Background

• High-throughput experiments and large-scale datasets
• Bioinformatics approaches to integration and interpretation
Gene ⇔ Genome ⇔ Genomics

↓

Post-genome

Biology ⇔ Bioinformatics

Informatics

Molecular Building Blocks of Life
Genomic and Chemical Spaces

Genomic Space

<table>
<thead>
<tr>
<th>DNA (Gene)</th>
<th>Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>Transcriptome</td>
</tr>
<tr>
<td>Protein</td>
<td>Proteome</td>
</tr>
</tbody>
</table>

Central dogma

DNA ⇔ RNA ⇔ Protein

Genetic code

Replication
Transcription
Translation

Chemical Space

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Metabolome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycan</td>
<td>Glycome</td>
</tr>
<tr>
<td>Lipid</td>
<td>Lipidome</td>
</tr>
</tbody>
</table>

(Endogenous)

<table>
<thead>
<tr>
<th>Compound</th>
<th>Chemical genome</th>
</tr>
</thead>
</table>

(Exogenous)

Biosynthetic code

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

**Genomic space**
- Repertoire of genes and proteins in individual organisms (genomics, transcriptomics, proteomics, etc.)
- Repertoire of genes in environmental samples (metagenomics)
- Enzyme gene families (biosynthetic potential)
- Receptor gene families (drug targets)
- Human genome ↔ Gut microbiome
- Human ↔ Vector ↔ Pathogen

**Chemical space**
- Repertoire of endogenous molecules (metabolomics, glycomics, lipidomics, etc.)
- Repertoire of exogenous molecules (chemical genomics)
- Natural products ↔ Environmental compounds
- Drug leads

Bioinformatics Approaches to Reconstructing Biological Systems from Building Blocks

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

**Gene-Disease Associations**

**Single-gene diseases**

Genetic disorders with Mendelian inheritance patterns

- Disease gene ➔ Disease
germine mutation

**Multifactorial (polygenic) diseases**

Common diseases such as cancers, heart disease, diabetes, etc.

- Genetic factors ➔ Disease
- Environmental factors

Cancer causing factors:
- somatic mutation, translocation, overexpression, etc.
- carcinogen, viral infection, etc.
Network-Disease Associations

**Single-gene diseases**

Single genetic perturbation → Molecular network → Disease

germline mutation

**Multifactorial diseases**

Multiple genetic perturbations → Molecular network → Disease

Environmental perturbations

somatic/germline mutations, translocation, overexpression, etc.
environmental chemicals, infections, human microbiome, etc.

**Infectious diseases**

Environmental perturbations → Molecular network → Disease

pathogens

---

Growth of sequence and 3D structure databases

Complete Genomes in KEGG

Eukaryotes (131)
- Animals (47)
- Plants (11)
- Fungi (43)
- Protists (30)

Prokaryotes (1124)
- Bacteria (1041)
- Archaea (83)

As of June 12, 2010


A History of Life Science Databases

Compounds
- 1879 Index Medicus
- 1907 CAS
- 1964 Medlars

DNA Sequences
- 1968 Dayhoff Atlas
- 1971 Medline
- 1979 Los Alamos
- 1982 GenBank
- 1982 GenBank
- 1982 EMBL
- 1984 DDBJ
- 1984 PDB (BNL)

Biological Systems
- 1992 GenomeNet
- 1995 KEGG
- 2003 UniProt
- 2004 PubChem

Protein Sequences
- 1971 PDB
- 1980 NBRF
- 1984 PIR
- 1986 SwissProt

3D Structures
- 1994 EBI
- 1999 PDB (RCSB)
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

<table>
<thead>
<tr>
<th>Database</th>
<th>Content</th>
<th>Data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG PATHWAY</td>
<td>Pathway maps, reference (total)</td>
<td>361 (112,000)</td>
</tr>
<tr>
<td>KEGG BRITE</td>
<td>Functional hierarchies, reference (total)</td>
<td>89 (29,483)</td>
</tr>
<tr>
<td>KEGG MODULE</td>
<td>Pathway modules</td>
<td>537</td>
</tr>
<tr>
<td>KEGG DISEASE</td>
<td>Human diseases</td>
<td>374</td>
</tr>
<tr>
<td>KEGG DRUG</td>
<td>Drugs</td>
<td>9,454</td>
</tr>
<tr>
<td>KEGG ORTHOLOGY</td>
<td>KEGG Orthology (KO) groups</td>
<td>13,597</td>
</tr>
<tr>
<td>KEGG GENOME</td>
<td>KEGG Organisms, manual/koala + kaas</td>
<td>1,255 + 99</td>
</tr>
<tr>
<td>KEGG GENES</td>
<td>Genes in high-quality genomes</td>
<td>5,568,602</td>
</tr>
<tr>
<td>KEGG SSDB</td>
<td>(131 eukaryotes + 1041 bacteria + 83 archaea)</td>
<td>41,546,146,182</td>
</tr>
<tr>
<td>KEGG DGENES</td>
<td>Bi-directional best hit relations within GENES</td>
<td>710,431,505</td>
</tr>
<tr>
<td>KEGG EGEnES</td>
<td>Genes in draft genomes (15 eukaryotes)</td>
<td>284,078</td>
</tr>
<tr>
<td>KEGG COMPOUND</td>
<td>Genes as EST contigs (84 eukaryotes)</td>
<td>3,133,980</td>
</tr>
<tr>
<td>KEGG GLYCAN</td>
<td>Metabolites and other small molecules</td>
<td>16,250</td>
</tr>
<tr>
<td>KEGG REACTION</td>
<td>Biochemical reactions</td>
<td>8,173</td>
</tr>
<tr>
<td>KEGG RPAIR</td>
<td>Reactant pair chemical transformations</td>
<td>12,265</td>
</tr>
<tr>
<td>KEGG ENZYME</td>
<td>Enzyme nomenclature</td>
<td>5,184</td>
</tr>
</tbody>
</table>

As of June 12, 2010

# KEGG Object Identifiers

<table>
<thead>
<tr>
<th>Database</th>
<th>Prefix</th>
<th>Example</th>
<th>Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG PATHWAY</td>
<td>map/ko/ec/rm/(org)</td>
<td>hsa04930</td>
<td>1995</td>
</tr>
<tr>
<td>KEGG BRITE</td>
<td>br/jp/ko/(org)</td>
<td>ko01003</td>
<td>2005</td>
</tr>
<tr>
<td>KEGG MODULE</td>
<td>M</td>
<td>M00008</td>
<td>2007</td>
</tr>
<tr>
<td>KEGG DISEASE</td>
<td>H</td>
<td>H00004</td>
<td>2008</td>
</tr>
<tr>
<td>KEGG DRUG</td>
<td>D</td>
<td>D01441</td>
<td>2005</td>
</tr>
<tr>
<td>KEGG ORTHOLOGY</td>
<td>K</td>
<td>K04527</td>
<td>2002</td>
</tr>
<tr>
<td>KEGG GENOME</td>
<td>T</td>
<td>T01009 (hsa)</td>
<td>2000</td>
</tr>
<tr>
<td>KEGG COMPOUND</td>
<td>C</td>
<td>C00031</td>
<td>1995</td>
</tr>
<tr>
<td>KEGG GLYCAN</td>
<td>G</td>
<td>G00109</td>
<td>2003</td>
</tr>
<tr>
<td>KEGG REACTION</td>
<td>R</td>
<td>R00259</td>
<td>2001</td>
</tr>
<tr>
<td>KEGG RPAIR</td>
<td>RP</td>
<td>RP04458</td>
<td>2004</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Database</th>
<th>Prefix</th>
<th>Example</th>
<th>Release</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG GENES</td>
<td>org:gene</td>
<td>hsa:3643</td>
<td>1995</td>
</tr>
<tr>
<td>KEGG ENZYME</td>
<td>ec:number</td>
<td>ec:2.7.10.1</td>
<td>1995</td>
</tr>
<tr>
<td>DBGET databases</td>
<td>db:entry</td>
<td>sp:P06213</td>
<td>1994</td>
</tr>
</tbody>
</table>
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 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.

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 map for type II diabetes mellitus

hsa04930

hsa:3643

KEGG Orthology (KO) System

KO (K number) entry for insulin receptor

KO04527

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

Top page (KEGG)

This icon will bring back to the KEGG top page.

Table of Contents (KEGG2)

PATHWAY

DISEASE

ORTHOLOGY

GENES

LIGAND

These icons will bring back to the KEGG2 page.

DBGET
BLAST/FASTA, KAAS, SIMCOMP/SUBCOMP, KCaM, PathPred, etc.

This icon will bring back to the GenomeNet top page.

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

GECS - Gene expression to chemical structure

Chemical Analysis

SIMCOMP / SUBCOMP - Chemical structure search

KCaM - Glycan structure search

PathPred - Reaction pathway prediction

E-zyme - Enzymatic reaction prediction
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

Diagram:
- Common ancestor
- B (Human)
- C (Mouse)
- Common ancestor
- A1
- B1 (Human)
- C1 (Mouse)
- B2 (Human)
- C2 (Mouse)
- C3
Computational Identification of Orthologs

Between two species

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

Among multiple species

1. COG
   Triangle of BBH relationships among three species

2. KEGG OC
   p-Quasi clique among multiple species

   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
- Partial annotation by automated KOALA
- Manual annotation with KOALA and GFIT tools

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

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

KEGG GENES in Oracle
- Genomes 1,255
- Genes 5,631,338
- Genes with KO 1,992,279
- KO assignment 35%

Genome/UniProt 0.2%
Genes/UniProt 50%

As of June 12, 2010
KOALA: KEGG Orthology And Links Annotation

Map 1
Map 2

Genome 1 Genome 2 Genome 3

Map-based annotation

Genome-based annotation

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

Automatic annotation for safe K numbers

KOALA’s suggestion
Current assignment
GFIT link
OC links

Ortholog table

Conserved gene cluster

Tools to check annotation quality
Based on reconstruction of functional units
Public version of KEGG annotation tools

SSDB query for orthologs and paralogs
Conserved gene cluster search (useful for bacterial genomes)
Read-only version of GFIT
Ortholog table via pathway entry
DB search against UniProt, etc. may be done on the fly

KAAS: KEGG Automatic Annotation Server
http://www.genome.jp/tools/kaas/

Pairwise genome comparison
Set of aa seq
Query genome
BLAST search in both directions
Annotated genomes in KEGG

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

\[ BHR = R_f \times R_r > 0.95 \]

K number assignment by a heuristic scoring

\[ S_{K} = S_{K} - \log_{2}(mn) - \log_{2} \left( \sum_{i=1}^{n} C_i p^i (1 - p)^{i-1} \right) \]

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

Molecular network (pathway map)

KEGG PATHWAY

Hierarchical list (ontology)

KEGG BRITE

Simple list (membership)

KEGG DISEASE

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

KEGG PATHWAY Database
Collection of KEGG pathway maps

Global Map
Metabolism
Carbohydrate Metabolism (15)
Energy Metabolism (8)
Lipid Metabolism (16)
Nucleotide Metabolism (2)
Amino Acid Metabolism (13)
Metabolism of Other Amino Acids (9)
Glycan Biosynthesis and Degradation (15)
Metabolism of Cofactors and Vitamins (12)
Metabolism of Terpenoids and Polyketides (20)
Biosynthesis of Secondary Metabolites (21)
Xenobiotics Biodegradation and Metabolism (25)
Overview (9)

Genetic Information Processing
Transcription (3)
Translation (2)
Folding, Sorting and Degradation (5)
Replication and Repair (6)

Environmental Information Processing
Membrane Transport (3)
Signal Transduction (14)
Signaling Molecules and Interaction (4)

Cellular Processes
Transport and Catabolism (4)
Cell Motility (3)
Cell Growth and Death (7)
Cell Communication (4)

Organisinal Systems
Immune System (15)
Endocrine System (7)
Circulatory System (2)
Excretory System (4)
Nervous System (3)
Sensory System (3)
Development (2)
Environmental Adaptation (4)

Human Diseases
Cancers (15)
Immune Disorders (6)
Neurodegenerative Diseases (5)
Cardiovascular Disease (4)
Metabolic Disorders (3)
Infectious Diseases (7)

Drug Development
Chronology: antibiotics (8)
Chronology: antineoplastics (5)
Chronology: nervous system agents (9)
Chronology: other drugs (9)
Target based classification (15)
Structure based classification (5)
Skeleton based classification (8)

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

As of June 12, 2010
KEGG BRITE Database
Collection of BRITE functional hierarchies

Pathways and Ontologies
  Pathways (1)
  Functional hierarchies (1)

Genes and Proteins
  Network hierarchy (1)
  Protein families: metabolism (9)
  Protein families: genetic information processing (10)
  Protein families: environmental information processing (6)
  Protein families: cellular processes (7)

and associated binary relations

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

Drugs and Diseases
  Drugs (10)
  Diseases (2)

Cells and Organisms
  Organisms (4)

http://www.genome.jp/kegg/brite.html
As of June 12, 2010

---

Drug-target relationship in KEGG PATHWAY and KEGG BRITE

Search BRITE hierarchies

Anatomical Therapeutic Chemical (ATC) Classifiers

Target gene classification

Drug classification

Target information in the context of pathway

Try brite join operations in the KEGG BRITE page.
Pathway mapping and BRITE mapping
Linking genomes to biological systems and the environment

KO assignment

Genes to K numbers

K numbers to C/D/G/R numbers

Convention of map number prefix

Reference pathway

Organism-specific pathway

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 r).

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

Try examples of pathway mapping and BRITE mapping from KEGG2 page.
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
Prediction of biosynthetic/biodegradation potentials
Linking genomes to endogenous/exogenous molecules

KO assignment

Enzyme repertoire

Chemical repertoire

Glycosyltransferases

PK/NRP synthases

Fatty acid desaturases/elongases

Prenyltransferases

Glycans

Polyketides / nonribosomal peptides

Polyunsaturated fatty acids

Terpenoid backbone structures

Knowledge on biosynthetic pathways

Knowledge on biodegradation pathways

Genes to K numbers for specific enzyme groups

…………

Endocrine disrupting compounds

…………

N-Glycan biosynthesis - Homo sapiens (human)

Pathway map

hsa00510

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

A. N-glycan Precursor

Hsz, Ath, Sec, Dll

Trypanosoma


B. GPI-anchor Core

Hsz, Sec, Dll

Ath, Thr

Entamoeba

Diatoms

Olu, Ota, Thr

Green algae

Trypanosoma

Dual Aspect of Metabolic Network

1. Genomic information network
   - Network of enzyme genes (or enzymes)

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

   ![Diagram](image)

   - K00928
   - K0133
   - K01714
   - K00480
   - R02291
   - R02292

2. Chemical information network
   - Network of small molecules (or chemical structure transformations)

   - Compounds are identified by C numbers

   ![Diagram](image)

   - C00049
   - C03082
   - C00441
   - C03340
Chemical structure transformation network

Biphenyl degradation

Styrene degradation

Chemical structure comparison based on atom typing

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


### 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

- **Alkyne**
  - C3a: R≡CH
  - C3b: R≡C-R

- **Aldehyde**
  - C4a: R-CH=O

- **Ketone**
  - C5a: R-C(=O)-R

- **Carboxylic acid**
  - C6a: R-C(=O)-OH

- **Carboxylic ester**
  - C7a: R-C(=O)-O-R

- **Aromatic ring**
  - C8a: ring-CH=ring

- **Undefined C**
  - C0

#### Nitrogen 16 types
- **Amine**
  - N1a: R-NH₂

- **Cyclic amine**
  - N1x: ring-NH-ring

- **Amide**
  - N2a: R-NH₂
  - N2b: R-NH-R

- **Imine**
  - N2x: ring-N=ring

- **Imide**
  - N2y: ring-N(-R)=+ring

- **Cyan**
  - N3a: R≡N

- **Aromatic ring**
  - N4a: R-NH-ring

- **Undefined N**
  - N0

#### Oxygen 18 types
- **Hydroxy**
  - O1a: R-OH

- **Ether**
  - O2a: R-O-R

- **Oxide**
  - O3a: R-OH

- **Ketone**
  - O5a: R-C(=O)-R

- **Carboxylic acid**
  - O6a: R-C(=O)-OH

- **Carboxylic ester**
  - O7a: R-C(=O)-O-R

- **Undefined O**
  - O0

#### Sulfur 7 types
- **Thiol**
  - S1a: R-SH

- **Thioether**
  - S2a: R-S-R

- **Disulfide**
  - S3a: R-S-S-R

- **Other sulfur compounds**
  - S4a: R-SO₃

- **Undefined S**
  - S0

#### Phosphorus 2 types
- **Other elements 2 types**
  - 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 + H₂O ⇌ ÅH + BOH

**EC4. Lyases**
- AB ⇌ A + B

**EC5. Isomerases**
- A ⇌ A'

**EC6. Ligases**
- A + B + ATP ⇌ AB + AMP + PP_i
Reactant pair transformation patterns (RDM patterns)

**KEGG REACTION Database**
- Reaction: acetyl-CoA + L-glutamate ⇌ CoA + N-acetyl-L-glutamate

**KEGG RPAIR Database**
- Reactant pair: L-glutamate → N-acetyl-L-glutamate

**RDM patterns**
- **KEGG atom changes**
  - Reaction center: \( \text{N1a} \rightarrow \text{N1b} \)
  - Difference atom: \( \text{H} \rightarrow \text{C5a} \)
  - Matched atom: \( \text{C1c} \rightarrow \text{C1c} \)


---

**Prediction of xenobiotics degradation pathways**

PathPred @ [http://www.genome.jp/tools/pathpred/](http://www.genome.jp/tools/pathpred/)

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

RDM patterns for xenobiotics biodegradation pathways

1,901 RDM patterns

Chemical structure transformation patterns in known enzymatic reactions


KEGG DISEASE and DRUG: Linking genomes to practical values

Developing ‘Computable’ Disease Information Resource

KEGG disease pathway maps
• molecular network representation
• limited number of diseases

KEGG DISEASE entries
• simple list representation
• diseases with known genes/genomes

Diseases with known genetic factors
Cancers
Nervous system diseases
Immune system diseases
Circulatory system diseases
Urinary system diseases
Metabolic diseases
Genetic disorders
Infectious diseases with known pathogen genomes
Bacterial infections
Viral infections
Mycosis
Parasitic infections

Diagnostic markers
Molecular network
Therapeutic drugs
Pharmacogenomic markers

Genetic perturbations
Environmental perturbations

Diseases with known genetic factors
Cancers
Nervous system diseases
Immune system diseases
Circulatory system diseases
Urinary system diseases
Metabolic diseases
Genetic disorders
Infectious diseases with known pathogen genomes
Bacterial infections
Viral infections
Mycosis
Parasitic infections
Congenital disorders of glycosylation (CDG)

Single-gene disease with various types

KEGG pathway map for N-Glycan biosynthesis

Colorectal cancer

KEGG disease pathway map for colorectal cancer

Cancers
55 DISEASE entries
14 pathway maps
1 global map
Chronic myeloid leukemia in the global cancer map

Drugs for chronic myeloid leukemia

- Molecular target drugs
- Alkylating agents
- Antimetabolic agents
- Protein kinase inhibitors

Drugs:
- Imatinib
- Busulfan
- Hydroxyurea
- Melphalan
- Busulfan
- Imatinib
- Hydroxyurea
- Melphalan

Diagram showing the relationships and interactions of these drugs within the context of chronic myeloid leukemia.
KEGG DRUG as a Chemical Structure Database

Chemical structures and/or chemical components of:
- All prescription drugs in Japan
- All OTC drugs in Japan
- Most prescription drugs in USA
- Many prescription drugs in Europe

Structure activity relationships organized in:
- ATC classification of drugs in Japan, USA, and Europe
- Therapeutic category of drugs in Japan
- Classification of OTC drugs in Japan
- TCM drugs in Japan
- Crude drugs in Japan

<table>
<thead>
<tr>
<th>Number of entries</th>
<th>KEGG DRUG: 9,454</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude drugs</td>
<td>531</td>
</tr>
<tr>
<td>TCM formulas</td>
<td>229</td>
</tr>
</tbody>
</table>

(As of June 12, 2010)

Imatinib mesylate (Gleevec)

KEGG DRUG as a Molecular Network Database

1. Molecular interaction network involving:
- Target molecules
- Drug metabolizing enzymes
- Drug transporters
- Adverse drug interactions
- Pharmacogenomic markers

2. Chemical structure transformation network for:
- Biosynthetic pathways of natural products (KEGG metabolic pathway maps)
- History of drug development (KEGG DRUG structure maps)
Chemical structure transformation networks
Linking genomes to natural products and drugs

Enzyme-catalyzed chemical transformations
Genome → Biosynthetic pathways → Natural products → Drugs

Human-made chemical transformations
Genome → Biodegradation pathways → Environmental compounds

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

Crude drugs and active ingredients

Stomachic and antidiarrheal drugs in the Traditional Chinese Medicine

- D01250 Berberine chloride
- D00092 Coptis rhizome
- D06689 Phellodendron bark
- D06686 Corydalis tuber
- D06720 Magnolia bark
- D06785 Sinomenium stem
- D06808 Epimedium herb
- D06725 Calumba
- D00092 Coptis rhizome
- D03868 Ipecac
- D09307 Colchicum seed
- D06686 Corydalis tuber
- D00092 Coptis rhizome
- D06689 Phellodendron bark
- D03444 Opium

Therapeutic efficacy of crude drugs:
- any active ingredients or
- synergistic action of multiple ingredients

KEGG pathway map:
Isoquinoline alkaloid biosynthesis
Isoquinoline alkaloids

Structured optimized drugs

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01C</td>
<td>Cardiac stimulants</td>
<td>Drugs for functional bowel disorders</td>
</tr>
<tr>
<td>A03A</td>
<td>Drugs for functional bowel disorders</td>
<td></td>
</tr>
<tr>
<td>R05D</td>
<td>Cough suppressants</td>
<td></td>
</tr>
<tr>
<td>N02A</td>
<td>Opioid analgesics</td>
<td></td>
</tr>
<tr>
<td>M03A</td>
<td>Muscle relaxants</td>
<td></td>
</tr>
<tr>
<td>M04A</td>
<td>Antigout preparations</td>
<td></td>
</tr>
<tr>
<td>N06D</td>
<td>Anti-dementia drugs</td>
<td>Anti-dementia drugs</td>
</tr>
</tbody>
</table>

Crude drugs

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01C</td>
<td>Cardiac stimulants</td>
<td>Drugs for functional bowel disorders</td>
</tr>
<tr>
<td>A03A</td>
<td>Drugs for functional bowel disorders</td>
<td></td>
</tr>
<tr>
<td>R05D</td>
<td>Cough suppressants</td>
<td></td>
</tr>
<tr>
<td>N02A</td>
<td>Opioid analgesics</td>
<td></td>
</tr>
<tr>
<td>M03A</td>
<td>Muscle relaxants</td>
<td></td>
</tr>
<tr>
<td>M04A</td>
<td>Antigout preparations</td>
<td></td>
</tr>
<tr>
<td>N06D</td>
<td>Anti-dementia drugs</td>
<td>Anti-dementia drugs</td>
</tr>
</tbody>
</table>

KEGG MEDICUS
Molecular network based information resource for human diseases and drugs

<table>
<thead>
<tr>
<th>URL</th>
<th>KEGG DISEASE</th>
<th>KEGG DRUG</th>
</tr>
</thead>
</table>

Content
Diseases with known genetic factors and infectious diseases with known pathogen genomes, represented as lists of disease genes, environmental factors, markers, drugs, etc.

Chemical structures and/or components of all approved drugs, together with information of interactions with targets, metabolizing enzymes, transporter, genomic biomarkers, etc.

Pathway
KEGG pathway maps for human diseases: cancers, immune system diseases, neurodegenerative diseases, cardiovascular diseases, metabolic disorders, and infectious diseases

KEGG pathway maps for drug metabolism and secondary metabolite and antibiotics biosynthesis
KEGG DRUG structure maps for the history of drug development

BRITE hierarchy
Disease classifications including: Human diseases

ICD-10 disease classification

Drug classifications including:

ATC classification (WHO)
Therapeutic category of drugs (Japan)
OTC drug classification in Japan
Crude drugs and TCM drugs in Japan

Summary

Three types of molecular networks in KEGG PATHWAY

1. Molecular reaction network
2. Chemical structure transformation network
3. Molecular interaction network

For understanding, for example, how chemical substances are biosynthesized, biodegraded, and used as signaling molecules

<table>
<thead>
<tr>
<th>Metabolic pathway maps</th>
<th>Drug structure maps</th>
<th>All other maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1+2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Molecular network-based approaches to
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

Molecular reaction network

- Perturbants
  - Chemical perturbants
    (environmental factors of
diseases, drugs)
  - Genetic perturbants
    (genetic factors of diseases)

- Perturbed system

- Chemical structure transformations
  - EDCs, etc.
  - Phytoestrogens,
etc.

Biodegradation
pathway in bacteria

Biosynthetic pathway
in plants

Changing roles of bioinformatics:
Basic research to practical values

- Capturing knowledge on molecular systems both in
  normal and perturbed (disease) states

- Capturing knowledge on drugs and environmental
  compounds as perturbants to molecular systems

- Generalizing knowledge on genes and proteins as
  ortholog groups

- Generalizing knowledge on chemical transformations
  in enzymatic reactions by RDM patterns

- Knowledge-based analysis of human disease mechanisms

- Drug discovery from the genomes of plants and microorganisms

- Knowledge-based prediction of xenobiotics degradation pathway
  and responsible genes