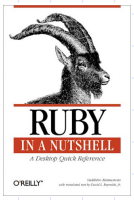
Bio**Ruby** + KEGG API + KEGG DAS = wiring knowledge for genome and pathway

Toshiaki Katayama <k@bioruby.org>
Human Genome Center, University of Tokyo, Japan
http://bioruby.org/
http://www.genome.jp/kegg/soap/
http://das.hgc.jp/

What is BioRuby?

- Yet another BioPerl written in Ruby
  - since Nov 2000
- Mainly developed in Japan
  - including supports for Japanese resources like KEGG

What is Ruby?



- Created by Japanese author 'matz'
- Scripting language
  - clean syntax, easy to learn, powerful enough
- Purely object oriented
  - Integer, String, Regexp, Exception etc. w/o exception
- Sufficient libraries
  - You can use most of the BioRuby functionallity without install additional libraries.

What you can do with BioRuby?

- Biological sequence manipulation
- Database entry retrieval and parsing
- Running and parsing usual applications
- Graph computation

Sequence manipulations

- Splicing
- Reverse complement
- Translation
- Composition
- Molecular weight
- Window search

```
# print in FASTA format
puts seq.to_fasta("foo", 60)
```

```
# for selenoproteins
ct =Bio::CodonTable.copy(1)
ct['tga'] = 'U'

puts seq.translate
puts seq.translate(1, ct)
```

```
#!/usr/bin/env ruby

require 'bio'

seq = Bio::Sequence::NA.new(ARGF.read)

puts seq.subseq(1,3)
puts seq.splicing("join(1..23,45..67)")
puts seq.complement
puts seq.translate
puts seq.gc_percent
puts seq.composition

seq.window_search(15, 3) do |subseq|
  peptide = subseq.translate
  puts peptide.molecular_weight
end
```


Entry parsers

- GenBank, GenPept, RefSeq, DDBJ
- EMBL, UniProt (TrEMBL, SwissProt)
- PDB
- KEGG/GenomeNet
  - GENES, GENOME, ENZYME, COMPOUND, KO, BRITE, CELL, Expression, Keggtab, AAindex
- GFF
- GO
- FANTOM
- Transfac, Prosite
- LITDB, MEDLINE
- NBRF, PIR
- FASTA format

```
#!/usr/bin/env ruby
# Usage:
# % auto.rb dbfile

require 'bio'

Bio::FlatFile.auto(ARGF) do |ff|
  ff.each do |entry|
    # do something
  end
end
```

Database access

- OBDA (Open Bio* Database Access)
  - BioRegistry
  - BioFlat
  - BioFetch
  - BioSQL
- PubMed
- DAS
  - Ensembl, WormBase etc.
  - KEGG DAS
- SOAP
  - KEGG API
  - DDBJ XML
  - NCBI ESOAP

OBDA configuration

- ~/.bioinformatics/seqdatabase.ini
- /etc/bioinformatics/seqdatabase.ini
- http://www.open-bio.org/registry/seqdatabase.ini

```
VERSION=1.00

[genbank]
protcol=flat
location=/export/database/
dbname=genbank

[swissprot]
protocol=biosql
location=db.bioruby.org
dbname=biosql
driver=mysql
biodbname=sp

[embl]
protocol=biofetch
location=http://bioruby.org/cgi-bin/biofetch.rb
dbname=embl
```

```
#!/usr/bin/env ruby

require 'bio'

reg = Bio::Registry.new

sp = reg.get_database('swissprot')
puts sp.get_by_id('CYC_BOVIN')

gb = reg.get_database('genbank')
puts gb.get_by_id('AA2CG')
```

Running applications

- Bio::Blast, Fasta, HMMER, EMBOSS
- Bio::ClustalW, MAFFT
- Bio::Genscan
- Bio::PSORT, TargetP
- Bio::SOSUI, TMHMM

```
#!/usr/bin/env ruby
require 'bio'

File.open("my_blast_output.xml") do |file|
  Bio::Blast.reports(file) do |report|
    report.hits do |hit|
      hit.each do |hsp|
        puts hsp.query_id, hsp.target_id,
        hsp.bit_score, hsp.evalue, hsp.overlap
    end
  end
end
```


What is KEGG? Kyoto Encyclopedia of Genes and Genomes

- http://www.genome.jp/kegg/
  - PATHWAY
    - LIGAND
  - GENES, GENOME
  - KO
  - SSDB
  - Expression
  - Glycan
  - etc.

KEGG PATHWAY

KEGG PATHWAY Database

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

KEGG PATHWAY Database

Current knowledge on molecular interaction networks, including metabolic pathways, regulatory pathways, and molecular complexes

KEGG2 PATHWAY GENES LIGAND EXPRESSION BRITE XML API D

Go to:
1. Metabolism
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Human Diseases

See also: KO (KEGG Orthology)

1. Metabolism

1.1 Carbohydrate Metabolism

Glycolysis / Gluconeogenesis	Ortholog, Oxidoreductases
Citrate cycle (TCA cycle)	Ortholog
Pentose phosphate pathway	Ortholog
Pentose and glucuronate interconversions	Ortholog
Fructose and mannose metabolism	Ortholog, PTS
Galactose metabolism	Ortholog
Ascorbate and aldarate metabolism	Ortholog
Pyruvate metabolism	Ortholog
Glyoxylate and dicarboxylate metabolism	Ortholog
Propanoate metabolism	Ortholog
Butanoate metabolism	Ortholog
C5-Branched dibasic acid metabolism	
Inositol metabolism	Ortholog

1.2 Energy Metabolism

Oxidative phosphorylation	Ortholog
ATP synthesis	Ortholog
Photosynthesis	Ortholog
Carbon fixation	
Reductive carboxylate cycle (CO2 fixation)	
Methane metabolism	Ortholog
Nitrogen metabolism	Ortholog
Sulfur metabolism	Ortholog

1.3 Lipid Metabolism

Fatty acid biosynthesis (path 1)	Ortholog
Fatty acid biosynthesis (path 2)	Ortholog

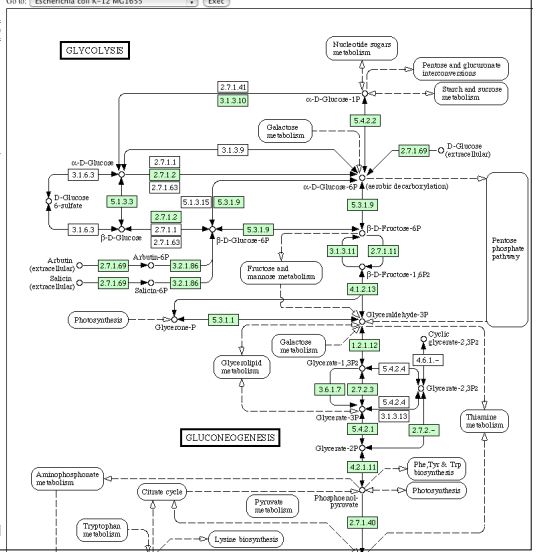
Glycolysis / Gluconeogenesis – Escherichia coli K-12 MG1655

http://www.genome.ad.jp/dbget-bin/get_pathway?org_name=eco&mapno=00010

Glycolysis / Gluconeogenesis - Escherichia coli K-12 MG1655

Go to: [LinkDB search | Ortholog Table]

Go to: Escherichia coli K-12 MG1655 Exec



KEGG GENES

DBGET Result: E.coli b0002

http://www.genome.ad.jp/dbget-bin/www_bget?eco:b0002

[LinkDB | SSDB | FASTA-genes | FASTA-sp | BLAST-nr | SOSUI | PSORT | MOTIF]

```
ENTRY        b0002            CDS       E.coli
NAME         thrA, thrA1, thrA2
DEFINITION   aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3]
ORTHOLOG     KO: K00003  homoserine dehydrogenase
             KO: K00928  aspartate kinase
CLASS        Metabolism; Amino Acid Metabolism; Glycine, serine and threonine
             metabolism [PATH:eco00260]
             Metabolism; Amino Acid Metabolism; Lysine biosynthesis
             [PATH:eco00300]
POSITION     337..2799
DBLINKS      Wisconsin: b0002
             Colibri: thrA
             NCBI: 1786183
             SPTR: P00561
CODON_USAGE          T              C              A              G
          T  11  19  10  13  11  10   6   9  12   8   0   0   3   9   1   4
          C   8  13   2  43   3   6   2  18   8   6  11  19  18  19   3   4
          A  30  15   1  23   5  19   2   8  22  16  22  12   3  12   0   2
          G  19  18   5  27  15  36  14  26  30  14  40  13  22  22   9  10
AASEQ        820
             MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDA
             LPNISDAERIFAELLTGLAAAQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINA
             ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP
             ADHMVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV
             PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD
             EDELPVKGISNLNNMAMFSVSGFPGMKGMVGMAARVFAAMSRARISVVLITQSSSEYSISF
             CVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIISVVGDGMRTLRGISAKFFAAL
             ARANINIVAIAQGSSERSISVVVNNDDATTGVRVTHQMLFNTDQVIEVFVIGVGGVGGAL
             LEQLKRQQSWLKNKHIDLRVCGVANSKALLTNVHGLNLENWQEELAQAKEPFNLGRLIRL
             VKEYHLLNFVIVDCTSSQAVADQYADFLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSR
             RKFLYDTNVGAGLFVIENLQNLLNAGDELMKFSGILSGSLSYIFGKLDEGMSFSEATTLA
             REMGYTEPDPRDDLSGMDVARKLLILARETGRELELADIEIEPVLPAEFNASGDVAAFMA
             NLSQLDDLFAARVAKARDEGKVLRYVGNIDEDGVCRVKIAEVDGNDPLFKVKNGENALAF
             YSHYYQPLPLVLRGYGAGNDVTAAGVFADLLRTLSWKLGV
NTSEQ         2463
             atgcgagtgttgaagttcggcggtacatcagtggcaaatgcagaacgttttctgcgtgtt
             gccgatattctggaaagcaatgccaggcaggggcaggtggccaccgtcctctctgccccc
             gccaaaatcaccaaccacctggtggcgatgattgaaaaaaccattagcggccaggatgct
             ttacccaatatcagcgatgccgaacgtatttttgccgaacttttgacgggactcgccgcc
             gcccagccggggttcccgctggcgcaattgaaaactttcgtcgatcaggaatttgcccaa
             ataaaacatgtcctgcatggcattagtttgttggggcagtgcccggatagcatcaacgct
             gcgctgatttgccgtggcgagaaaatgtcgatcgccattatggccggcgtattagaagcg
             cgcggtcacaacgttactgttatcgatccggtcgaaaaactgctggcagtggggcattac
```

LinkDB Se

http://www.genome.ad.jp/dbget-bin/get_

LinkDB Search Result

Database: LinkDB

```
Database of Link Information
Release 04-04-16, Apr 04
Institute for Chemical Research, Kyoto University
110,095,443 entries
```

From: E.coli:b0002
To : All

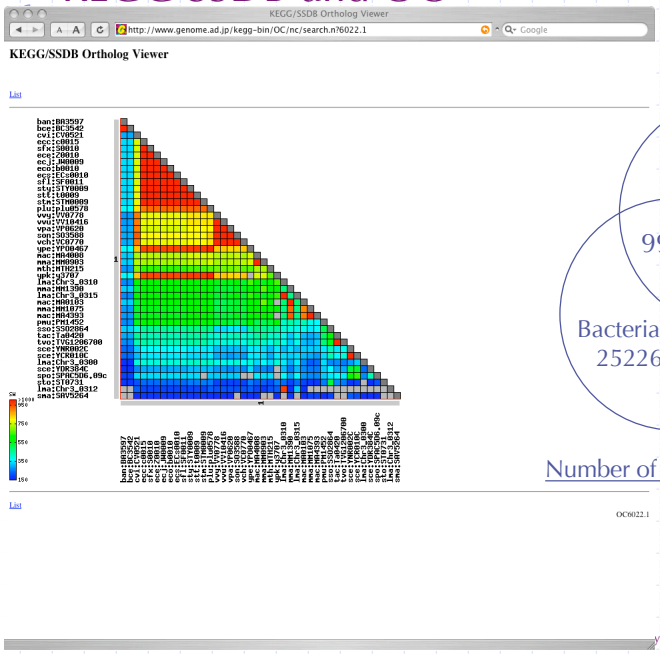
291 hits from 12+ databases

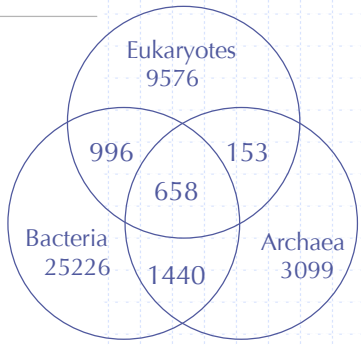
1. GenBank (159)
2. EMBL (22)
3. PIR (50)
4. PRF (25)
5. SPTrEMBL (1)
6. PDB (3)
7. PDBSTR (8)
8. KO (2)
9. COMPOUND (11)
10. ENZYME (2)
11. PATHWAY (4)
12. GENOME (1)
13. OTHERS (3)
14. All databases (291)

Link table for E.coli

DBGET integrated database retrieval system, GenomeNet

KEGG SSDB and OC



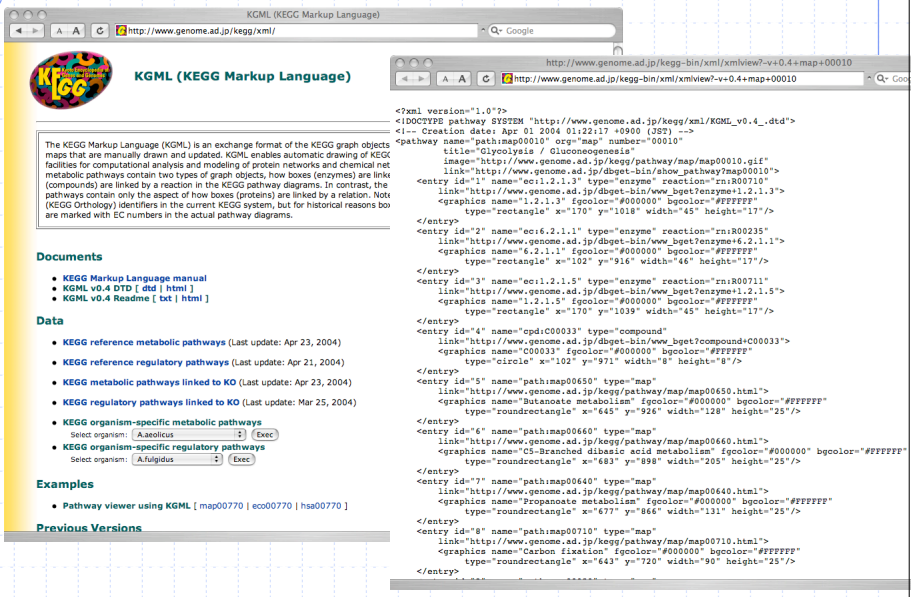


Number of OCs across 3 kingdoms

KEGG standardization

- KGML
  - ISMB2004 poster C-45
- KEGG API
  - ISMB2004 poster G-17
- KEGG DAS
  - ISMB2004 poster E-29

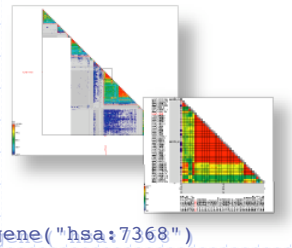
KGML



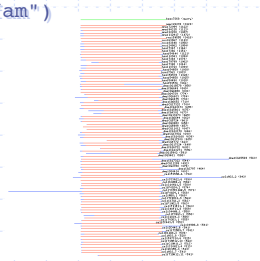
KEGG API

- http://www.genome.jp/kegg/soap/
- SOAP/WSDL based web service
  - XML, HTTP
  - Can be accessed by any language
- Proteome and pathway analysis
  - KEGG GENES, SSDB, PATHWAY
  - DBGET, LinkDB
- Updated to v3.0 (May 2004)

Example: obtain homologs and motifs by SSDB



```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

homologs = serv.get_all_best_neighbors_by_gene("hsa:7368")

homologs.each do |hit|
  gene = hit.genes_id2
  if motifs = serv.get_motifs_by_gene(gene, "pfam")
    motifs.each do |motif|
      name = motif.motif_id
      desc = motif.definition
      puts "#{gene}: #{name} #{desc}"
    end
  end
end
```



Example: KO, OC, PC

```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

ko_list = serv.get_ko_by_gene("eco:b0002")

# list of genes assigned same KO (KEGG orthology)
ko_genes = serv.get_ko_members(ko_list.first)

# list of genes assigned to same OC (ortholog cluster)
oc_genes = serv.get_all_oc_members_by_gene("hsa:7368")

# list of genes assigned to same PC (paralog cluster)
pc_genes = serv.get_all_pc_members_by_gene("hsa:7368")

puts "# KO", ko_genes, "# OC", oc_genes, "# PC", pc_genes
```

Example: getting database info/entries

```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

# lists available organisms in KEGG
orgs = serv.list_organisms
orgs.each do |entry|
  puts "#{entry.entry_id} #{entry.definition}"
end

# lists available pathways in KEGG
list = serv.list_pathways("hsa")
list.each do |entry|
  puts "#{entry.entry_id} #{entry.definition}"
end

# getting EMBL entry
puts serv.bget("embl:BUM")

# getting entries from KEGG GENES
puts serv.bget("hsa:7368 hsa:7369")
```


Example: using PATHWAY

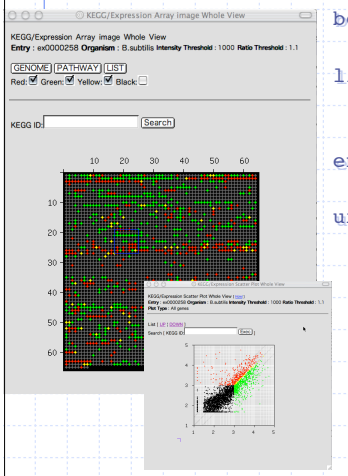
```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

# lists gene ids on pathway
genes = serv.get_genes_by_pathway("path:hsa00020")
puts "# genes on human's pathway 00020"
genes.each do |gene|
  puts gene
end

# converts EC numbers to genes
list = ["ec:1.1.1.1", "ec:1.2.1.1"]
list.each do |ec|
  puts "# E. coli genes for #{ec}"
  puts serv.get_genes_by_enzyme(ec, "eco")
end
```

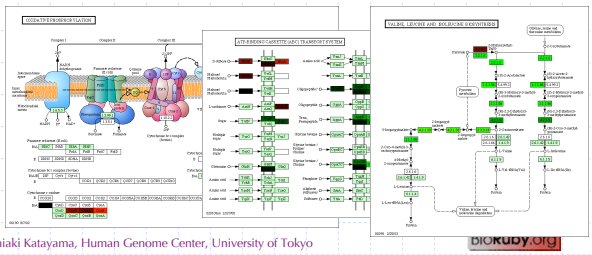
Example: coloring PATHWAY

```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

# mark pathway
objs = ['eco:b0002', 'cpd:C00263']
url1 = serv.mark_pathway_by_objects('path:eco00260', objs)

# color pathway
fg_list = ['blue', 'green']
bg_list = ['#ff0000', 'yellow']
url2 = serv.color_pathway_by_objects('path:eco00260',
        objs, fg_list, bg_list)

# save the result images
serv.save_image(url1, "marked_pathway.gif")
serv.save_image(url2, "colored_pathway.gif")
```


Example: gene expression and pathway analysis

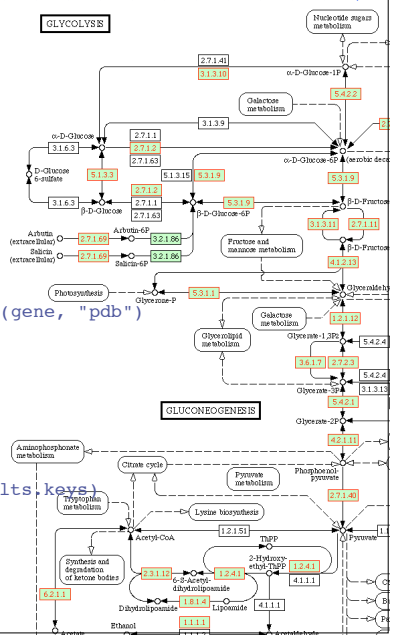
```
serv = Bio::KEGG::API.new

list = serv.get_genes_by_pathway("path:bsu00020")
fg_colors = Array.new
bg_colors = Array.new

list.each do |gene|
  fg_colors << "black"
  bg_colors << ratio2rgb(gene)
end

url = serv.color_pathway_by_objects(
  "path:bsu00020", list, fg_colors, bg_colors)
```





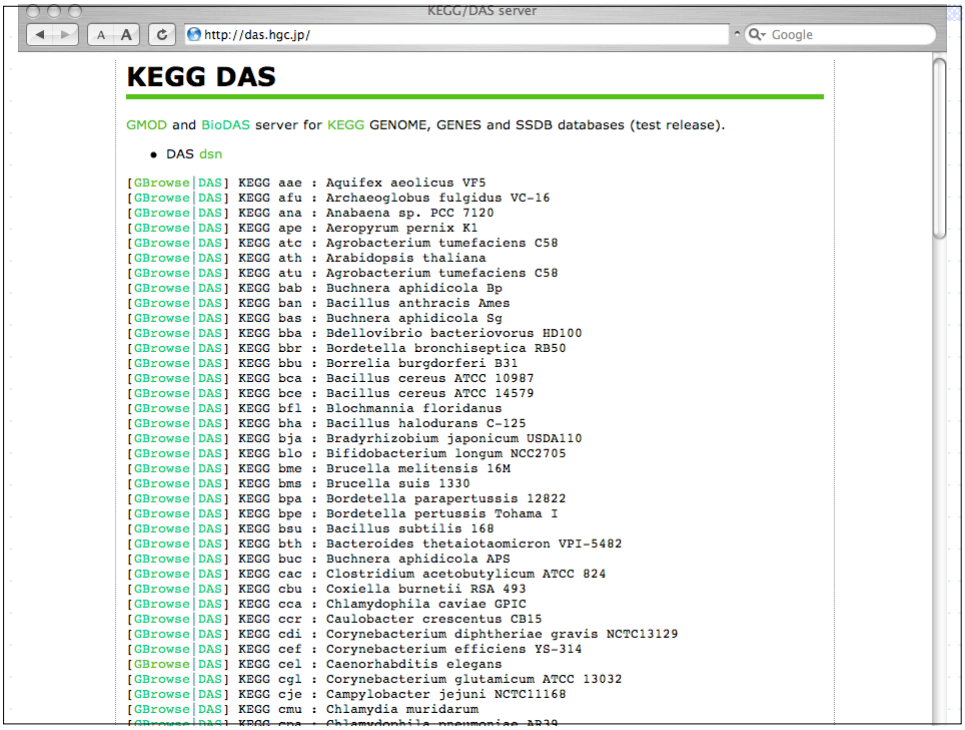
Example: mapping PDB on pathway

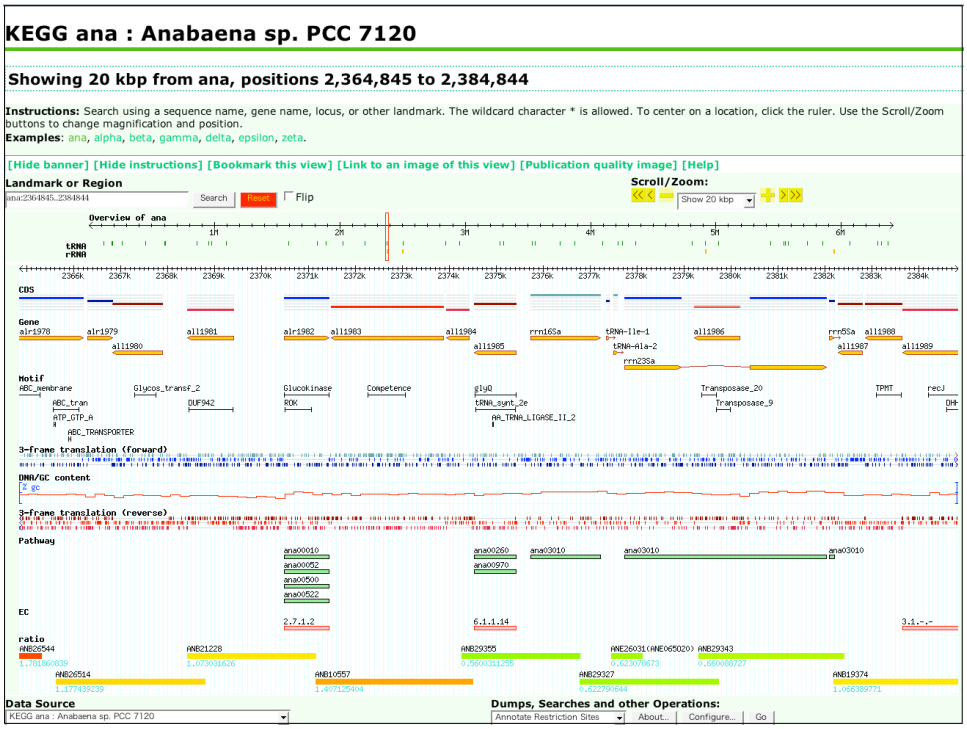
```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::KEGG::API.new

# list of genes on specified pathway
path = ARGV.shift || "path:eco00010"
genes = serv.get_genes_by_pathway(path)

# search DB links to PDB using LinkDB
results = Hash.new
genes.each do |gene|
  if pdb_links = serv.get_all_linkdb_by_entry(gene, "pdb")
    pdb_links.each do |link|
      results[gene] = true
    end
  end
end

# generates colored image
url = serv.mark_pathway_by_objects(path, results.keys)

# save the image
serv.save_image(url, "linked_to_pdb.gif")
```

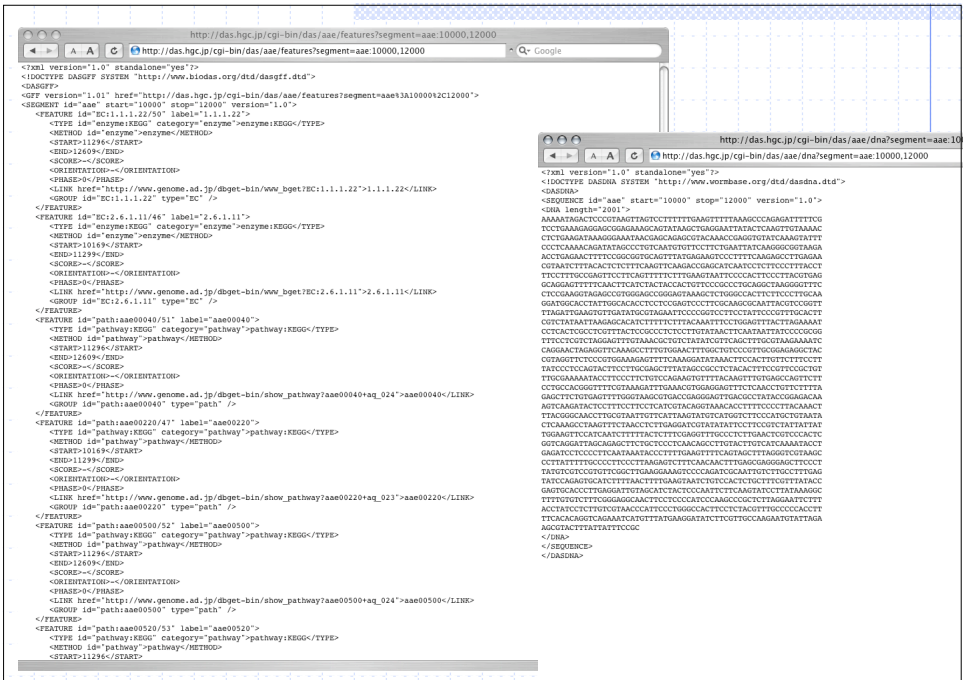


KEGG DAS

- http://das.hgc.jp/
  - Currently including 188 organisms
  - Data from KEGG GENOME and GENES
  - Build on to of the GMOD/GBrowse


KEGG/DAS server
http://das.hgc.jp/
Google
KEGG DAS
GMOD and BioDAS server for KEGG GENOME, GENES and SSDB databases (test release).
DAS dsn
[GBrowse|DAS] KEGG aae : Aquifex aeolicus VF5
[GBrowse|DAS] KEGG afu : Archaeoglobus fulgidus VC-16
[GBrowse|DAS] KEGG ana : Anabaena sp. PCC 7120
[GBrowse|DAS] KEGG ape : Aeropyrum pernix K1
[GBrowse|DAS] KEGG atc : Agrobacterium tumefaciens C58
[GBrowse|DAS] KEGG ath : Arabidopsis thaliana
[GBrowse|DAS] KEGG atu : Agrobacterium tumefaciens C58
[GBrowse|DAS] KEGG bab : Buchnera aphidicola Bp
[GBrowse|DAS] KEGG ban : Bacillus anthracis Ames
[GBrowse|DAS] KEGG bas : Buchnera aphidicola Sg
[GBrowse|DAS] KEGG bba : Bdellovibrio bacteriovorus HD100
[GBrowse|DAS] KEGG bbr : Bordetella bronchiseptica RB50
[GBrowse|DAS] KEGG bbu : Borrelia burgdorferi B31
[GBrowse|DAS] KEGG bca : Bacillus cereus ATCC 10987
[GBrowse|DAS] KEGG bce : Bacillus cereus ATCC 14579
[GBrowse|DAS] KEGG bfl : Blochmannia floridanus
[GBrowse|DAS] KEGG bha : Bacillus halodurans C-125
[GBrowse|DAS] KEGG bja : Bradyrhizobium japonicum USDA110
[GBrowse|DAS] KEGG blo : Bifidobacterium longum NCC2705
[GBrowse|DAS] KEGG bme : Brucella melitensis 16M
[GBrowse|DAS] KEGG bms : Brucella suis 1330
[GBrowse|DAS] KEGG bpa : Bordetella parapertussis 12822
[GBrowse|DAS] KEGG bpe : Bordetella pertussis Tohama I
[GBrowse|DAS] KEGG bsu : Bacillus subtilis 168
[GBrowse|DAS] KEGG bth : Bacteroides thetaiotaomicron VPI-5482
[GBrowse|DAS] KEGG buc : Buchnera aphidicola APS
[GBrowse|DAS] KEGG cac : Clostridium acetobutylicum ATCC 824
[GBrowse|DAS] KEGG cbu : Coxiella burnetii RSA 493
[GBrowse|DAS] KEGG cca : Chlamydophila caviae GPIC
[GBrowse|DAS] KEGG ccr : Caulobacter crescentus CB15
[GBrowse|DAS] KEGG cdi : Corynebacterium diphtheriae gravis NCTC13129
[GBrowse|DAS] KEGG cef : Corynebacterium efficiens YS-314
[GBrowse|DAS] KEGG cel : Caenorhabditis elegans
[GBrowse|DAS] KEGG cgl : Corynebacterium glutamicum ATCC 13032
[GBrowse|DAS] KEGG cje : Campylobacter jejuni NCTC11168
[GBrowse|DAS] KEGG cmu : Chlamydia muridarum


KEGG ana : Anabaena sp. PCC 7120
Showing 20 kbp from ana, positions 2,364,845 to 2,384,844
Instructions: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.
Examples: ana, alpha, beta, gamma, delta, epsilon, zeta.
[Hide banner] [Hide instructions] [Bookmark this view] [Link to an image of this view] [Publication quality image] [Help]
Landmark or Region
Search
Reset
Flip
Scroll/Zoom:
Show 20 kbp
Overview of ana
tRNA
rRNA
CDS
Gene
Motif
3-frame translation (forward)
DNA/GC content
3-frame translation (reverse)
Pathway
EC
ratio
Data Source
KEGG ana : Anabaena sp. PCC 7120
Dumps, Searches and other Operations:
Annotate Restriction Sites
About...
Configure...
Go



Example: sequence and annotations by DAS

```
#!/usr/bin/env ruby

require 'bio'

serv = Bio::DAS.new("http://das.hgc.jp/cgi-bin/")

segment = Bio::DAS::SEGMENT.region("I", 1001, 2000)

# get DNA sequence from S. cerevisiae
list = serv.get_dna("sce", segment)
list.each do |dna|
  puts dna.sequence
end

# get features
list = serv.get_features("sce", segment)
list.segments.each do |segment|
  segment.features.each do |feature|
    puts feature.entry_id
    puts feature.start
  end
end
```

Acknowledgements

- BioRuby developers
  - Naohisa Goto, Mitsuteru Nakao, Yoshinori Okuji, Shuichi Kawashima, Masumi Itoh, Alex Gutteridge and some other contributors on the net.
- KEGG curators and KEGG API developers
  - Bioinformatics center, Human genome center
  - Yoko Sato, Miho Matsubayashi, Satoshi Miyazaki
- KEGG DAS
  - Mayumi Takashio, Mari Watanabe
- Open Bio* community

URL

- BioRuby
  - http://bioruby.org/
  - CVS, ML - hosted at open-bio.org
- KEGG API
  - http://www.genome.jp/kegg/soap/
- KEGG DAS
  - http://das.hgc.jp/
- ISMB posters at E-29 G-17 C-45