bioRxiv preprint doi: https://doi.org/10.1101/416131; this version posted September 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

KPGminer: A tool for retrieving pathway genes from KEGG pathway database

A. K. M. Azad

School of Biotechnology and Biomolecular Sciences, University of NSW, Chancellery Walk, Kensington, 2033, Australia

Abstract

Pathway analysis is a very important aspect in computational systems biology as it serves as a crucial component in many computational pipelines. KEGG is one of the prominent databases that host pathway information associated with various organisms. In any pathway analysis pipelines, it is also important to collect and organize the pathway constituent genes for which a tool to automatically retrieve that would be a useful one to the practitioners. In this article, I present KPGminer, a tool that retrieves the constituent genes in KEGG pathways for various organisms and organizes that information suitable for many downstream pathway analysis pipelines. We exploited several KEGG web services using REST APIs, particularly GET and LIST methods to request for the information retrieval which is available for developers. Moreover, KPGminer can operate both for a particular pathway (single mode) or multiple pathways (batch mode). Next, we designed a crawler to extract necessary information from the response and generated outputs accordingly. KPGminer brings several key features including organism-specific and pathway-specific extraction of pathway genes from KEGG and always up-to-date information. Thus, we hope KPGminer can be a useful and effective tool to make downstream pathway analysis easier and faster. KPGminer is freely available for download from https://sourceforge.net/projects/kpgminer/.

Keywords: GSEA, KEGG pathway, pathway analysis, pathway genes, information retrieval, Web API, REST

bioRxiv preprint doi: https://doi.org/10.1101/416131; this version posted September 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

1. Introduction

Biological pathway is defined as a collection of genes or proteins that are functionally related to each others to perform some biological activities such as signaling or regulatory activities. Some of the on-line pathway databases are KEGG [1], Reactome [2], Wikipathways [3] etc. where pathways related to signaling, metabolomic, cellular processes, diseases, genetic information are stored for various organism.

Pathway analysis is an important downstream component for many bioinformatics pipelines. One of the important aspects of a pathway analysis task set is to conduct enrichment test with already annotated pathways. This enrichment analysis include evaluating the enrichment of *de novo* gene sets (either computationally predicted or experimentally determined) with those already annotated pathways. Azad *et al.* designed a method called VToD [4] for identifying cancer-related gene modules, which were validated with known pathways from databases including KEGG [1] and GO terms [5] using gene set enrichment test. Another example of gene set enrichment analysis is to check the overlap of a particular set of interest e.g. differentially expressed genes with those annotated pathways. All of these enrichment tests require a set annotated pathways presented in the databases such as KEGG [1]. One of the sources to collect such annotated pathway sets is the Molecular Signatures Database (MSigDB) [6] which stores gene sets from well known pathway databases like KEGG [1] and Reactome [2]. But these gene sets are static and the pathway annotations are always updating. Hence, it is required to have a tool for retrieving up-to-date gene collections for users. Moreover, those collections aren't organism-specific.

In this article, I present a standalone tool called *KPGminer* that retrieves the pathway genes from KEGG [1] for all the organisms every time it runs. This provide always up-to-date and organisms specific information which can be stored in the local machine for conducting downstream pathway analysis using some statistical methods such as hyper-geometric tests. I hope, this tool can be very useful for the researchers and contribute to their bioinformatics pipelines.

2. Implementation

Figure 1 shows the main KPGminer interface. When user opens KPGminer tool it loads all the available organisms in KEGG database by making

bioRxiv preprint doi: https://doi.org/10.1101/416131; this version posted September 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

HTTP web request via a web API call using REST protocol. This REST API protocol is available in the KEGG website for the developers' use. Once, all the organisms are loaded successfully, the main interface of KPGminer pops-up will a dropdown box populated with all the organism name. Next, user has to select a particular organism for the that list, another HTTP web request takes place for retrieving all the pathways currently available in KEGG database for that particular organism. The response of that request is then parsed to get the list of those pathways and a listbox gets populated with them. User can pick one or more pathways from that list which will be shown in another listbox (called selection listbox).

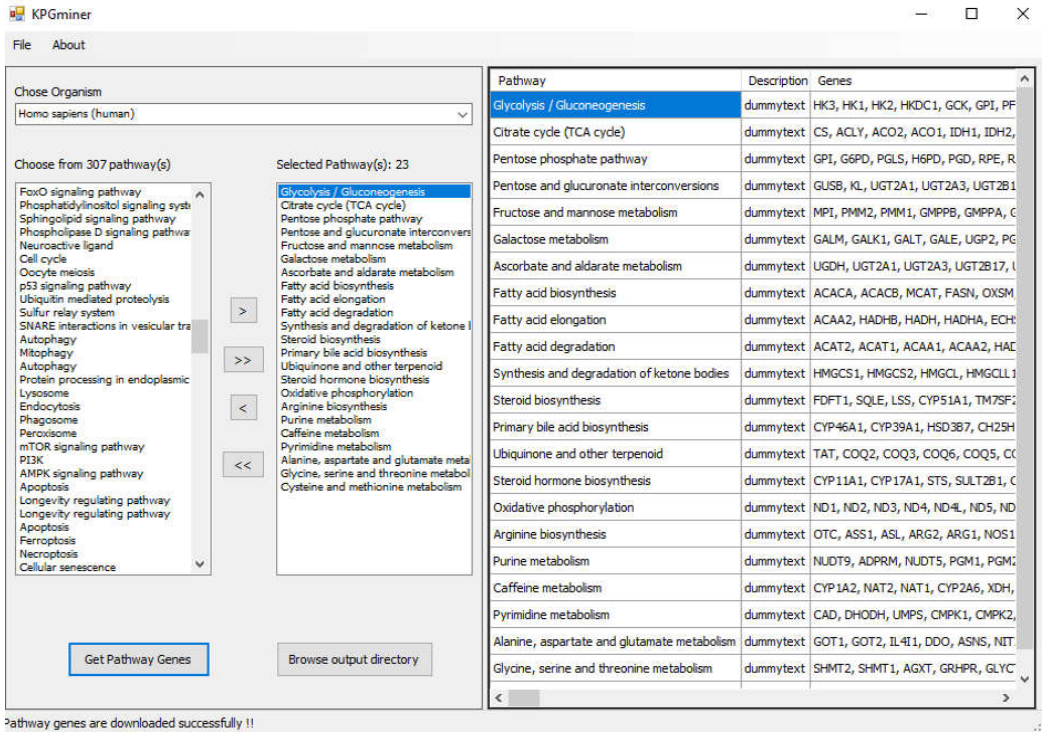


Figure 1: Main interface of KPGminer tool

To get all the pathway genes for those selected pathways, user press a button which makes another HTTP web request. KPGminer reports pathway genes in single or batch mode depending on the number of selected pathways. Once loaded all the pathway genes the results are shown in the right panel on the main KPGminer interface. Finally, to all of these pathway genes are

bioRxiv preprint doi: https://doi.org/10.1101/416131; this version posted September 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

can be saved in a file by clicking a button which asks a place to save that file in the local directory. The file is saved with a *.gmt* extension just as similar to the MSigDB for the convenience of users. A tooltip label keeps providing messages for every stages of KPGminer in retrieving pathway genes for selected pathway(s) for a particular organism. Table 1 shows the KPGminer metadata.

	Technology used
Version	v1.0.0
Language	C#.Net
Platform	Microsoft .Net platform
Operative systems	Windows
HTTP web request	REST API
Information retrieval technique	In-house built crawler

Table 1: KPGminer Metadata

3. Discussion

KPGminer has several useful features. First, even though KEGG provided necessary APIs for retrieving those information, a single platform to facilitate organism-specific and pathway-specific (single or batch mode) information retrieval may be advantageous for practitioners by abstracting their corresponding lower-level implementations. Second, while loading, KPGminer starts requesting KEGG databases for pathway information, which indicates that it always brings the up-to-date information. Third, KPGminer is a open source and free software that can help scientific communities to conduct pathway analysis required with KEGG pathway databases.

In this version of KGPminer, there is one limitation which is in batch-mode (for multiple pathways) operation, it creates HTTP web request for each pathways separately, which is a time consuming. But this limitation can be overcome by exploiting multi-threading approach by making each HTTP web request running in a single thread, which can be implemented in future versions of KPGminer. In future I also hope to extend this tool for retrieving information from other pathway databases including Reactome [2], Wikipathways [3] or GO [5] database. I hope KPGminer can be a very useful tool for the researchers in their pathway analysis.

bioRxiv preprint doi: https://doi.org/10.1101/416131; this version posted September 13, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

4. References

[1] M. Kanehisa, The KEGG database, Novartis Found. Symp. 247 (2002) 91–101.

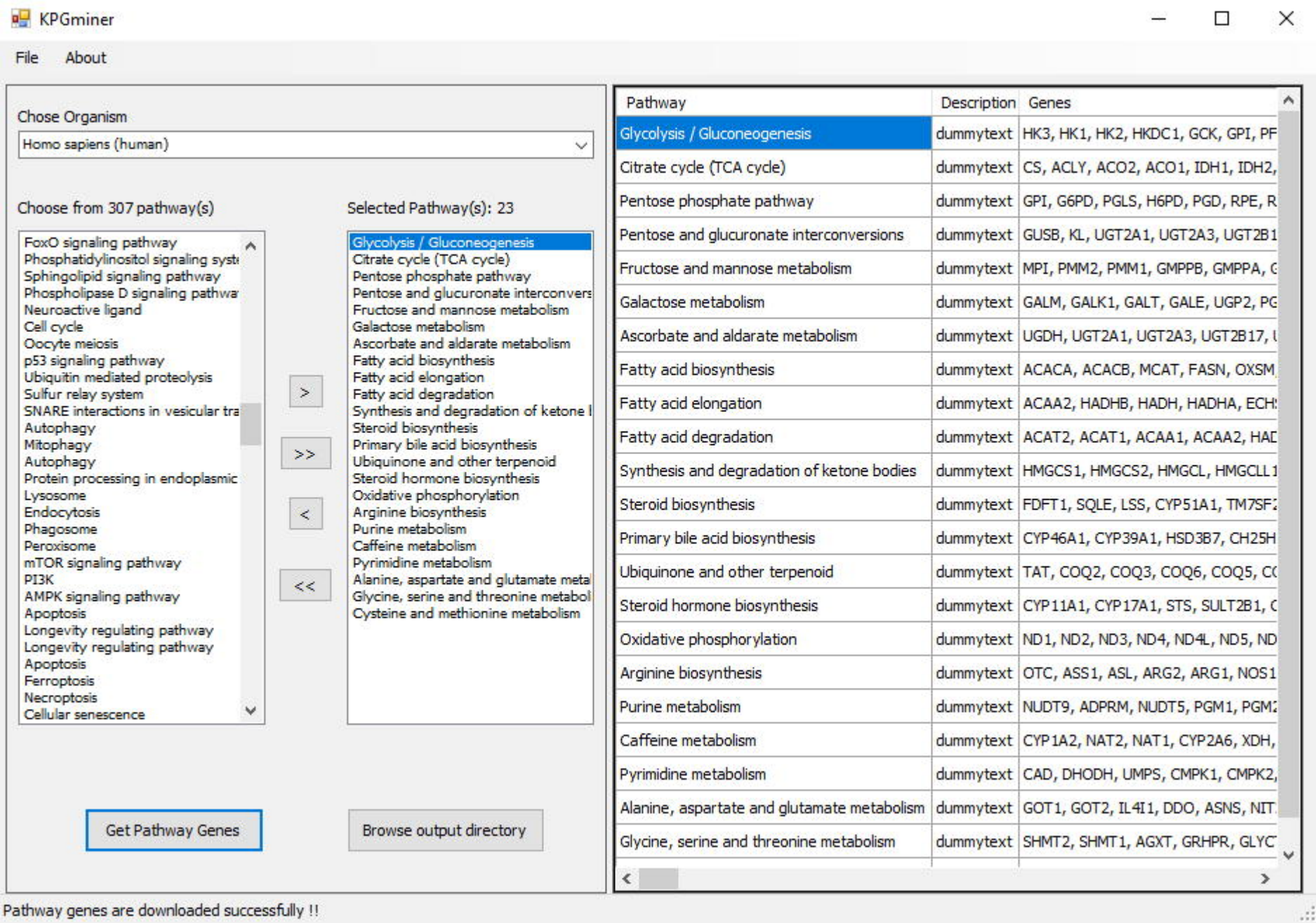
[2] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, P. D'Eustachio, The Reactome pathway knowledgebase, Nucleic Acids Res. 42 (2014) D472–477.

[3] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, A. R. Pico, WikiPathways: building research communities on biological pathways, Nucleic Acids Res. 40 (2012) D1301–1307.

[4] A. K. Azad, H. Lee, Voting-based cancer module identification by combining topological and data-driven properties, PLoS ONE 8 (2013) e70498.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29.

[6] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences 102 (2005) 15545–15550.


KPGminer
File
About
Chose Organism
Homo sapiens (human)
Choose from 307 pathway(s)
FoxO signaling pathway
Phosphatidylinositol signaling syste
Sphingolipid signaling pathway
Phospholipase D signaling pathwa
Neuroactive ligand
Cell cycle
Oocyte meiosis
p53 signaling pathway
Ubiquitin mediated proteolysis
Sulfur relay system
SNARE interactions in vesicular tra
Autophagy
Mitophagy
Autophagy
Protein processing in endoplasmic
Lysosome
Endocytosis
Phagosome
Peroxisome
mTOR signaling pathway
PI3K
AMPK signaling pathway
Apoptosis
Longevity regulating pathway
Longevity regulating pathway
Apoptosis
Ferroptosis
Necroptosis
Cellular senescence
>
>>
<
<<
Selected Pathway(s): 23
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronate interconvers
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Fatty acid biosynthesis
Fatty acid elongation
Fatty acid degradation
Synthesis and degradation of ketone
Steroid biosynthesis
Primary bile acid biosynthesis
Ubiquinone and other terpenoid
Steroid hormone biosynthesis
Oxidative phosphorylation
Arginine biosynthesis
Purine metabolism
Caffeine metabolism
Pyrimidine metabolism
Alanine, aspartate and glutamate meta
Glycine, serine and threonine metabol
Cysteine and methionine metabolism
Get Pathway Genes
Browse output directory
Pathway
Description
Genes
Glycolysis / Gluconeogenesis
dummytext
HK3, HK1, HK2, HKDC1, GCK, GPI, PF
Citrate cycle (TCA cycle)
dummytext
CS, ACLY, ACO2, ACO1, IDH1, IDH2,
Pentose phosphate pathway
dummytext
GPI, G6PD, PGLS, H6PD, PGD, RPE, R
Pentose and glucuronate interconversions
dummytext
GUSB, KL, UGT2A1, UGT2A3, UGT2B1
Fructose and mannose metabolism
dummytext
MPI, PMM2, PMM1, GMPPB, GMPPA, G
Galactose metabolism
dummytext
GALM, GALK1, GALT, GALE, UGP2, PG
Ascorbate and aldarate metabolism
dummytext
UGDH, UGT2A1, UGT2A3, UGT2B17,
Fatty acid biosynthesis
dummytext
ACACA, ACACB, MCAT, FASN, OXSM
Fatty acid elongation
dummytext
ACAA2, HADHB, HADH, HADHA, ECH
Fatty acid degradation
dummytext
ACAT2, ACAT1, ACAA1, ACAA2, HAD
Synthesis and degradation of ketone bodies
dummytext
HMGCS1, HMGCS2, HMGCL, HMGCLL1
Steroid biosynthesis
dummytext
FDFT1, SQLE, LSS, CYP51A1, TM7SF2
Primary bile acid biosynthesis
dummytext
CYP46A1, CYP39A1, HSD3B7, CH25H
Ubiquinone and other terpenoid
dummytext
TAT, COQ2, COQ3, COQ6, COQ5, CO
Steroid hormone biosynthesis
dummytext
CYP11A1, CYP17A1, STS, SULT2B1, C
Oxidative phosphorylation
dummytext
ND1, ND2, ND3, ND4, ND4L, ND5, ND
Arginine biosynthesis
dummytext
OTC, ASS1, ASL, ARG2, ARG1, NOS1
Purine metabolism
dummytext
NUDT9, ADPRM, NUDT5, PGM1, PGM2
Caffeine metabolism
dummytext
CYP1A2, NAT2, NAT1, CYP2A6, XDH,
Pyrimidine metabolism
dummytext
CAD, DHODH, UMPS, CMPK1, CMPK2,
Alanine, aspartate and glutamate metabolism
dummytext
GOT1, GOT2, IL4I1, DDO, ASNS, NIT
Glycine, serine and threonine metabolism
dummytext
SHMT2, SHMT1, AGXT, GRHPR, GLYC
Pathway genes are downloaded successfully !!