

Name	<b>KEGG</b> (Kyoto Encyclopedia of Genes and Genomes)
URL	<a href="http://www.genome.ad.jp/kegg/">http://www.genome.ad.jp/kegg/</a>
Web service	<a href="http://www.genome.jp/kegg/soap/">http://www.genome.jp/kegg/soap/</a>
Description	KEGG is a set of databases storing information of three categories: systems, genomic and chemical information.
Usage	Given an EC number, web service is used to query KEGG ENZYME database and obtain the following data for the given enzyme: <ul style="list-style-type: none"> <li>▪ synonyms</li> <li>▪ related compounds (i.e. substrates and products of the reaction catalysed by the enzyme)</li> <li>▪ gene encoding the enzyme (in the organism of interest, e.g. <i>Saccharomyces cerevisiae</i>)</li> </ul> Further, for each compound retrieved, KEGG COMPOUND is queried to obtain their synonyms and cross-references to related external databases (ChEBI and PubChem). Similarly, for each gene retrieved, KEGG GENES is queried to obtain the gene names and cross-references to related external databases (SGD and CYGD – note that these databases are relevant for <i>S. cerevisiae</i> only).
Java class	EnzymeKEGG.java
Name	<b>ChEBI</b> (Chemical Entities of Biological Interest)
URL	<a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a>
Web service	<a href="http://www.ebi.ac.uk/chebi/webServices.do">http://www.ebi.ac.uk/chebi/webServices.do</a>
Description	ChEBI is a database of molecular entities focused on ‘small’ chemical compounds, which are either products of nature or synthetic products used to intervene in the processes of living organisms.
Usage	Given a ChEBI identifier, a web service is used to query this database for different names used synonymously for the given compound.
Java class	EnzymeKEGG.java
Name	<b>PubChem</b>
URL	<a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>
Download	<a href="ftp://ftp.ncbi.nih.gov/pubchem/Compound/Extras">ftp://ftp.ncbi.nih.gov/pubchem/Compound/Extras</a>
Description	PubChem is a set of databases, which provide information on the biological activities of small molecules including compounds (Compound), chemical substances (Substance) and bioassays (BioAssay).
Usage	PubChem is accessible through Entrez and its web service. However, at the time of implementation the mode of its operating was to upload a resulting file at a dynamically produced URL within 24 hours, which was not suitable for our application. We opted to install a local database with a relevant subset of data from PubChem databases. The local database was used to store data from files available at the given URLs. The structure of the database tables follows that of the given files (the database schema is available), which makes upload of the data straightforward. These data are used to map a given SID (substance identifier in PubChem Substance) to the corresponding CID (compound identifier in PubChem Compound), which is then mapped to synonyms available for the given compound.
Java class	PubChemLocal.java, EnzymeKEGG.java
Name	<b>SGD</b> (Saccharomyces Genome Database)
URL	<a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a>
Description	SGD is a database of the molecular biology and genetics of <i>Saccharomyces cerevisiae</i> .
Usage	Given an SGD identifier (e.g. S000000001) for a gene, its SGD web page (URL: <a href="http://db.yeastgenome.org/cgi-bin/reference/litGuide.pl?dbid=S000000001">http://db.yeastgenome.org/cgi-bin/reference/litGuide.pl?dbid=S000000001</a> ) is scrapped for the gene names.
Java class	EnzymeKEGG.java
Name	<b>CYGD</b> (Comprehensive Yeast Genome Database)
URL	<a href="http://mips.gsf.de/genre/proj/yeast/">http://mips.gsf.de/genre/proj/yeast/</a>
Description	CYGD a database of the molecular structure and functional network of the entirely sequenced <i>Saccharomyces cerevisiae</i> .
Usage	Given the ORF (e.g. YAL001C) of a gene, its CYGD web page (URL: <a href="http://mips.gsf.de/genre/proj/yeast/searchEntryAction.do?text=YAL001C">http://mips.gsf.de/genre/proj/yeast/searchEntryAction.do?text=YAL001C</a> ) is scrapped for the gene names.
Java class	EnzymeKEGG.java
Name	<b>SBO</b> (Systems Biology Ontology)
URL	<a href="http://www.ebi.ac.uk/sbo/">http://www.ebi.ac.uk/sbo/</a>
Web service	<a href="http://www.ebi.ac.uk/sbo/SBOWSLib/ws.html">http://www.ebi.ac.uk/sbo/SBOWSLib/ws.html</a>
Description	SBO is an ontology describing the concepts used in systems biology, especially in the context of computational modeling.
Usage	Given the SBO identifier for a concept of interest, SBO ontology is queried to obtain the synonyms used to name the concept.
Java class	SBO.java

Name	<b>GO</b> (Gene Ontology)
URL	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
Web service	<a href="http://www.ebi.ac.uk/ontology-lookup/WSDLDocumentation.do">http://www.ebi.ac.uk/ontology-lookup/WSDLDocumentation.do</a>
Description	GO encodes knowledge of gene and protein roles in cells that can be applied to all organisms. It provides three structured networks of defined terms to describe gene product attributes in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.
Usage	Given the GO identifier for a concept of interest, GO ontology is queried to obtain the synonyms used to name the concept.
Java class	GO.java
Name	<b>UMLS</b> (Unified Medical Language System)
URL	<a href="http://umlsks.nlm.nih.gov/">http://umlsks.nlm.nih.gov/</a>
Web service	<a href="http://kswebpl.nlm.nih.gov/DocPortlet/html/dGuide/Guide.htm">http://kswebpl.nlm.nih.gov/DocPortlet/html/dGuide/Guide.htm</a>
Description	UMLS is a set of databases and tools aimed at facilitating the development of information systems for text processing in the domain of biomedicine. The role of UMLS in such systems is to provide a formal representation of the domain-specific knowledge in order to process, retrieve, integrate and aggregate biomedical data and information contained in the relevant literature.
Usage	Given a term, UMLS is queried to obtain its synonyms.
Java class	ExpertClientUMLS.java
Name	<b>MeSH</b> (Medical Subject Headings)
URL	<a href="http://www.nlm.nih.gov/mesh">http://www.nlm.nih.gov/mesh</a>
Download	<a href="http://www.nlm.nih.gov/mesh/filelist.html">http://www.nlm.nih.gov/mesh/filelist.html</a>
Description	MeSH is a hierarchically structured controlled vocabulary used to index and subsequently search the literature in PubMed and PubMed Central at various levels of specificity.
Usage	KiPar makes no direct access to the content of MeSH, hence no MeSH files need to be downloaded. However, Entrez (the retrieval system used to query literature databases) utilizes MeSH to enhance the search results. When a literature search request is issued to Entrez, KiPar passes an Entrez query annotated with search tags, some of which indicate in which way MeSH should be utilized (see <a href="http://nlm.nih.gov/training/resources/meshtri.pdf">http://nlm.nih.gov/training/resources/meshtri.pdf</a> for detailed information).
Java class	KiPar.java
Name	<b>PubMed</b>
URL	<a href="http://www.ncbi.nlm.nih.gov/PubMed/">http://www.ncbi.nlm.nih.gov/PubMed/</a>
Description	PubMed is a bibliographic database that contains over 17 million article references from approximately 5,000 journals in biomedical and life sciences. It offers free access to abstracts of these articles.
Usage	PubMed is initially searched for abstracts relevant for individual concepts of interest, i.e. enzymes, compounds, genes, pathways and kinetics. Each concept is mapped to an OR query consisting of all synonyms retrieved earlier from public databases, e.g. the query used to search for information on enzyme with EC number 3.1.3.12 is:

"3.1.3.12"[TEXT:noexp] OR  
"trehalose 6-phosphatase"[TEXT:noexp] OR  
"trehalose 6-phosphate phosphatase"[TEXT:noexp] OR  
"trehalose-phosphatase"[TEXT:noexp]\*

These search results are used to index PubMed with individual concepts. This information is then stored in a local database and used to formulate more complex pathway-related queries, which are encoded as SQL queries over the local database. These queries score abstracts with respect to their relevance for a pathway of interest using the following weighting formula:

$$S = \frac{100}{\omega_{PATH} + \omega_{SBO}} \cdot \left[ \frac{\omega_{PATH}}{\omega_{GO} + \omega_{RN}} \cdot \left( \omega_{GO} \cdot hits(GO) + \frac{\omega_{RN}}{\omega_{EC} + \omega_{CPD} + \omega_{SCE}} \cdot \max_{e \in EC} \left\{ \begin{array}{l} \omega_{EC} \cdot hit(e) + \\ \omega_{CPD} \cdot hits(CPD_e) + \\ \omega_{SCE} \cdot hits(SCE_e) \end{array} \right\} \right) + \omega_{SBO} \cdot hits(SBO) \right]$$

where:

- *EC* is a set of enzymes for individual reactions from the pathway,

- $e$  is an enzyme from  $EC$ ,
- $CPD_e$  is a set of compounds (substrates and products) involved in the reaction catalysed by the enzyme  $e$ ,
- $SCE_e$  is a set of *S. cerevisiae* genes encoding the enzyme  $e$ ,
- $GO$  is a set of a pathway-related concepts from the Gene Ontology,
- $SBO$  is a set of a kinetics-related concepts from the Systems Biology Ontology,
- $hits(S)$  is the percentage of concepts in the set  $S$  matching the given document,  
 $\omega_C$  ( $0 \leq \omega_C \leq 100$ ) is the weight chosen for the concept  $C$  (e.g. we used these weights for enzymes ( $\omega_{EC} = 60.0$ ), compounds ( $\omega_{CPD} = 30$ ), genes ( $\omega_{SCE} = 10$ ), reactions ( $\omega_{RN} = 80$ ), GO terms ( $\omega_{GO} = 20$ ), pathway ( $\omega_{PATH} = 50$ ), kinetics ( $\omega_{SBO} = 50$ )).

\* See information on MeSH for an explanation of the search tags such as [TEXT:noexp].

Java class	KiPar.java, Entrez.java
Name	<b>PubMed Central</b>
URL	<a href="http://www.pubmedcentral.nih.gov/">http://www.pubmedcentral.nih.gov/</a>
Description	PubMed Central is a database that contains over 1 million full-text research articles from approximately 400 journals in biomedical and life sciences.
Usage	PubMed Central is initially searched for full-text articles relevant for individual concepts of interest. This is done in the same way described for PubMed above.
Java class	KiPar.java, Entrez.java
Name	<b>Entrez</b>
URL	<a href="http://www.ncbi.nlm.nih.gov/Entrez/">http://www.ncbi.nlm.nih.gov/Entrez/</a>
Web service	<a href="http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html">http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html</a>
Description	Entrez is an integrated search and retrieval system that provides access to a collection of NLM (National Library of Medicine) databases (including PubMed, PubMed Central and PubChem) simultaneously with a single query and user interface. Entrez query supports Boolean operators and search term tags to limit parts of the search statement to particular fields. Entrez search request returns a unified results page, that shows the number of hits in each database together with the links to actual search results for that particular database.
Usage	Entrez is used to access the literature stored in PubMed and PubMed Central. It is used to search and retrieve the literature via its web service.
Java class	Entrez.java