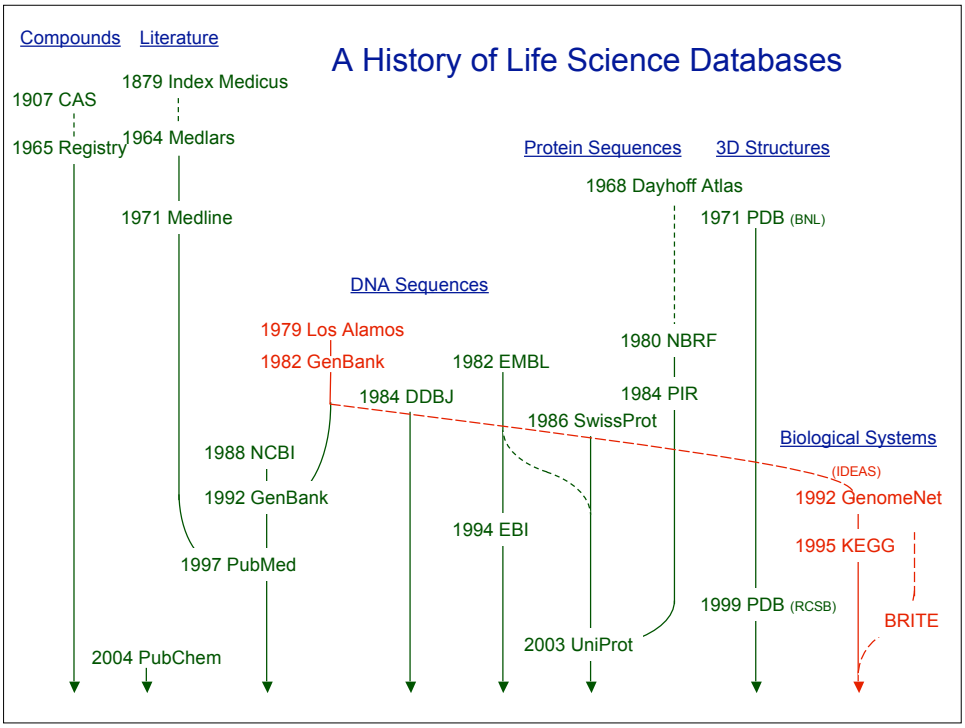
KEGG: Kyoto Encyclopedia of Genes and Genomes

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

June 2009

Background

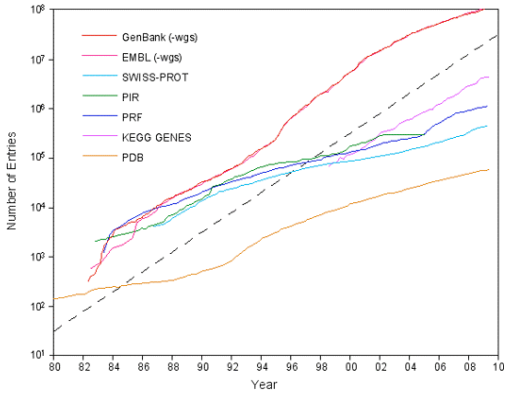
- Integration of life science databases
- Reconstruction of biological systems



IDEAS

Integrated Database and Extended Analysis System for nucleic acids and proteins

Growth of sequence and 3D structure databases



- Originally developed in 1980 as part of the pre-GenBank project (Los Alamos Sequence Library)
- GenBank awarded to Los Alamos in September 1982
- Package rewritten at NIH in 1983 for integration of sequence and 3D structure databases
- Relational database named "Cell"

Compiled from:

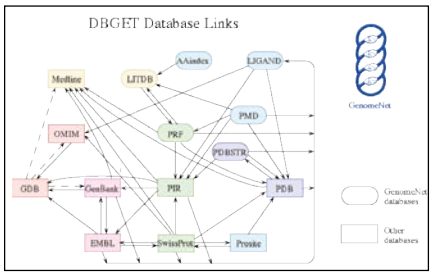
1980-1983	Los Alamos sequence analysis system
1983-1986	IDEAS system at the National Institutes of Health
1986-1992	IDEAS system at Kyoto University
1992-	DBGET system in GenomeNet database service

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

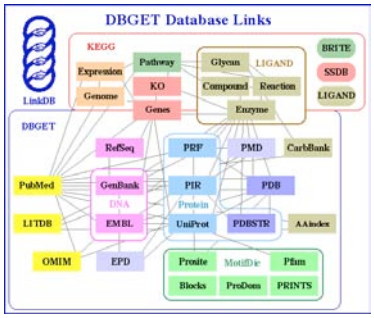
DBGET

Integrated Database Retrieval System for the Web of Molecular Biology Data

Link-based integration
Identifier database:entry
Relation database1:entry1 → database2:entry2



15 databases in September 1994



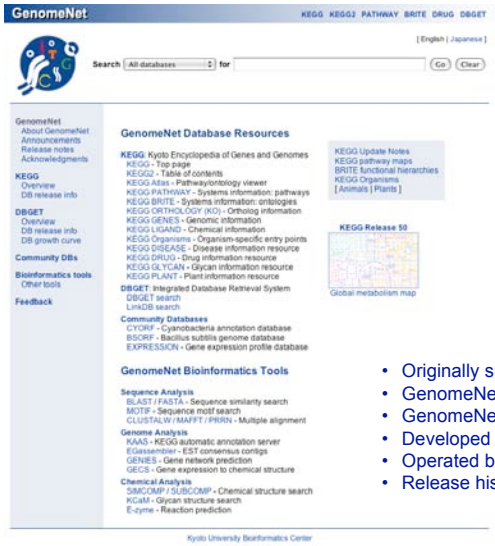
32 databases in September 2004

Category	bget	bfind	blink	brite	#DB
1. KEGG databases in DBGET	yes	yes	yes	yes	20
2. Other DBGET databases	yes	yes	yes	no	19
3. Searchable databases on the web	no	yes	yes	no	17
4. Link-only databases on the web	no	no	yes	no	94
5. PubMed database	yes	no	yes	no	1

151 databases in June 2009

GenomeNet

http://www.genome.jp/



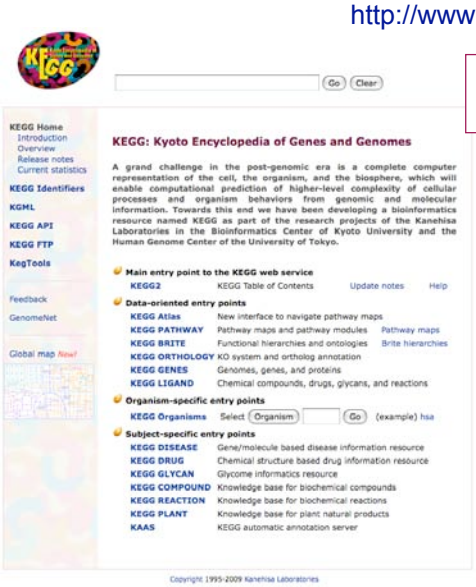
Original Services
KEGG
DBGET
MOTIF
KAAS
SIMCOMP / SUBCOMP
KCaM

Other Services
BLAST / FASTA
CLUSTALW / MAFFT / PRRN
CYORF / BSORF

- Originally supported by Japanese Human Genome Project
- GenomeNet e-mail service launched in September 1992
- GenomeNet web service launched in July 1994
- Developed by Kanehisa Laboratories
- Operated by Kyoto University Bioinformatics Center
- Release history at: http://www.genome.jp/en/release.html

KEGG

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



Organisms viewed as databases
Identifier organism:gene
Relation organism1:gene1 → organism2:gene2

- Collect only completely sequenced genomes for computer representation of cell/organism functions
- Organize higher-level biological and chemical knowledge to complement with the existing databases
- Depend on internal team members to develop high-quality databases (accept feedbacks, but no submissions)
- Develop associated research programs that utilize the new databases and demonstrate their utility

- Supported by multiple research grants to Minoru Kanehisa
- Project initiated in May 1995
- First release in December 1995

What is Database?

	NCBI	KEGG
Database	Information infrastructure	Computer representation of biological systems
Content	Comprehensive repositories	Building blocks and wiring-diagrams
Integration	Linking	Reconstruction
Implementation	Entrez	KEGG
Main use	Individual data retrieval	Pathway mapping and Brite mapping

Overview of KEGG

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

Goal

An ultimate goal of bioinformatics is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity, such as

- molecular interaction networks involving various cellular processes and
- phenotypes (morphological, physiological, and behavioral aspects) of entire organisms

from genomic information.

Kanehisa, M. and Bork, P.; Bioinformatics in the post-sequence era. *Nat. Genet.* **33**, 305-310 (2003).

Building Blocks of Life
Genomic and Chemical Spaces

Genomic Space

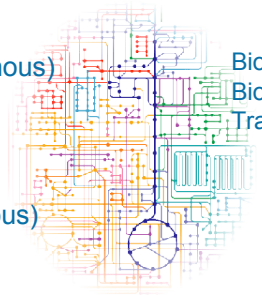
DNA (Gene)	Genome
RNA	Transcriptome
Protein	Proteome

Replication
Transcription
Translation

DNA ⇄ RNA → Protein

Chemical Space

Metabolite	Metabolome	(Endogenous)
Glycan	Glycome	
Lipid	Lipidome	
Compound	Chemical genome	(Exogenous)



Biosynthesis
Biodegradation
Transport

Genomic and Chemical Spaces

Genomic Space

- Contains genetic building blocks of life, i.e., DNA, RNA, and proteins
- Represents all possible sequences
- Uncovered by genomics, transcriptomics, and proteomics

Chemical Space

- Contains chemical building blocks of life, e.g., small molecules, glycans, and lipids
- Represents all possible chemical structures
- Uncovered by metabolomics, glycomics, lipidomics, etc. for endogenous molecules
- Uncovered by chemical genomics for exogenous molecules

Integration of Genomic and Chemical Spaces

Genomic space	Chemical space
Repertoire of genes and proteins (genomics, proteomics, etc.)	Repertoire of endogenous molecules (metabolomics, glycomics, etc.)

Glycosyltransferase genes ⟺ Glycan structures

Other enzyme gene groups ⟺ Lipids, secondary metabolites, etc.

Repertoire of genes in the biosphere (metagenomics)	Repertoire of exogenous molecules (chemical genomics)

Genetic factors of diseases ⟺ Environmental factors of diseases

Drug targets ⟺ Drug leads

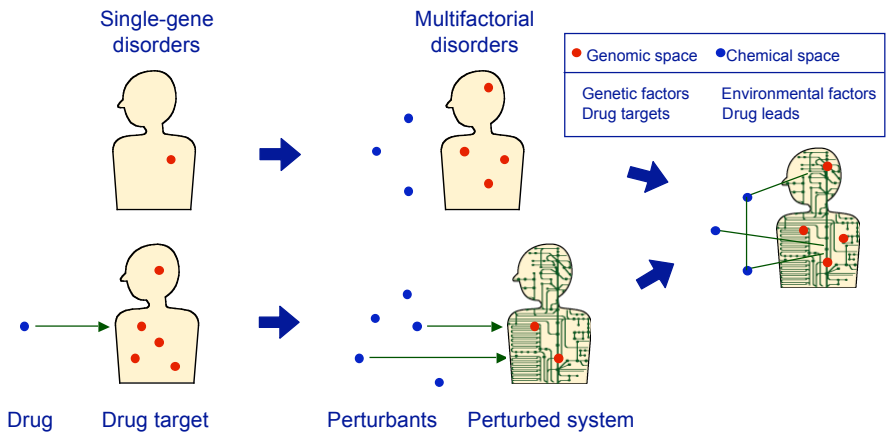
Human ⟺ Vector ⟺ Pathogen

Human gut ⟺ Microbial community

Integration of Genomic and Chemical Spaces

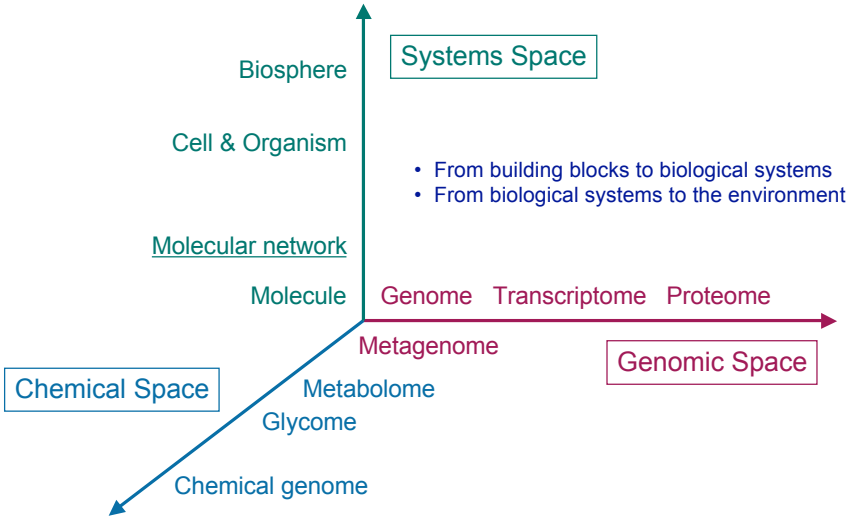
Medical and Pharmaceutical Implications

Diseases viewed as perturbed states of molecular systems
Drugs viewed as perturbants to molecular systems



Integration of Genomic and Chemical Spaces

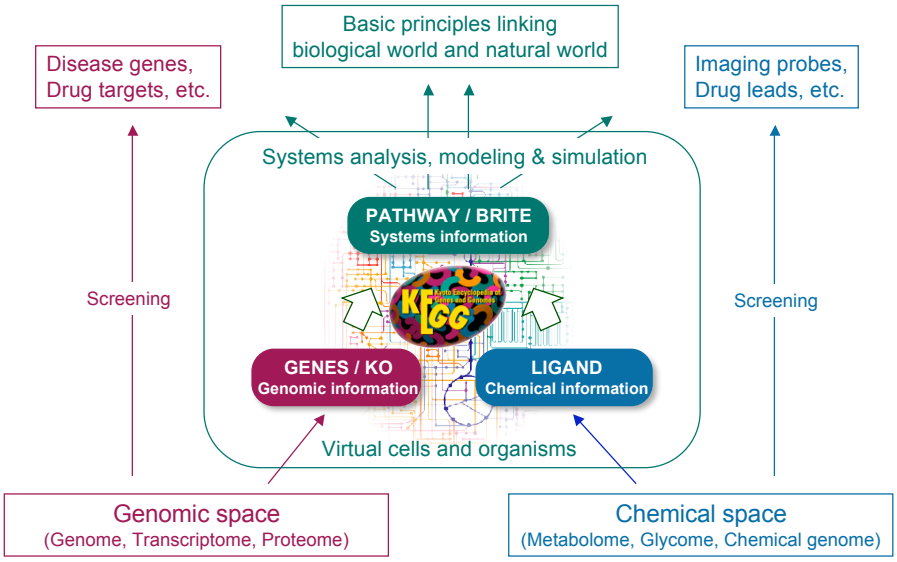
Bioinformatics Approaches to the Systems Space



- From building blocks to biological systems
- From biological systems to the environment

Kanehisa, M. et al.; From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res.* **34**, D354-D357 (2006).

KEGG: Computer representation of biological systems



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

KEGG Databases

Database	Objects	Data size
KEGG PATHWAY	Pathway maps	330 (95,271)
KEGG BRITE	Functional hierarchies and ontologies	61 (21,684)
KEGG MODULE	Pathway modules and complexes	692
KEGG DISEASE	Diseases	104
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	12,178
KEGG GENES	Genes in high-quality genomes	4,553,978
KEGG DGENES	Genes in draft genomes	151,662
KEGG EGENES	Genes as EST contigs	2,919,245
KEGG GENOME	Organisms	1,008 + 67
KEGG SSDB	Sequence similarities and best hit relations	-
KEGG COMPOUND	Metabolites and other small molecules	15,477
KEGG DRUG	Drugs	8,912
KEGG GLYCAN	Glycans	10,969
KEGG ENZYME	Enzyme nomenclature	5,061
KEGG REACTION	Biochemical reactions	7,915
KEGG RPAIR	Reactant pair chemical transformations	11,356
KEGG DPAIR	Drug pair chemical transformations	496

As of June 12, 2009

KEGG Object Identifiers

Prefix + 5-digit number

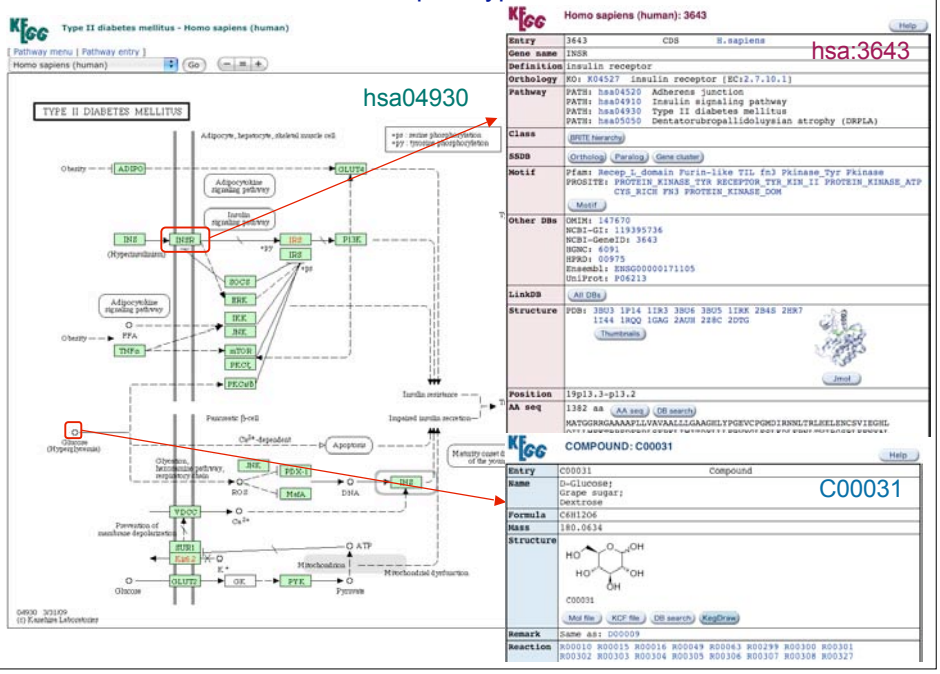
Prefix	Content	Database	Example
K	Ortholog group	KEGG ORTHOLOGY	K04527 for insulin receptor
C	Chemical compound	KEGG COMPOUND	C00031 for D-glucose
D	Drug	KEGG DRUG	D01441 for Gleevec
G	Glycan	KEGG GLYCAN	G00109 for GM2
R	Reaction	KEGG REACTION	R00259 for EC 2.3.1.1
RP	Reactant pair	KEGG RPAIR	RP04458 for C00025_C00624
map/ko/ec/rn/(org)	Pathway map	KEGG PATHWAY	hsa04930 for type II diabetes
br/jp/ko/(org)	Brite hierarchy	KEGG BRITE	ko01003 for glycosyltransferases
M	Pathway module	KEGG MODULE	M00008 for Entner-Doudoroff pathway
H	Human disease	KEGG DISEASE	H00004 for Chronic myeloid leukemia

Other identifiers

Identifier	Content	Database	Example
org	KEGG organism code • Three letter code • Prefix d for draft genome • Prefix e for EST contig	KEGG GENOME	hsa for Homo sapiens
org:gene	Gene entry	KEGG GENES	hsa:3643 for human insulin receptor
db:entry	Database entry	Any database	gn:hsa for Homo sapiens up:P06213 for UniProt insulin receptor

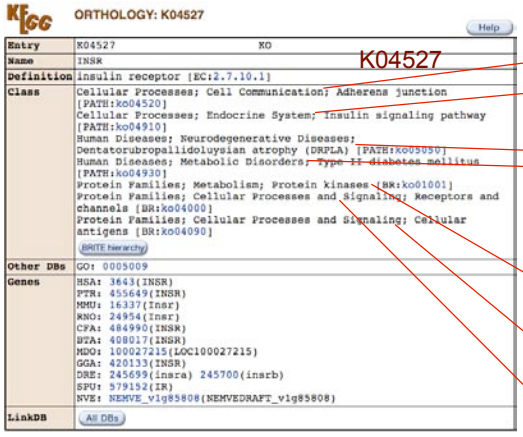
Try these examples by entering them in the search box of the KEGG top page.

KEGG PATHWAY map for type II diabetes mellitus



The KEGG Orthology (KO) System

KO (K number) entry for insulin receptor



KO identifiers (K numbers) represent manually defined ortholog groups corresponding to the KEGG pathway nodes and the BRITE hierarchy nodes (bottom leaves)

KEGG Orthology (ko00001)

- Metabolism
- Genetic Information Processing
- Environmental Information Processing
- Cellular Processes
  - .....
  - Cell Communication
  - .....
  - Endocrine System
  - .....
- Human Diseases
  - .....
  - Neurodegenerative Diseases
  - Metabolic Disorders
  - .....

Protein Kinases (ko01001)

- Serine/threonine protein kinases: AGC group
  - .....
- Tyrosine protein kinases
  - .....
  - InsR family
  - .....
- Histidine protein kinases

Receptors and Channels (ko04000)

- G-Protein Coupled Receptors
- 1 TM Signaling Receptors
  - Receptor tyrosine kinase
    - .....
    - RTK class II
    - .....
- Nuclear Receptors
- Ion Channels
- Other Channels

Cellular antigens (ko04090)

- CD molecules
- Non-CD molecules

Color Coding and Icons

KEGG Web Service

PATHWAY / BRITE

ORTHOLOGY

GENES

LIGAND

DISEASE, DRUG, GLYCAN, COMPOUND, REACTION, PLANT

GenomeNet, DBGET, BLAST / FASTA, MOTIF, ALIGN SIMCOMP / SUBCOMP, KCaM

KEGG Oracle Database (internal database)

Reference DB
Genome, Pathway, Brite, Module, Disease, KO

Genes DB
Gene annotation

Ligand DB
Chemical annotation

Java / Cocoa Applications

KegSketch (internal only)

KegDraw

KegHier

KegArray

Web Page Organization

Top page (KEGG)

This search box accepts both keywords and KEGG identifiers

This search box accepts keywords only

PATHWAY / BRITE ORTHOLOGY GENES LIGAND

Table of Contents (KEGG2)

These icons will bring back to the KEGG2 page.

DISEASE, DRUG, GLYCAN, COMPOUND, REACTION, PLANT

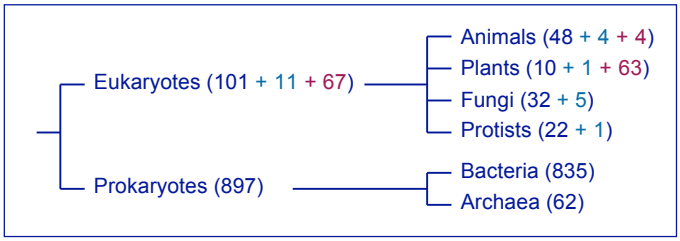
DBGET
BLAST / FASTA
SIMCOMP / SUBCOMP
KCaM

This icon will bring back to the KEGG top page.

This icon will bring back to the GenomeNet top page.

KEGG GENES
and ortholog annotation

KEGG Organisms



As of June 12, 2009

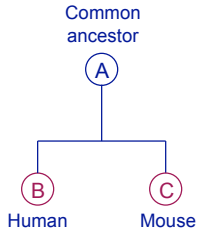
998 organisms in GENES - High-quality genomes with manual (koala) annotation
11 organisms in DGENES - Draft genomes with automatic (kaas) annotation
67 organisms in EGENES - EST contigs with automatic (kaas) annotation

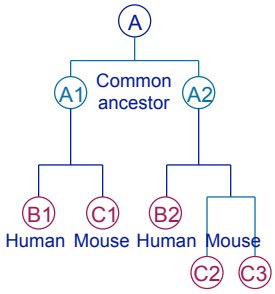
Total number of genes/proteins in GENES	4,553,978
Total number of proteins in UniProt	8,385,695 (468,851 in SwissProt)
Fraction of UniProt covered by GENES	about 90% (in terms of protein families)

KEGG GENES already covers most of the known protein universe

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.



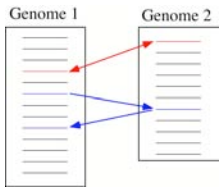


Orthologs: B–C, B1–C1
Co-orthologs: B2–(C2,C3)
Inparalogs: C2–C3
Outparalogs: B1–B2, B1–(C2,C3), B2–C1

Computational Identification of Orthologs

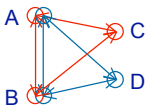
<u>Between two species</u>

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



<u>Among multiple species</u>

1. COG
   Triangle of BBH relationships among three species



Superposition of ABC and ABD

2. KEGG OC
   p-Quasi clique among multiple species

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: KO (K number) assignment

KEGG GENES

- Gene information for completely sequenced genomes
- Computationally generated from RefSeq and other public resources
- Manual annotation with KOALA and GFIT tools
- Automatic annotation by KAAS for draft genomes

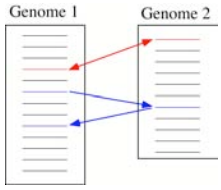
KEGG GENES in Oracle	
Genomes	997
Genes	4,553,977
Genes with KO	1,504,505
KO assignment	33%

As of June 12, 2009

KEGG ORTHOLOGY (KO)

- Manually defined ortholog groups that correspond to KEGG pathway nodes and BRITE hierarchy nodes
- Identified by K numbers

Bidirectional best hit



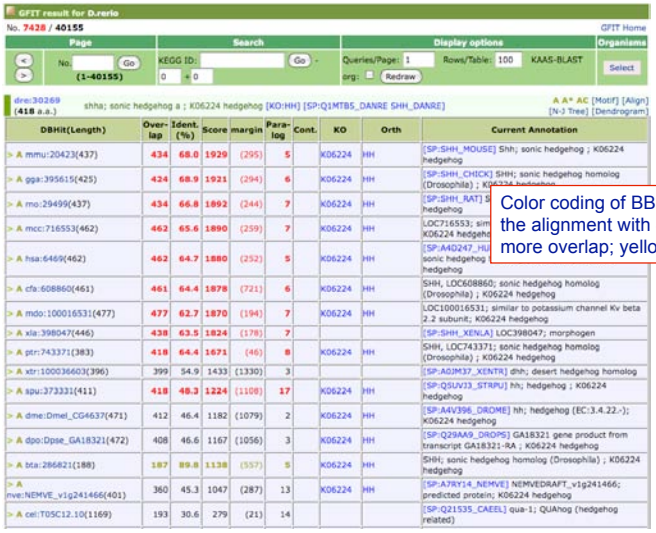
KEGG SSDB

- Sequence similarity scores and best hit relations
- Computationally generated from GENES by pairwise genome comparisons using SSEARCH

KEGG OC

- Ortholog clusters
- Computationally generated from SSDB by a quasi-clique finding algorithm

GFIT: Gene Function Identification Tool



GFIT result for D.rerio

No. 7428 / 40155 — GFIT Home

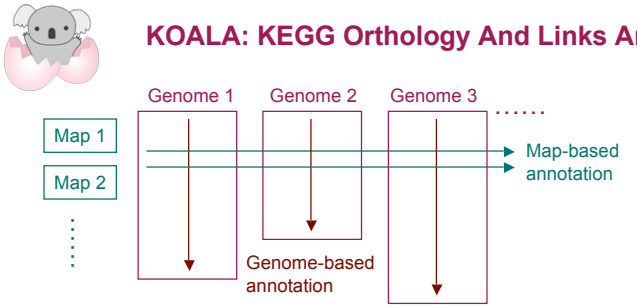
Page: No. (1-40155) Go — Search: KEGG ID: Go — Display options: Queries/Page: 1, Rows/Table: 100, KAAS-BLAST, org: Redraw — Organisms: Select

dre:30269 (418 a.a.) shha; sonic hedgehog a ; K06224 hedgehog [KO:HH] [SP:Q1MTB5_DANRE SHH_DANRE] — A A* AC [Motif] [Align] [N-J Tree] [Dendrogram]

DBHit(Length)	Over-lap	Ident. (%)	Score	margin	Para-log	Cont.	KO	Orth	Current Annotation
A mmu:20423(437)	434	68.0	1929	(295)	5		K06224	HH	[SP:SHH_MOUSE] Shh; sonic hedgehog ; K06224 hedgehog
A gga:395615(425)	424	68.9	1921	(294)	6		K06224	HH	[SP:SHH_CHICK] SHH; sonic hedgehog homolog (Drosophila) ; K06224 hedgehog
A rno:29499(437)	434	66.8	1892	(244)	7		K06224	HH	[SP:SHH_RAT] S hedgehog
A mcc:716553(462)	462	65.6	1890	(259)	7		K06224	HH	LOC716553; sim K06224 hedgeho
A hsa:6469(462)	462	64.7	1880	(252)	5		K06224	HH	[SP:A4D247_HU sonic hedgehog hedgehog
A cfa:608860(461)	461	64.4	1878	(721)	6		K06224	HH	SHH, LOC608860; sonic hedgehog homolog (Drosophila) ; K06224 hedgehog
A mdo:100016531(477)	477	62.7	1870	(194)	7		K06224	HH	LOC100016531; similar to potassium channel Kv beta 2.2 subunit; K06224 hedgehog
A xla:398047(446)	438	63.5	1824	(178)	7				[SP:SHH_XENLA] LOC398047; morphogen
A ptr:743371(383)	418	64.4	1671	(46)	8		K06224	HH	SHH, LOC743371; sonic hedgehog homolog (Drosophila) ; K06224 hedgehog
A xtr:100036603(396)	399	54.9	1433	(1330)	3				[SP:A0JM37_XENTR] dhh; desert hedgehog homolog
A spu:373331(411)	418	48.3	1224	(1108)	17		K06224	HH	[SP:Q5UV13_STRPU] hh; hedgehog ; K06224 hedgehog
A dme:Dmel_CG4637(471)	412	46.4	1182	(1079)	2		K06224	HH	[SP:A4V396_DROME] hh; hedgehog (EC:3.4.22.-); K06224 hedgehog
A dpo:Dpse_GA18321(472)	408	46.6	1167	(1056)	3		K06224	HH	[SP:Q29AA9_DROPS] GA18321 gene product from transcript GA18321-RA ; K06224 hedgehog
A bta:286821(188)	187	89.8	1138	(557)	5		K06224	HH	SHH; sonic hedgehog homolog (Drosophila) ; K06224 hedgehog
A nve:NEMVE_v1g241466(401)	360	45.3	1047	(287)	13		K06224	HH	[SP:A7RY14_NEMVE] NEMVEDRAFT_v1g241466; predicted protein; K06224 hedgehog
A cel:T05C12.10(1169)	193	30.6	279	(21)	14				[SP:Q21535_CAEEL] qua-1; QUAhog (hedgehog related)

Color coding of BBH: red for the alignment with 80% or more overlap; yellow otherwise

- GFIT table is a summary of best hit relations for individual genes in a genome generated from the KEGG SSDB database. It has been used to manually assign K numbers (KO identifiers).
- KOALA processes this table and automatically assigns K numbers when considered safe to do so.



KOALA: KEGG Orthology And Links Annotation

Annotation of K numbers and links through K numbers (EC numbers, GO terms, etc.)

GRP	ORG	KEGG ID	KO (KOALA)		KO (GENES)	Orth (GENES)	OC	Score	Memo	Ann
E.Ani	hsa	hsa:5664(448)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	18611		A
E.Ani	ptr	ptr:457790(442)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	16345		A
E.Ani	mcc	mcc:698770(654)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	10464		A
E.Ani	mmu	mmu:19165(448)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	17031		A
E.Ani	rno	rno:81751(448)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	16204		A
E.Ani	cfa	cfa:490382(717)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	7851		A
E.Ani	bta	bta:282010(449)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	16358		A
E.Ani	ssc	ssc:780410(448)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	18558		A
E.Ani	mdo	mdo:100026453(426)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	15132		A
E.Ani	oaa	oaa:100076299(455)	K04522	>			62906/60206 /Animals.27398	14136		
E.Ani	gga	gga:374188(451)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	10574		A
E.Ani	xla	xla:397713(449)	K04522		K04522	PSEN2	62906/60206 /Animals.27398	11327		A

KEGG ID	KO	Orth	Definition
hsa:5663	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
hsa:5664	K04522	PSEN2	presenilin 2 (Alzheimer di
ptr:457790	K04522	PSEN2	presenilin 2 (Alzheimer di
mcc:698770	K04522	PSEN2	presenilin 2
mmu:19164	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
mmu:19165	K04522	PSEN2	presenilin 2
rno:29192	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
rno:81751	K04522	PSEN2	presenilin 2
cfa:403408	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
cfa:490382	K04522	PSEN2	presenilin 2 (Alzheimer di
bta:282010	K04522	PSEN2	presenilin 2 (Alzheimer di
bta:282705	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
ssc:780410	K04522	PSEN2	presenilin 2
ssc:780411	K04505	PSEN1	presenilin 1 (EC:3.4.23.-)
mdo:100019628	K04505	PSEN1	similar to presenilin 1
mdo:100026453	K04522	PSEN2	similar to presenilin 2
oaa:100076299			similar to presenilin 2 (Alz
oaa:100091250			similar to presenilin 1
gga:373977	K04505	PSEN1	presenilin 1 (Alzheimer di
gga:374188	K04522	PSEN2	presenilin 2 (Alzheimer di
xla:397713	K04522	PSEN2	presenilin-beta
xla:399258	K04505	PSEN1	presenilin-alpha

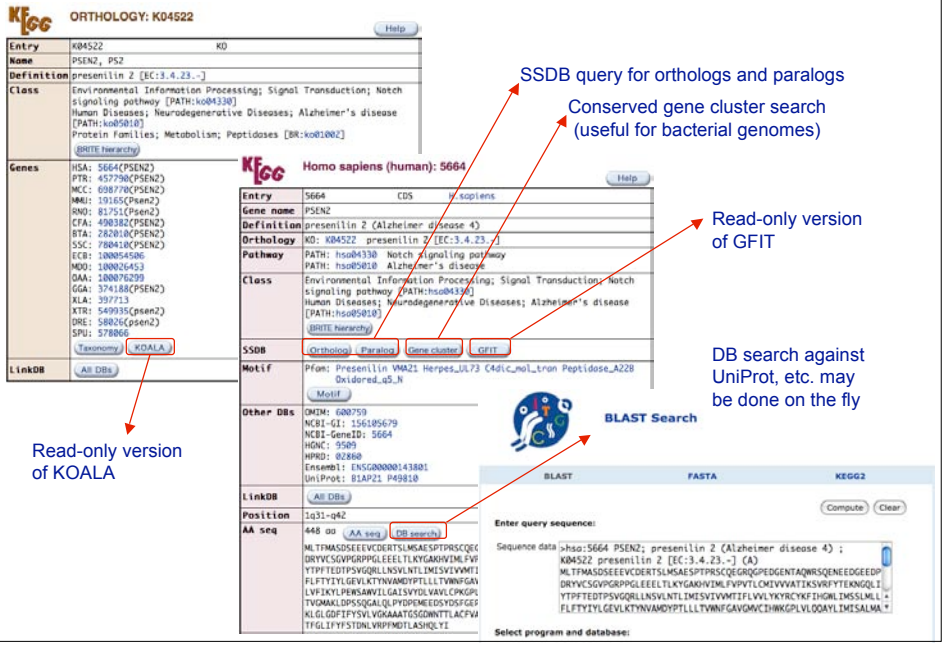
GFIT link

KOALA's suggestion

Current assignment

OC links

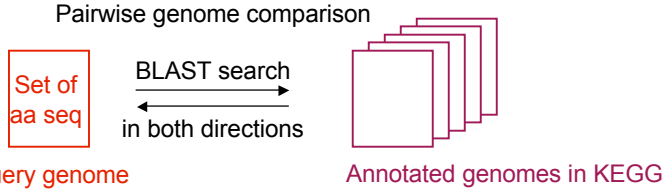
Public version of KEGG annotation tools



KAAS: KEGG Automatic Annotation Server

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

Pairwise genome comparison

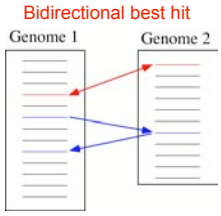


BLAST result screening by bi-directional best hit rate (*BHR*)

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

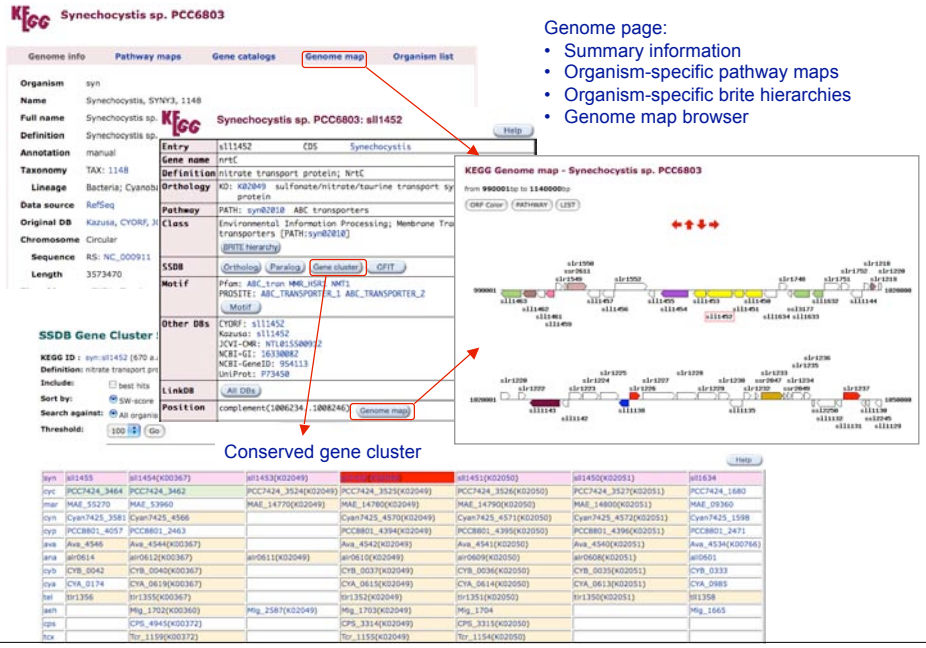
K number assignment by a heuristic scoring

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



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).

Genome Map Tools



Genome page:

- Summary information
- Organism-specific pathway maps
- Organism-specific brite hierarchies
- Genome map browser

KEGG PATHWAY and BRITE: Reference knowledge base

Data objects for computer representation of molecular systems

Element
gene, protein, small molecule, etc.

Pair (binary relation)
protein-protein interaction, drug-target relationship, etc.

Graph (wiring diagram)
pathway, complex, etc.

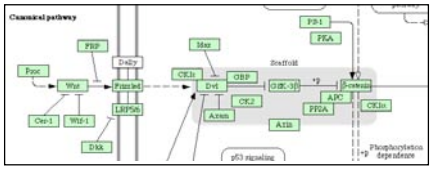
Simple list (membership)
pathway, complex, etc.

Hierarchical list
hierarchical classification, ontology, etc.

Knowledge Representation of Systemic Functions

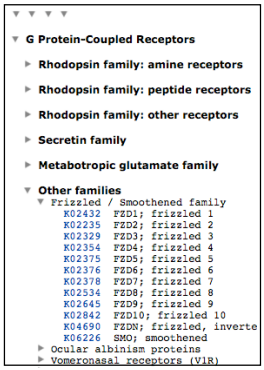
Graph (pathway map)

KEGG PATHWAY



Hierarchical list (ontology)

KEGG BRITE



Simple list (membership)

KEGG MODULE

List of molecules that constitute a pathway, a complex, etc.

KEGG DISEASE

List of disease genes, environmental factors, markers, drugs, etc.

Data source: textbooks, review articles, other publications, specialists' websites

KEGG PATHWAY Database

Collection of KEGG pathway maps

Global Map
- Metabolism Map (1)

Metabolism
- Carbohydrate Metabolism (16)
- Energy Metabolism (8)
- Lipid Metabolism (16)
- Nucleotide Metabolism (2)
- Amino Acid Metabolism (16)
- Metabolism of Other Amino Acids (9)
- Glycan Biosynthesis and Metabolism (15)
- Biosynthesis of Polyketides and Nonribosomal Peptides (9)
- Metabolism of Cofactors and Vitamins (12)
- Biosynthesis of Secondary Metabolites (27)
- Xenobiotics Biodegradation and Metabolism (26)
- Overview (9)

Genetic Information Processing
- Transcription (2)
- Translation (2)
- Folding, Sorting and Degradation (8)
- Replication and Repair (6)

Environmental Information Processing
- Membrane Transport (2)
- Signal Transduction (14)
- Signaling Molecules and Interaction (4)

Cellular Processes
- Cell Motility (3)
- Cell Growth and Death (4)
- Cell Communication (5)
- Circulatory System (2)
- Endocrine System (7)
- Immune System (9)
- Nervous System (2)
- Sensory System (2)
- Development (2)
- Behavior (3)

Human Diseases
- Cancers (15)
- Immune Disorders (6)
- Neurodegenerative Diseases (4)
- Metabolic Disorders (3)
- Infectious Diseases (4)

Drug Development
- Chronology: Antibiotics (8)
- Chronology: Antineoplastics (5)
- Chronology: Nervous System Agents (9)
- Chronology: Other Drugs (12)
- Target Based Structure Classification (12)
- Skeleton Based Structure Classification (8)

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

http://www.genome.jp/kegg-bin/get_htext?br08901.keg

KEGG BRITE Database

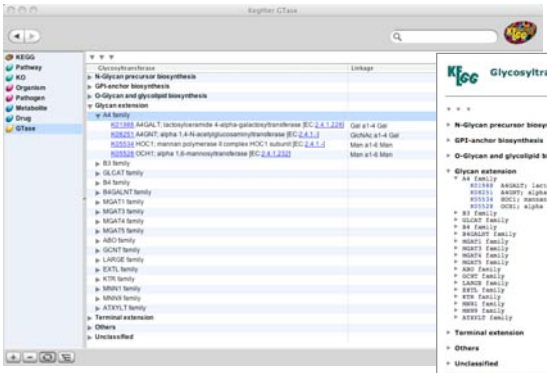
Collection of BRITE functional hierarchies

Genes and Proteins
- Network hierarchy (4)
- Protein families: metabolism (7)
- Protein families: genetic information processing (9)
- Protein families: environmental information processing (5)
- Protein families: cellular processes (7)

Compounds and Reactions
- Compounds (5)
- Reactions (3)
- Compound interactions (6)

Drugs and Diseases
- Drugs (15)
- Diseases (5)

Cells and Organisms
- Organisms (4)

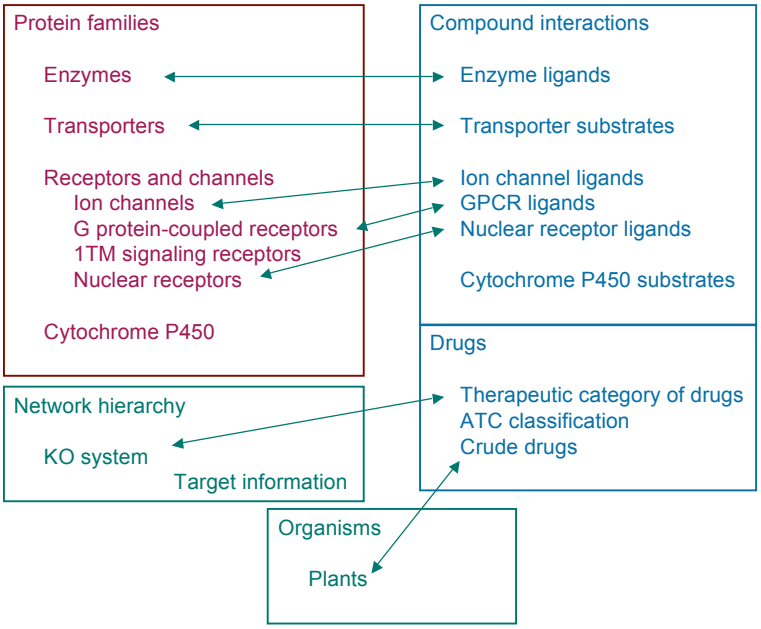


KegHier browser

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

http://www.genome.jp/kegg-bin/get_htext?br08902.keg

Interaction/relation data linking BRITE hierarchies



Pathway mapping and BRITE mapping

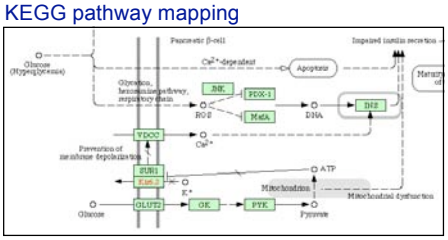
Linking genomes to biological systems and the environment

KO assignment

Gene	KO
10000	K04456
10003	K01301
10004	K01301
10005	K01068
10007	K02564
10008	K04897
10020	K00885
10020	K01791
10021	K04957
10023	K03069
10026	K05290
10038	K00774
10039	K00774
10050	K08193
10051	K06675
10052	K07616
10056	K01890
10057	K05668
10058	K05661
10059	K01528
.....	

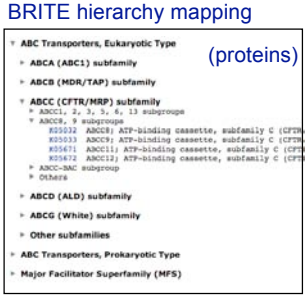
Genes to K numbers

KEGG pathway mapping

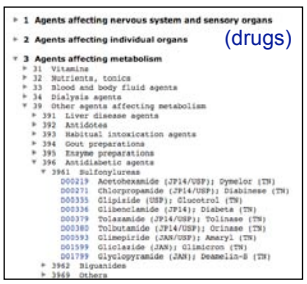


Disease pathway

BRITE hierarchy mapping



(proteins)

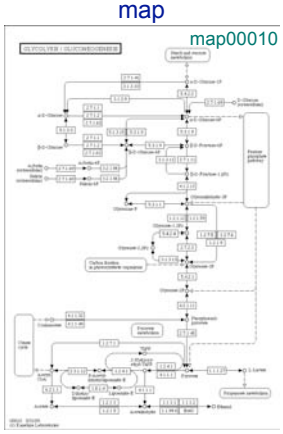


(drugs)

K numbers to C/D/G/R numbers

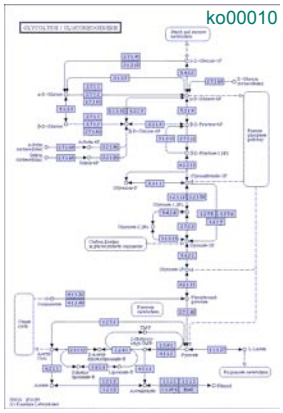
Convention of map number prefix

map



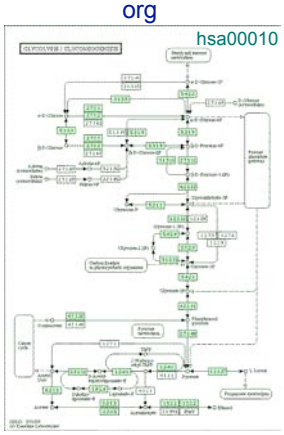
Each box is manually associated with KO identifier (K number), EC number, and reaction identifier (R number)

ko / ec / rn



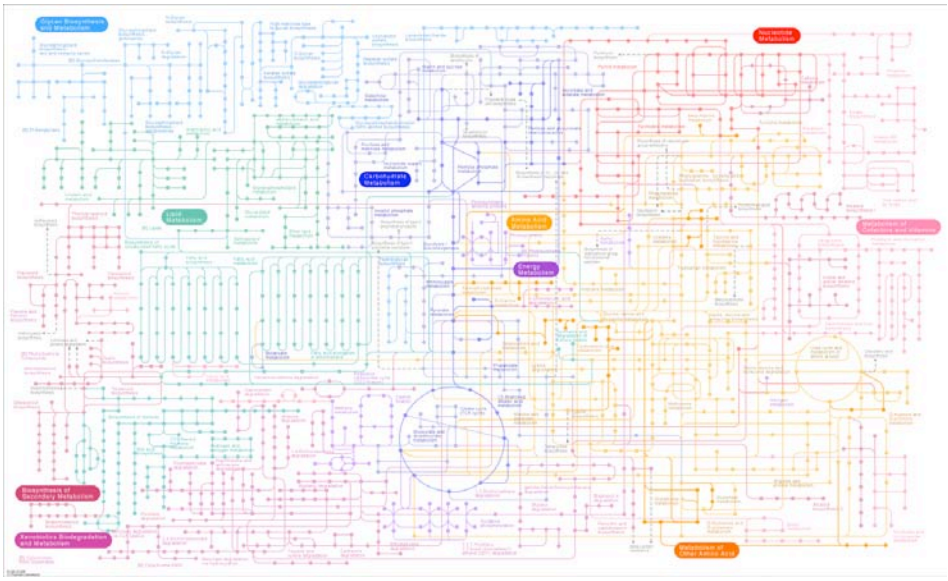
Blue boxes represent selection of K numbers, EC numbers, or R numbers

org

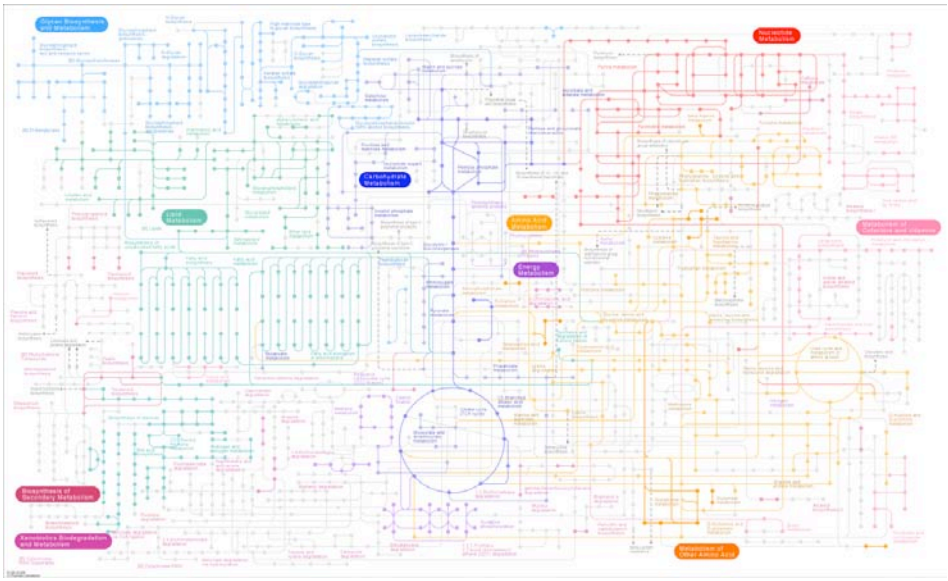


Green boxes correspond to gene identifiers in an organism that are computationally converted from K numbers

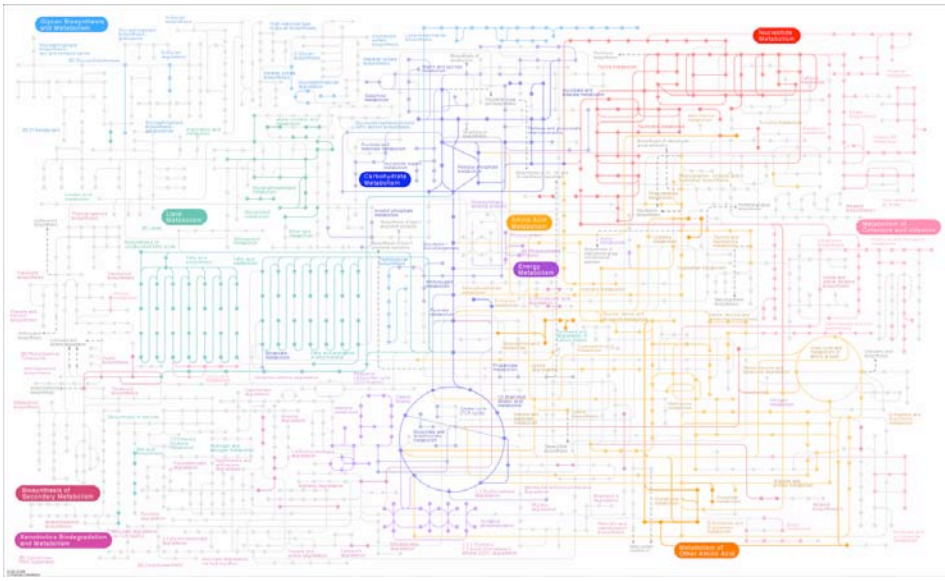
Global metabolism map - Reference pathway



Global metabolism map - Homo sapiens

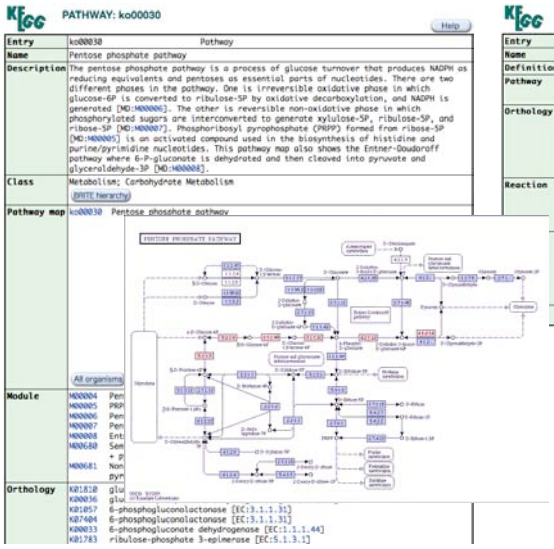


Global metabolism map - Escherichia coli



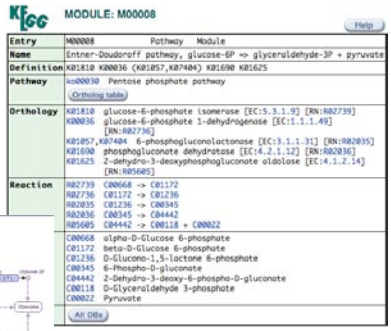
Pathway Entry

Text-based representation of KEGG pathway maps



Pathway MODULE

Tighter functional units in KEGG pathway maps



KEGG-NCBI Collaboration

KEGG	NCBI
GENES	RefSeq (Gene)
PATHWAY	BioSystems
COMPOUND	PubChem
DRUG	PubChem

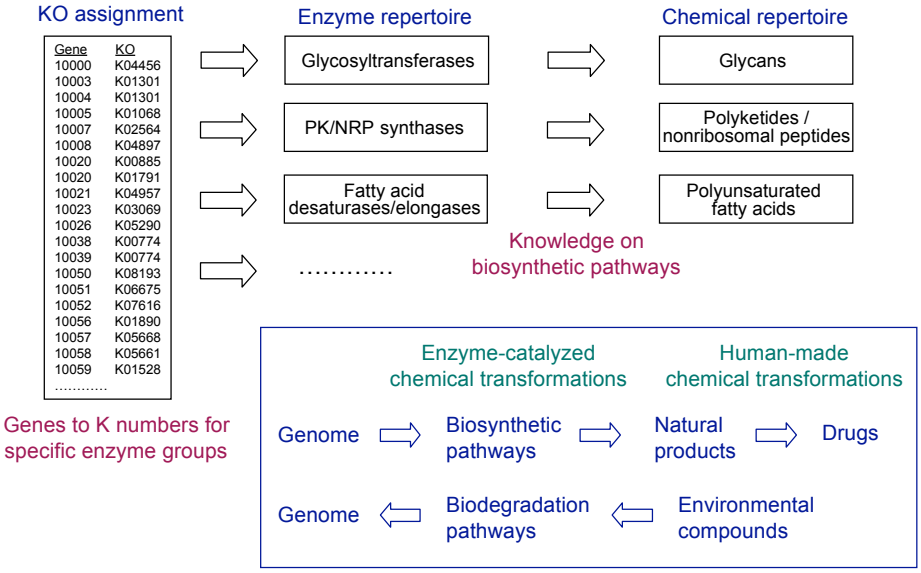
KEGG LIGAND
and chemical bioinformatics

Bioinformatics for Small Molecules

1. Chemical structure similarity
   - Comparison of bit-represented vectors (fingerprints)
   - Comparison of graph objects
2. Chemical building blocks
   - Conserved substructures as building blocks of compounds/drugs
   - Variable substructures as building blocks of reactivity/efficacy
3. Network modules
   - Genomic module, e.g. operon
   - Chemical module, e.g. overall reaction
4. Predictive methods
   - Interaction prediction, e.g. toxicity
   - Reaction prediction, e.g. metabolic fate
5. Examples
   - Plant/fungi/bacterial genomes and secondary metabolites via biosynthetic pathways
   - Bacterial genomes and environmental compounds via biodegradation pathways

Chemical structure prediction

Linking genomes to small molecules



Linking genomes to chemical structures

Glycans

- Hashimoto, K., Goto, S., Kawano, S., Aoki-Kinoshita, K.F., Ueda, N., Hamajima, M., Kawasaki, T., and Kanehisa, M.; KEGG as a glycome informatics resource. *Glycobiology* 16, 63R-70R (2006).
- Kawano, S., Hashimoto, K., Miyama, T., Goto, S., and Kanehisa, M.; Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* 21, 3976-3982 (2005).
- 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).

Polyketides & nonribosomal peptides

- Minowa, Y., Araki, M., and Kanehisa, M.; Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.* 368, 1500-1517 (2007).

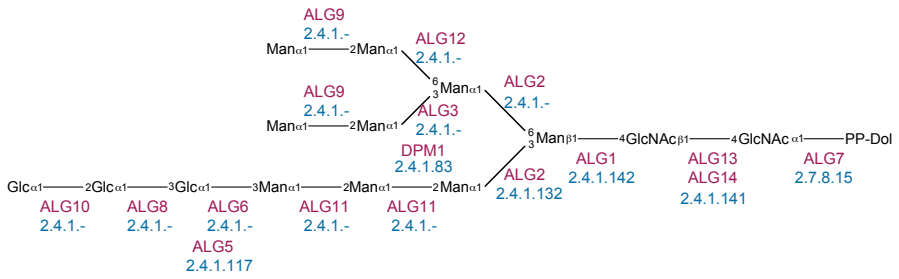
Polyunsaturated fatty acids

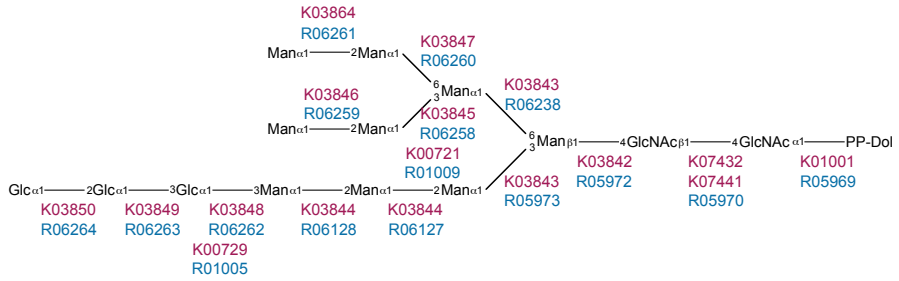
- Hashimoto, K., Yoshizawa, A.C., Okuda, S., Kuma, K., Goto, S., and Kanehisa, M.; The repertoire of desaturases and elongases reveals fatty acid variations in 56 eukaryotic genomes. *J. Lipid Res.* 49, 183-191 (2008).

Environmental compounds

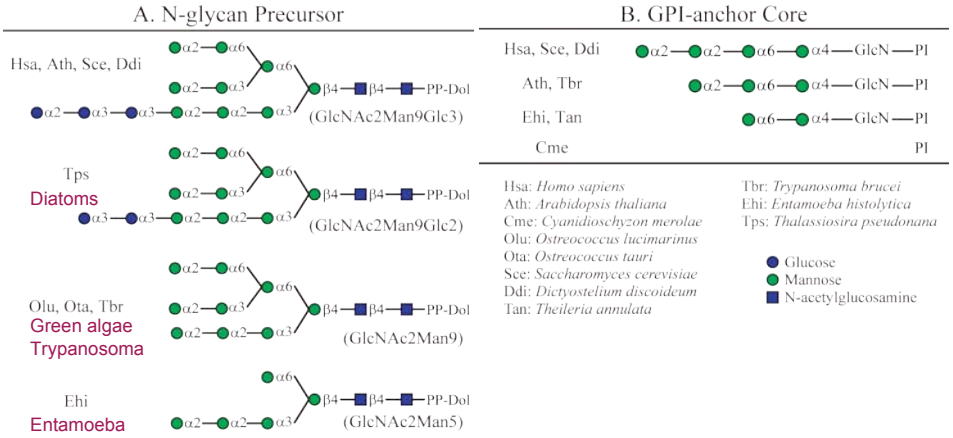
- Oh, M., Yamada, T., Hattori, M., Goto, S., and Kanehisa, M.; Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Info. Model.* 47, 1702-1712 (2007).

N-Glycan Precursor Structure



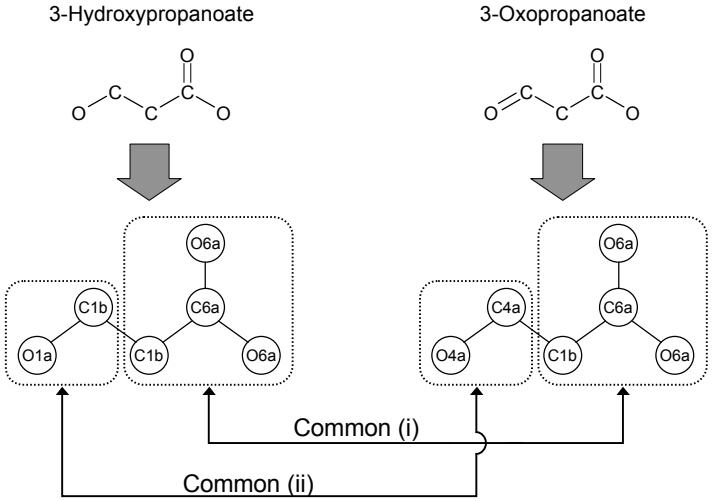


Truncated glycan structures in parasitic protists and algae predicted from genomic information



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).

Chemical structure comparison based on atom typing



Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M.; Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.* 125, 11853-11865 (2003).

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	

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	

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	

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	

Reactant pairs extracted from enzymatic reactions

EC1. Oxidoreductases

A.red + C.ox ⇔ A.ox + C.red

EC2. Transferases

AB + C ⇔ A + BC

EC3. Hydrolases

AB + H_2O ⇔ AH + BOH

EC4. Lyases

AB ⇔ A + B

EC5. Isomerases

A ⇔ A'

EC6. Ligases

A + B + ATP ⇔ AB + AMP + PP_i

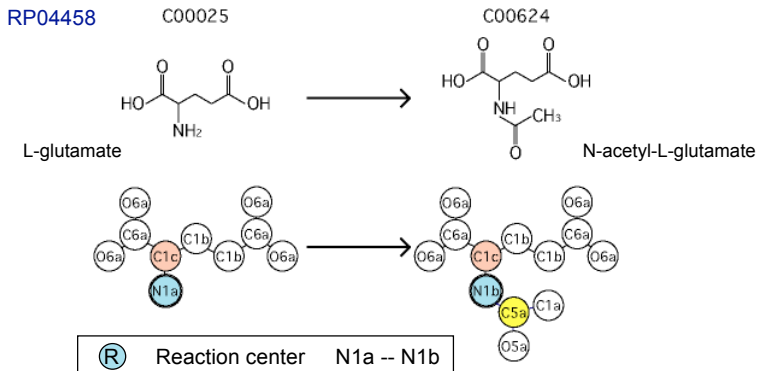
EC (Enzyme Commission) number

- EC numbers represent classification of enzymatic reactions (and enzymes)
- Manually assigned by IUBMB-IUPAC Biochemical Nomenclature Committee
- Published experimental evidence on the enzyme and the reaction is required
- Four numbers separated by periods, representing reaction specificity (class, sub-class, and sub-subclass numbers) and substrate specificity (serial number).

Reactant pair transformation patterns (RDM patterns)

R00259 C00024 + C00025 <=> C00010 + C00624 [EC:2.3.1.1]
Acetyl-CoA + L-Glutamate <=> CoA + N-Acetyl-L-glutamate

RP04458

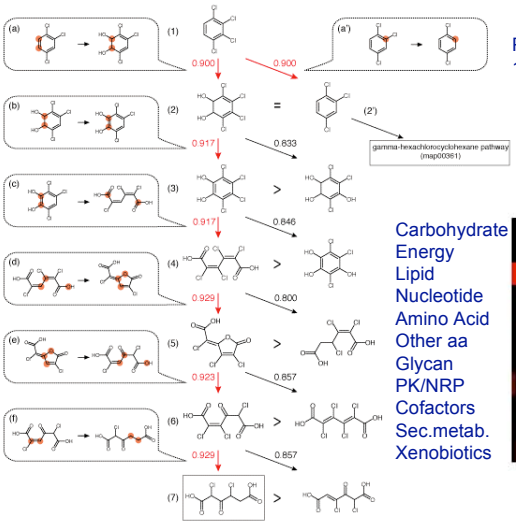


R	Reaction center	N1a -- N1b
D	Difference atom	(H) -- C5a
M	Matched atom	C1c -- C1c

RDM pattern

Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M.; Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* 126, 16487-16498 (2004).

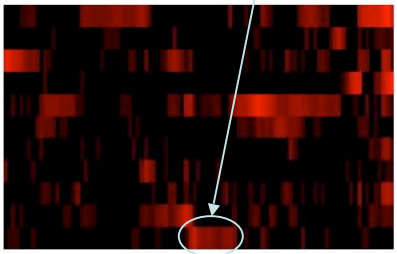
Prediction of xenobiotics degradation pathways



Predicted pathways for 1,2,3,4-tetrachlorobenzene

RDM patterns for xenobiotics biodegradation pathways

Carbohydrate
Energy
Lipid
Nucleotide
Amino Acid
Other aa
Glycan
PK/NRP
Cofactors
Sec.metab.
Xenobiotics

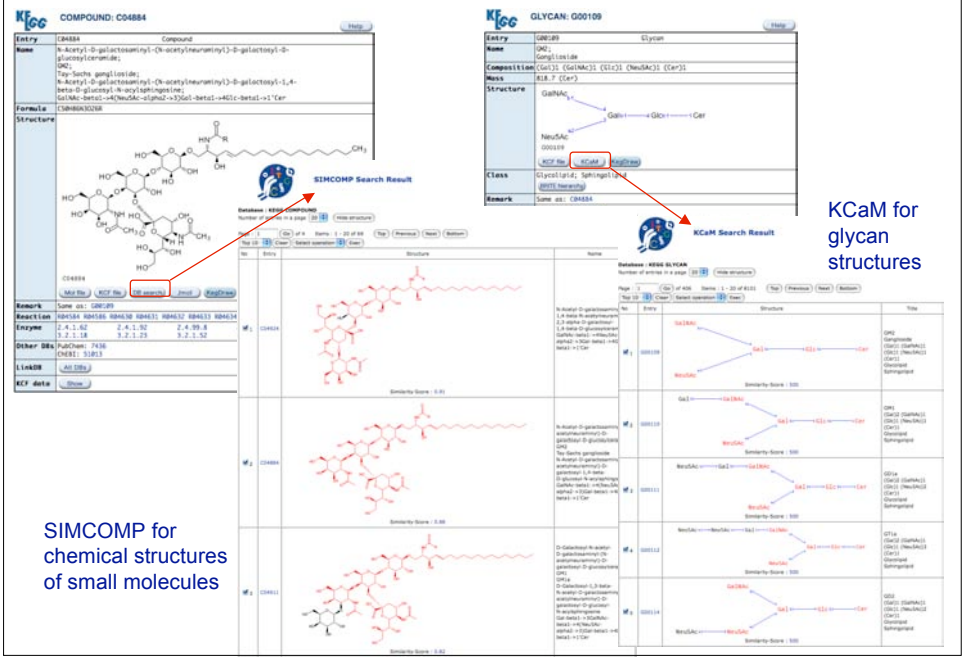


RDM patterns

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).

Structural Similarity Search



KCaM for glycan structures

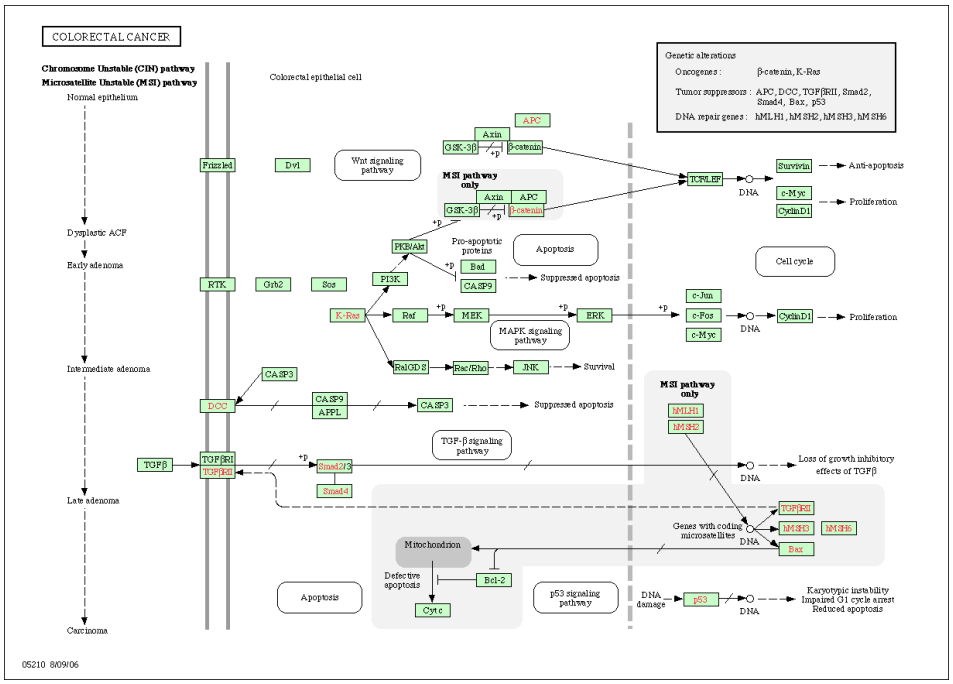
SIMCOMP for chemical structures of small molecules

KEGG DISEASE and DRUG: An ultimate knowledge base

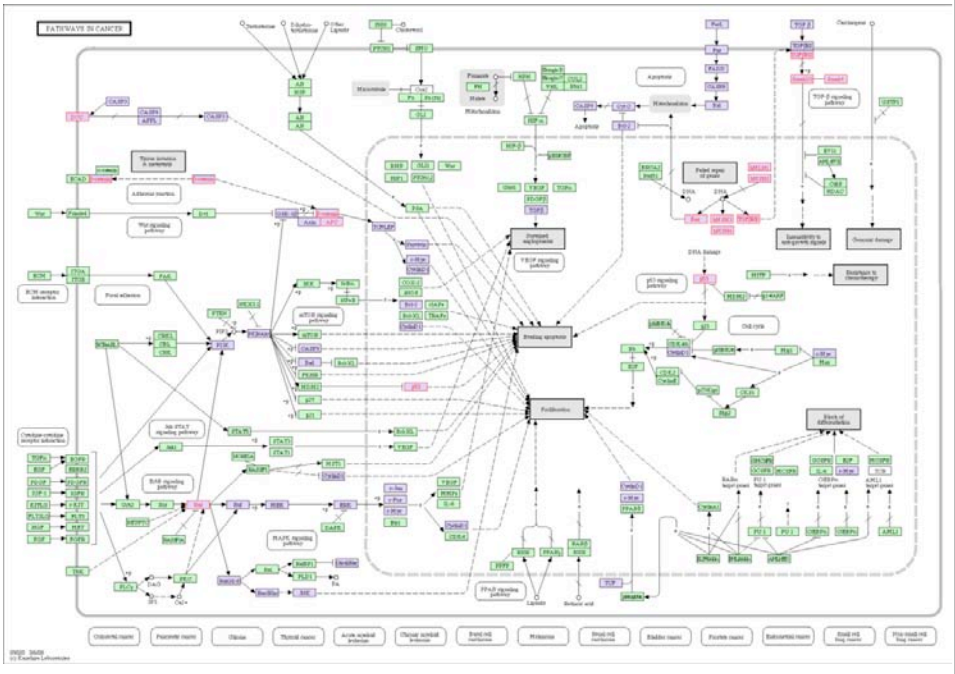
Disease and Drug Information in KEGG

	KEGG DISEASE	KEGG DRUG
URL	http://www.genome.jp/kegg/disease/	http://www.genome.jp/kegg/drug/
Content	Lists of disease genes and molecular factors	Chemical structure based collection of all approved drugs in Japan, USA, and Europe
Pathway	KEGG pathway maps for human diseases: cancers, immune disorders, neurodegenerative diseases, metabolic disorders, and infectious diseases	KEGG DRUG structure maps for drug development
BRITE hierarchy	Disease classifications including: Pathogens and infectious diseases Human diseases ICD-10 disease classification	Drug classifications including: Therapeutic category of drugs (Japan) USP drug classification (USA) ATC classification (WHO) TCM (Traditional Chinese Medicine) drugs in Japan Crude drugs

Colorectal cancer



Colorectal cancer in the global cancer map



KEGG DISEASE

H00004
Chronic myeloid leukemia (CML)

Disease gene
- BCR-ABL (translocation)
- EVI1 (overexpression)
- AML1 (translocation)
- p16/INK4A (mutation)
- p53 (mutation)
- RB1 (mutation)

Carcinogen
- 1,3-Butadiene
- Rubber industry

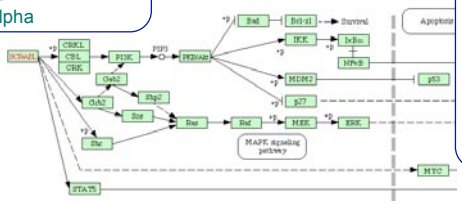
Marker
- BCR-ABL (translocation)
- WT1

Drug
- Imatinib mesylate (Gleevec)
- Hydroxyurea
- Interferon-alpha

Cancers

55 DISEASE entries
14 pathway maps
1 global map

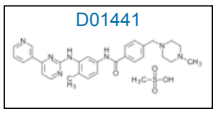
KEGG PATHWAY



KEGG DRUG

Chemical structure

D01441



Generic name

Imatinib mesilate (JAN)
Imatinib mesylate (NF)

Trade name

Gleevec (TN)
Glivec (TN)
……

Efficacy

Therapeutic categories as BRITE hierarchies

Target

Target information as a context of KEGG pathways

Drug structure map

Chemical modifications in drug development

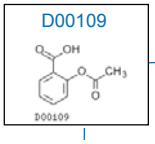
KEGG DRUG

KEGG DRUG

Chemical structure based information resource for all approved drugs

Chemical structure

D00109



Generic names

Aspirin (JP15/USP)

Trade names

Aspirin Bayer (TN)
……

Efficacy,
Target information,
Drug classification,
History of drug development,
……

KEGG DRUG

CAS: 50-78-2
PubChem: 4594
ChEBI: 15365

Research community

Number of entries	
KEGG DRUG:	8,912
Crude drugs	263
TCM formulas:	229
JAPIC:	12,763
JAPIC-OTC:	12,080

(As of June 12, 2009)

JAPIC code

00051357
……

JAPIC

Package insert information, including side effects, drug interactions, etc.

Approval number

21700AMY00243
……

Regulatory agency

JAN code (barcode)

4987222735865
……

Industry

YJ code

1143001X1228
……

Doctors
Pharmacists
Patients
Consumers

Herbal Medicine in KEGG

Traditional Chinese Medicine in Japan

- **Crude Drugs**
  - Drugs for relieving the exterior (diaphoretics)
    - Diaphoretics pungent in flavor and warm in property
    - Diaphoretics pungent in flavor and cool in property
      - D06693 Pueraria root (JP15); Pueraria root (TN)
      - D06794 Chrysanthemum flower (JP15)
      - D06723 Burdock fruit (JP15)
      - D06727 Bupleurum root (JP15); Bupleurum root (TN)
      - D06745 Cimicifuga rhizome (JP15); Cimicifuga rhizome
      - D05431 Peppermint (JP15/NF); Peppermint (TN)
  - Drugs for clearing heat
  - Purgative drugs
  - Drugs for dispeling wind-damp (antirheumatics)
  - Aromatic drugs for resolving dampness
  - Drugs for inducing diuresis and excreting dampness (diuret
  - Drugs for warming the interior
  - Drugs for regulating Qi
  - Drugs for removing food stagnation (digestants)
  - Drugs for expelling parasites (anthelmintics)
  - Hemostatic drugs
  - Drugs for promoting blood circulation and removing blood stasis
  - Expectorant, antitussive and anti-asthmatic drugs
  - Drugs for calming the spirit (tranquilizers)
  - Drugs for calming the liver and extinguishing wind
  - Drugs for resuscitation
  - Drugs for supplementing efficiency (tonics)
  - Astringent drugs
  - Emetic drugs
  - Drugs for external use
- **Formulas**
  - Formulas for relieving the exterior
    - Formulas for relieving the exterior with pungent warmnes
      - D06698 Kakkonto extract (JP15); Kakkonto
      - D06928 Kakkontokasenkyushin'i
      - D06939 Keishikaogito
      - D06940 Keishikakakkonto
      - D06941 Keishikakobokukyoninto
      - D06944 Keishikajutsubuto
      - D06947 Keishito
      - D06953 Keimakakuhanto
      - D06954 Kososan
      - D06987 Shoseiryuto
      - D07042 Maoto
      - D07045 Makyoyokukanto
    - Formulas for relieving the exterior with pungent coolness
    - Formulas for relieving the exterior by supporting the normal
  - Purgative formulas
  - Formulas for mediation

Crud drug: combination of compounds

Kakkon (Pueraria root)

DRUG: D06693

Entry	D06693 Crude Drug
Name	Pueraria root (JP15); Pueraria root (TN)
Component	Starch [DR:D06507] [CPD:C00369], Daidzin [CPD:C10216], Daidzein [CPD:C10208], Puerarin [CPD:C10524], Puerarinxyloside, Methyl palmitate, Dimethyl suberate
Source	Pueraria lobata [TAX:3893]
Remark	Therapeutic category: 5100 (BRITE hierarchy)
Comment	Pueraria root
Other DBs	PubChem: 47208344
LinkDB	All DBs

Biosynthetic pathways

Plant genomes

Formula: combination of crude drugs

Kakkonto

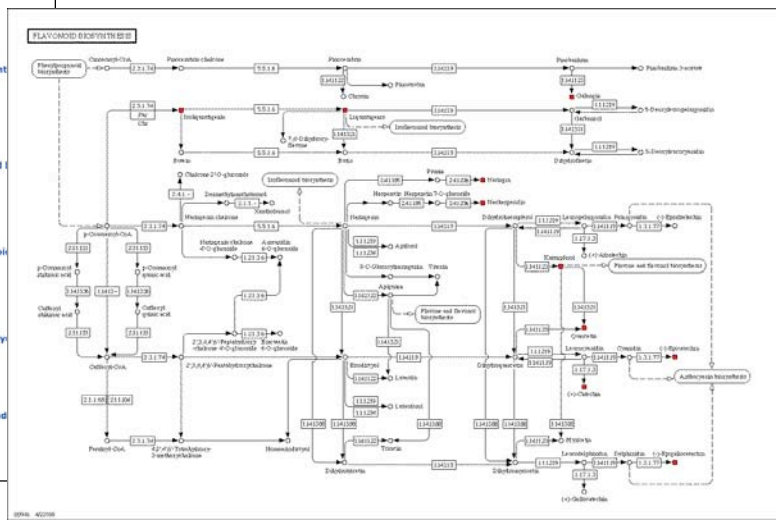
DRUG: D06698

Entry	D06698 Formula Drug
Name	Kakkonto extract (JP15); Kakkonto
Component	Pueraria root [DR:D06693], Ephedra herb [DR:D06791], Jujube [DR:D06758], Peony root [DR:D06739 D06850], Glycyrrhiza [DR:D04365 D06828], Ginger [DR:D06744 D06852], Cinnamon bark [DR:D06712 D06841] or Cinnamon quassia [DR:D06898]
Activity	Cold; Coryza
Remark	Therapeutic category: 5200 (BRITE hierarchy)
Other DBs	PubChem: 47208349
LinkDB	All DBs

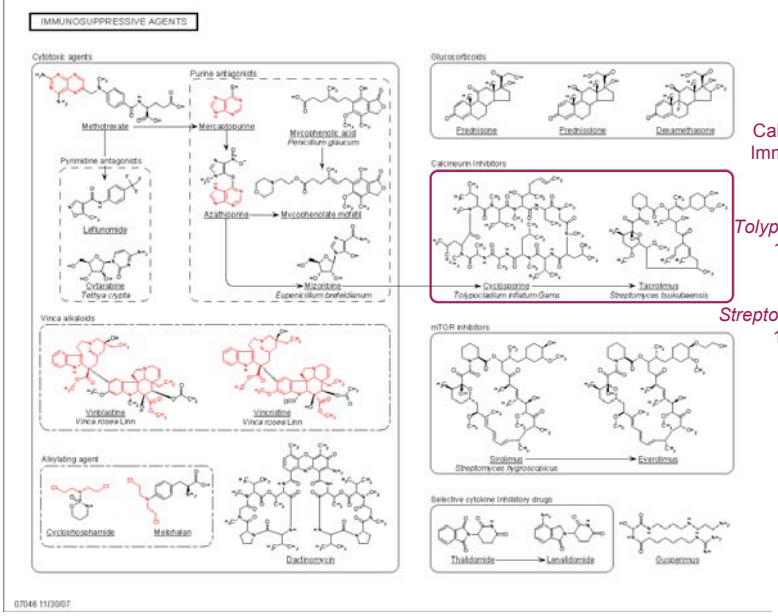
KEGG pathway mapping of chemical components in herbal medicine

Pathway Search Result

- map00941 Flavonoid biosynthesis
  - C00389
  - C05903
  - C06592
  - C08650
  - C09727
  - C09742
  - C09789
  - C09806
  - C10044
  - C12136
- map00950 Alkaloid biosynt
  - C00737
  - C01795
  - C02105
  - C02890
  - C04118
  - C05176
  - C05189
  - C05243
  - C05315
- map00902 Monoterpenoid
  - C00400
  - C00843
  - C06074
  - C06076
  - C09844
  - C09893
  - C11951
  - C11952
- map00940 Phenylpropanoi
  - C00811
  - C00852
  - C00903
  - C01494
  - C01527
  - C01533
  - C01752
  - C05451
- map00061 Fatty acid biosy
  - C00249
  - C00712
  - C01530
  - C01571
  - C02679
  - C04423
  - C08362
- map00281 Geraniol degrad
  - C01499
  - C01500
  - C09847
  - C09848
  - C09871
  - C11386



KEGG DRUG structure maps for the history of drug development



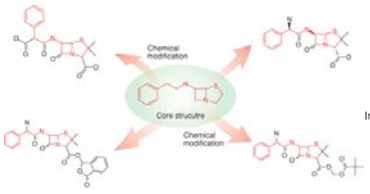
Calcineurin inhibitor
Immunosuppressant

Ciclosporin
Tolypocladium inflatum
1972 Sandoz

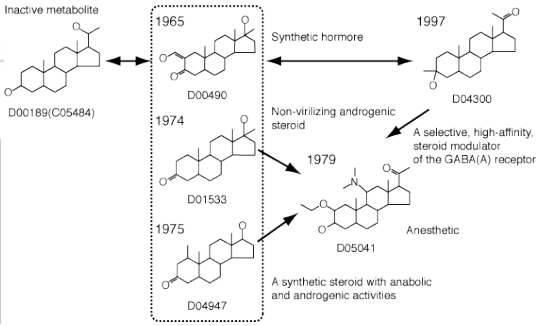
Tacrolimus
Streptomyces tsukubaensis
1984 Fujisawa

Chemical modification patterns in drug development

Towards understanding chemical architecture of marketed drugs



Reconstructed drug structure map



3,745 drugs from KEGG DRUG
236 core structures
506 peripheral fragments

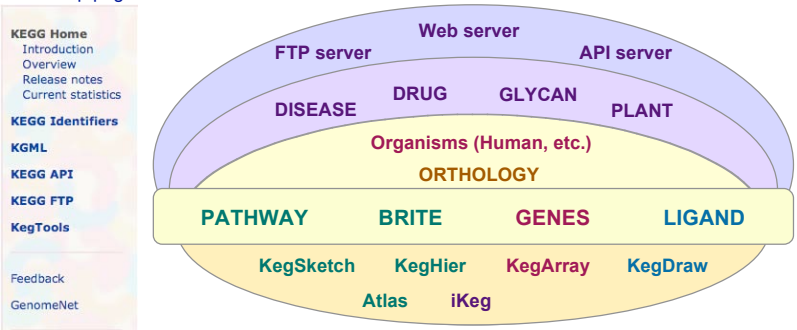
255 drug pairs from 28 maps
204 modification patterns

Shigemizu, D., Araki, M., Okuda, S., Goto, S., and Kanehisa, M.; Extraction and analysis of chemical modification patterns in drug development. *J. Chem. Info. Model.* 49, 1122-1129 (2009).

Summary

KEGG Services

KEGG top page



KEGG2 page

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

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

A new XML version of KEGG pathway maps will be available in both KGML and BioPAX Level 2 formats.

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

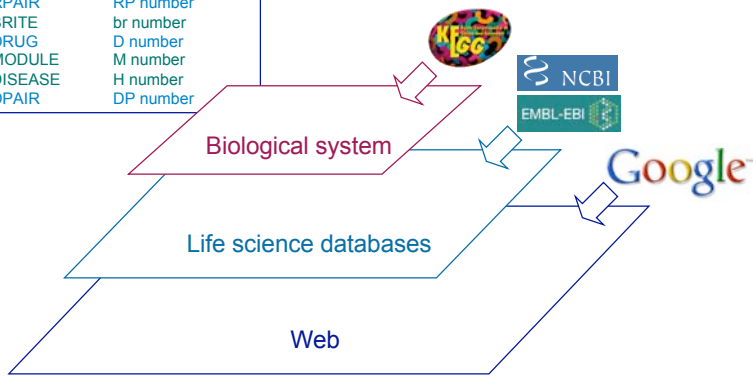
REST http://www.kegg.jp/entry/keggid
http://www.kegg.jp/pathway/keggid
http://www.kegg.jp/brite/keggid

KEGG Integrated with Other Resources

Namespace for KEGG objects

Release	Database	Object identifier
1995	KEGG PATHWAY	map number
	KEGG GENES	locus_tag / GeneID
	KEGG ENZYME	EC number
	KEGG COMPOUND	C number
2000	KEGG GENOME	organism code
2001	KEGG REACTION	R number
2002	KEGG ORTHOLOGY	K number
2003	KEGG GLYCAN	G number
2004	KEGG RPAIR	RP number
2005	KEGG BRITE	br number
	KEGG DRUG	D number
2007	KEGG MODULE	M number
	KEGG DISEASE	H number
2009	KEGG DPAIR	DP number

- KEGG is a computer representation of the biological systems.
- KEGG objects (database entries) are highly integrated.
- 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.



KEGG for biological interpretation of large-scale datasets

Genes and proteins

Genome sequences
Metagenomics data
EST data
Gene expression data
etc.

Chemical substances

Metabolomics data
Glycomics data
etc.

Large-scale data sets

Identical K numbers

Pathway mapping
BRITE mapping
Structure mapping

Identical C/D/G numbers

Structural similarity

Mapping as a set operation

KEGG PATHWAY
Pathway maps
Global maps
Structure maps

KEGG BRITE
Functional hierarchies

KEGG MODULE
Gene/molecule lists

KEGG DISEASE
Gene/molecule lists

Reference knowledge base of systemic biological functions

Binary relations

Protein-protein interactions
Protein-small molecule interactions
etc.

+

KEGG PATHWAY
Pathway maps
Global maps

Extended interaction network

Changing roles of bioinformatics: molecules to molecular systems

More divergent datasets
- Biopolymers: DNA, RNA, protein, glycan, lipid, polyketide, nonribosomal peptide, etc.
- Small molecules: metabolite, environmental compound, drug, etc.
- Molecular systems: pathway, complex, ontology, etc.

More complex data types
- Biopolymers: linear sequence, circular sequence, tree structure, 3D structure
- Small molecules: chemical structure, 3D structure
- Molecular systems: graph, tree, list, etc.

Hierarchy of data
- atom - molecule - molecular system
- monomer - biopolymer
- molecule - molecular interaction - molecular system
- cell - cellular interaction - cellular system

Changing roles of bioinformatics: basic research to practical values

Diseases viewed as perturbed states of molecular systems

- Capturing knowledge on molecular systems both in normal and perturbed states ⇨ KEGG PATHWAY, KEGG DISEASE
- Capturing knowledge on drugs as perturbants to molecular systems ⇨ KEGG DRUG
- Capturing knowledge on natural product biosynthetic pathways and human-made chemical structure transformations ⇨ KEGG RPAIR, KEGG DPAIR
- Generating knowledge from genome sequencing and other high-throughput data ⇨ KEGG ORTHOLOGY
- Knowledge based analysis of human diseases
- Drug discovery from the genomes of plants and microorganisms