

KEGG/GenomeNetの利用法

京都大学化学研究所バイオインフォマティクスセンター

守屋 勇樹

2013/11/6



KEGG/GenomeNetへのアクセス

本実習では主にブラウザを用いたKEGG (Kyoto Encyclopedia of Genes and Genomes) データベース、および計算ツールの利用法を紹介します。使い慣れたブラウザを使ってGenomeNetにアクセスしましょう。

GenomeNet

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

GenomeNet

GenomeNetは、ゲノム情報を基盤とした新しい生命科学研究と創薬・医療・環境保全への応用を推進するために、京都大学化学研究所バイオインフォマティクスセンターが提供するインターネットサービスです。KEGGを主幹とするデータベース群と、様々な生命情報データを解析するための計算ツール群からなっています。

KEGG

KEGGはゲノムや分子レベルの情報から細胞、個体、エコシステムといった高次生命システムの機能や有用性を理解するためのリソースです。生命システムのコンピュータ表現として、遺伝子やタンパク質（ゲノム情報）と化合物など（ケミカル情報）の分子部品の情報を、分子間の相互作用・反応・関係ネットワーク（システム情報）の知識で統合した生命システム情報統合データベースです。



ゲノム情報

KEGG Organisms

KEGG GENOMEにはゲノム配列の決定したほぼ全ての生物種が登録されています。配列データは主にNCBIで作成されているRefSeqデータベースや各シークエンスセンターから集めてきています。生物種はNCBIのTaxonomyに従い分類され、KEGG独自の3~4文字からなる生物種コードが付加されています。例えば、大腸菌K-12 MG1655株には”eco”というコードが定義されています。また、サブカテゴリとして、完全長ゲノムは決定していないが、生物学的に重要と考えられる真核生物（ドラフトゲノム）を集めたDGENOME、生物種の他にも環境サンプルから直接遺伝子をシークエンスしたメタゲノムの配列を集めたMGENOMEがあります。他にも主に植物のESTから構築したEGENOMEがありますが、これは廃止予定です。

KEGG GENOME

各エントリには生物種コードや種名、学名、系統情報、データソースなどの情報が記載されています。また、KEGGでは新規にKEGGに登録された遺伝子に自動で機能アノテーションを行ったあとに手作業による修正を行うことで、精度の高い機能アノテーションを提供しています。アノテーション作業の進み具合もGENOMEエントリに記載されています。manualはGENESに登録されています。

KEGG Homo sapiens (human)

Genome info Pathway map Brite hierarchy Module Genome map

Search genes: Go Clear

Genome information

T number	T01001
Org code	hsa
Aliases	HUMAN, 9606
Full name	Homo sapiens (human)
Definition	Homo sapiens (human)
Annotation	manual
Taxonomy	TAX: 9606
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Data source	RefSeq (Project:168)

る生物種で、手作業による修正が完了しています。KOALAは新規にGENESに登録された生物種に、KOALAというツールによってSW (Smith-Waterman) スコアに基づいた自動アノテーションが行われています。KAASはDGENES、EGENES、MGENESの遺伝子に対し、

KAASというツールによってBLASTの類似性スコアに基づいた自動アノテーションが行われています。また、KAASはGenomeNetから誰でも利用できます。

エントリの検索

KEGG及びGenomeNetのWebページでは、ゲノム情報に限らず、至る所から様々なデータベースに対してキーワード検索が行えるようになっています。

Search GENES for Go Clear
 bfind mode bget mode

基本的な使用方法は共通で、データベースを選択し、キーワードを入力し、”Go”をクリックするだけです。エントリ名があらかじ判っている場合には”bget mode”を用いることで直接エントリが表示されます。生物種を限定して、遺伝子を検索する場合には生物種コードが必要です。生物種コードは”Organism”をクリックして開くウインドウから検索することができます。

Search Organism hsa for Go Clear
 bfind mode bget mode

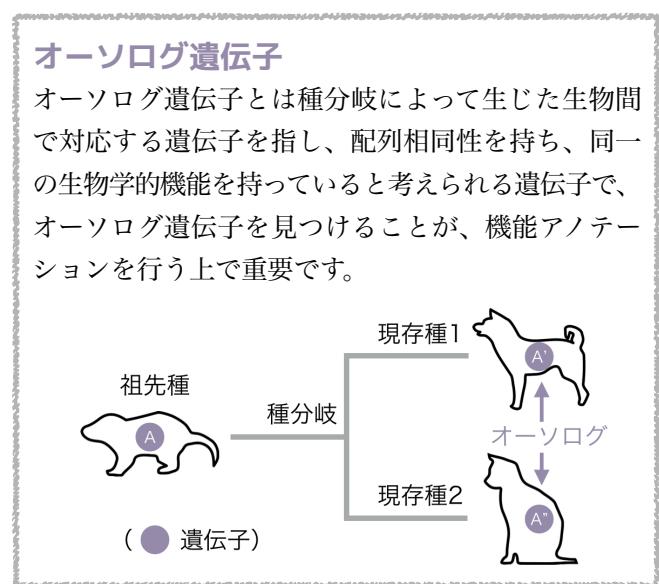
KEGG GENES

各エントリには遺伝子ID (Entry)、遺伝子・タンパク質名 (Gene name)、データソースオリジナルの機能アノテーション及びその他の記述 (Definition)、KEGGにおける機能アノテーション (Orthology)、種名

KEGG Homo sapiens (human): 6883 Help

Entry	6883 CDS T01001
Gene name	TAF12, TAF2J, TAPII20
Definition	TAF12 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 20kDa
Orthology	K03126 transcription initiation factor TFIID subunit 12
Organism	hsa Homo sapiens (human)
Pathway	hsa03022 Basal transcription factors
Class	Genetic Information Processing; Transcription; Basal transcription factors [PATH:hsa03022] [BRITE hierarchy]
SSDB	Ortholog Paralog GFIT
Motif	Pfam: TFIID_20kDa_CBFID_NFYB_HMF Histone TFIID-31kDa TAF4 Condensin2nSMC Motif
Other DBs	NCBI-GI: 206725450 NCBI-GenoID: 6883 OMIM: 600773 HGNC: 11545 HPRD: 15981 Ensembl: ENSG00000120656 Vega: OTTHUMG000000003655 UniProt: Q16514
Structure	PDB: 1H3O Thumbnails Jmol
Position	1p35.3
AA seq	161 aa AA seq DB search MNQFGPSALLNLNFSSIKPEPASTPPQGSMANSTAVVKIPGTPGAGGRILSPENNQVLTK KKLQLDVLREVDPNEQLDEDVVEEMLLQIADDPIESVVTAACQLARHRRKSSTLEVKDQLHL ERQWNWMWIPFGSEEIRPYKKACTTEAHKQRMALIRKTKK
NT seq	486 nt NT seq atgaaccaggttggccctcagccctaatacacttcacccatggccaaatgtactgcagtgtttaaagata gaaccaggcagcccccaccaaggctccatggccaaatgtactgcagtgtttaaagata

(Organism)、KEGG PATHWAYへのリンク (Pathway)、KEGG BRITEへのリンク (Class)、KEGG SSDB からの情報抽出のためのリンク (SSDB)、遺伝子が持つPfamモチーフへのリンク (Motif)、他のデータベースへのリンク (Other DBs)、構造情報 (Structure)、ゲノム上での位置もしくは染色体番号 (Position)、アミノ酸配列 (AA seq)、塩基配列 (NT seq)などの遺伝子に付随する情報が記載されています。KEGGでは現在3,000を超える生物種、メタゲノムサンプルを蓄積し、その配列数は1億本を超えていきます。これらを扱いやすくするために、KEGGではオーソログ遺伝子毎に分類したKEGG ORTHOLOGYデータベースを作成しています。



KEGG ORTHOLOGY (KO)

KEGGでは代謝ネットワーク (PATHWAY) や機能の階層分類 (BRITE) に従って分類したKEGG ORTHOLOGY (KO) という遺伝子のオーソロググループを作成し、KEGG GENESに登録された遺伝子を配列類似性やゲノム上での遺伝子の並び、文献情報などに基づきオーソロググループに割り当てることで、独自に機能アノテーションを行っています。各エントリにはK番号でできたID (Entry)、KOグループの遺伝子の名前 (Name)、KOグループの機能アノテーション (Definition)、KEGG PATHWAYへのリンク (Pathway)、KEGG BRITEへのリンク (Brite)、他のデータベースへのリンク (Other DBs)、KOグループに含まれる遺伝子 (Genes)、KOグループ作成の基となつた文献情報 (Reference) などが記載されています。

KEGG ORTHOLOGY: K03126

Entry K03126 KO

Name TAF12

Definition transcription initiation factor TFIID subunit 12

Pathway ko03022 Basal transcription factors

Brite KEGG Orthology (KO) [BR:ko00001] Genetic Information Processing Transcription 03022 Basal transcription factors K03126 TAF12; transcription initiation factor TFIID subunit 1 Transcription machinery [BR:ko03021] TAF12; transcription initiation factor TFIID subunit 1 TAF12; transcription initiation factor TFIID subunit 1 SAGA complex K03126 TAF12; transcription initiation factor TFIID subunit 1 SLIK complex K03126 TAF12; transcription initiation factor TFIID subunit 1 [BRITE hierarchy]

Other DBs GO: 0016251

Genes HSA: 6883(TAF12)
PTR: 456686(TAF12)
PPS: 100991350(TAF12)
GGO: 101128154(TAF12)
PON: 100460824(TAF12)
MCC: 717135(TAF12)
MMU: 66464(Taf12)
RNO: 682902(Taf12)
CCE: 100769367
HGL: 101719733(Taf12)
» show all
[Taxonomy] [KOALA] [UniProt]

Reference PMID:19308322
Authors Cler E, Papai G, Schultz P, Davidson I
Title Recent advances in understanding the structure and function of general transcription factor TFIID.
Journal Cell Mol Life Sci 66:2123-34 (2009)

Reference PMID:173380162
Authors Lee KK, Workman JL
Title Histone acetyltransferase complexes: one size doesn't fit all.
Journal Nat Rev Mol Cell Biol 8:284-95 (2007)

Reference PMID:17967894
Authors Liu X, Vorontchikhina M, Wang YL, Faiola F, Martinez E
Title STAGA recruits Mediator to the MYC oncprotein to stimulate transcription and cell proliferation.
Journal Mol Cell Biol 28:108-21 (2008)



3つのタイプのデータベースの多くは、KOを介してリンクされているため、KEGGを利用する上で非常に重要

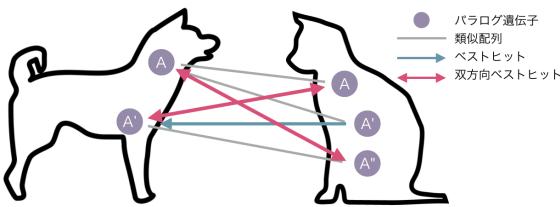
KAAS

KAAS (KEGG Automatic Annotation Server) は、配列類似性やベストヒットの情報、経験則を用いることで、KEGGに登録されていない未知の遺伝子に対しても、KOを自動的にアサインできるようにしたインターネットサービスです。マルチFASTA形式の問い合わせ配列を入力とし、配列とKOの対応関係が outputされます。KAASは問い合わせ配列の種類によってあらかじめ適切なパラメータがセットされています。“Complete or Draft Genome”では入力配列が特定の種の遺伝子で、ある程度網羅性が期待できるとき双方向ベストヒット (BBH) を指標に、より精度の高いアノテーションを行います。双方向の相同性スコアを計算するため計算量が増えます。“Partial

Genome”では入力が網羅的でないときには双方向の意味がないので片方向ベストヒット（SBH）を用いてアノテーションを行います。”ESTs”では入力配列がESTの場合に用います。スクレオチドを全パターン翻訳し、スコアを計算するため計算時間がかかります。また、医科学研究所ヒトゲノム解析センターでサービスしているEGassemblerを利用することで、ESTをアセンブルした配列をアノテーションすることもできます。

ベストヒットとオーソログ

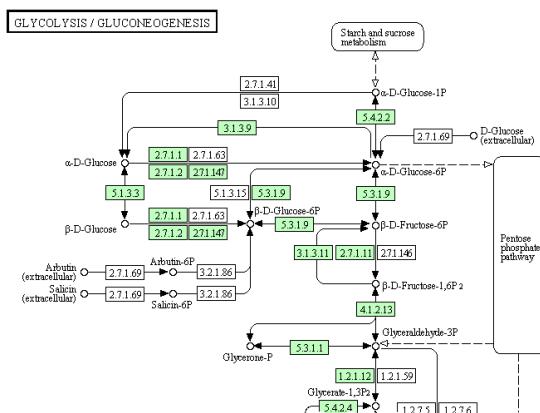
複数パラログがある場合に正確なオーソログ関係を同定することは難しいですが、生物種間でもっとも似ている配列、ベストヒットを手がかりにすることで、オーソログである可能性の高い遺伝子を探し出すことができます。



システム情報

KEGG PATHWAY

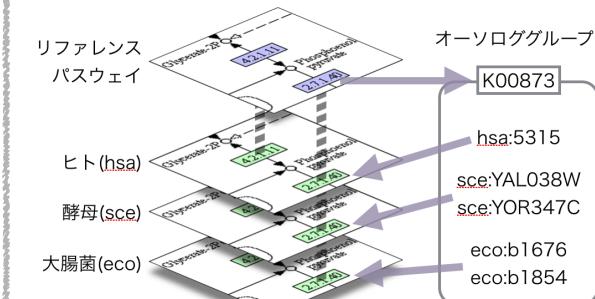
KEGG PATHWAYは文献情報から構築した生体内、生体外の分子と遺伝子のネットワークです。現在は代謝系、遺伝子制御、環境シグナル、細胞プロセス、生体システム、ヒト疾患、薬剤の開発の7つのサブカテゴリからなっています。それぞれのパスウェイマップは化合物を示す丸と、遺伝子を示す四角のネットワーク図として表現されています。また、四角はKOグループや、反応、酵素などを表している場合もあります。プルダウンメニューか



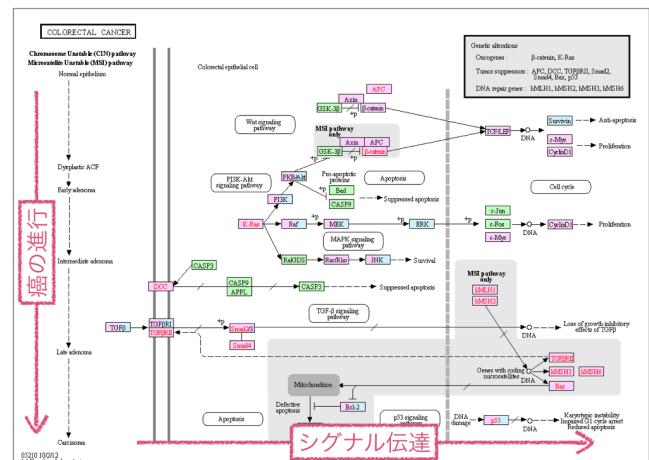
らリファレンスパスウェイや生物毎のパスウェイを表示することができます。

KOとPATHWAY

各生物のパスウェイマップを遺伝子と化合物の相互作用ネットワークとすると、リファレンスパスウェイはKOと化合物の相互作用ネットワークと考えることができます。



ヒトの疾患マップの一部では、疾患の進行と、各段階で細胞内に起こるシグナル伝達の流れが、縦軸と横軸で示されています。また、疾患の原因とされる遺伝子や、薬剤の標的となる遺伝子などを色分けによって表すことができます。



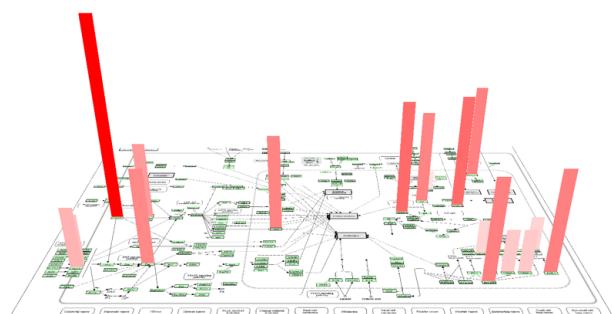
KEGG BRITE

BRITEは様々なデータを、階層的に編集された言葉によって分類した、階層分類群です。現在は5つのサブカテゴリからなり、“Pathway and Ontologies”ではPATHWAYデータベースやBRITEデータベース自身の一覧を、“Genes and Proteins”では遺伝子やタンパク質の機能やネットワークの分類を、“Drugs and Diseases”では薬物や疾患の分類を、“Compounds and Reactions”では化合物の機能による分類や酵素反応の種類による分類を、“Organisms and Cells”では生物種の系統分類や細胞や器官の分類を行っており、現在100を超える階層分

類が文献情報を基に構築されています。Webブラウザでは矢印をクリックすることで、各階層の展開、収納が可能になっています。

KEGG Mapper

KEGG Mapperでは多数のパスウェイマップからKOや遺伝子、化合物や反応などを検索し、自由に色を塗り分けることができます。”Search&Color Pathway”ではKEGGで用いられるIDの他、NCBI-GeneID、NCBI-gi、UniProtのIDが使用できます。遺伝子の発現量などをパスウェイマップ上で可視化するのに便利な”Color Pathway 3D”もあります。またBRITE階層分類ファイルに対しても色によって目印をつけることが可能です。



ケミカル情報

KEGG LIGAND

生化学的な情報を収録したもので、6つのデータベースからなっています。主に中間代謝、二次代謝産物などの化合物を収録したKEGG COMPOUNDデータベース、糖

鎖分子の構造を収録したKEGG GLYCANデータベース、生化学反応を収録したKEGG REACTIONデータベース、反応における基質と生成物間の変化を収録したKEGG RPAIRデータベース、変化のパターンを分類したKEGG RCLASSデータベース、酵素番号を収録したKEGG ENZYMEデータベースがあります。

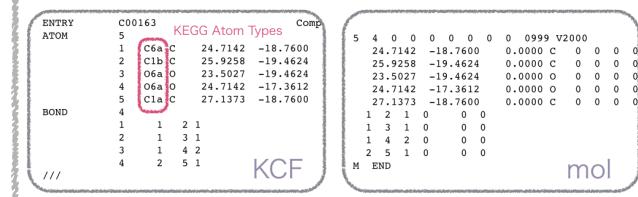
KEGG COMPOUND

各エントリにはC番号でできたID (Entry)、化合物の名前 (Name)、組成式 (Formula)、質量 (Exact mass)、分子量 (Mol weight)、構造 (Structure) の他、内外のデータベースへのリンクなどが記載されています。構造は二次元グラフで収録されており、立体構造は含まれていませんが、立体異性の判明している場合は、立体異性情報が記載されています。

Compound: C00163		Help
Entry	C00163	Compound
Name	Propanoate; Propionate; Propanoic acid; Propionic acid	
Formula	C3H6O2	
Exact mass	74.0368	
Mol weight	74.0785	
Structure	 C00163	
	Mol file KCF file DB search Jmol KegDraw	
Remark	Same as: D02310	
Reaction	R00920 R00925 R00928 R01353 R01354 R01355 R01449 R05366	
Pathway	map00640 Propanoate metabolism map00642 Ethylbenzene degradation	
Enzyme	2.7.2.1 2.7.2.15 2.8.3.1 3.7.1.- 4.1.3.32 6.2.1.1 6.2.1.13 6.2.1.17	
Brite	Compounds with biological roles [BR:br08001] Organic acids Carboxylic acids Monocarboxylic acid C00163 Propionate; Propanoate BRITE hierarchy	
Other DBs	CAS: 79-09-4 PubChem: 3463 ChEBI: 30768	
LinkDB	All DBs	
KCF data	Show	

KCFフォーマットとKEGG Atom Types

KEGGでは構造をKCF (KEGG Chemical Function) フォーマットで記述しています。MDL molフォーマットとの大きな違いは元素を周辺環境の情報を基にタイプ分けしたKEGG Atom Typesを用いて記述されていることです。



KEGG REACTION

各エントリにはR番号でできたID (Entry) 、反応の名前 (Name) 、化合物名による反応式 (Definition) 、C番号による反応式 (Equation) 、画像による反応式及び他のデータベースへのリンクが記載されています。

KEGG REACTION: R01353

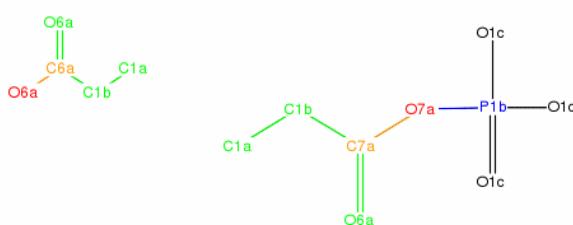
Entry: R01353 Reaction
Name: ATP:propanoate phosphotransferase
Definition: ATP + Propanoate <=> ADP + Propanoyl phosphate
Equation: C00002 + C00163 <=> C00008 + C02876

Reaction diagram showing the transfer of a propanoyl group from ATP to a substrate. The reaction involves the conversion of ATP (C00002) and Propanoate (C00163) to ADP (C00008) and Propanoyl phosphate (C02876). The diagram illustrates the movement of atoms: R atoms (red), D atoms (blue), M atoms (yellow), and other atoms (black).

RPair: RP00003_C00008 main
RP01451_C00163_C02876 main
RP06540_C00002_C02876 trans
Enzyme: 2.7.2.1 2.7.2.15
Pathway: rnr00640 Propanoate metabolism
Orthology: K00925 acetate kinase [EC:2.7.2.1]
K00932 propionate kinase [EC:2.7.2.15]
LinkDB: All DBs

KEGG RPAIR

KEGG RPAIRは生化学反応における基質と生成物の組み合わせにおける、構造の変化を収録したデータベースです。2つの化合物間のグラフマッチを行っており、生化学反応の前後での原子の追跡が可能になっています。画像では構造変化の中心となった原子 (R atom) が赤で示され、R atomに隣接し変化した原子 (D atom) は青色で示されています。また、反応の前後で変化の無い原子は緑色で示され、特にR atomに隣接する原子 (M atom) は黄色で示されています。その他の原子は黒で示されています。



KEGG RCLASS

KEGG RCLASSはKEGG RPAIRデータベースで定義されたR atom、D atom、M atomだけを抽出したRDMパターンによって、RPAIRの各エントリを分類したものです。RDMパターンは"::"で区切られた3つの部分からなっていて、R atomの変化、D atomの変化、M atomの変化で表されています。

O6a-O7a:*-P1b:C6a-C7a

SIMCOMP/SUBCOMP

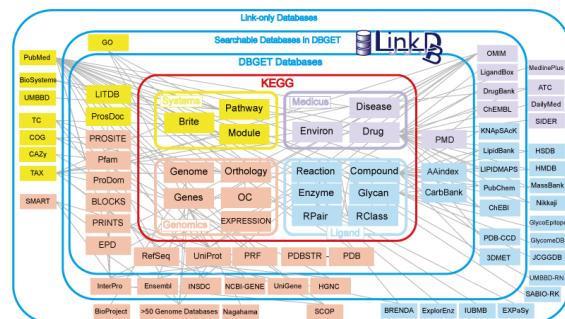
化合物の二次元グラフ構造からグラフマッチを行うことでKEGG COMPOUNDデータベース及びKEGG DRUGデータベースから、構造の類似した化合物を探し出すことができます。



LinkDB

KEGG/GenomeNetは内外の様々なデータベースとつながっています。LinkDBでは様々なデータベース間のエントリのリンクを二項関係で表したテキスト形式で取得することができます。

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



KEGG SSDB

GenomeNetでは様々な配列データベースに対してBLASTやFASTAといった高速な配列相同性検索をサービスしていますが、KEGG GENESに蓄積されている遺伝子に対しては、より高精度のSmith-Watermanアルゴリズムを用いた相同性スコア (SWスコア) があらかじめ計算され、KEGG SSDBに蓄積されています。そのため、GENESに蓄積されている遺伝子・タンパク質については、まだオーソロググループが定義されていない場合であっても、SSDBに蓄積されたスコアを基にオーソログ配列を推定することが可能になっています。

<http://www.kegg.jp/kegg/ssdb/>

KEGG OC

KEでは機能の判明した遺伝子を基にした遺伝子群の分類を行っているため、未だ機能の判らない多くの遺伝子はオーソロググループを定義できていません。そこでSSDBに蓄積されたスコア情報を基に、KEGG GENESに登録されている全遺伝子のクラスタリングを行うことで、

オーソロググループを推定したものがKEGG OCです。系統樹に従った多段階のクラスタリングを行うことで、階層的な推定オーソロググループが定義されています。

<http://www.genome.jp/tools/oc/>

ID	Taxon	Sequence	Annotation
84398	Bacteria, Firmicutes, Clostridia, Clostridiales, Clostridium, 101348	haloarobium, 1512	has:364 has:Halsa_0367 prephenate dehydrogenase
84398	Bacteria, Firmicutes, Clostridia, Clostridiales, Clostridium, 185200		chv:494 chv:CHY_0474 tyra; prephenate dehydrogenase
84398	Bacteria, Firmicutes, Clostridia, Clostridiales, Clostridium, 126262		hor:1014 hor:Hore_10360 prephenate dehydrogenase
84398	Bacteria, Actinobacteria, 25218		cwo:3552 cwo:Cwoe_3737 prephenate dehydrogenase
84398	Bacteria, Actinobacteria, 25218		aym:2431 aym:YM304_24990 tyra; putative prephenate dehydrogenase [EC:1.3.1.12]
84398	Bacteria, Actinobacteria, 44303		ccu:653 ccu:Ccur_06830 prephenate dehydrogenase
84398	Bacteria, Actinobacteria, Eggerthellia, 145797		ele:1273 ele:Elen_1331 prephenate dehydrogenase
84398	Bacteria, Actinobacteria, Eggerthellia, 145797	183	eyy:1573 eyy:EGYY_17670 hypothetical protein
84398	Bacteria, Bacteroidetes, Salinibacter, 29249		srw:1563 srw:SRU_1616 tyra; prephenate dehydrogenase
84398	Bacteria, Bacteroidetes, Salinibacter, 29249	879	srw:1866 srw:SRM_01816 tyrA; prephenate dehydrogenase
84398	Bacteria, Bacteroidetes, Rhodothermus, 29249	719	rmm:1769 rmm:Rmar_1803 prephenate dehydrogenase
84398	Bacteria, Bacteroidetes, Rhodothermus, 29249	719	rmg:1715 rmg:Rhomb172_1759 prephenate dehydrogenase [EC:1.3.1.121]

KEGG MEDICUS

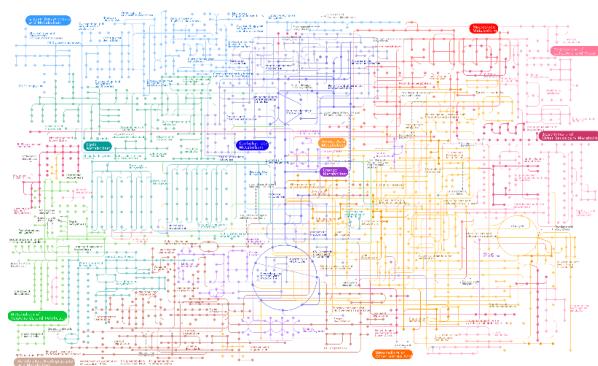
疾患・医薬品・環境物質など社会的ニーズの高いデータを、ゲノム情報を基盤とした生体システム情報として統合したリソースです。医薬品情報や成分の検索、JAPIC 医薬品添付文書に記載された併用禁忌、併用注意といった医薬品間の相互作用を調べることができます。

<http://www.kegg.jp/kegg/medicus/>

KEGG Atlas

パスウェイマップをつなぎ合わせたグローバルマップを見るためのビュアです。

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



E-zyme

反応前後の2つの化合物から、SIMCOMPを用いたグラフマッチを行うことでRDMパターンを推定し、EC番号を予測するツールです。

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

PathPred

微生物における生体異物の分解および、植物における二次代謝産物の合成経路を予測するツールです。

<http://www.genome.jp/tools/pathsearch/>

GENIES

遺伝子発現データなどの実験データや、系統プロファイルといったゲノムデータなどの様々な種類の大規模データから遺伝子相互作用ネットワークを推定するツールです。

<http://www.genome.jp/tools/genies/>

KEGG API

KEGGデータベースのREST APIサービスです。各データベースのエントリの検索や取得が可能です。

<http://www.kegg.jp/kegg/rest/>

GenomeNet API

APIでのSIMCOMP、SUBCOMPの利用及び、KEGG OC エントリの検索や取得、LinkDBの全てのリンク情報の取得ができます。

http://www.genome.jp/tools/gn_tools_api.html

http://www.genome.jp/tools/gn_ga_tools_api.html

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

KGML

KGML (KEGG Markup Langage) はパスウェイマップを XML形式で記述したものです。KEGG APIにて取得できます。

<http://www.kegg.jp/kegg/xml/>

KEGG Tools

KegHierはBRITE階層分類ファイルビュアです。

KegArrayはマイクロアレイデータ解析ソフトウェアです。

KegDrawは化合物の構造を作図するソフトウェアです。

<http://www.kegg.jp/kegg/download/kegtools.html>

Feedback

KEGG/GenomeNetに関する全ての事柄について、質問やコメントをお待ちしております。日本語で問題ありません。

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