KEGG: Kyoto Encyclopedia of Genes and Genomes

Minoru Kanehisa
Bioinformatics Center, Institute for Chemical Research, Kyoto University
Human Genome Center, Institute of Medical Science, University of Tokyo

June 2011

Background

- High-throughput experiments and large-scale datasets
- Bioinformatics approaches to integration and interpretation

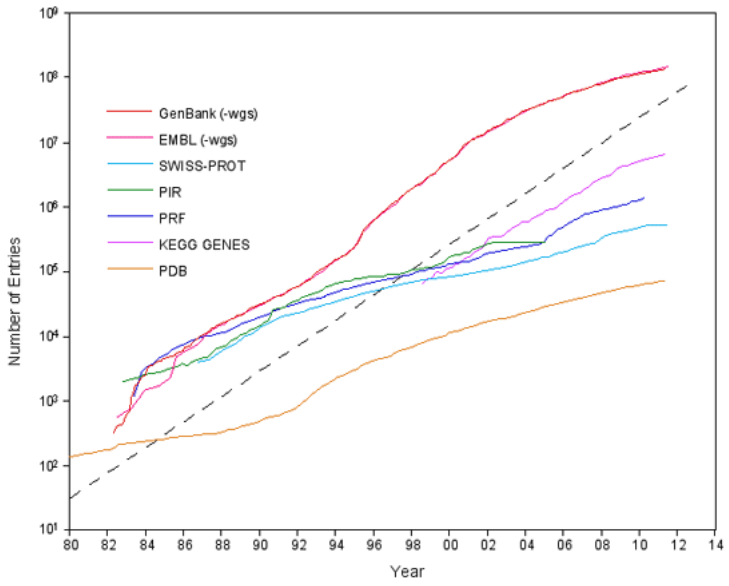
Seeds and Needs: Changing Roles of Bioinformatics

Period	Seeds	Needs
1990s	Informatics technologies	Large-scale sequence data in the Human Genome Project
2000s	Bioinformatics technologies (basic)	Large-scale molecular data in all areas of biological sciences
2010s	Bioinformatics technologies (applied)	Scientific data (personal genome data, etc.) in society

Current Roles of Bioinformatics

1. Meeting different needs
   - Scientific community (scientists, software engineers, etc.)
   - Practitioners (doctors, pharmacists, etc.)
   - Society (patients, consumers, etc.)
2. Developing computational methods
   - Especially for large-scale data processing
3. Developing databases
   - Especially 'knowledge-intensive' databases
4. Maintaining information infrastructure
   - Database and computational services meeting individual needs

Growth of sequence and 3D structure databases

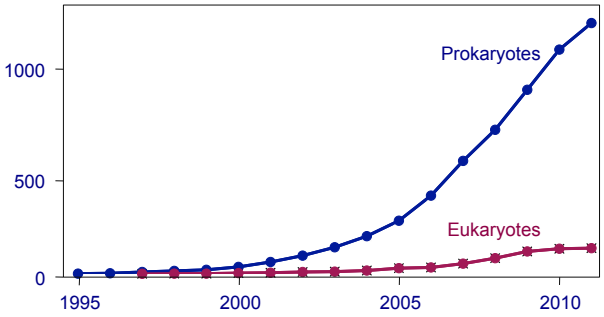


More details at: http://www.genome.jp/en/db_growth.html

Complete Genomes in KEGG

- Eukaryotes (142)
  - Animals (50)
  - Plants (14)
  - Fungi (47)
  - Protists (31)
- Prokaryotes (1380)
  - Bacteria (1275)
  - Archaea (105)

As of June 17, 2011



More details at: http://www.genome.jp/kegg/docs/upd_genome.html

Genetic and Chemical Blueprints of Life

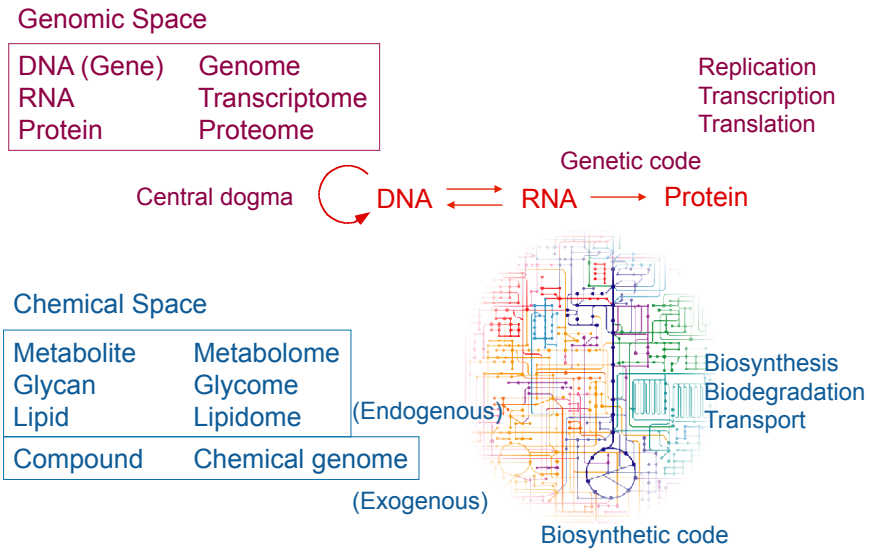
Genetic Blueprint of Life (Genome)

- DNA sequence
- Information about genes, gene regulatory elements, proteins, RNAs, etc.
- Partial information about transcription network

Chemical Blueprint of Life (Chemical Logic)

- Molecular reaction/interaction network
- Information about small molecules, biopolymers, etc. (biosynthetic codes)
- Information about interactions with the environment

Molecular Building Blocks of Life
Genomic and Chemical Spaces

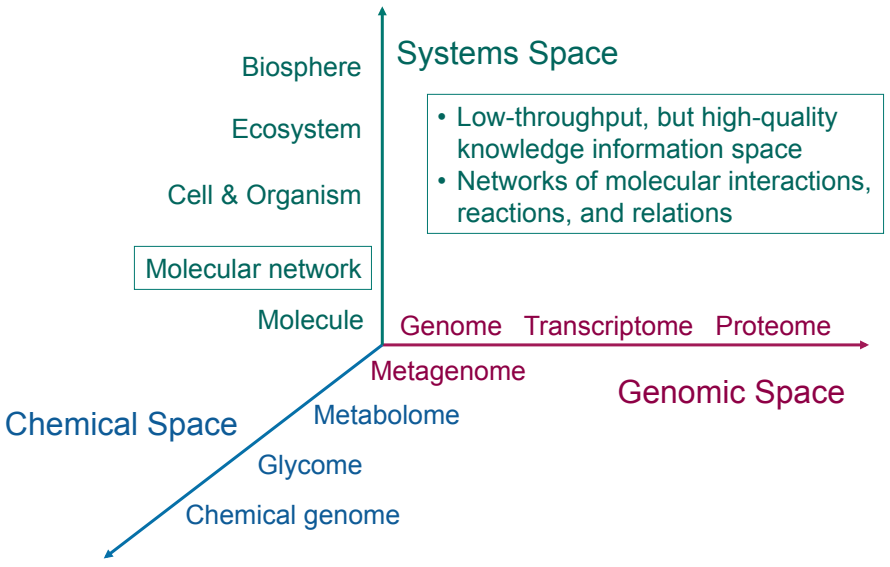


High-throughput Experimental Projects to Uncover Molecular Building Blocks of Life

Genomic space		Chemical space
Genes and proteins in individual organisms (genomics, transcriptomics, proteomics)	⟺	Endogenous molecules (metabolomics, glycomics, lipidomics, etc.)
	⟺	Exogenous molecules (chemical genomics)
Genes and proteins in environmental samples (metagenomics, etc.)	⟺	Metabolites in environmental samples (meta-metabolomics)

- Integration of genomic and chemical spaces
- Interpretation of higher-level systemic functions

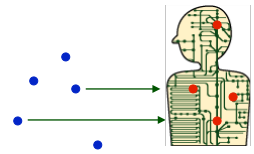
Bioinformatics Approaches to Reconstructing Biological Systems from Building Blocks



Molecular network-based analysis of diseases, drugs, and environmental compounds

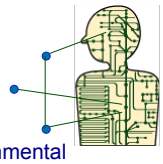
- Diseases viewed as perturbed states of the molecular system
- Drugs and environmental compounds viewed as perturbants to the molecular system

Molecular interaction network



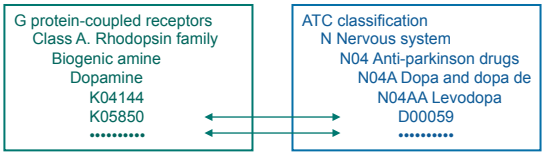
Perturbants Perturbed system

Molecular reaction network



Environmental compounds Biodegradation

Molecular relation network



Network-Disease Associations

Single-gene disease

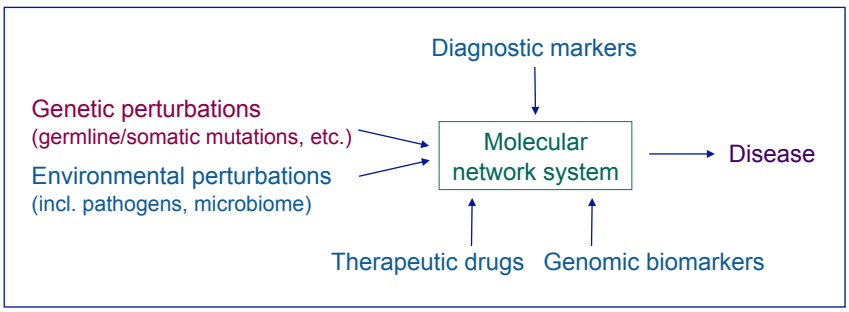
Disease gene → Molecular network → Disease

Multifactorial disease

Genetic factors
Environmental factors
→ Molecular network → Disease

Infectious disease

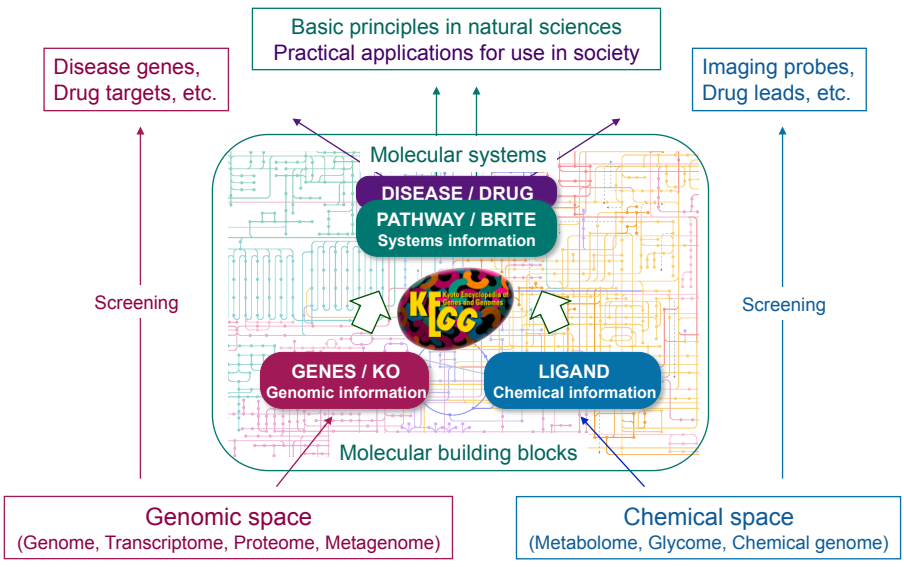
Pathogen → Molecular network → Disease



Overview of KEGG

- From building blocks to biological systems
- Integration of genomics and chemistry

KEGG: Computer Representation of Biological Systems



http://www.kegg.jp/ http://www.genome.jp/kegg/

KEGG Databases

Database	Content	Data size
KEGG PATHWAY	Pathway maps, reference (total)	398 (140,607)
KEGG BRITE	Functional hierarchies, reference (total)	124 (41,364)
KEGG MODULE	KEGG modules, reference (total)	371 (83,523)
KEGG DISEASE	Human diseases	375
KEGG DRUG	Drugs	9,402
KEGG EDRUG	Crude drugs and health-related substances	836
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	14,715
KEGG GENOME	KEGG Organisms, manual/koala + kaas	1,522 + 117
KEGG GENES	Genes in high-quality genomes (142 eukaryotes + 1275 bacteria + 105 archaea)	6,667,326
KEGG SSDB	Best hit relations within GENES Bi-directional best hit relations within GENES	56,310,107,051 1,040,663,735
KEGG DGENES	Genes in draft genomes (18 eukaryotes)	398,519
KEGG EGENES	Genes as EST contigs (99 eukaryotes)	3,792,883
KEGG MGENES	Genes in metagenomes (139 samples)	3,590,397
KEGG COMPOUND	Metabolites and other small molecules	17,641
KEGG GLYCAN	Glycans	10,978
KEGG REACTION	Biochemical reactions	8,494
KEGG RPAIR	Reactant pair chemical transformations	12,652
KEGG RCLASS	Reaction class	2,336
KEGG ENZYME	Enzyme nomenclature	5,419

As of June 17, 2011

KEGG Object Identifiers

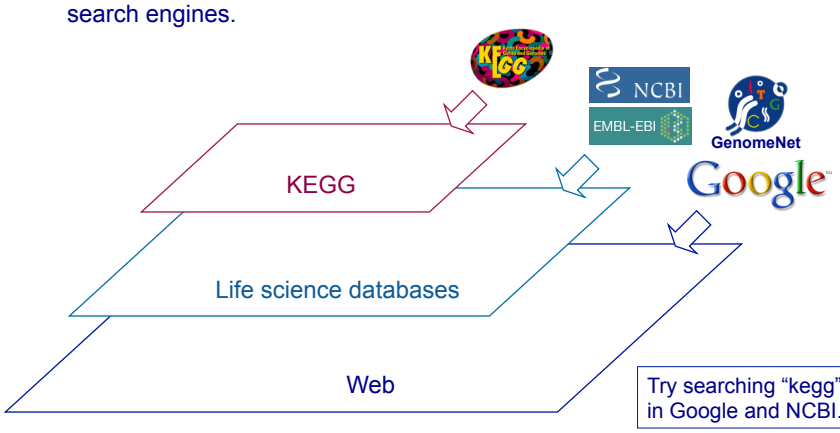
Prefix + 5-digit number

Database	Prefix	Example	Release
KEGG PATHWAY	map/ko/ec/rn/(org)	hsa04930	1995
KEGG BRITE	br/jp/ko/(org)	ko01003	2005
KEGG MODULE	M	M00008	2007
KEGG DISEASE	H	H00004	2008
KEGG DRUG	D	D01441	2005
KEGG ENVIRON	E	E00048	2010
KEGG ORTHOLOGY	K	K04527	2002
KEGG GENOME	T	T01000 (hsa)	2000
KEGG COMPOUND	C	C00031	1995
KEGG GLYCAN	G	G00109	2003
KEGG REACTION	R	R00259	2001
KEGG RPAIR	RP	RP04458	2004
KEGG RCLASS	RC	RC00046	2010
db:name			
KEGG GENES	org:gene	hsa:3643	1995
KEGG ENZYME	ec:number	ec:2.7.10.1	1995
DBGET databases	db:entry	sp:P06213	1994

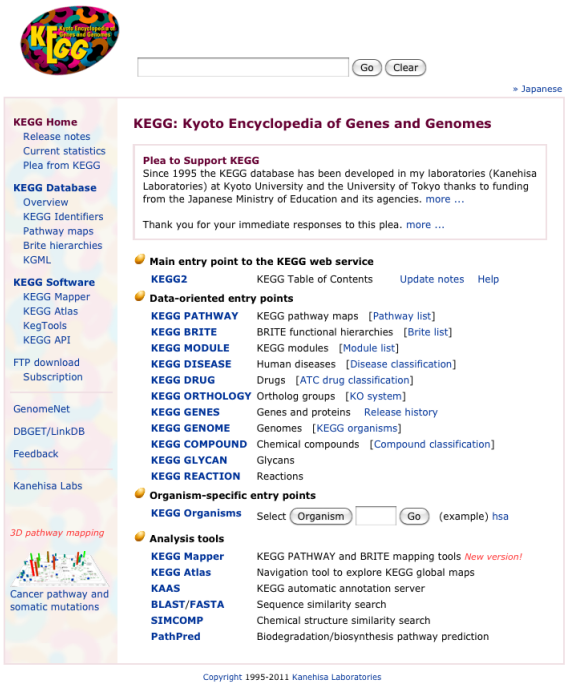
org KEGG Organism code
gene gene identifier (locus_tag, Gene ID, etc.)

KEGG Objects

- KEGG is a computer representation of the biological systems consisting of molecular objects and higher-level objects.
- KEGG objects (database entries) are highly integrated representing inherent biological relationships.
- KEGG objects are linked to/from major life science databases.
- KEGG objects are part of the Web; they can be found by Web search engines.



Try searching "kegg" in Google and NCBI.

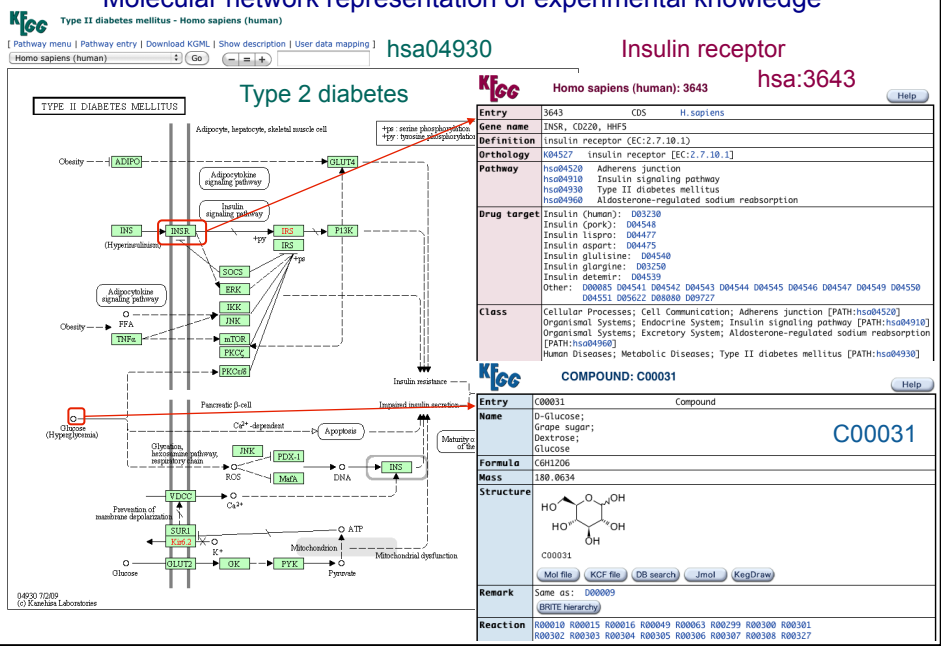


Access the KEGG top page: http://www.genome.jp/kegg/ and enter keywords or any KEGG object identifier in the search box.

Try, for example, hsa04930 to retrieve the KEGG pathway map for type 2 diabetes.

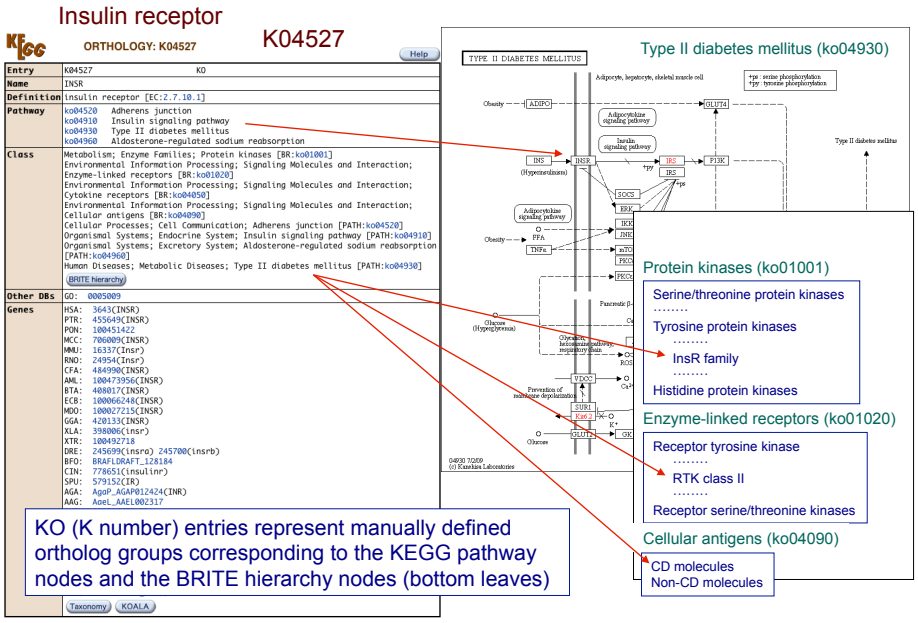
KEGG Pathway Maps

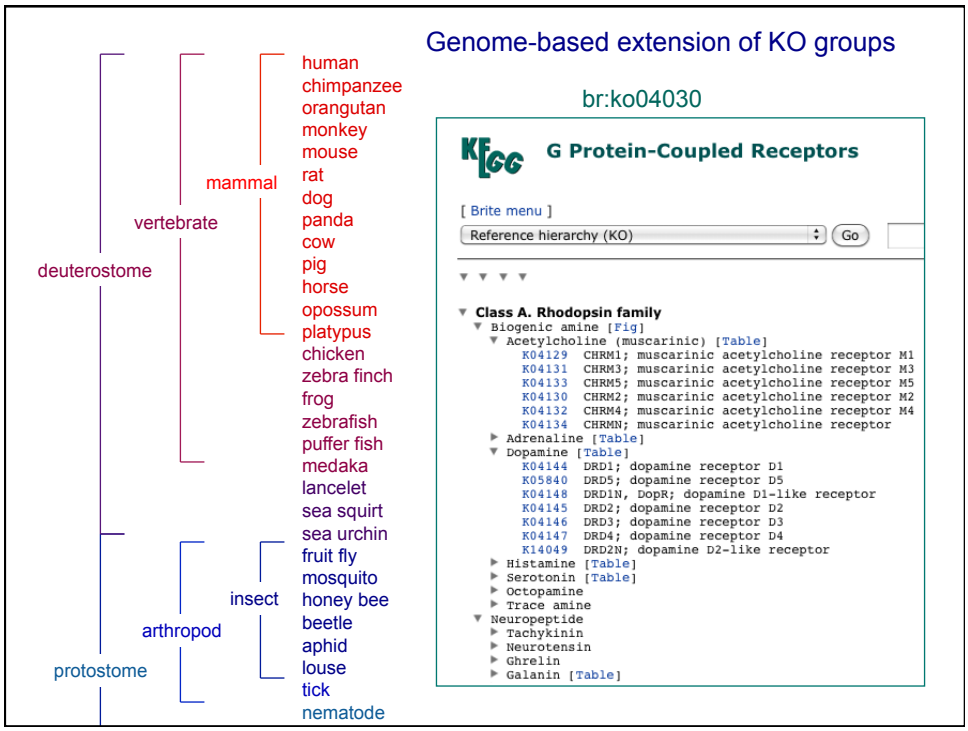
Molecular network representation of experimental knowledge

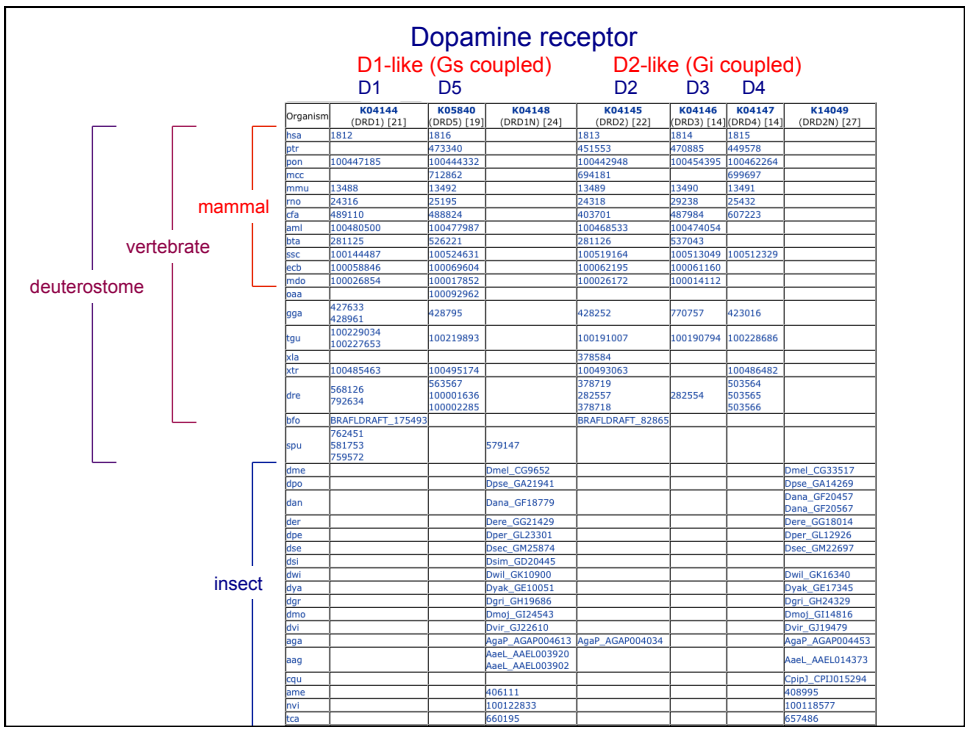


KEGG Orthology (KO)

Genome-based generalization of experimental knowledge



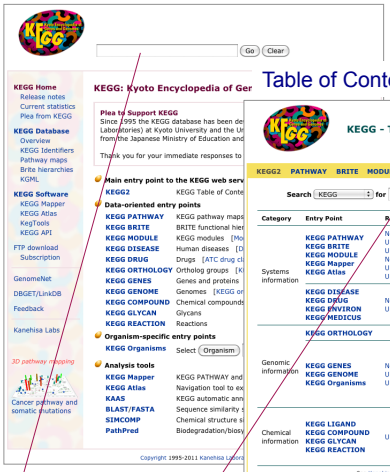




Organism	K04144 (DRD1) [21]	K05840 (DRD5) [19]	K04148 (DRD1N) [24]	K04145 (DRD2) [22]	K04146 (DRD3) [14]	K04147 (DRD4) [14]	K14049 (DRD2N) [27]
hsa	1812	1816		1813	1814	1815	
ptr		473340		451553	470885	449578	
pon	100447185	100444332		100442948	100454395	100462264	
mcc		712862		694181		699697	
mmu	13488	13492		13489	13490	13491	
rno	24316	25195		24318	29238	25432	
cfa	489110	488824		403701	487984	607223	
aml	100480500	100477987		100468533	100474054		
bta	281125	526221		281126	537043		
ssc	100144487	100524631		100519164	100513049	100512329	
ecb	100058846	100069604		100062195	100061160		
mdo	100026854	100017852		100026172	100014112		
oaa		100092962					
gga	427633 428961	428795		428252	770757	423016	
tgu	100229034 100227653	100219893		100191007	100190794	100228686	
xla				378584			
xtr	100485463	100495174		100493063		100486482	
dre	568126 792634	563567 100001636 100002285		378719 282557 378718	282554	503564 503565 503566	
bfo	BRAFLDRAFT_175493			BRAFLDRAFT_82865			
spu	762451 581753 759572		579147				
dme			Dmel_CG9652				Dmel_CG33517
dpo			Dpse_GA21941				Dpse_GA14269
dan			Dana_GF18779				Dana_GF20457 Dana_GF20567
der			Dere_GG21429				Dere_GG18014
dpe			Dper_GL23301				Dper_GL12926
dse			Dsec_GM25874				Dsec_GM22697
dsi			Dsim_GD20445				
dwi			Dwil_GK10900				Dwil_GK16340
dya			Dyak_GE10051				Dyak_GE17345
dgr			Dgri_GH19686				Dgri_GH24329
dmo			Dmoj_GI24543				Dmoj_GI14816
dvi			Dvir_GJ22610				Dvir_GJ19479
aga			AgaP_AGAP004613	AgaP_AGAP004034			AgaP_AGAP004453
aag			AaeL_AAEL003920 AaeL_AAEL003902				AaeL_AAEL014373
cqu							CpipJ_CPIJ015294
ame			406111				408995
nvi			100122833				100118577
tca			660195				657486

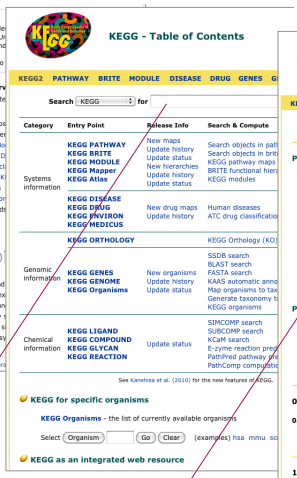
Web Page Organization

Top page (KEGG)

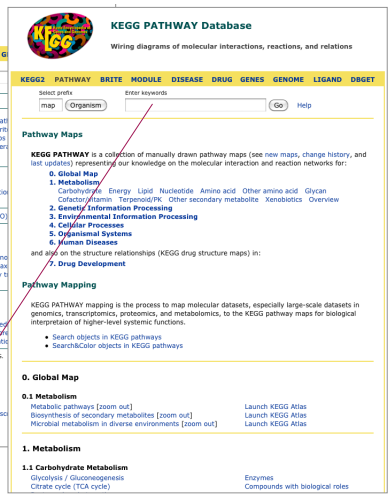


PATHWAY BRITE | DISEASE DRUG | KO | GENES GENOME | LIGAND

Table of Contents (KEGG2)



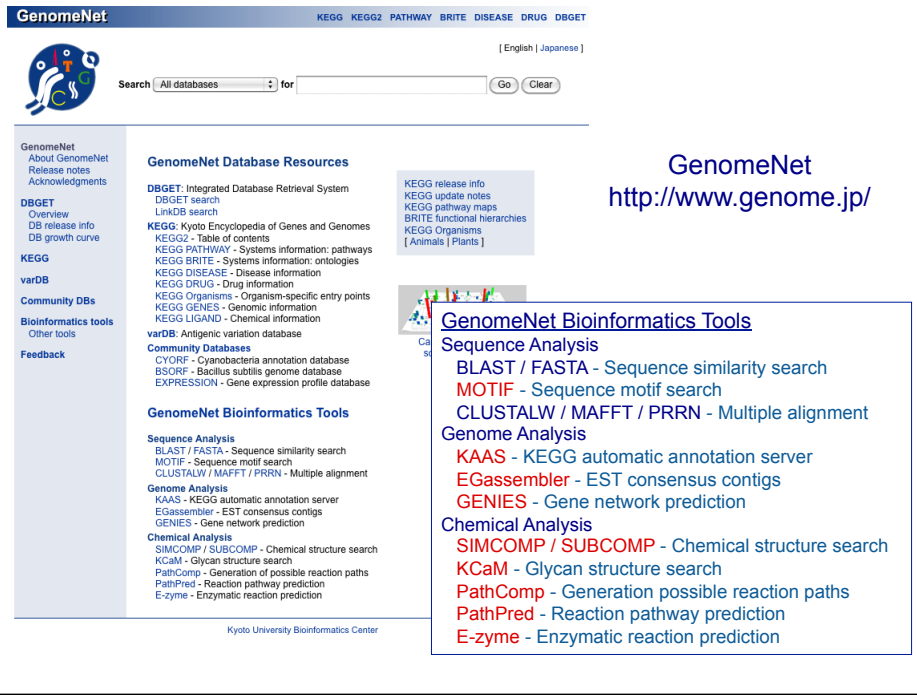
Database top page (eg, PATHWAY)



KEGG identifiers are specially treated

Normal keyword search

Keyword search against KEGG pathway maps as well



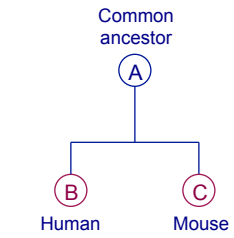
GenomeNet
http://www.genome.jp/

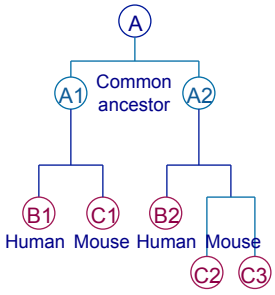
GenomeNet Bioinformatics Tools
Sequence Analysis
BLAST / FASTA - Sequence similarity search
MOTIF - Sequence motif search
CLUSTALW / MAFFT / PRRN - Multiple alignment
Genome Analysis
KAAS - KEGG automatic annotation server
EGassembler - EST consensus contigs
GENIES - Gene network prediction
Chemical Analysis
SIMCOMP / SUBCOMP - Chemical structure search
KCaM - Glycan structure search
PathComp - Generation possible reaction paths
PathPred - Reaction pathway prediction
E-zyme - Enzymatic reaction prediction

KEGG GENES
and ortholog annotation

Orthologs and Paralogs

- Sequence similarity between two genes (or proteins) may imply ortholog or paralog relationship.
- Orthologs are genes in different species evolved from a common ancestral gene by speciation and tend to have the same function.
- Paralogs are generated by gene duplication within a species and often represent diversified functions in a broader functional category.
- Identification of ortholog relationships is the basis for genome annotation (assigning gene functions), and it requires distinction from paralog relationships.



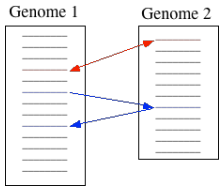


Orthologs: B–C, B1–C1
Co-orthologs: B2–(C2,C3)
Inparalogs: C2–C3
Outparalogs: B1–B2, B1–(C2,C3), B2–C1

Computational Identification of Orthologs

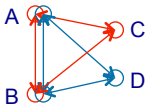
Between two organisms

Bi-directional best hit (BBH)
(Reciprocal best hit)



Among multiple organisms

1. COG
   Triangle of BBH relationships among three organisms



Superposition of
ABC and ABD

2. KEGG OC
   p-Quasi clique among multiple organisms

Clique
(completely connected subgraph)

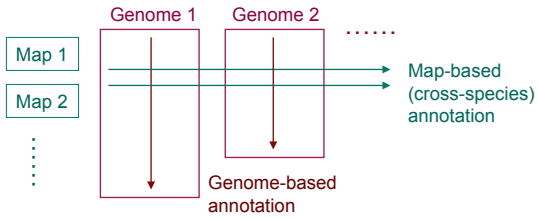
p-Quasi clique is an almost complete subgraph, where the degree of completeness is represented by p.

Genome annotation in KEGG

Separate function database
- Experimental evidence is stored in the KEGG ORTHOLOGY (KO) database
- KO entries identified by K numbers are manually defined in the context of molecular networks; i.e., as KEGG pathway nodes and BRITE hierarchy nodes

Ortholog annotation
- KO (K number) assignment; i.e., it establishes links from the KO database
- Gene definitions in the original database (mostly RefSeq) are not rewritten
- But the definitions of KO entries are frequently updated to follow guidelines

Cross-species annotation
- Molecular network-based annotation; i.e., starting from a pathway map, etc. orthologs are searched in all available genomes
- Genome-based annotation is also done



Genomes	1,522
Genes	6,667,326
Genes with KO	2,665,692
KO assignment	40%

KO	14,715
Safe KO	8,051
KOALA automation	55%

As of June 17, 2011

KOALA: KEGG Orthology And Links Annotation

Integrating computationally generated SSDB graph
and manually defined KO groups

RefSeq, etc.

Literature, etc.

KEGG GENES

KEGG PATHWAY
KEGG BRITE

SSEARCH

SSDB graph

Manual grouping
of KO entries

KEGG Oracle
Genes | KO

GFIT tables

KOALA program

KOALA tool
GFIT tool

SSDB is a huge graph weighted by similarity scores and directed by best-hit relations

Original membership of each KO group (subgraph of SSDB) is manually defined

Computational annotation

Manual annotation

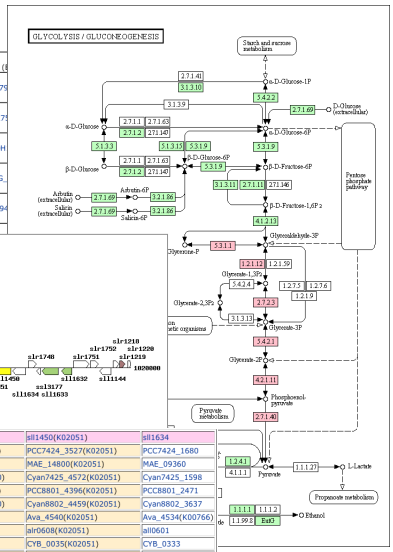
Manual re-grouping of KO entries

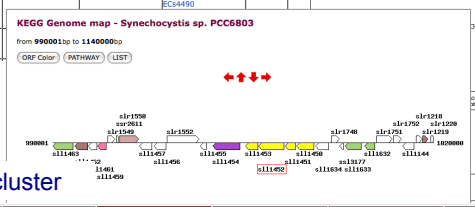
Tools to check annotation quality

Based on reconstruction of functional units

Ortholog table

Organism	K01803 (TPI) [1170]	K00134 (GAPDH) [1469]	K00150 (gap2) [127]	K00927 (PGK) [1108]	K01834 (PGAM) [2711]	K01837 (BPGM) [1]
eco	b3919	b1779		b2926	b0755 b3612 b4395	
ecj	JW3890	JW1413 JW1768		JW2893	JW0738 JW3587 JW4358	
ecd	ECDH10B_4108	ECDH10B_1917		ECDH10B_3101	ECDH10B_0822 ECDH10B_3794 ECDH10B_4553	
ebw	BWG_3588	BWG_1592		BWG_2649	BWG_3303 BWG_0607 BWG_4087	
ece	Z5464	Z2304 Z2818		Z4265	Z0925 Z5039 Z5997	
ecs	ECs4844	ECs2022 ECs2488			ECs4490	
ecf	ECH74115_5374	ECH74115_2503 ECH74115_2022				

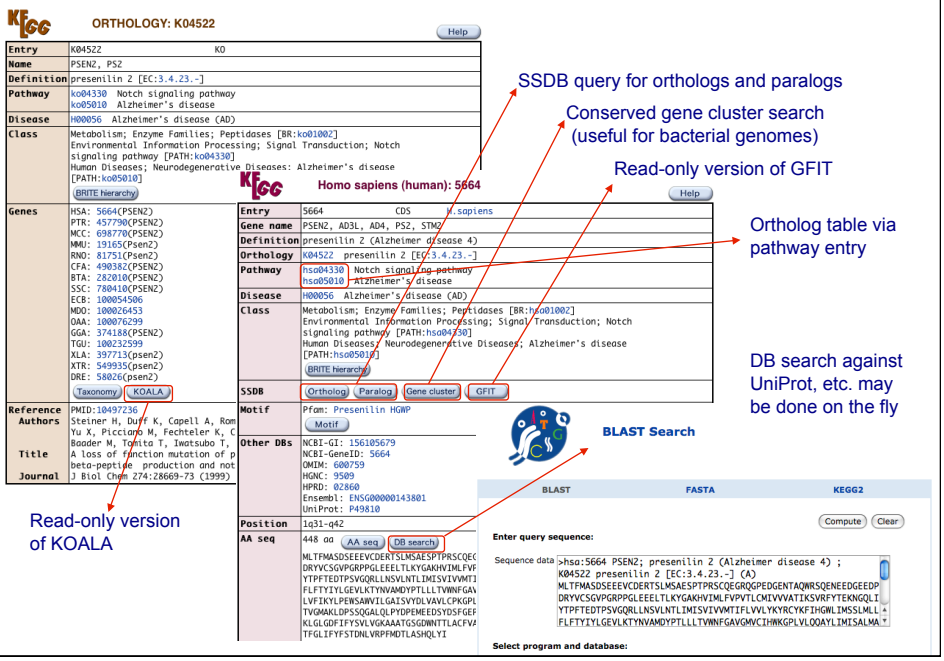




Conserved gene cluster

syn	sll1455	sll1454(K00367)	sll1453(K02049)	sll1452(K02050)	sll1451(K02050)	sll1450(K02051)	sll1634
cyc	PCC7424_3464	PCC7424_3462	PCC7424_3524(K02049)	PCC7424_3525(K02049)	PCC7424_3526(K02050)	PCC7424_3527(K02051)	PCC7424_1680
mar	MAE_55270	MAE_53960	MAE_14770(K02049)	MAE_14780(K02049)	MAE_14790(K02050)	MAE_14800(K02051)	MAE_09360
cyn	Cyan7425_3581	Cyan7425_4566		Cyan7425_4570(K02049)	Cyan7425_4571(K02050)	Cyan7425_4572(K02051)	Cyan7425_1598
cyp	PCC8801_4057	PCC8801_2463		PCC8801_4394(K02049)	PCC8801_4395(K02050)	PCC8801_4396(K02051)	PCC8801_2471
cyh	Cyan8802_4096	Cyan8802_3646		Cyan8802_4457(K02049)	Cyan8802_4458(K02050)	Cyan8802_4459(K02051)	Cyan8802_3637
ava	Ava_4546	Ava_4544(K00367)		Ava_4542(K02049)	Ava_4541(K02050)	Ava_4540(K02051)	Ava_4534(K00766)
ana	alr0614	alr0612(K00367)	alr0611(K02049)	alr0610(K02049)	alr0609(K02050)	alr0608(K02051)	all0601
cyb	CYB_0042	CYB_0040(K00367)		CYB_0037(K02049)	CYB_0036(K02050)	CYB_0035(K02051)	CYB_0333
cya	CYA_0174	CYA_0619(K00367)		CYA_0615(K02049)	CYA_0614(K02050)	CYA_0613(K02051)	CYA_0985
tel	tlr1356	tlr1355(K00367)		tlr1352(K02049)	tlr1351(K02050)	tlr1350(K02051)	tll1358
aeh		Mlg_1702(K00360)	Mlg_2587(K02049)	Mlg_1703(K02049)	Mlg_1704(K02050)		Mlg_1665
cps		CPS_4945(K00372)		CPS_3314(K02049)	CPS_3315(K02050)		
tcx		Tcr_1159(K00372)		Tcr_1155(K02049)	Tcr_1154(K02050)		

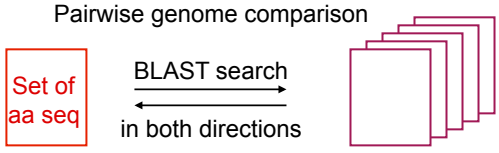
Public version of KEGG annotation tools



KAAS: KEGG Automatic Annotation Server

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

Pairwise genome comparison



Query genome

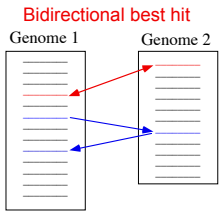
Annotated genomes in KEGG

BLAST result screening by bi-directional best hit rate (*BHR*)

$BHR = R_f \times R_r \quad > 0.95$

K number assignment by a heuristic scoring

$$S_{KO} = S_h - \log_2(mn) - \log_2\left(\sum_{k=N}^{x} {}_xC_k p^k (1-p)^{x-k}\right)$$



Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M.; KAAS: an automatic genome annotation and pathway reconstruction server. *Nucl. Acids Res.* 35, W182-W185 (2007).

KEGG PATHWAY and BRITE: Reference knowledge base

Data objects for computer representation of molecular systems

Data object	Database
Element gene, protein, small molecule, etc.	KEGG GENES KEGG COMPOUND KEGG GLYCAN KEGG DRUG
Pair (binary relation) reactant pair, drug-target relationship, etc.	KEGG RPAIR KEGG BRITE
Graph (wiring diagram) pathway, complex, etc.	KEGG PATHWAY KEGG SSDB
Hierarchical list (tree) hierarchical classification, ontology, etc.	KEGG BRITE
Simple list (membership) disease gene list, etc.	KEGG MODULE KEGG DISEASE KEGG ORTHOLOGY

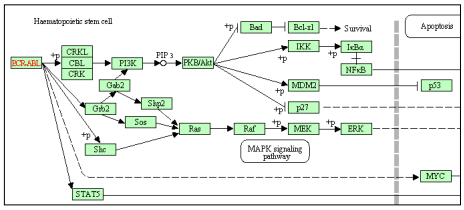
Set operations
for biological interpretation of large-scale datasets

Large scale data + Reference knowledge	Tool
KEGG pathway mapping (pathway enrichment) elements + KEGG pathway map	KEGG Mapper
KEGG pathway mapping (network extension) binary relations + KEGG pathway map	——
BRITE mapping (ontology enrichment) elements + BRITE functional hierarchy	KEGG Mapper
BRITE mapping (tree manipulation) binary relations + BRITE functional hierarchy	KEGG Mapper
KEGG annotation SSDB graph + KO group	KOALA

Knowledge Representation of Systemic Functions

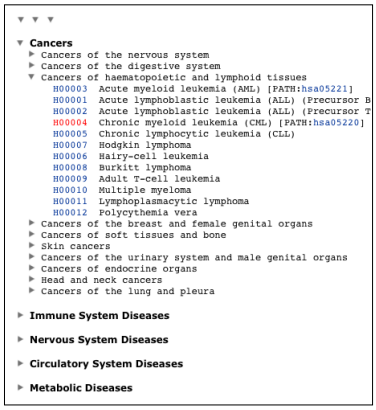
Pathway map (interaction/reaction network)

KEGG PATHWAY



Hierarchical list (ontology)

KEGG BRITE



Simple list (membership)

KEGG DISEASE

Gene	BCR-ABL (translocation) [HSA:613 25] EVI1 (overexpression) [HSA:2122] AML1 (translocation) [HSA:861] p16/INK4A (mutation) [HSA:1029] p53 (mutation) [HSA:7157] RB1 (mutation) [HSA:5925]
Carcinogen	1,3-Butadiene [CPD:C16450] Rubber industry
Marker	BCR-ABL (translocation) [HSA:613 25] WT1 [HSA:7490]
Drug	Imatinib mesylate (Gleevec) [DR:D01441] Hydroxyurea [DR:D00341] Interferon-alpha [DR:D00745 D02745 D03305 D04552 D04553]

Data source: review articles, other publications, specialists' websites, etc.

KEGG Pathway Maps

- **Global Map**
- **Metabolism**
- **Genetic Information Processing**
  - Transcription
  - Translation
  - Folding, Sorting and Degradation
  - Replication and Repair
    - 03030 DNA replication
    - 03410 Base excision repair
    - 03420 Nucleotide excision repair
    - 03430 Mismatch repair
    - 03440 Homologous recombination
    - 03450 Non-homologous end-joining
- **Environmental Information Processing**
- **Cellular Processes**
  - Transport and Catabolism
  - Cell Motility
  - Cell Growth and Death
    - 04110 Cell cycle
    - 04111 Cell cycle - yeast
    - 04112 Cell cycle - Caulobacter
    - 04113 Meiosis - yeast
    - 04114 Oocyte meiosis
    - 04210 Apoptosis
    - 04115 p53 signaling pathway
  - Cell Communication
- **Organismal Systems**
- **Human Diseases**

BRITE Functional Hierarchies

- **Pathways and Ontologies**
- **Genes and Proteins**
  - Network hierarchy
  - Protein families: metabolism
  - Protein families: genetic information processing
    - ko03000 Transcription factors
    - ko03041 Spliceosome
    - ko03011 Ribosome
    - ko03009 Ribosome biogenesis
    - ko03012 Translation factors
    - ko03110 Chaperones and folding catalysts
    - ko04131 SNAREs
    - ko04121 Ubiquitin system
    - ko03051 Proteasome
    - ko03032 DNA replication proteins
    - ko03036 Chromosome
    - ko03400 DNA repair and recombination proteins
  - Protein families: environmental information proce
  - Protein families: cellular processes
- **Diseases and Drugs**
- **Compounds and Reactions**
- **Organisms and Cells**

124 brite hierarchies

398 pathway maps

As of June 17, 2011

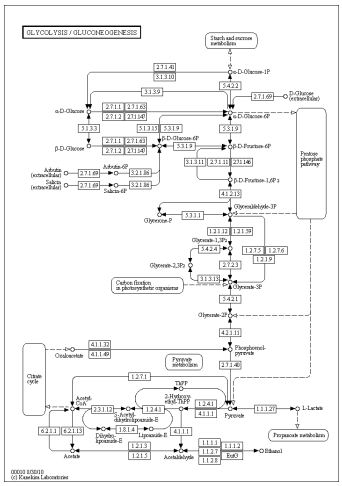
KEGG PATHWAY mapping

Human gene list with KO assignment

Gene	KO
3098	K00844
2645	K12407
2821	K01810
5211	K00850
2203	K03841
226	K01623
7167	K01803
2597	K00134
26330	K10705
5230	K00927
2023	K01689
5315	K00873
5313	K12406
5160	K00161
5162	K00162
1737	K00627
1738	K00382
3939	K00016
124	K13951
130	K13952
10327	K00002
217	K00128
218	K00129
55902	K01895
5236	K01835
2538	K01084
83440	K08074
.....	

+

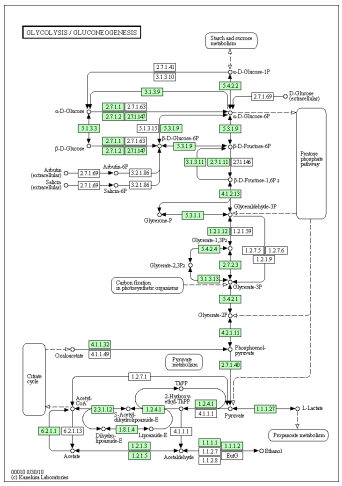
Reference pathway map



Glycolysis / Gluconeogenesis
map00010

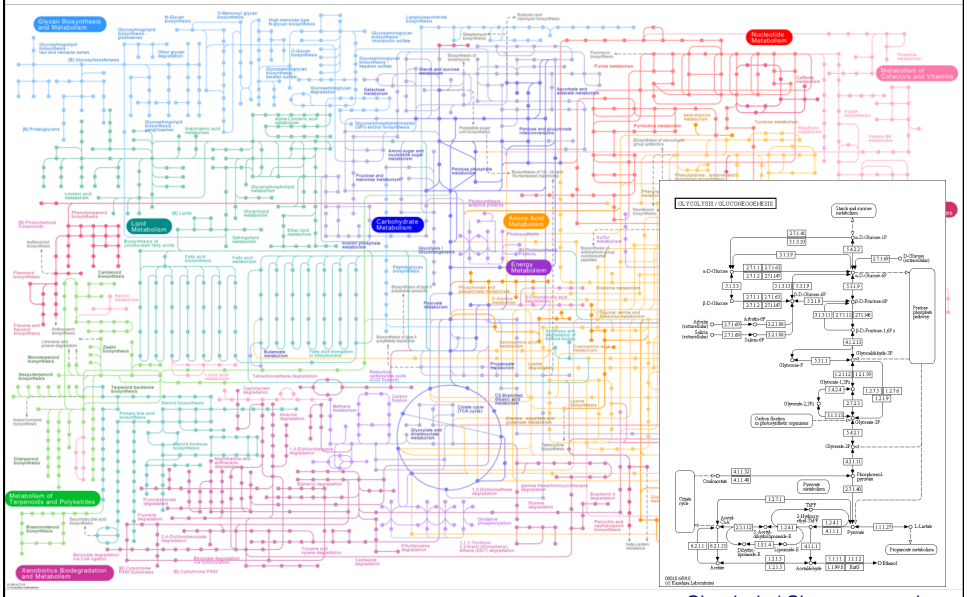
=

Human pathway map



Glycolysis / Gluconeogenesis
hsa00010

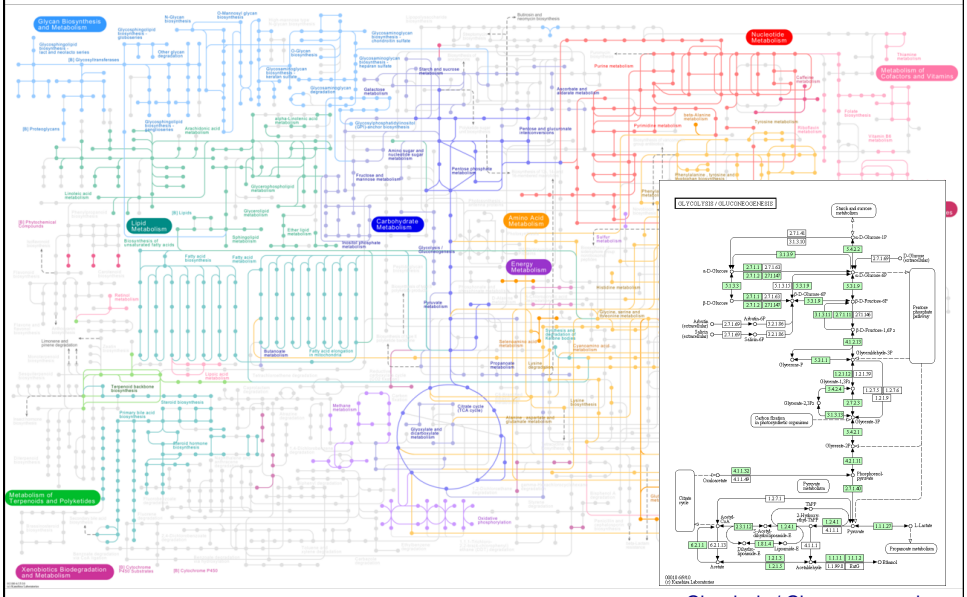
Reference pathway



Metabolic pathways (global map)
map01100

Glycolysis / Gluconeogenesis
map00010

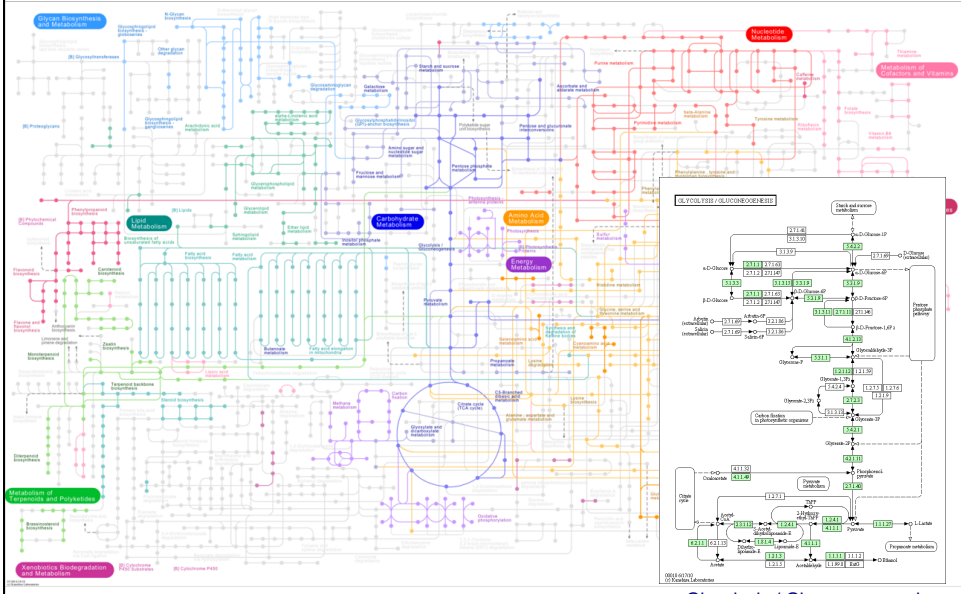
Organism-specific pathway – *Homo sapiens* (human)



Metabolic pathways (global map)
hsa01100

Glycolysis / Gluconeogenesis
hsa00010

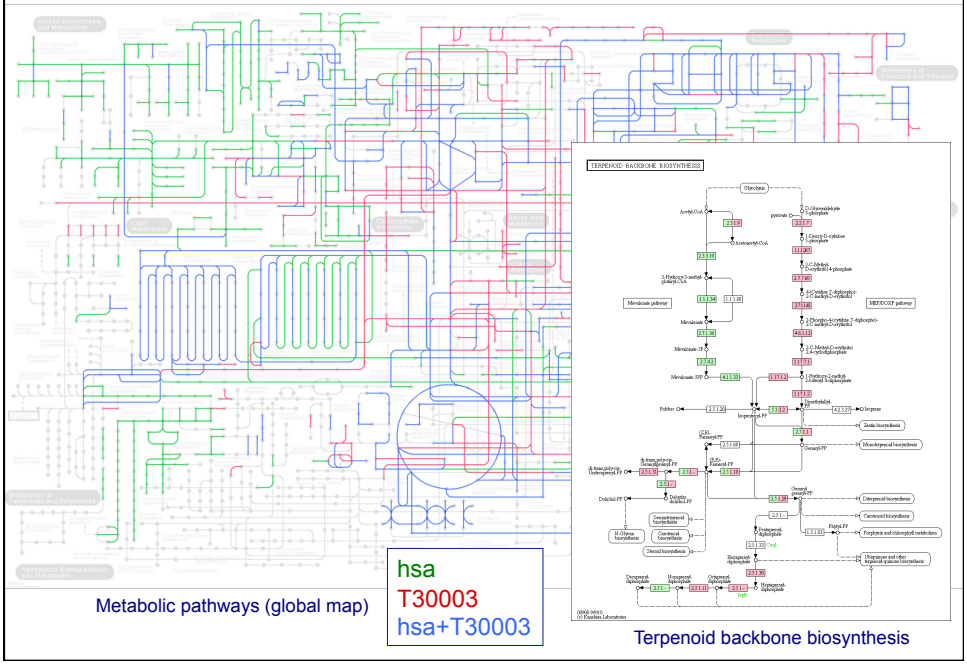
Organism-specific pathway - *Arabidopsis thaliana*



Metabolic pathways (global map)
ath01100

Glycolysis / Gluconeogenesis
ath00010

Metagenome analysis: Combined pathway of human and gut microbiome



Metabolic pathways (global map)

Terpenoid backbone biosynthesis

KEGG Module

M00095: Mevalonate pathway

C5 isoprenoid biosynthesis, mevalonate pathway

[Module menu | Module entry | Hide definition | Ortholog table]

Entry hsa_M00095
Name C5 isoprenoid biosynthesis, mevalonate pathway
Definition K00626 K01641 K00021 K00869 (K00938,K13273) K01597 K01823
Type Pathway module
Organism Homo sapiens (human)
Pathway hsa00900

Homo sapiens (human) Go

K00626
K01641
K00021
K00869
K00938 K13273
K01597
K01823

M00096: Non-mevalonate pathway

C5 isoprenoid biosynthesis, non-mevalonate pathway

[Module menu | Module entry | Hide definition | Ortholog table]

Entry eco_M00096
Name C5 isoprenoid biosynthesis, non-mevalonate pathway
Definition K01662 K00099 (K00991,K12506) K00919 (K01770,K12506) K03526 K03527 K01823
Type Pathway module
Organism Escherichia coli K-12 MG1655
Pathway eco00900

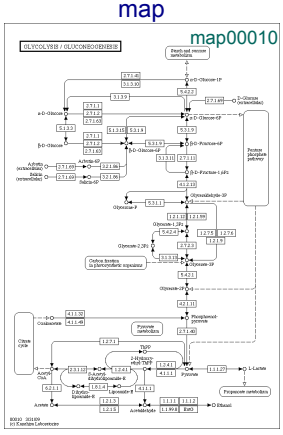
Escherichia coli K-12 MG1655 Go

K01662
K00099
K00991 K12506
K00919
K01770 K12506
K03526
K03527
K01823

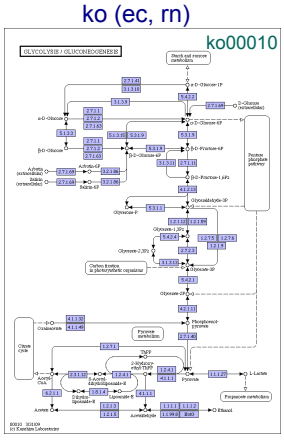
KEGG MODULE
- Pathway module
- Structural complex
- Functional set
- Genotypic signature

Convention of map number prefix

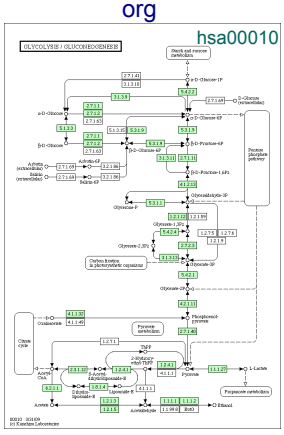
Reference pathway		Organism-specific pathway
map	ko (ec, rn)	org
map00010	ko00010	hsa00010



Each box is manually associated with KO identifier (K number), EC number, and reaction identifier (R number)



Blue boxes represent selection of K numbers (prefix ko), as well as EC numbers (prefix ec) and R numbers (prefix rn)



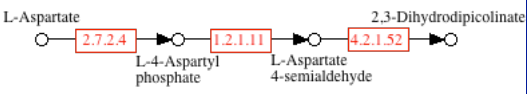
Green boxes correspond to gene identifiers in an organism that are computationally converted from K numbers

KEGG LIGAND
and chemical bioinformatics

Dual Aspect of Metabolic Network

1. Genomic information network

Network of enzyme genes (or enzymes)

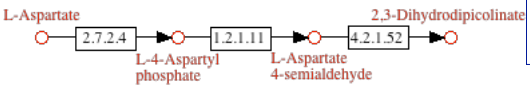


K00928 K00133 K01714

- Genes and proteins are identified by K numbers
- EC numbers are displayed as node names, but they are not used as identifiers in KEGG

2. Chemical information network

Network of small molecules
Chemical structure transformation network

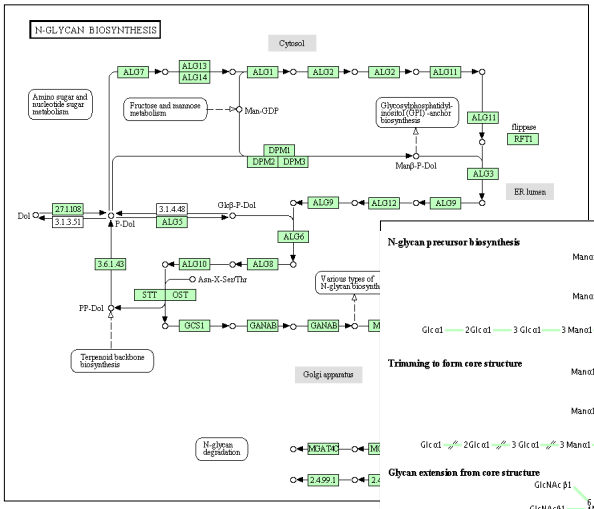


C00049 C03082 C00441 C03340

R00480 R02291 R02292

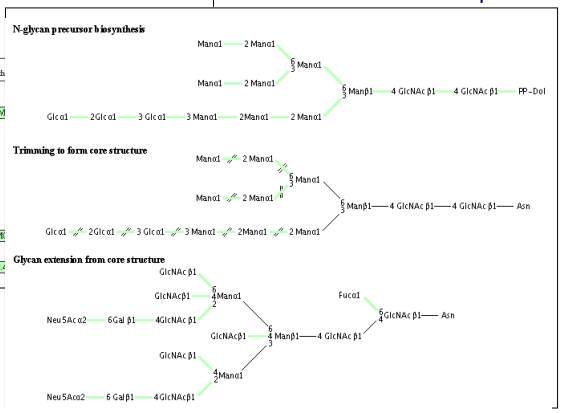
- Compounds are identified by C numbers
- Reactions are identified by R numbers

N-Glycan biosynthesis - Homo sapiens (human)



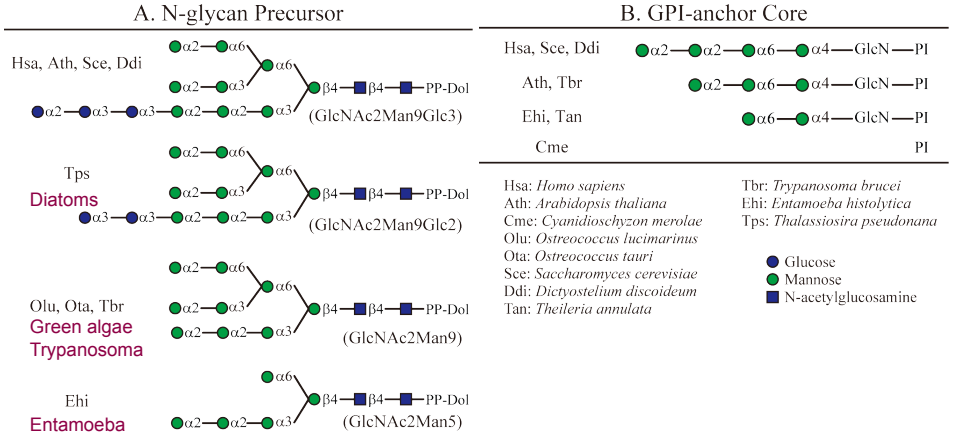
Pathway map

Chemical structure map



hsa00510

Truncated glycan structures in parasitic protists and algae predicted from genomic information



Hashimoto, K., Tokimatsu, T., Kawano, S., Yoshizawa, A.C., Okuda, S., Goto, S., and Kanehisa, M.; Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans. *Carbohydrate Res.* 344, 881-887 (2009).

Predictive annotation of paralogous genes using chemical logic

KEGG Orthology (KO)
Sequence similarity groups that can be mapped to pathway and brite nodes

Reaction Class (RC)
Chemical structure transformation patterns extracted from known reactions

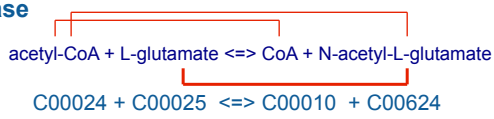
GENES database		REACTION database
⇩ Known orthologs		⇩ Known reactions
KO system	⟺	RC system
⇩ Unknown paralogs		⇩ Unknown reactions
Repertoire of enzyme genes		Repertoire of biochemical reactions

Genome ⇨ Set of K numbers ⇨ Set of RC numbers ⇨ Biosynthetic potential Biodegradation potential

Reaction Class defined by RDM patterns

KEGG REACTION Database

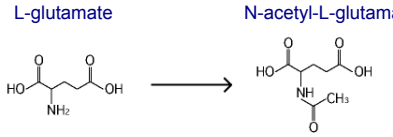
Reaction



R00259

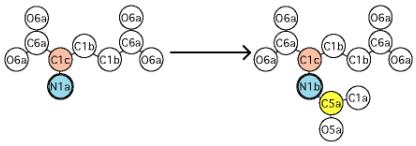
KEGG RPAIR Database

Reactant pair



RP04458

Reactant pair in the KEGG atom type representation



KEGG RCLASS Database

KEGG atom changes for R, D, M atoms (RDM patterns)

R	Reaction center	N1a -- N1b
D	Difference atom	(H) -- C5a
M	Matched atom	C1c -- C1c

RC00064

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

KEGG atom types

Carbon 23 types

Alkane	C1a	R-CH3
	C1b	R-CH2-R
	C1c	R-CH(-R)-R
	C1d	R-C(-R)2-R
Cyclic alkane	C1x	ring-CH2-ring
	C1y	ring-CH(-R)-ring
	C1z	ring-CH(-R)2-ring
Alkene	C2a	R=CH2
	C2b	R=CH-R
	C2c	R=C(-R)2
Cyclic alkene	C2x	ring-CH=ring
	C2y	ring-C(-R)=ring
		ring-C(=R)-ring
Alkyne	C3a	R≡CH
	C3b	R≡C-R
Aldehyde	C4a	R-CH=O
Ketone	C5a	R-C(=O)-R
Cyclic ketone	C5x	ring-C(=O)-ring
Carboxylic acid	C6a	R-C(=O)-OH
Carboxylic ester	C7a	R-C(=O)-O-R
	C7x	ring-C(=O)-O-ring
Aromatic ring	C8x	ring-CH=ring
	C8y	ring-C(-R)=ring
Undefined C	C0	

Nitrogen 16 types

Amine	N1a	R-NH2
	N1b	R-NH-R
	N1c	R-N(-R)2
	N1d	R-N(-R)3+
Cyclic amine	N1x	ring-NH-ring
	N1y	ring-N(-R)-ring
Imine	N2a	R=N-H
	N2b	R=N-R
Cyclic imine	N2x	ring-N=ring
	N2y	ring-N(-R)+=ring
Cyan	N3a	R≡N
Aromatic ring	N4x	ring-NH-ring
	N4y	ring-N(-R)-ring
	N5x	ring-N=ring
	N5y	ring-N(-R)+=ring
Undefined N	N0	

Sulfur 7 types

Thiol	S1a	R-SH
Thioether	S2a	R-S-R
	S2x	ring-S-ring
Disulfide	S3a	R-S-S-R
	S3x	ring-S-S-ring
Sulfate	S4a	R-SO3
Undefined S	S0	

Oxygen 18 types

Hydroxy	O1a	R-OH
	O1b	N-OH
	O1c	P-OH
	O1d	S-OH
Ether	O2a	R-O-R
	O2b	P-O-R
	O2c	P-O-P
	O2x	ring-O-ring
Oxo	O3a	N=O
	O3b	P=O
	O3c	S=O
Aldehyde	O4a	R-CH=O
Ketone	O5a	R-C(=O)-R
	O5x	ring-C(=O)-ring
Carboxylic acid	O6a	R-C(=0)-OH
Ester	O7a	R-C(=O)-O-R
	O7x	ring-C(=O)-O-ring
Undefined O	O0	

Phosphorus 2 types

Attatched to other elements	P1a	P-R
Attatched to oxygen	P1b	P-O

Other elements 2 types

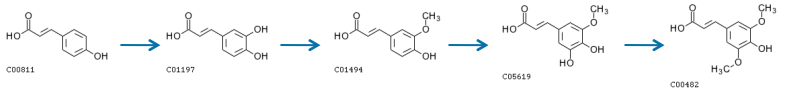
Halogens	X	F, Cl, Br, I
Others	Z	

Chemical structure transformation network

Biosynthetic reaction patterns represented by RC numbers

Phenylpropanoid biosynthesis

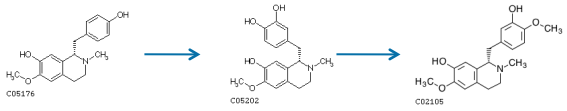
monooxygenase — RC00046 → methyltransferase — RC00392 → monooxygenase — RC00046 → methyltransferase — RC00392



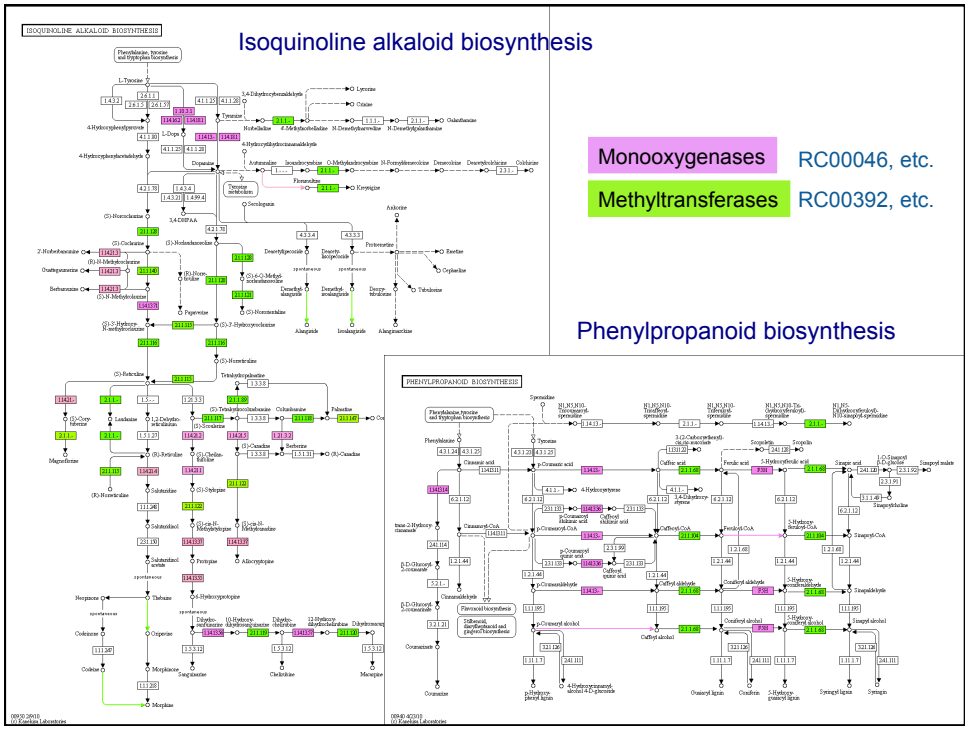
4-Coumarate → Caffeate → Ferulate → 5-Hydroxyferulate → Sinapate

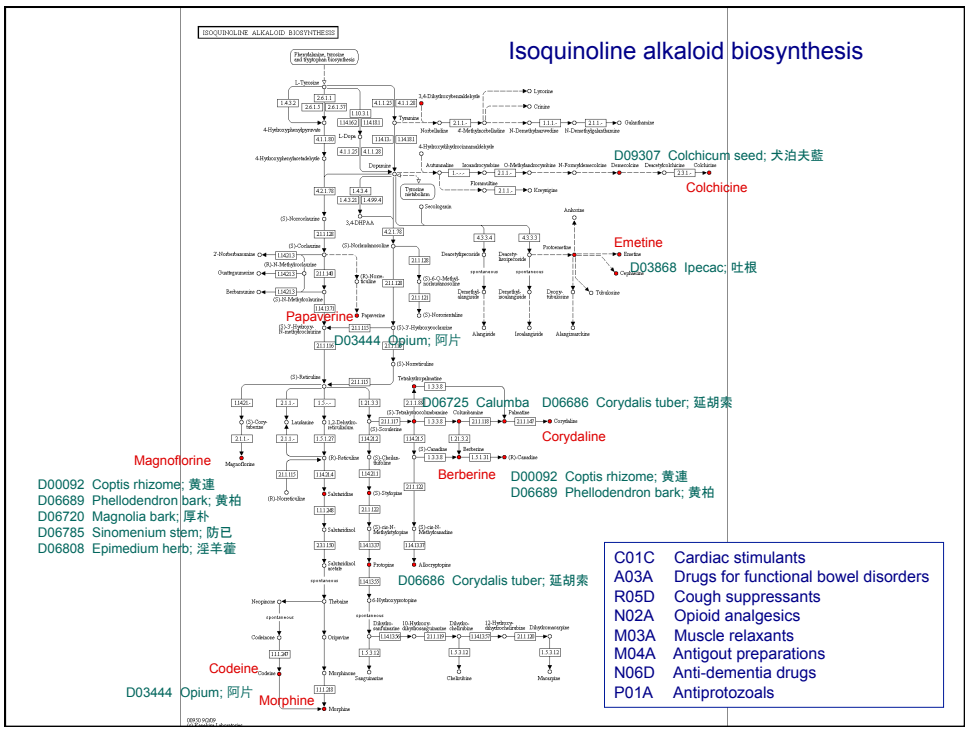
Isoquinoline alkaloid biosynthesis

monooxygenase — RC00046 → methyltransferase — RC00392



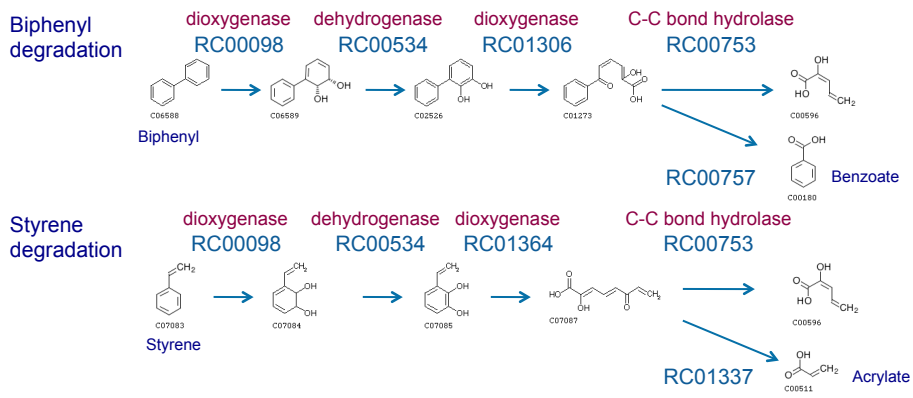
(S)-N-Methylcoclaurine → (S)-3'-Hydroxy-N-methylcoclaurine → (S)-Reticuline


ISOQUINOLINE ALKALOID BIOSYNTHESIS
Isoquinoline alkaloid biosynthesis
Monooxygenases RC00046, etc.
Methyltransferases RC00392, etc.
Phenylpropanoid biosynthesis
PHENYLPROPANOID BIOSYNTHESIS


ISOQUINOLINE ALKALOID BIOSYNTHESIS
Isoquinoline alkaloid biosynthesis
D09307 Colchicum seed; 犬泊夫藍
Colchicine
Emetine
D03868 Ipecac; 吐根
Papaverine
D03444 Opium; 阿片
D06725 Calumba
D06686 Corydalis tuber; 延胡索
Corydaline
Magnoflorine
Berberine
D00092 Coptis rhizome; 黃連
D06689 Phellodendron bark; 黃柏
D00092 Coptis rhizome; 黃連
D06689 Phellodendron bark; 黃柏
D06720 Magnolia bark; 厚朴
D06785 Sinomenium stem; 防己
D06808 Epimedium herb; 淫羊藿
D06686 Corydalis tuber; 延胡索
Codeine
D03444 Opium; 阿片
Morphine
C01C Cardiac stimulants
A03A Drugs for functional bowel disorders
R05D Cough suppressants
N02A Opioid analgesics
M03A Muscle relaxants
M04A Antigout preparations
N06D Anti-dementia drugs
P01A Antiprotozoals

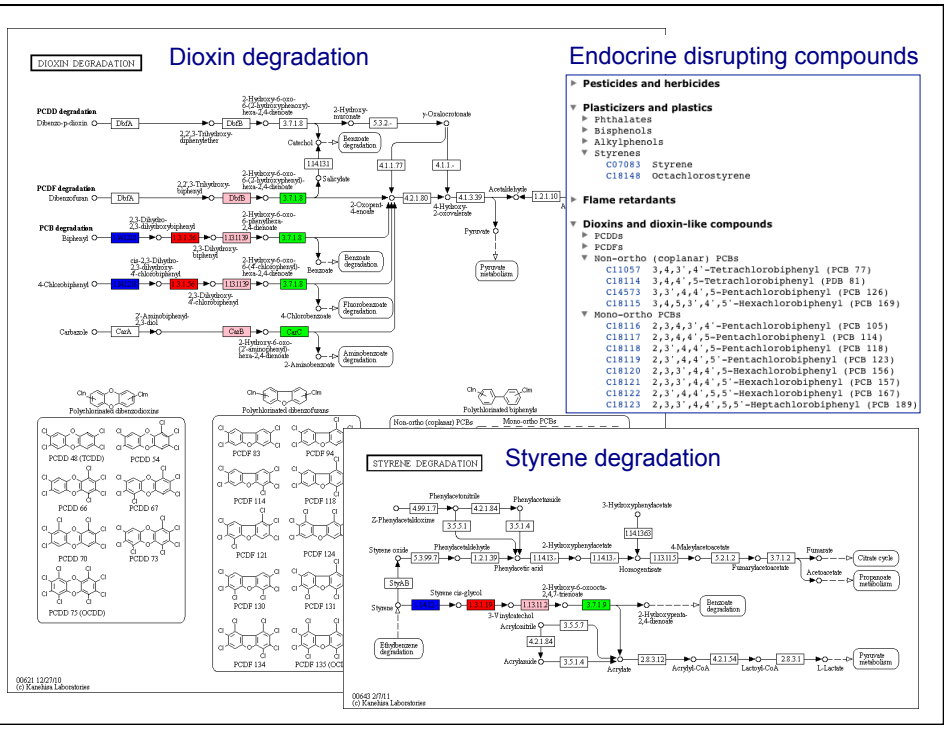
Chemical structure transformation network

Biodegradation reaction patterns represented by RC numbers



Burkholderia xenovorans
Polaromonas naphthalenivorans
Rhodococcus sp. RHA1

bphA	bphB	bphC	bphD
K08689	K08690	K00462	K10222
R05263	R05239	R03462	R02606
RC00098	RC00534	RC01306	RC00753
			RC00757

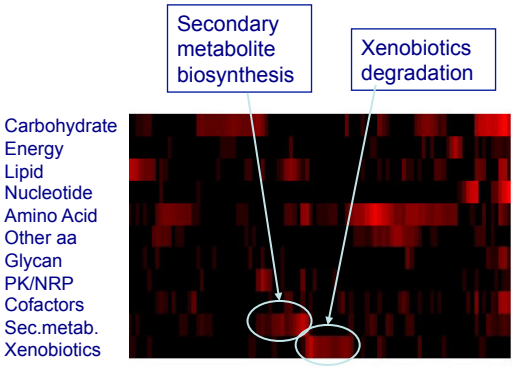


Prediction of biodegradation and biosynthetic pathways

PathPred @ http://www.genome.jp/tools/pathpred/

Prediction procedure

1. Global similarity search against KEGG COMPOUND
2. Local pattern match against KEGG RCLASS
3. Structure transformation using matched pattern

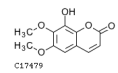


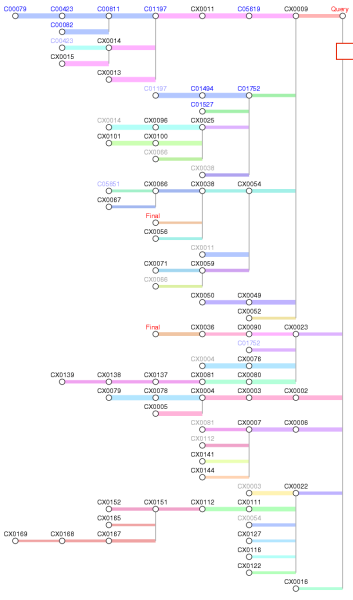
1,901 RDM patterns

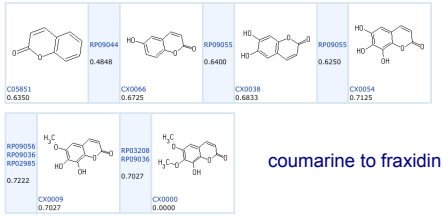
Chemical structure transformation patterns in known enzymatic reactions

Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M.; Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Info. Model.* 47, 1702-1712 (2007).
Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., and Kanehisa, M.: PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.* 38, W138-W143 (2010).

Prediction of fraxidin biosynthesis pathway

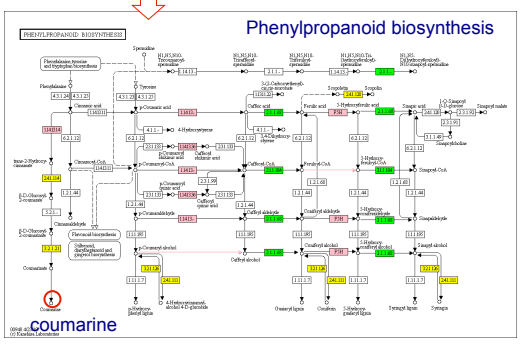






coumarine to fraxidin

RC00046 RC00392 RC00171

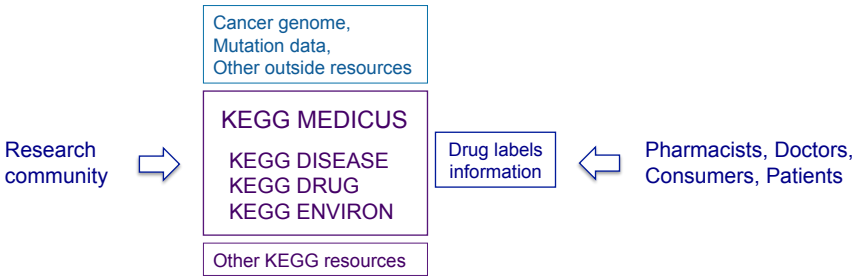


KEGG MEDICUS:
Bringing genomic revolution to society

KEGG MEDICUS

Molecular network based information resource for human diseases and drugs

KEGG DISEASE	KEGG DRUG
http://www.genome.jp/kegg/disease/	http://www.genome.jp/kegg/drug/
Diseases with known genetic factors and infectious diseases with known pathogen genomes, represented as gene/molecule lists	Drugs approved in Japan, USA,and Europe, containing chemical structures and interaction information including targets, metabolizing enzymes, transporters, genomic biomarkers, etc.



KEGG DISEASE entry for Alzheimer's disease

DISEASE: H00056

Help

Entry	H00056 Disease
Name	Alzheimer's disease (AD)
Description	Alzheimer's disease (AD) is the most prevalent neurodegenerative disease. Three genes have been implicated in the pathogenesis of early onset autosomal dominant AD; namely, the genes encoding amyloid precursor protein (APP), presenilin-1 (PSEN1) and presenilin-2 (PSEN2). On the hand, the apolipoprotein E4 allele is a genetic risk factor related to an increased risk of late-onset AD development. The neuropathological hallmarks of AD are extracellular amyloid plaques (senile plaques) and intraneuronal neurofibrillary tangles (NFT). The main symptoms of AD are progressive memory impairment.
Category	Neurodegenerative disease BRITE hierarchy
Pathway	hsa05010 Alzheimer's disease
Gene	(AD1) APP; amyloid beta (A4) protein (mutation) [HSA:351] [KO:K04520] (AD2) APOE; apolipoprotein E (mutation) [HSA:348] [KO:K04524] (AD3) PSEN1; presenilin 1 (mutation) [HSA:5663] [KO:K04505] (AD4) PSEN2; presenilin 2 (mutation) [HSA:5664] [KO:K04522]
Marker	N-acetylaspartate (reduction)
Drug	Donepezil [DR:D00670] acetylcholinesterase (AChE) inhibitor Galantamine [DR:D04292] acetylcholinesterase (AChE) inhibitor Rivastigmine [DR:D03822] acetylcholinesterase (AChE) inhibitor Memantine [DR:D04905] NMDA receptor antagonist
Comment	Disease class: tauopathy Affected region: hippocampus, cerebral cortex Microscopic lesion: amyloid plaques, neurofibrillary tangles, Lewy bodies (seen in Lewy body variant)
Other DBs	OMIM: 104300 104310 607822 606889 ICD-10: G30
Reference	PMID:19679070 (AD2)
Authors	Kim J, Basak JM, Holtzman DM
Title	The role of apolipoprotein E in Alzheimer's disease.
Journal	Neuron 63:287-303 (2009)
Reference	PMID:19524503
Authors	Kim D, Tsai LH
Title	Bridging physiology and pathology in AD.
Journal	Cell 137:997-1000 (2009)
Reference	PMID:18802446
Authors	Bertram L, Tanzi RE
Title	Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses.
Journal	Nat Rev Neurosci 9:768-78 (2008)
Reference	PMID:18414205
Authors	Bird TD
Title	Genetic aspects of Alzheimer disease.
Journal	Genet Med 10:231-9 (2008)
Reference	PMID:18370236 (drug)
Authors	Cacabelos R
Title	Pharmacogenomics in Alzheimer's disease.
Journal	Methods Mol Biol 448:213-357 (2008)

Four known disease genes

APP (amyloid beta protein)
APOE (apolipoprotein E)
PSEN1 (Presenilin 1)
PSEN2 (Presenilin 2)

Two therapeutic drugs

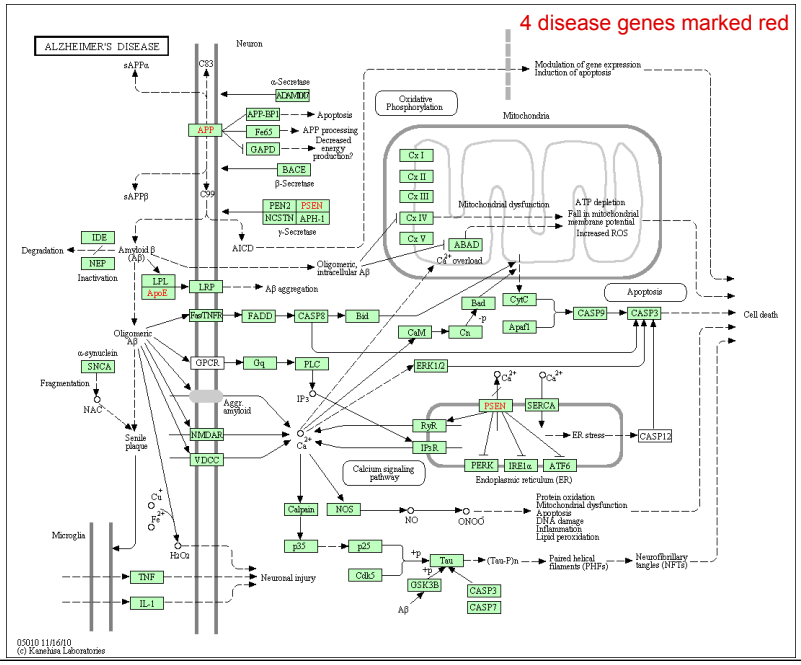
NMDA receptor antagonist (in the map)
Memantine hydrochloride
AChE inhibitor (not in the map)
Donepezil hydrochloride (Aricept)

Related neurodegenerative diseases

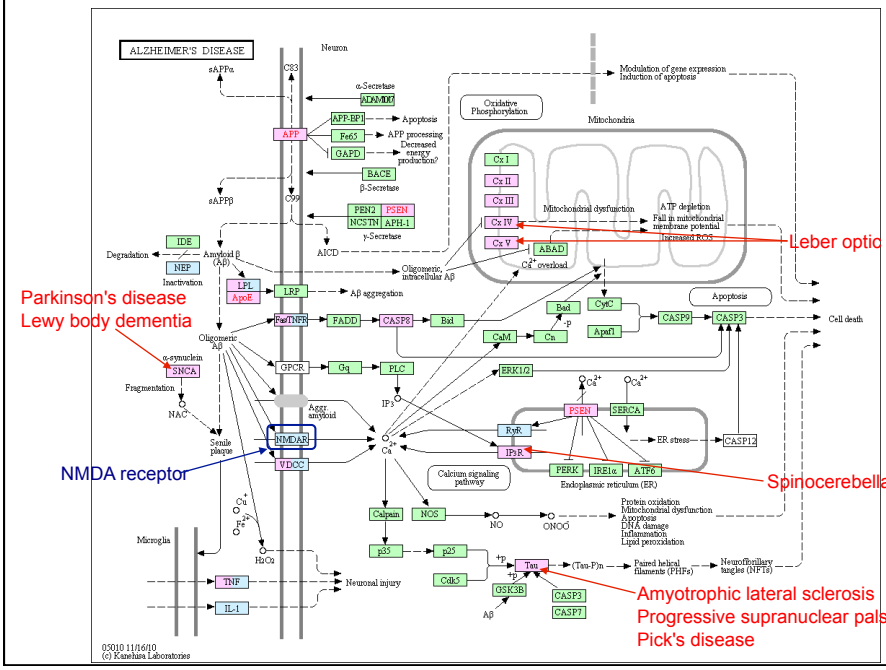
- **Nervous system diseases**
  - Neurodegenerative diseases
    - H00056 Alzheimer's disease [PATH:hsa05010]
    - H00066 Lewy body dementia
    - H00057 Parkinson's disease [PATH:hsa05012]
    - H00058 Amyotrophic lateral sclerosis [PATH:hsa05014]
    - H00059 Huntington's disease [PATH:hsa05016]
    - H00060 Dentatorubropallidoluysian atrophy (DRPLA)
    - H00062 Kennedy's disease
    - H00063 Spinocerebellar ataxia [PATH:hsa04730]
    - H00061 Prion diseases [PATH:hsa05020]
    - H00064 Ataxia telangiectasia
    - H00065 Alexander disease
    - H00075 Refsum disease [PATH:hsa04146]
    - H00067 Friedreich ataxia
    - H00068 Leber optic atrophy [PATH:hsa00190]
    - H00076 Cockayne syndrome
    - H00077 Progressive supranuclear palsy
    - H00078 Pick's disease
    - H00074 Canavan disease [PATH:hsa00250]

KEGG pathway map for Alzheimer's disease

4 disease genes marked red



Disease genes and drug targets in the Alzheimer map



KEGG DISEASE entry for Parkinson's disease

DISEASE: H00057

Help

Entry	H00057 Disease
Name	Parkinson's disease (PD)
Description	Parkinson's disease (PD) is a progressive degenerative movement disorder, characterized by motor symptoms such as tremor, postural imbalance, slowness of movement and rigidity. The main pathological features are the selective degeneration of the dopaminergic neurons of the substantia nigra pars compacta and the presence of intraneuronal proteinacious cytoplasmic inclusions termed "Lewy bodies". Seven genes have been implicated in the pathogenesis of familial PD; namely, alpha-synuclein (SNCA), parkin, ubiquitin carboxylterminal hydrolase L1 (UCHL1), DJ-1, leucine-rich repeat kinase 2 (LRRK2) and PTEN-induced kinase 1 (PINK1), and a nuclear receptor family protein (NR4A2). Several genes associated with susceptibility to Parkinson disease have been also identified.
Category	Neurodegenerative disease BRITE hierarchy
Pathway	hsa05012 Parkinson's disease
Gene	(PARK1/PARK4) SNCA; alpha-synuclein (mutation/duplication, triplication) [HSA:6622] [KO:K04528] (PARK2) Parkin (mutation, genomic rearrangement) [HSA:5071] [KO:K04556] (PARK5) UCHL1; ubiquitin carboxylterminal hydrolase L1 (mutation) [HSA:7345] [KO:K05611] (PARK6) PINK1;PTEN-induced kinase 1 (mutation) [HSA:65018] [KO:K05688] (PARK7) DJ1 (mutation) [HSA:11315] [KO:K05687] (PARK8) LRRK2; leucine-rich repeat kinase 2 (mutation) [HSA:120892] [KO:K08844] (PARK9) ATP13A2; ATPase type 13A2 (mutation) [HSA:23400] [KO:K13526] (PARK13) HTRA2; HtrA serine peptidase 2 (mutation) [HSA:27429] [KO:K08669] (PARK14) PLA2G6; phospholipase A2, group VI (mutation) [HSA:8398] [KO:K01047] (PARK15) FBXO7; F-box protein 7 (mutation) [HSA:25793] [KO:K10293] NR4A2 (mutation) [HSA:4929] [KO:K08558]
Env factor	MPTP [cpd:C04599] Rotenone [cpd:C07593] Maneb [cpd:C15231] Paraquat [cpd:C14701]
Marker	Dopamine [CPD:C03758] Homovanillic acid [CPD:C05582] Tyrosine hydroxylase [HSA:7054] DOPA decarboxylase [HSA:1644]
Drug	Carbidopa [DR:D00558] DOPA decarboxylase inhibitor Levodopa [DR:D00059] dopamine precursor Biperiden [DR:D00779] Amantadine hydrochloride [DR:D00777]
Comment	Disease class: synucleinopathy Affected region: substantia nigra, putamen, caudate nucleus, hypothalamus Microscopic lesion: Lewy bodies
Other DBs	OMIM: 168600 168601 600116 191342 605909 606324 607060 606693 610297 260300 ICD-10: G20
Reference **Authors** **Title** **Journal**	PMID:19729209 (Env_factor) Cicchetti F, Drouin-Ouellet J, Gross RE Environmental toxins and Parkinson's disease: what have we learned from pesticide-induced animal models? Trends Pharmacol Sci 30:475-83 (2009)

Seven known disease genes

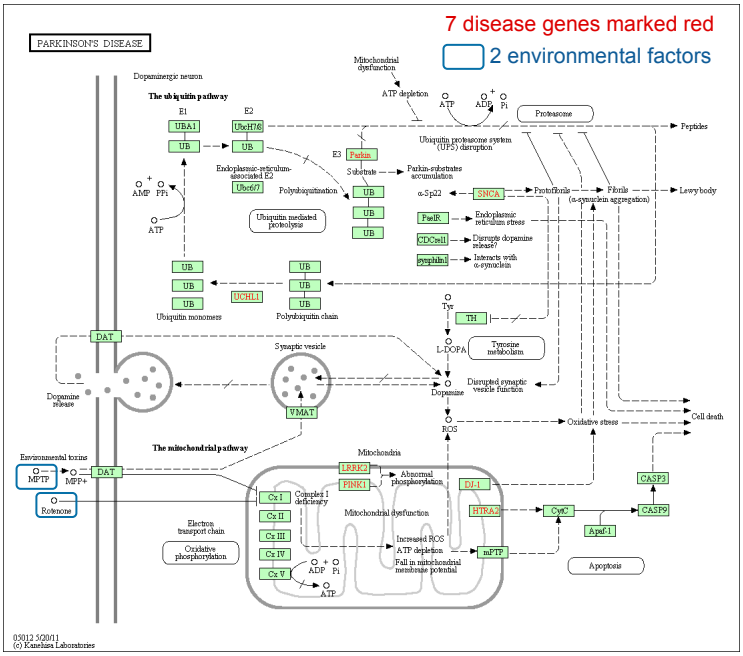
SNCA (alpha-synuclein)
Parkin (ubiquitin ligase)
UCHL1 (ubiquitin carboxylterminal hydrolase L1)
PINK1 (PTEN-induced kinase 1)
DJ1
LRRK2 (leucine-rich repeat kinase 2)
NR4A2 (nuclear receptor family protein)

Two environmental factors

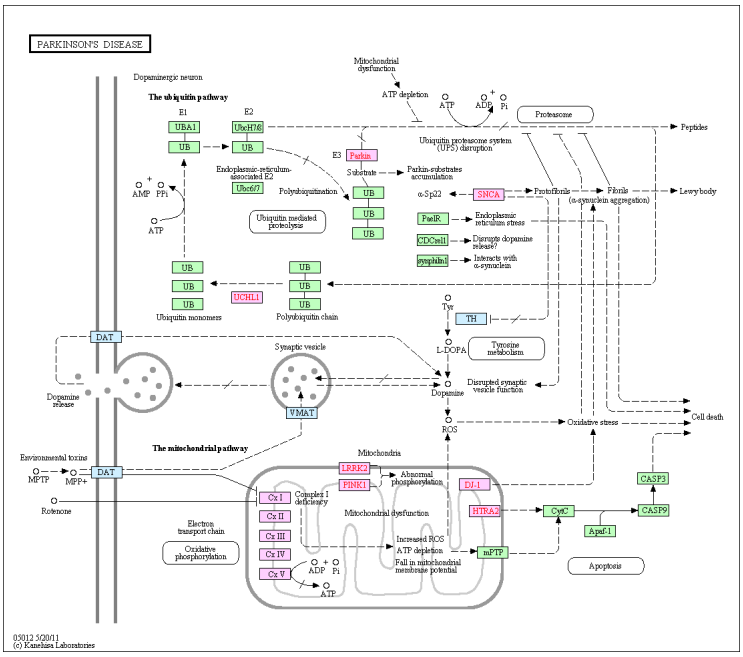
MTPT (neurotoxin)
Rotenone (isoflavonoid)

and other genetic and environmental factors

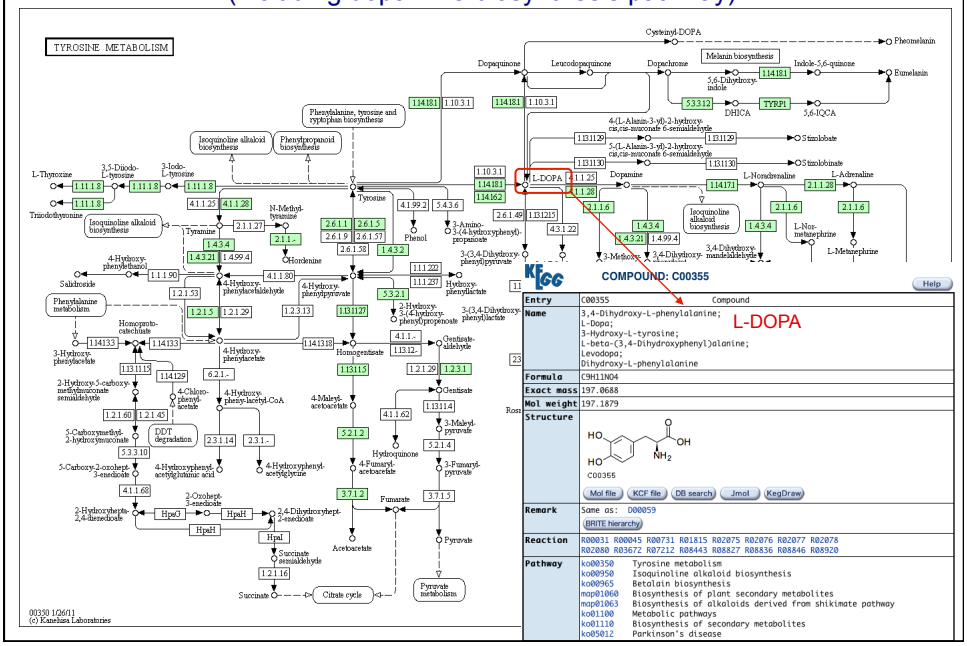
KEGG pathway map for Parkinson's disease



Disease genes and drug targets in the Parkinson map

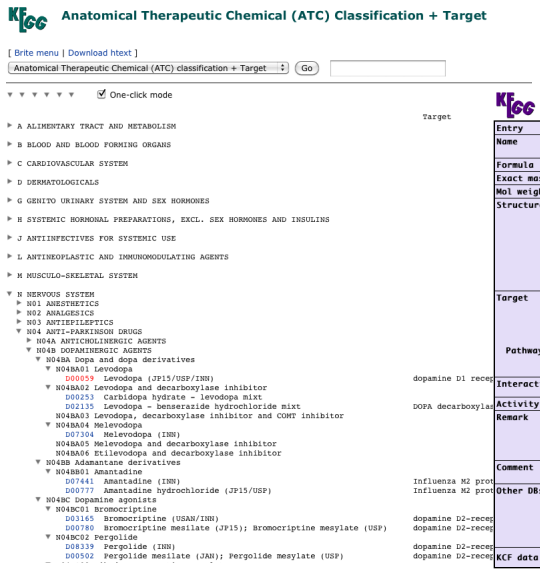


KEGG pathway map for tyrosine metabolism (including dopamine biosynthesis pathway)

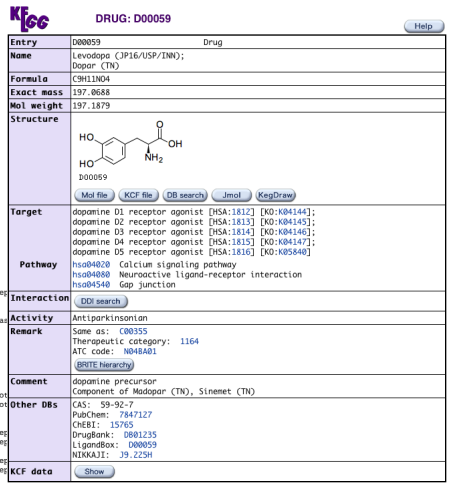


Anti-Parkinson drugs

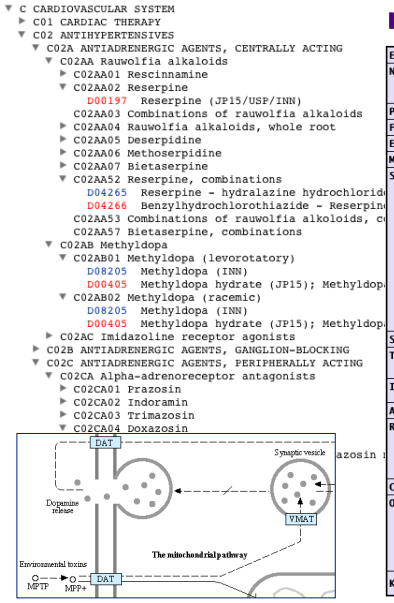
KEGG BRITE hierarchy for ATC classification

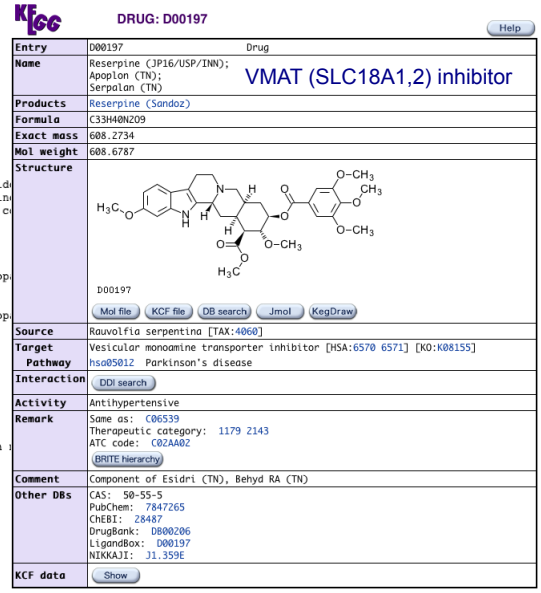


KEGG DRUG entry for Levodopa



Adverse drug interaction with antihypertensives



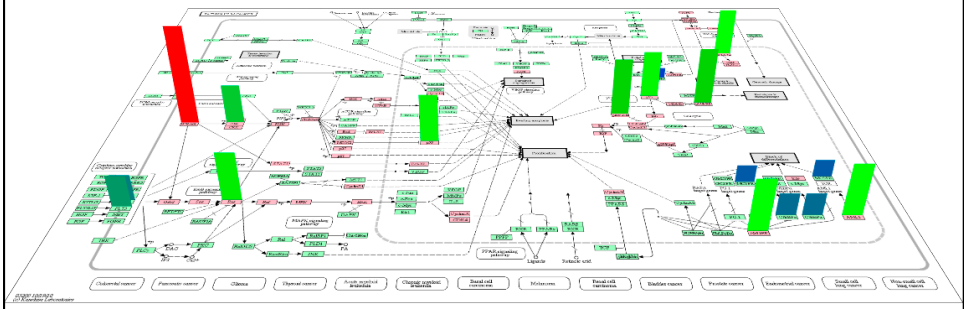


KEGG CANCER Resource

55 KEGG DISEASE entries
14 KEGG pathway maps
1 KEGG pathway map for 'Pathways in Cancer' (combined map of 14 cancers)
1 BRITE hierarchy for 'Cancer stages' (of 14 cancers)
1 BRITE hierarchy for 'Antineoplastics'
1 BRITE hierarchy for 'Carcinogens'

Additional data from outside resouces:

Somatic mutations in cancer genomes (from COSMIC)



Chronic myeloid leukemia in the overview map

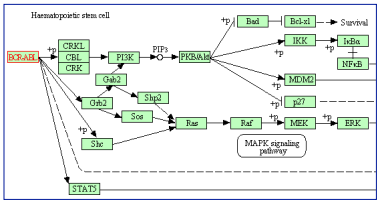
Molecular target drug for chronic myeloid leukemia

KEGG **DRUG: D01441**

Help

Entry	D01441 Drug
Name	Imatinib mesilate (JAN); Imatinib mesylate; Gleevec (TN); Glivec (TN)
Products	GLEEVEC (Novartis Pharma), GLEEVEC (Physicians Total Care)
Formula	C29H31N7O. CH4SO3
Exact mass	589.2471
Mol weight	589.7084
Structure	D01441 Mol file · KCF file · DB search · Jmol · KegDraw
Target	Bcr-Abl tyrosine kinase inhibitor [HSA:25] [KO:K06619]; PDGFRA tyrosine kinase inhibitor [HSA:5156] [KO:K04363]; PDGFRB tyrosine kinase inhibitor [HSA:5159] [KO:K05089]; c-KIT tyrosine kinase inhibitor [HSA:3815] [KO:K05091]
Pathway	hsa04010 MAPK signaling pathway hsa04020 Calcium signaling pathway hsa04060 Cytokine-cytokine receptor interaction hsa04640 Hematopoietic cell lineage hsa05200 Pathways in cancer hsa05220 Chronic myeloid leukemia
Metabolism	Enzyme: CYP3A4 [HSA:1576], CYP3A5 [HSA:1577], CYP2D6 [HSA:1565], CYP2 [HSA:1559] Transporter: ABCG2 [HSA:9429]
Interaction	Genomic biomarker: c-KIT expression [HSA:3815], Philadelphia chromoso CYP inhibition: CYP3A4 [HSA:1576], CYP3A5 [HSA:1577], CYP3A7 [HSA:155 DDI search
Activity	Antineoplastic
Remark	Therapeutic category: 4291 ATC code: L01XE01 BRITE hierarchy
Comment	Treatment of chronic myelogenous leukemia (CML) and gastrointestinal tumors (GIST)
Structure map	map07045 Antineoplastics - protein kinases inhibitors
Other DBs	CAS: 220127-57-1 PubChem: 7848504 ChEBI: 31690 DrugBank: DB00619 LigandBox: D01441 NIKKAJI: J1.337.144J
KCF data	Show

Gleevec

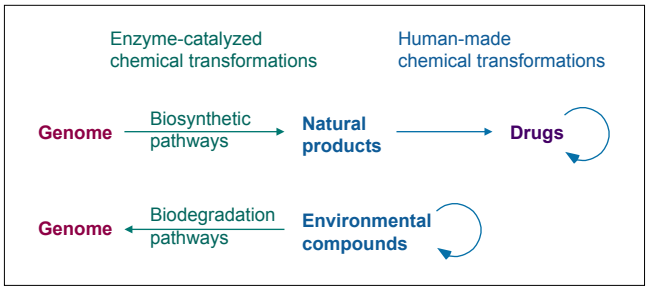


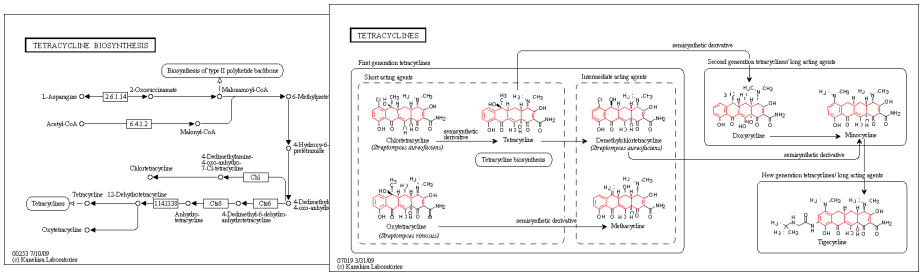
Type

```
A ALIMENTARY TRACT AND METABOLISM
  A01 STOMATOLOGICAL PREPARATIONS
    A01A STOMATOLOGICAL PREPARATIONS
      A01AA Caries prophylactic agents
      A01AB Antiinfectives and antiseptics for local oral treatment
      A01AC Corticosteroids for local oral treatment
        A01AC01 Triamcinolone
        A01AC02 Dexamethasone
          D00292  Dexamethasone (JP15/USP/INN)                          Enzyme: CYP3A4
          D00975  Dexamethasone sodium phosphate (JAN/USP)              Enzyme: CYP3A4
          D01510  Dexamethasone metasulfobenzoate sodium (JAN)          unclassified
          D01615  Dexamethasone palmitate (JAN)                        Enzyme: CYP3A4
          D01632  Dexamethasone propionate (JAN); Dexamethasone dip     unclassified
          D01948  Dexamethasone valerate (JAN)                         unclassified
          D02174  Dexamethasone acetate (JAN/USP)
          D02591  Dexamethasone acefurate (USAN/INN)
          D02592  Dexamethasone beloxil (USAN)
          D07073  Dexamethasone cipecilate (JAN/INN)                   unclassified
          D07796  Dexamethasone 21-acetate
          D07797  Dexamethasone isonicotinate
          D07798  Dexamethasone sodium hemisulfate
          D07799  Dexamethasone 21-tebutate
          D07800  Dexamethasone 21-valerate
          D07801  Dexamethasone phenylpropionate
          D07802  Dexamethasone phosphate
```

Chemical structure transformation networks

Linking genomes to natural products, drugs, and xenobiotics

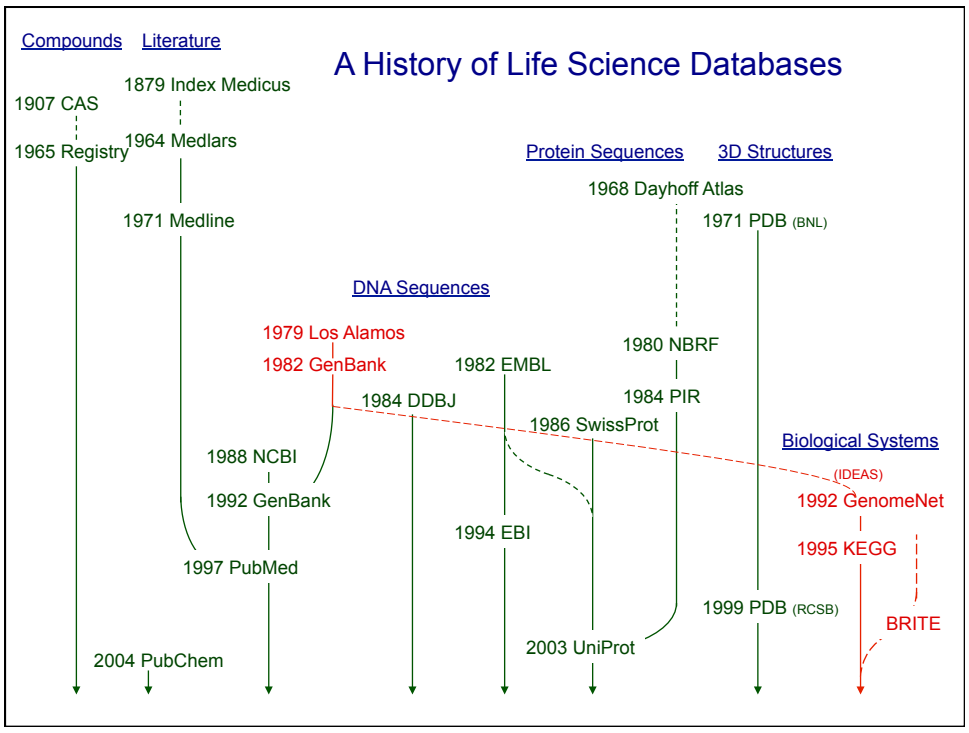




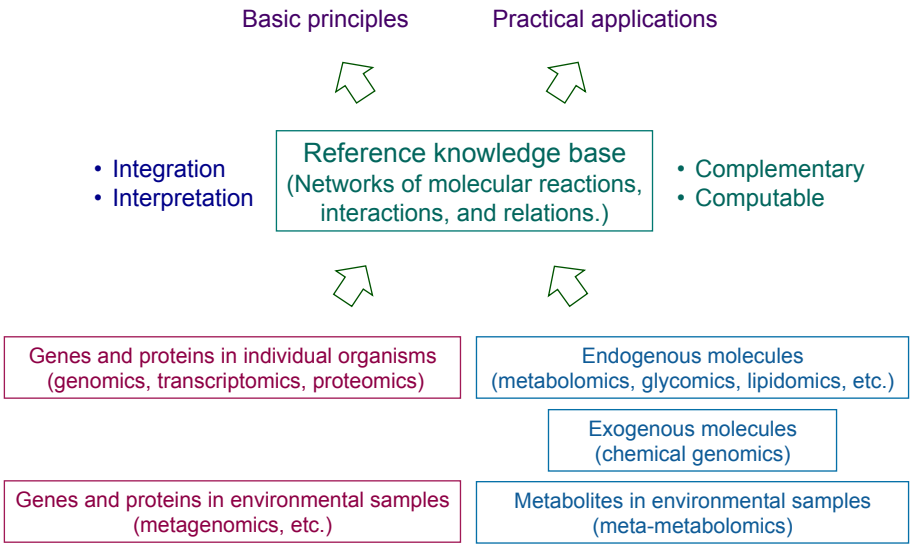
KEGG metabolic pathway map

KEGG DRUG structure map

Summary


A History of Life Science Databases
Compounds
Literature
1907 CAS
1879 Index Medicus
1965 Registry
1964 Medlars
1971 Medline
1997 PubMed
2004 PubChem
Protein Sequences
3D Structures
1968 Dayhoff Atlas
1971 PDB (BNL)
1980 NBRF
1984 PIR
1986 SwissProt
2003 UniProt
1999 PDB (RCSB)
DNA Sequences
1979 Los Alamos
1982 GenBank
1982 EMBL
1984 DDBJ
1988 NCBI
1992 GenBank
1994 EBI
Biological Systems
(IDEAS)
1992 GenomeNet
1995 KEGG
BRITE

KEGG: Reference knowledge base for biological interpretation of large-scale datasets



KEGG: Reference knowledge base for use in practice and in society

• Capturing experimental knowledge on molecular systems both in normal and perturbed (disease) states	KEGG PATHWAY KEGG BRITE KEGG DISEASE
• Capturing knowledge on drugs and environmental compounds as perturbants to molecular systems	KEGG DRUG KEGG COMPOUND
• Generalizing knowledge on genes and proteins by KEGG Orthology	KEGG ORTHOLOGY KEGG GENES
• Generalizing knowledge on chemical transformations in enzymatic reactions by reaction class	KEGG RCLASS KEGG REACTION

Possible uses of KEGG in practice and in society

- Realization of personalized medicine based on personal genomes
- Prevention and therapy of infectious diseases based on pathogen genomes
- Discovery of useful natural products and drug leads based on plant genomes
- Biodegradation of environmental compounds based on microbial genomes