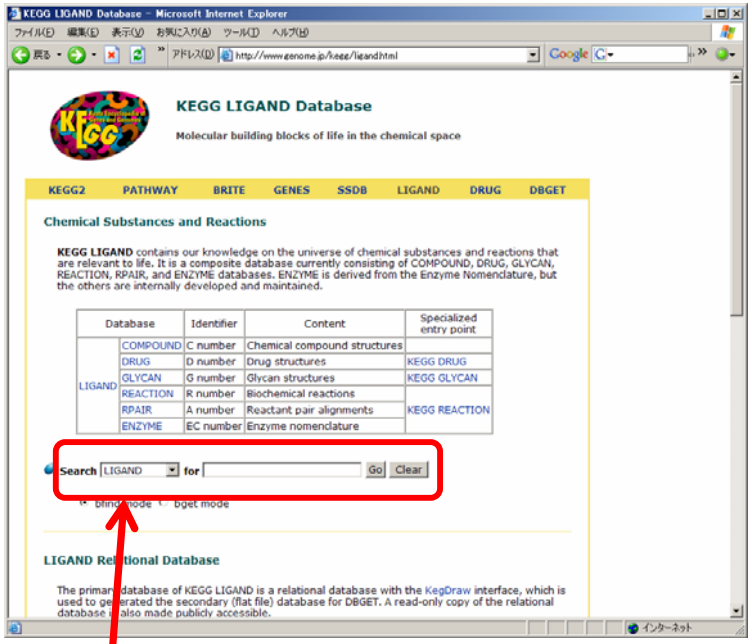


Kotera et al., J Am Chem Soc 126:16487-16498 (2004)

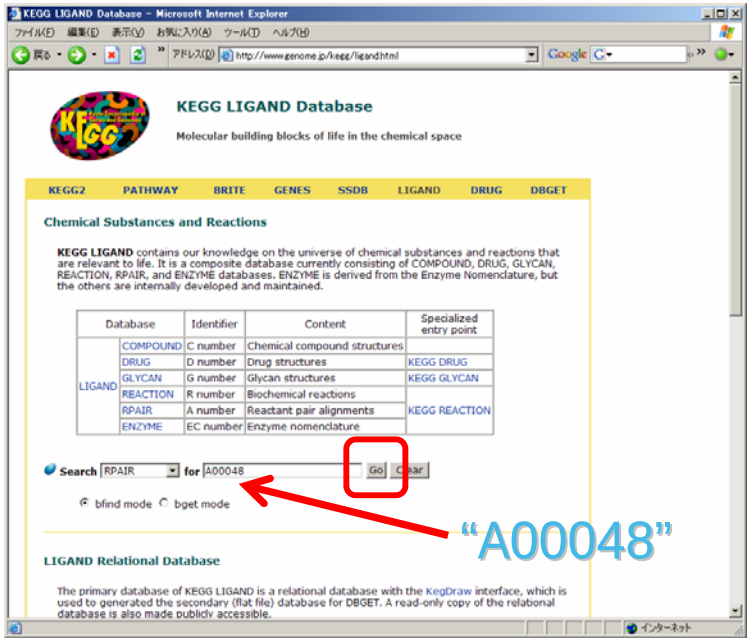
LIGAND Database

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

Oh et al., J Chem Inf Model 47:1702-1712 (2007)

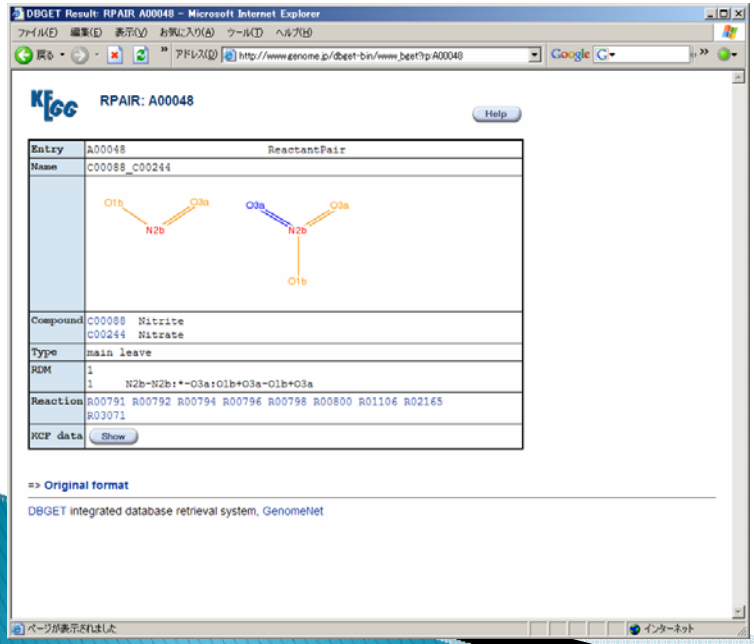


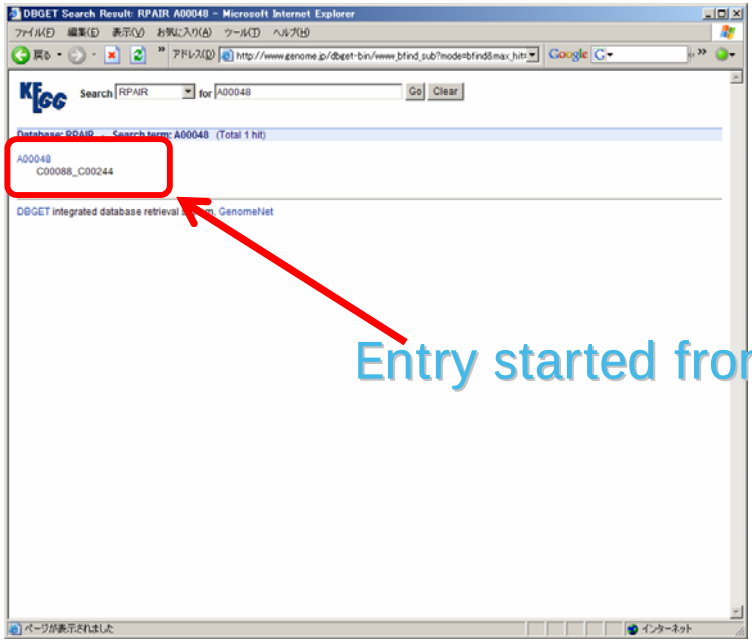
Select “RPAIR”



“A00048”

Search example of RPAIR entry



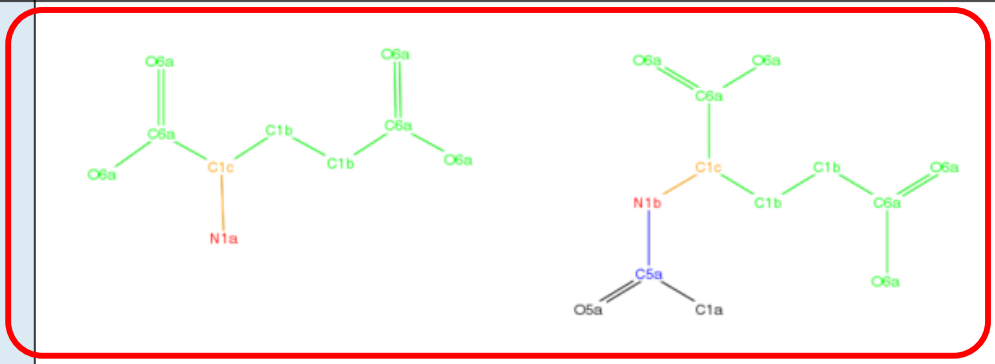


Entry started from “A”

Example of REACTION Entry

KEGG **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	Show

=> Original format

DBGET integrated database retrieval system, GenomeNet

Alignment result

Character string notation of RDM

Link to other databases (COMPOUND, REACTION)

Alignment result（KCF）

http://www.genome.jp/dbget-bin/www_bget?rp+A04458

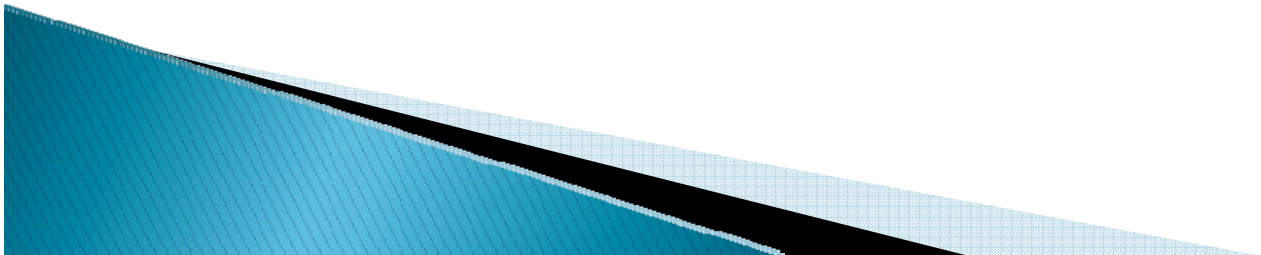
LIGAND Database

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

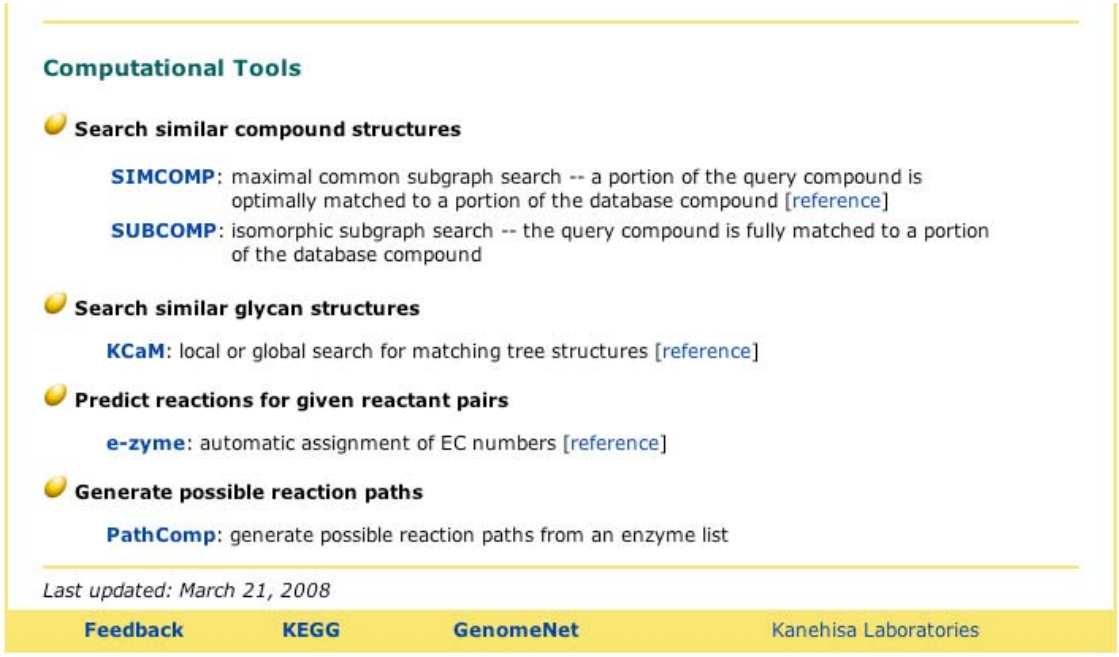
ENZYME Database

- 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 Computaional 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

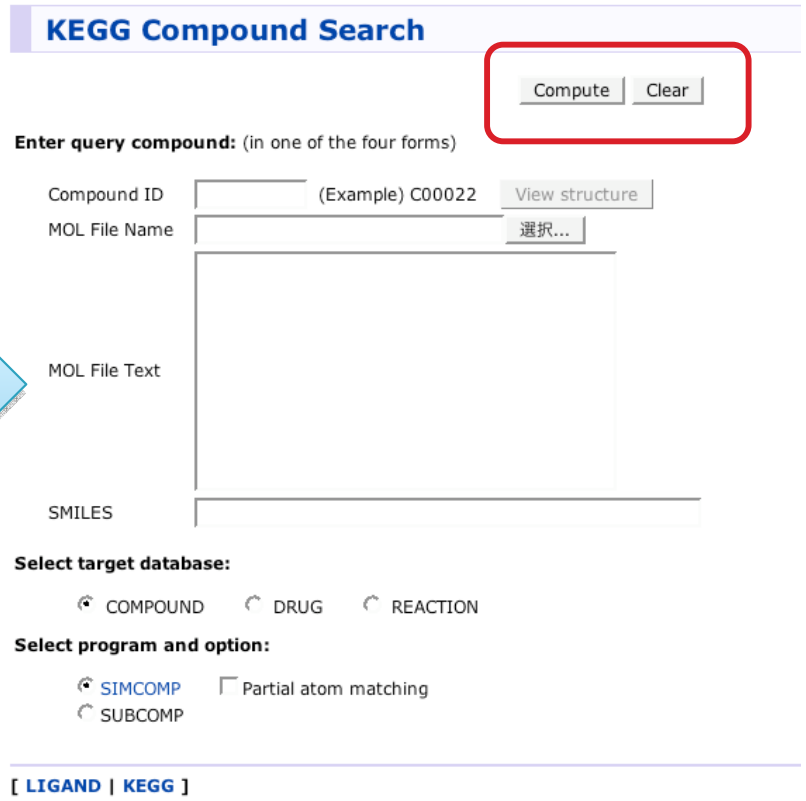
- Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- Desktop application for compounds and glycans
  - KegDraw Java based software to input 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.

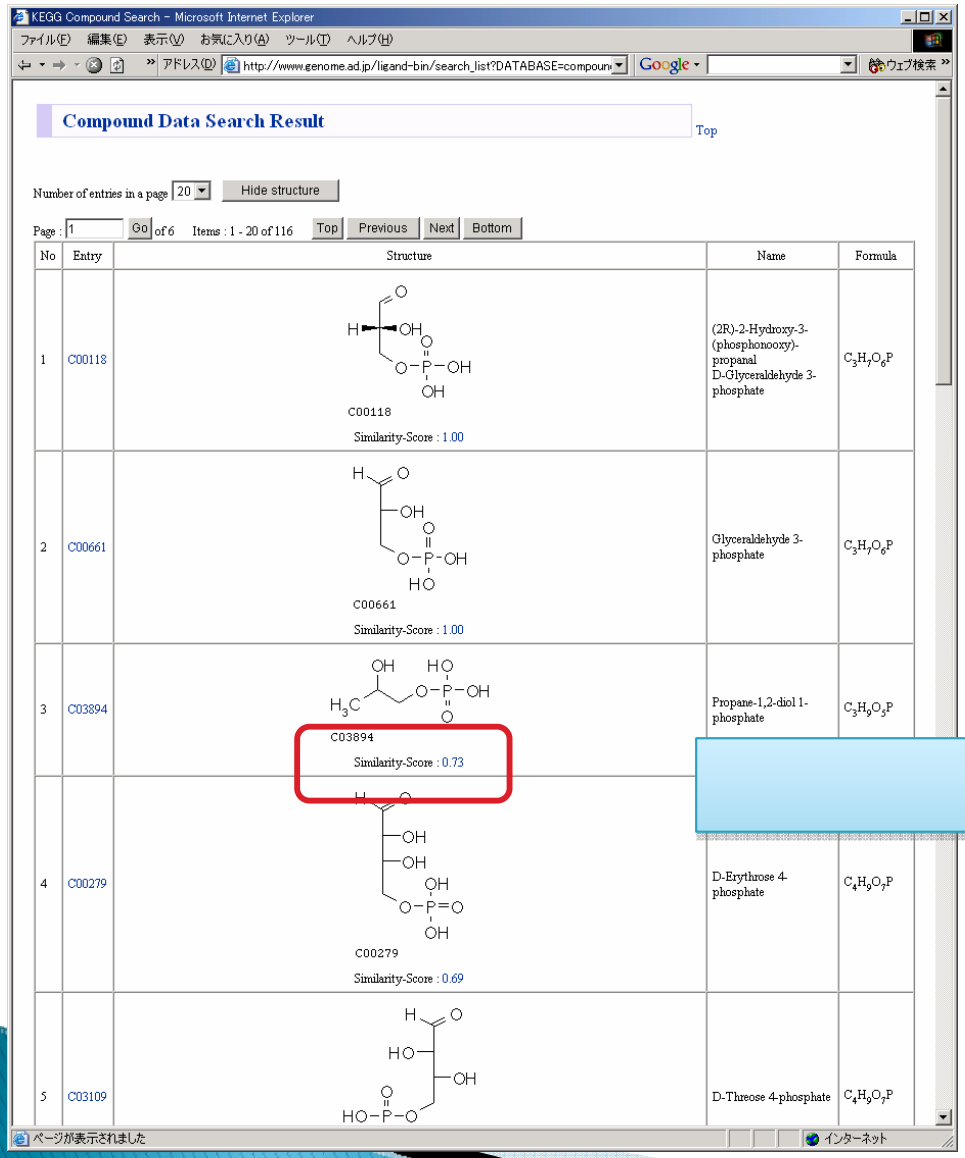
Hattori et al., J Am Chem Soc 125:11853-65 (2003)

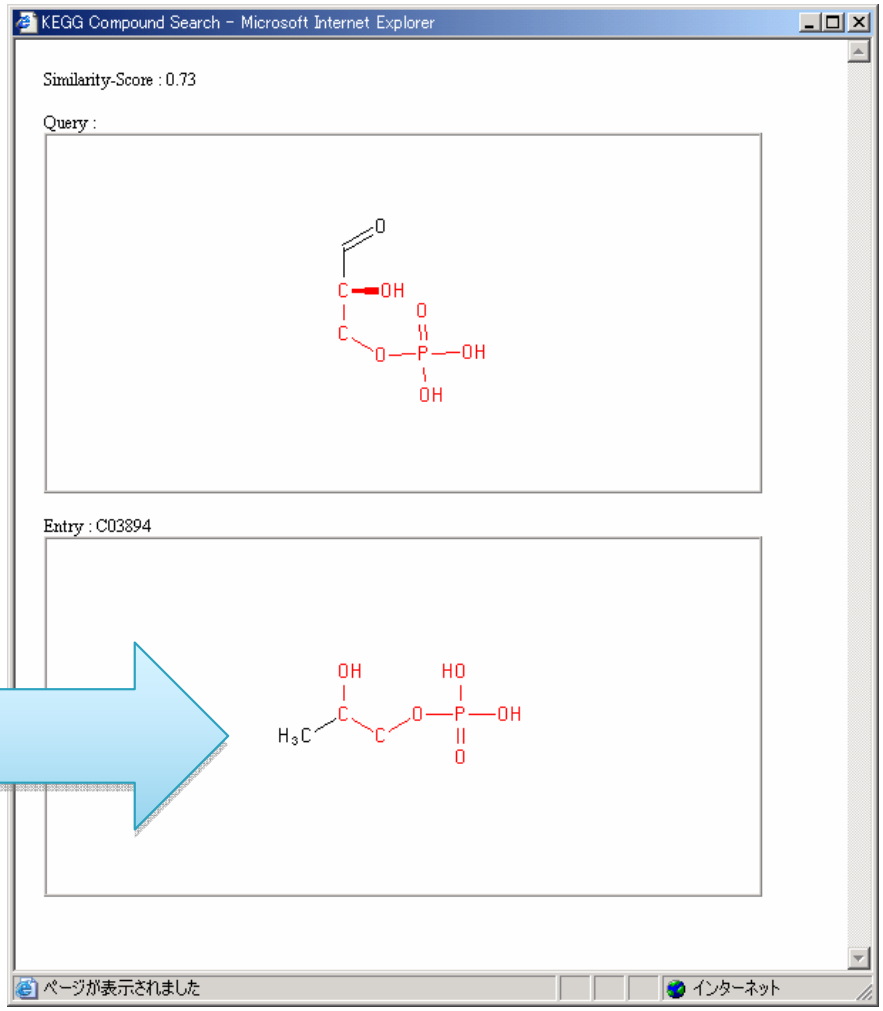
SIMCOMP: Start page



http://www.genome.jp/tools/simcomp/

SIMCOMP: Result example

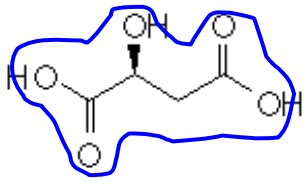




Similarity Score in SIMCOMP

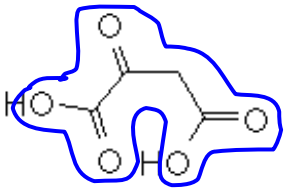
$$JC(G_1, G_2) \equiv \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} = \frac{|MCS(G_1, G_2)|}{|G_1 + G_2 - MCS(G_1, G_2)|} = \frac{|MCS(G_1, G_2)|}{|G_1| + |G_2| - |MCS(G_1, G_2)|}$$

(s)-malate



C00149

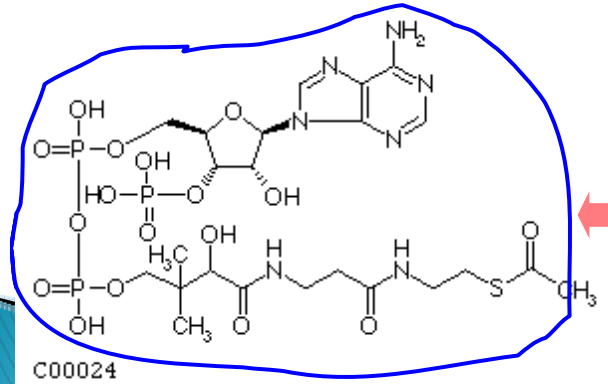
oxaloacetate



C00036

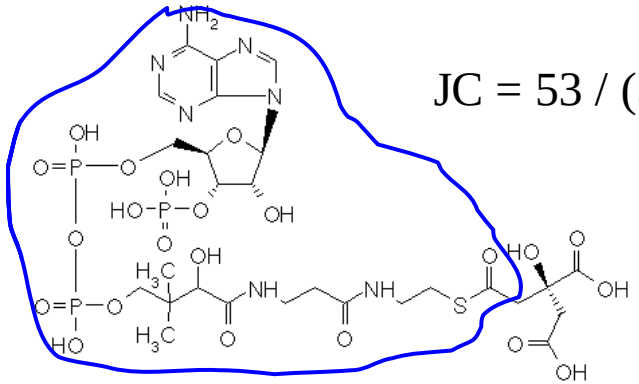
JC = 9 / (9 + 9 – 9) = 1.000

Acetyl-CoA



C00024

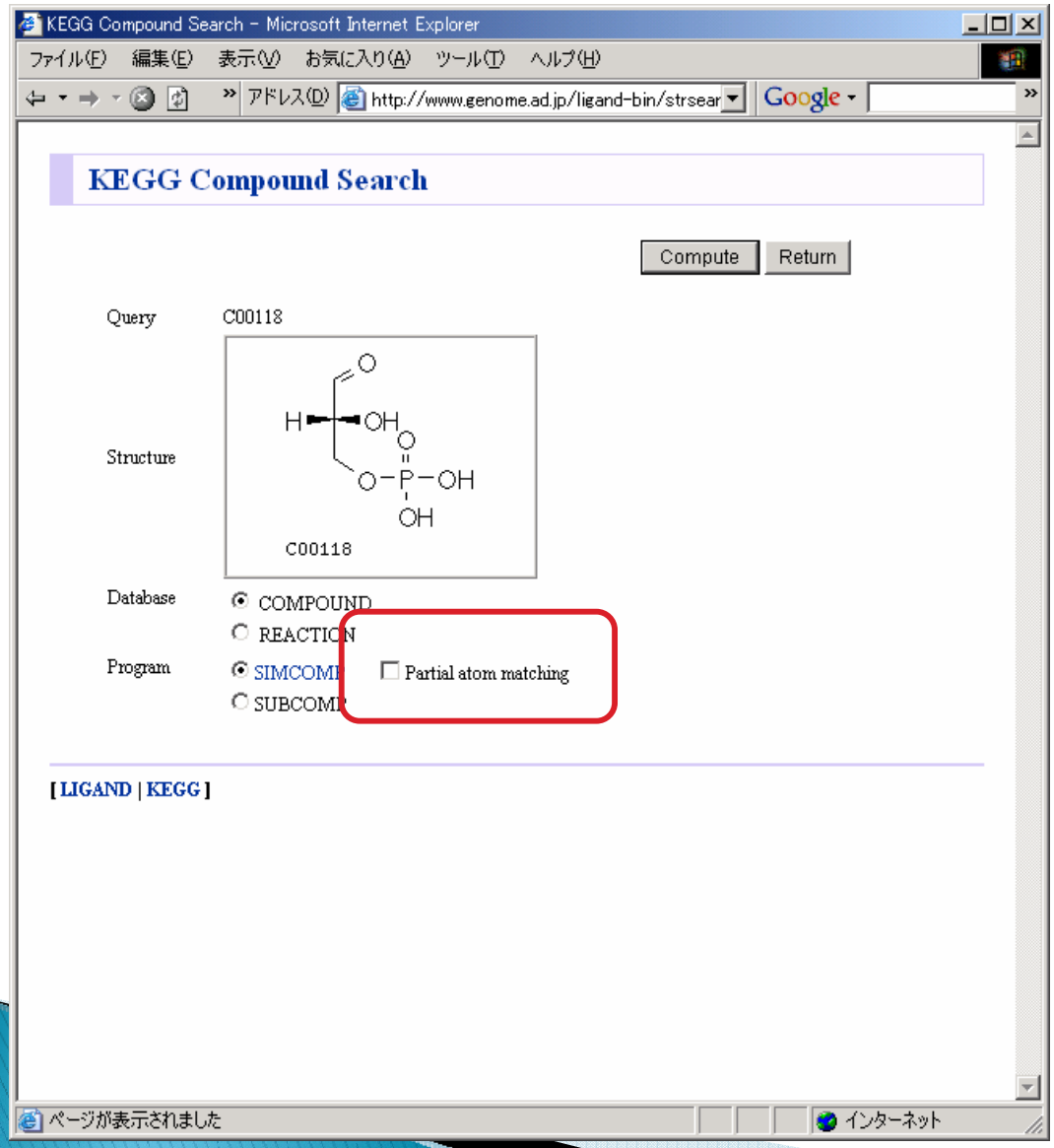
(3S)-citryl-CoA



C00566

JC = 53 / (53 + 62 – 53) = 0.854

About the Option in SIMCOMP



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

↓

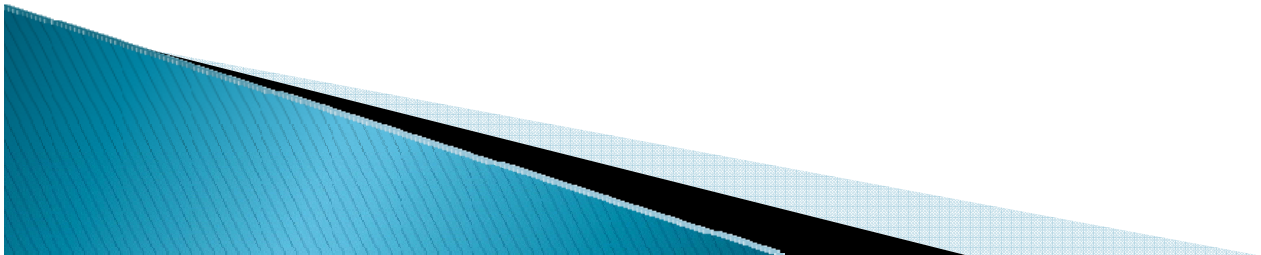
The resulting alignment will become larger and thus the score will also become larger.

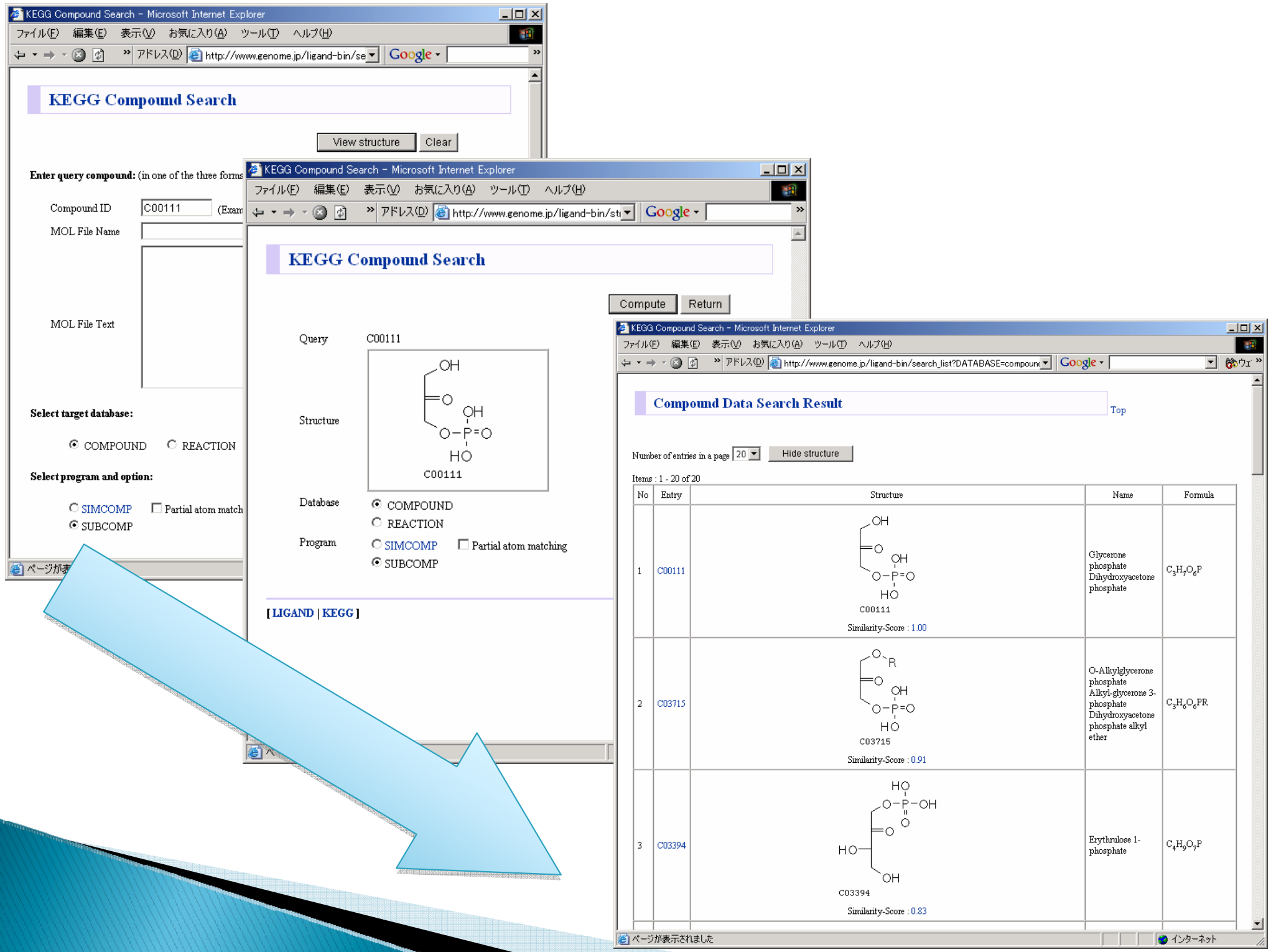
Provided Tools in KEGG

- 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
KEGG Compound Search
View structure
Clear
Enter query compound: (in one of the three forms
Compound ID
C00111
MOL File Name
MOL File Text
Select target database:
COMPOUND
REACTION
Select program and option:
SIMCOMP
Partial atom match
SUBCOMP
KEGG Compound Search - Microsoft Internet Explorer
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
Compound Data Search Result
Top
Number of entries in a page
20
Hide structure
Items : 1 - 20 of 20
No
Entry
Structure
Name
Formula
1
C00111
C00111
Similarity-Score : 1.00
Glycerone phosphate
Dihydroxyacetone phosphate
C3H7O6P
2
C03715
C03715
Similarity-Score : 0.91
O-Alkylglycerone phosphate
Alkyl-glycerone 3-phosphate
Dihydroxyacetone phosphate alkyl ether
C3H6O6PR
3
C03394
C03394
Similarity-Score : 0.83
Erythrulose 1-phosphate
C4H9O7P
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

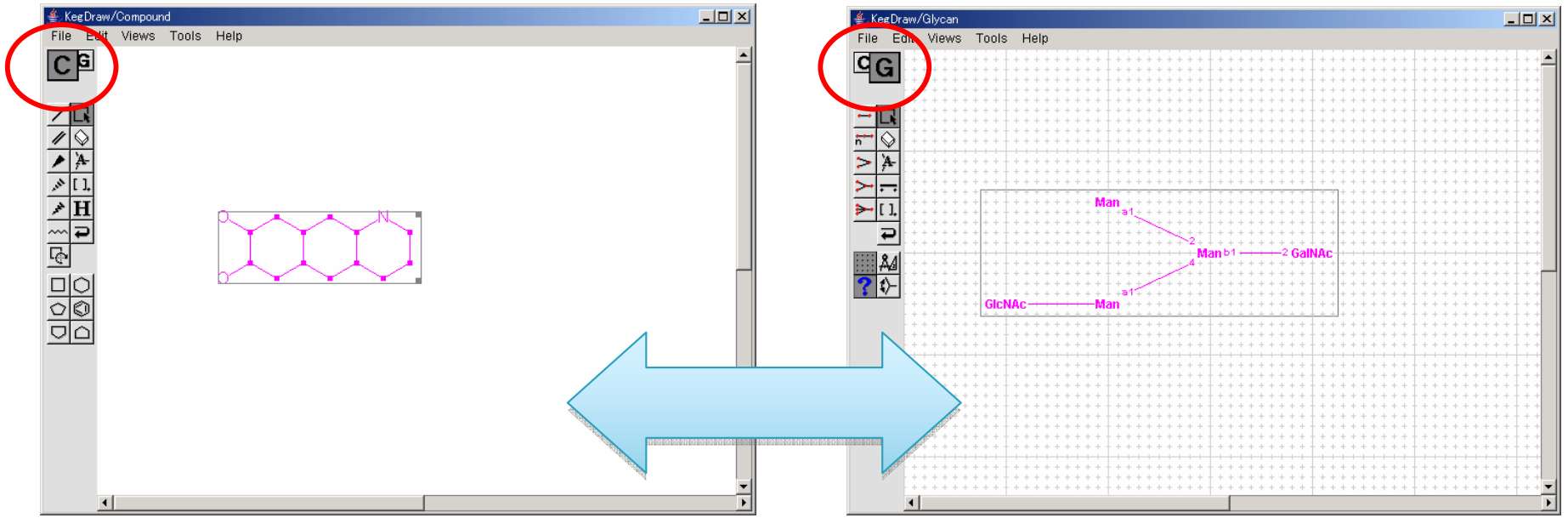
- Desktop application to edit chemical compound structures, like the ChemDraw or ISIS/Draw.
- KegDraw can also edit the glycan structures.

- A user can perform the structural search against KEGG or PubChem database via the Tools menu of KegDraw.

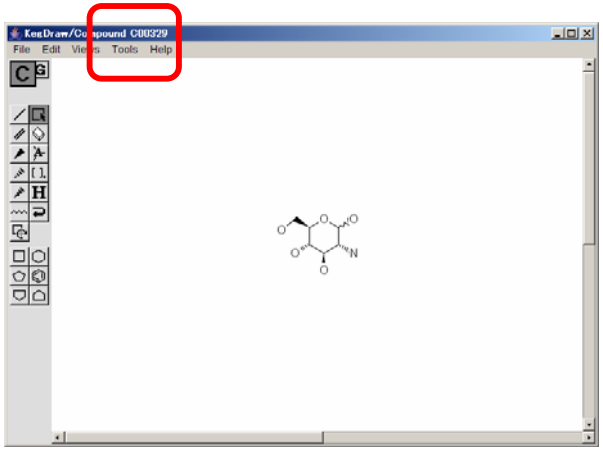
http://www.genome.jp/download/

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.

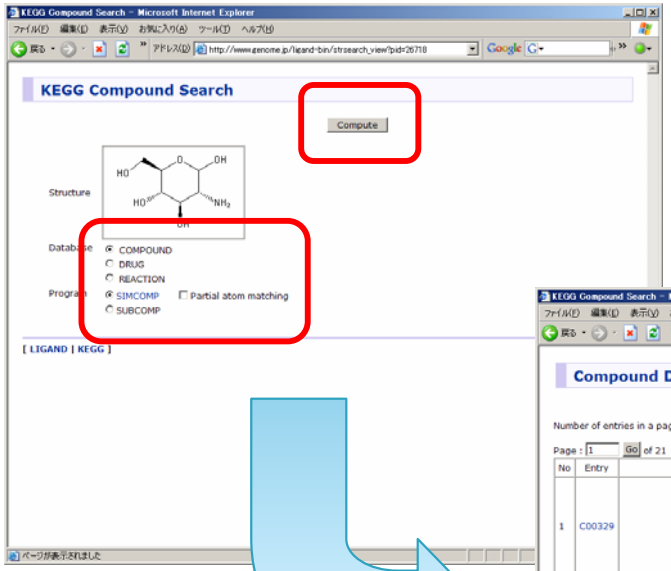


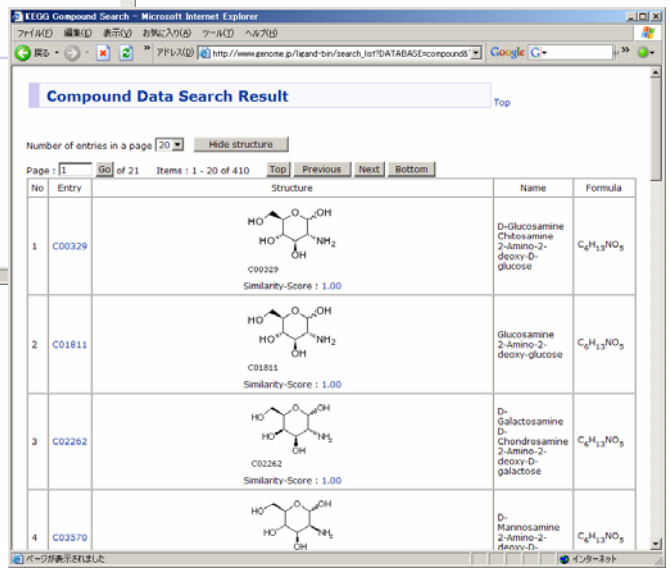
LIGAND (KEGG)

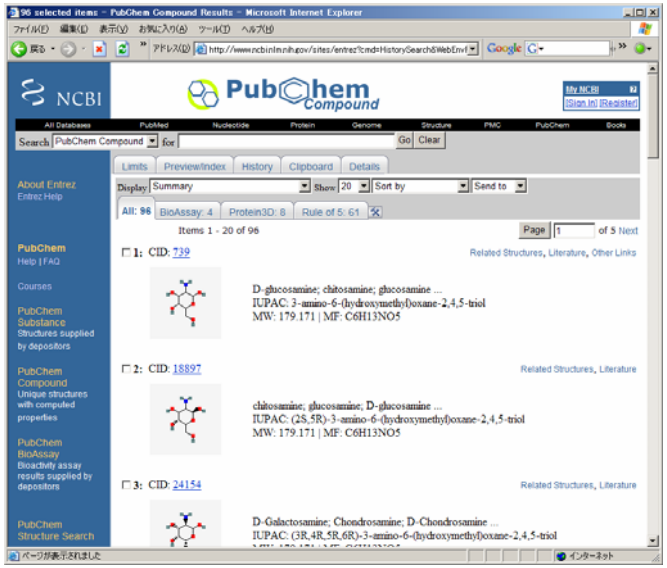


PubChem(NCBI)

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







Provided Tools in KEGG

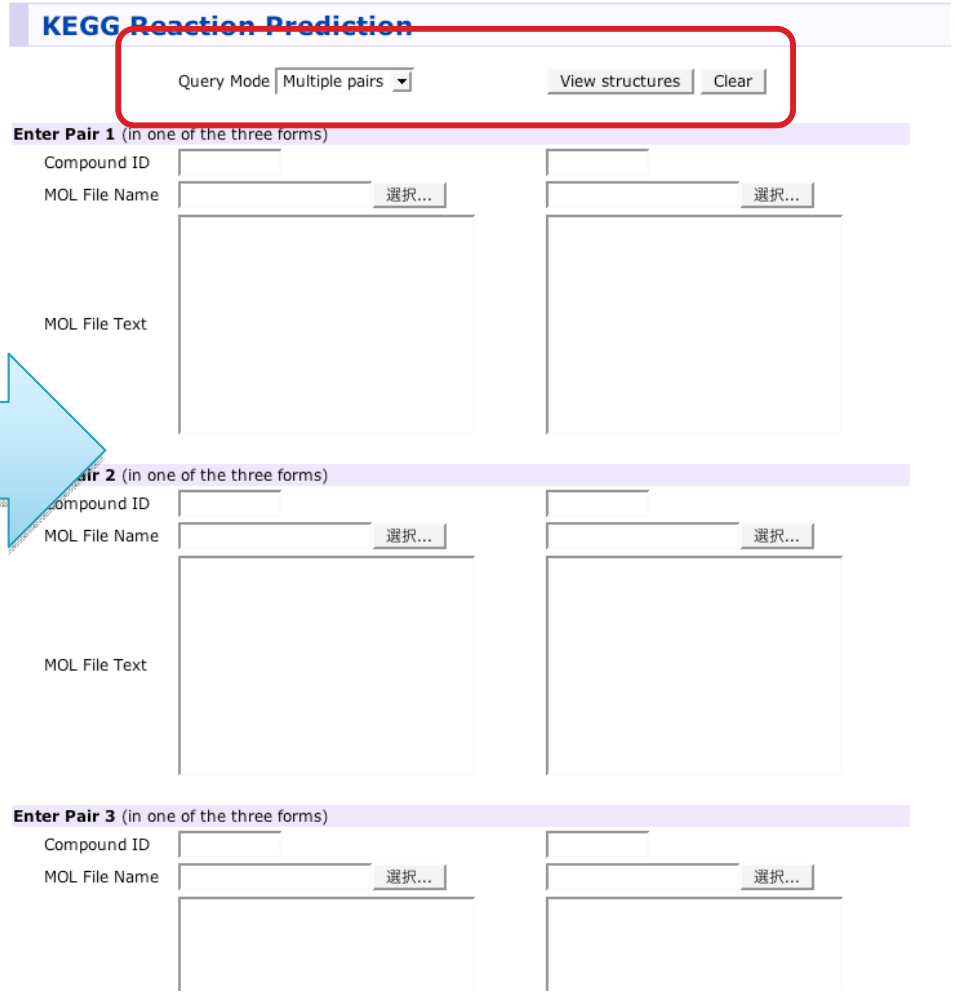
- Computation of chemical compound graphs
  - SIMCOMP to get the maximal common subgraph
  - SUBCOMP to solve the subgraph isomorphism
- Computation of glycan structures
  - KCaM Developed by Kiyoko Aoki-Kinoshita.
- Desktop application for compounds and glycans
  - KegDraw Java based software to input 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

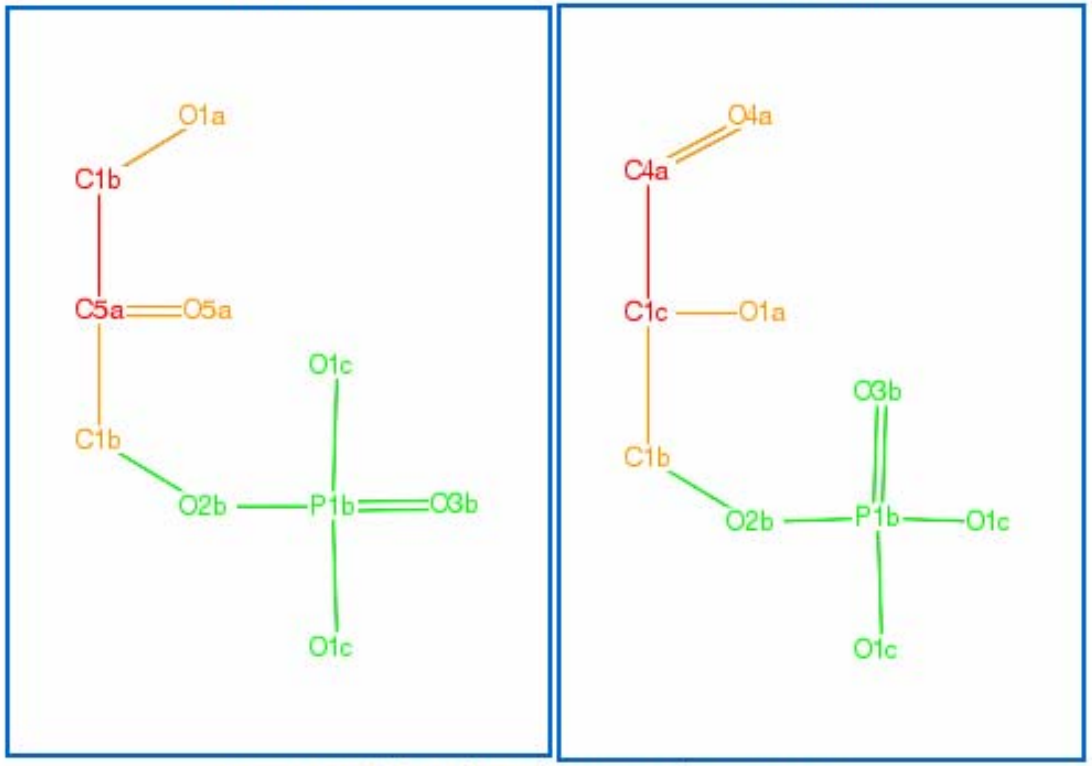


http://www.genome.jp/tools/e-zyme/

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	

[LIGAND | KEGG]

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)
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)

correlation coefficient :

$corr(\mathrm{RDM}i, \mathrm{RDM}j)$

$(i, j = 1,2,....,n)$

e-zyme: Weighted scoring

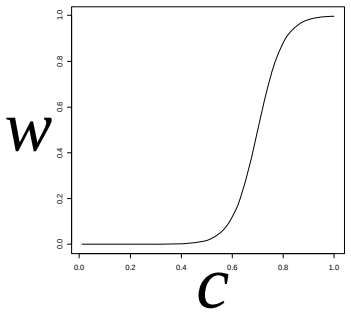
- Assume that the RDM pattern of the query reaction is computed (denoted by RDMquery)

  - The old method used in Kotera's paper: select EC*k* with the highest score by

$$S(EC_k) = f_k, \text{ where } f_k : \text{frequency of } \mathrm{EC}_k \text{ in } \mathrm{RDM}_{\text{query}}$$

  - New method: select EC*k* with the highest score by

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

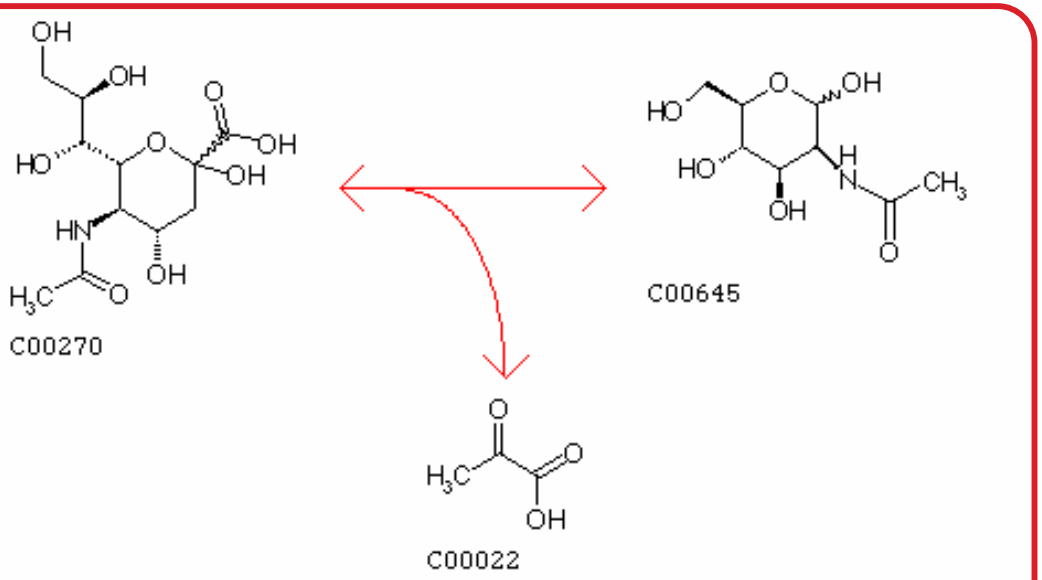


e-zyme: Multiple mode

KEGG

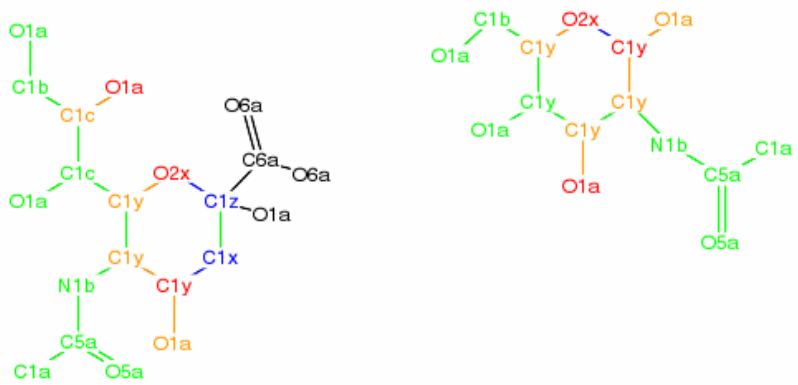
REACTION: R01811

Help

Entry	R01811 Reaction
Name	N-Acetylneuraminate pyruvate-lyase
Definition	N-Acetylneuraminate <=> N-Acetyl-D-mannosamine + Pyruvate
Equation	C00270 <=> C00645 + C00022
	
RPair	RP: A00421 C00270_C00645 main RP: A01773 C00022_C00270 main
Pathway	PATH: rn00530 Aminosugars metabolism
Enzyme	4.1.3.3
Orthology	KO: K01639 N-acetylneuraminate lyase
LinkDB	All DBs

RPAIR: A00421

Help

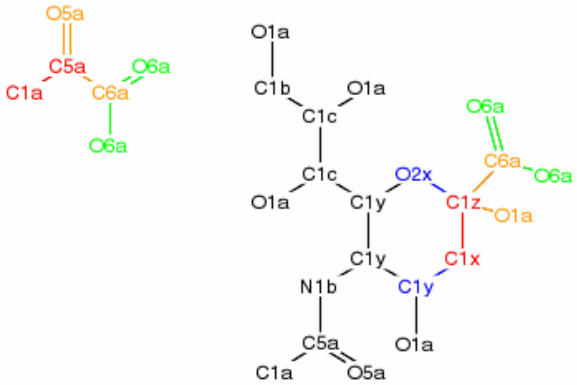
Entry	A00421 ReactantPair
Name	C00270_C00645
	
Compound	C00270 N-Acetylneuraminate C00645 N-Acetyl-D-mannosamine
Type	main
RDM	3 1 O2x-O1a:C1z-*:C1y-C1y 2 C1y-C1y:C1x+*-*+O2x:C1y+O1a-C1y+O1a 3 O1a-O2x:*-C1y:C1c-C1y
Related pair	A00295 A00429 A00430 A00548 A00945 A01132 A01262 A01269

This is the real solution for pair 1.

This is pair 2.

RPAIR: A01773

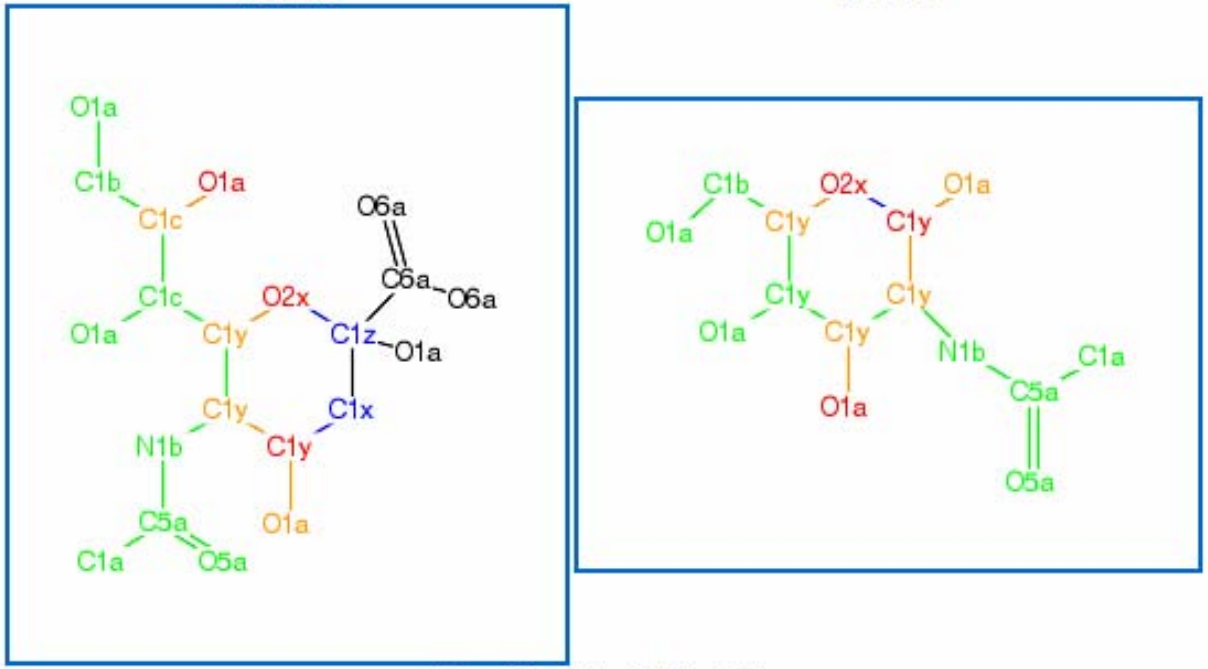
Help

Entry	A01773 ReactantPair
Name	C00022_C00270
	
Compound	C00022 Pyruvate C00270 N-Acetylneuraminate
Type	main
RDM	2 1 C1a-C1x:*-C1y:C5a-C1z 2 C5a-C1z:*-O2x:C1a+C6a+O5a-C1x+C6a+O1a
Reaction	R01811
KCF data	Show

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.

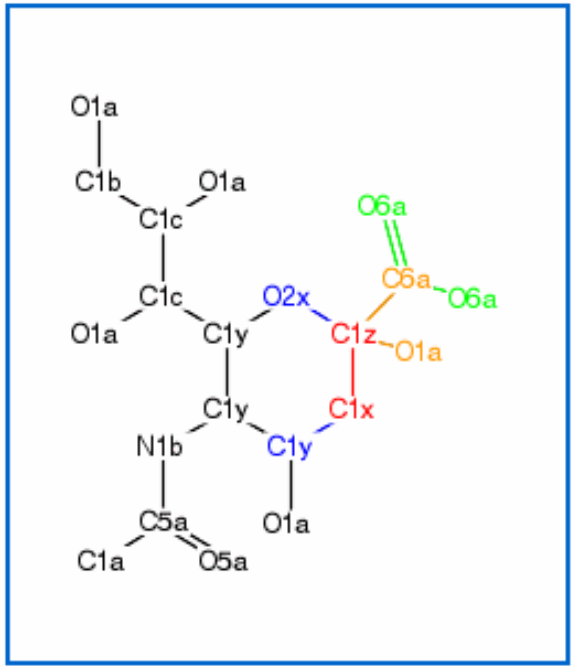
EC assignments

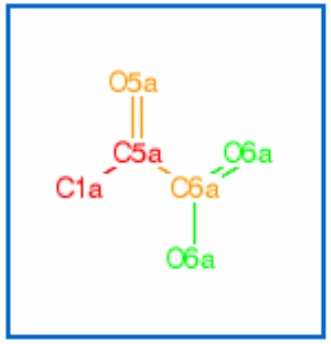
EC number	Weighted score	Observed freq.
2.5.1	71.6	1.0
4.1.3	15.7	1.0
3.1.1	2.0	
4.1.99	1.3	
4.4.1	0.9	
1.5.99	0.9	
3.3.1	0.5	

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

Pair 2 Edit

C00270 C00022





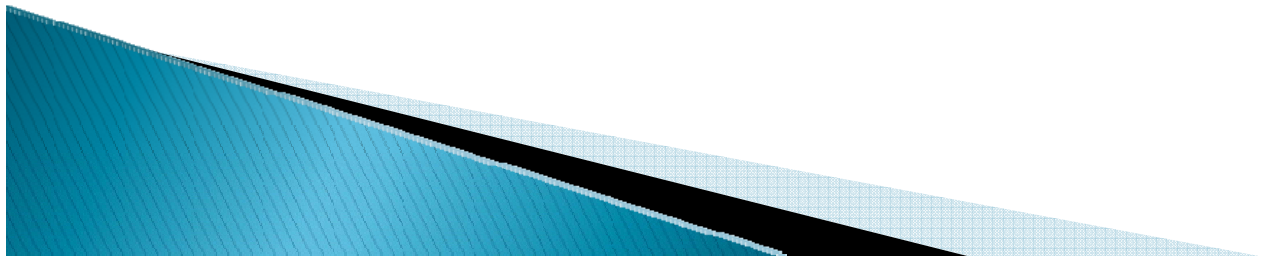
C1z-C5a:O2x-*:C1x+C6a+O1a-C1a+C6a+O5a
C1x-C1a:C1y-*:C1z-C5a

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

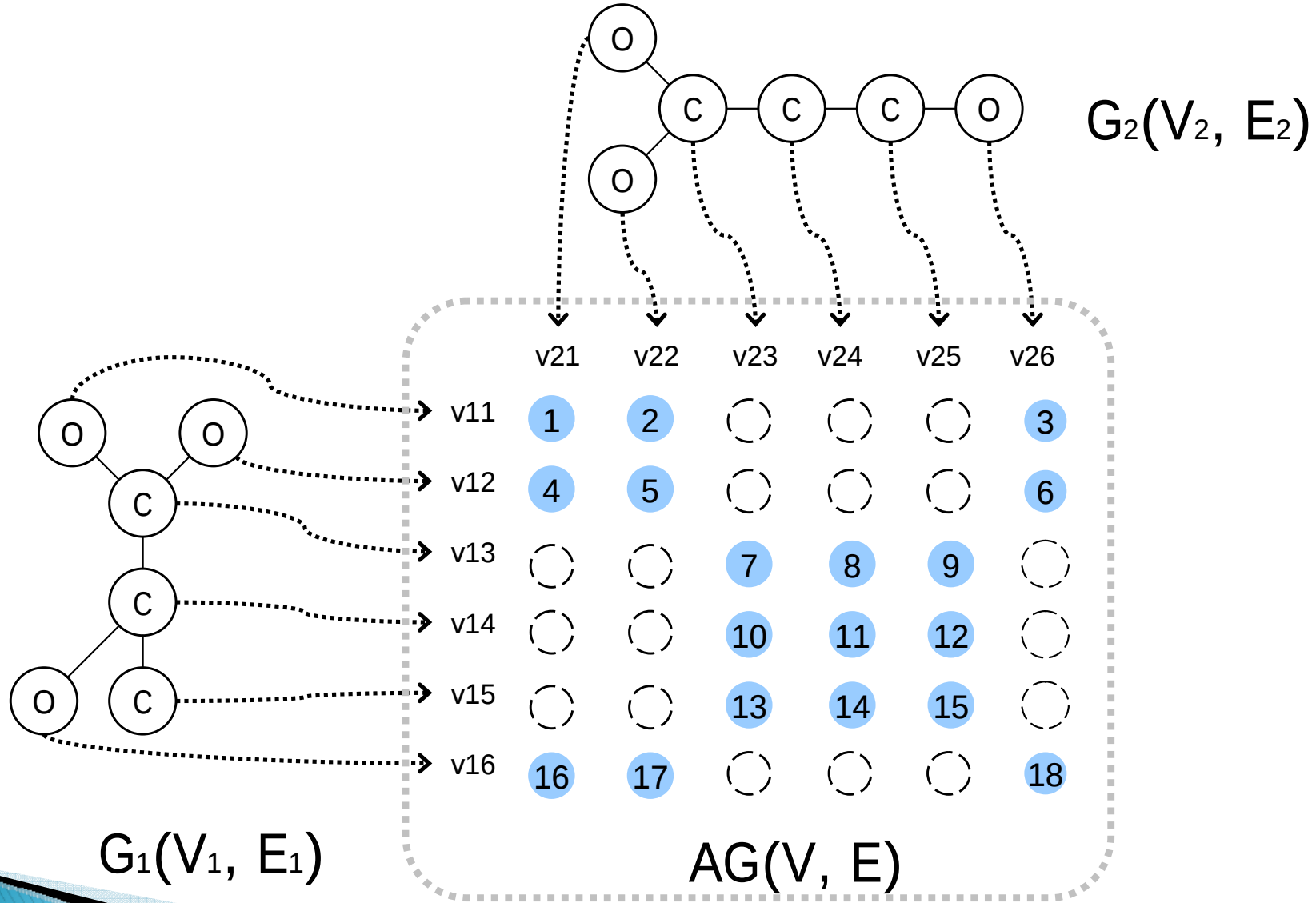
EC assignments

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	

[LIGAND | KEGG]

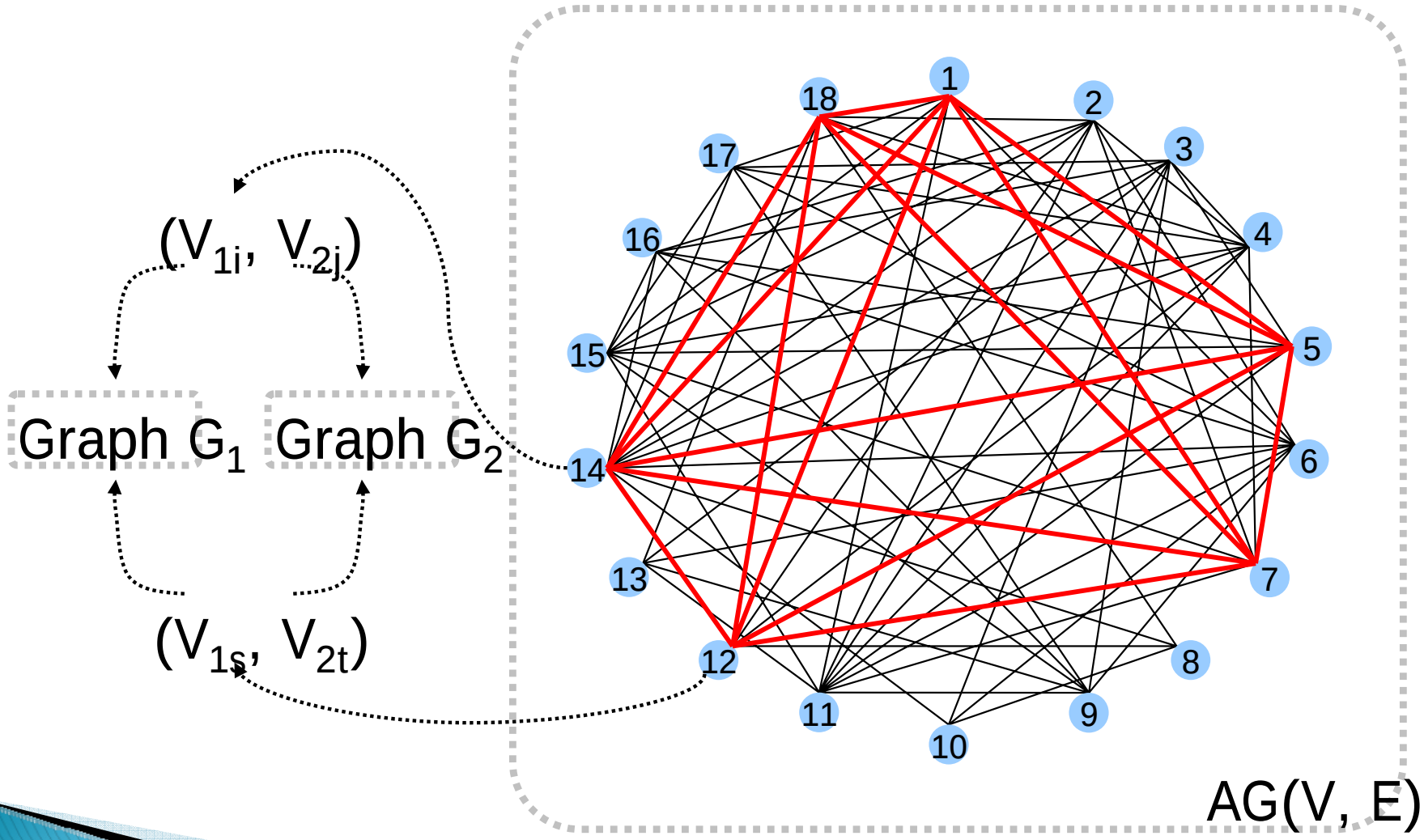


SIMCMOP: Appendix



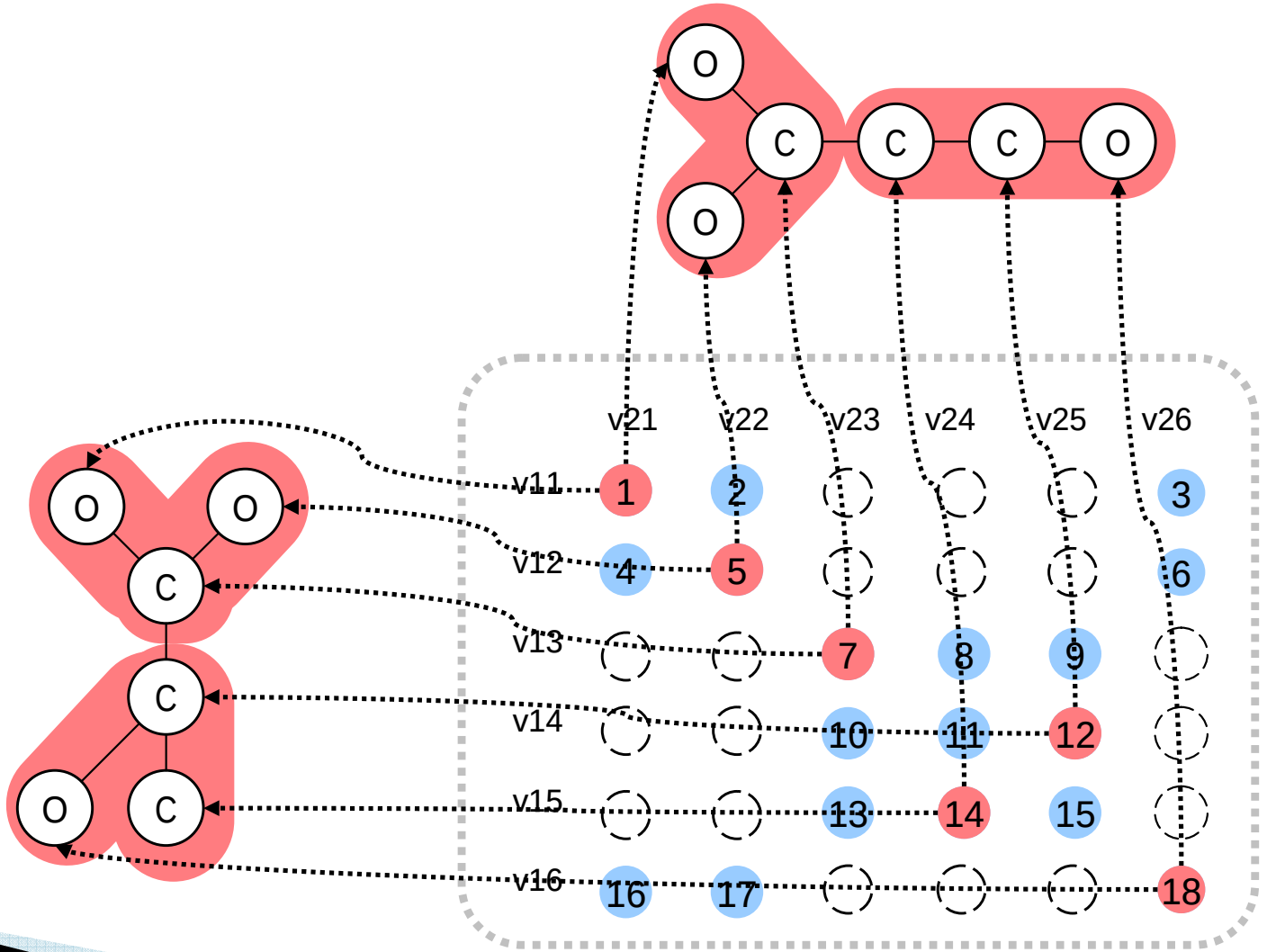
Hattori et al., J Am Chem Soc 125:11853-65 (2003)

SIMCMOP: Appendix



Hattori et al., J Am Chem Soc 125:11853-65 (2003)

SIMCMOP: Appendix



Hattori et al., J Am Chem Soc 125:11853-65 (2003)

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)

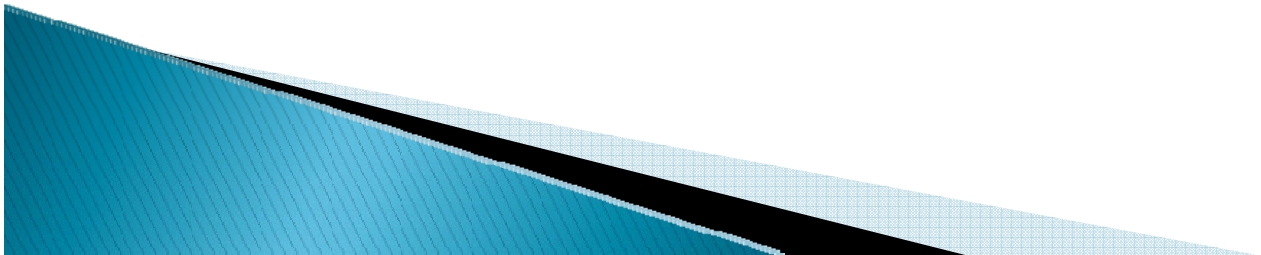
e-zyme: Appendix

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:

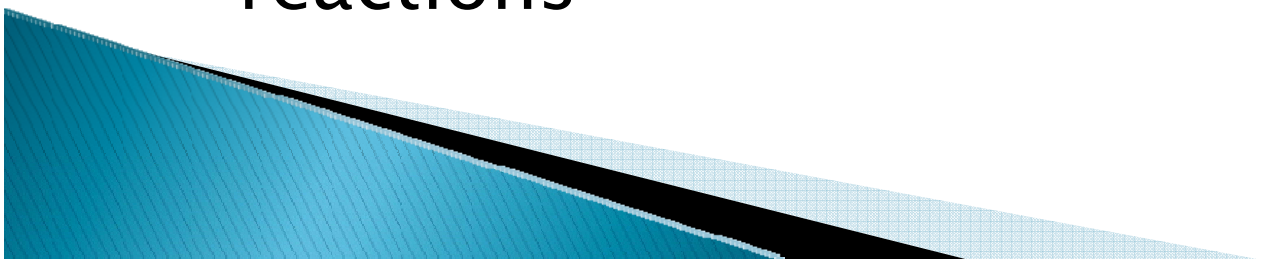
1. Use the full RDM for matched reactions
2. Use RD only when the query RDM is a singleton (no hits in the database)



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