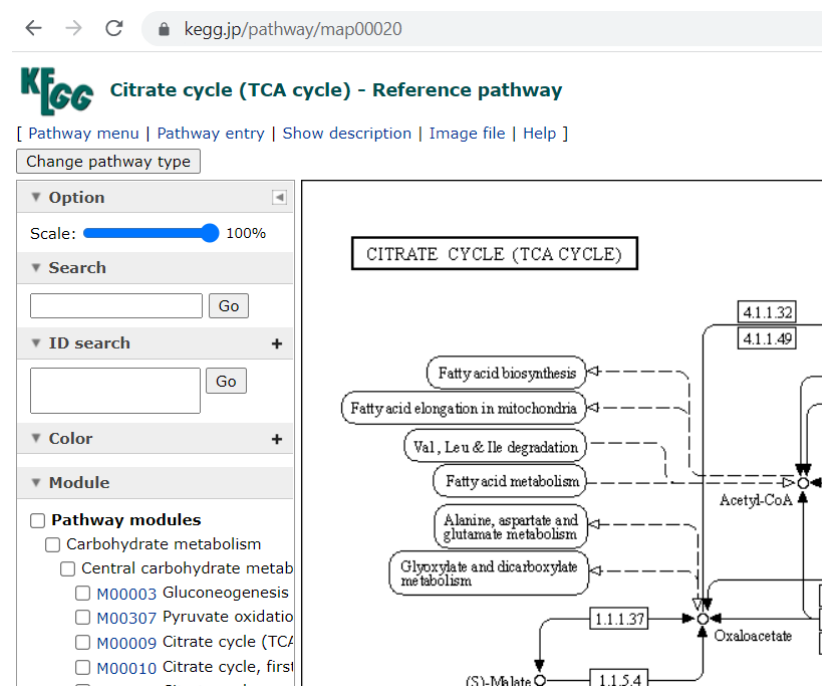


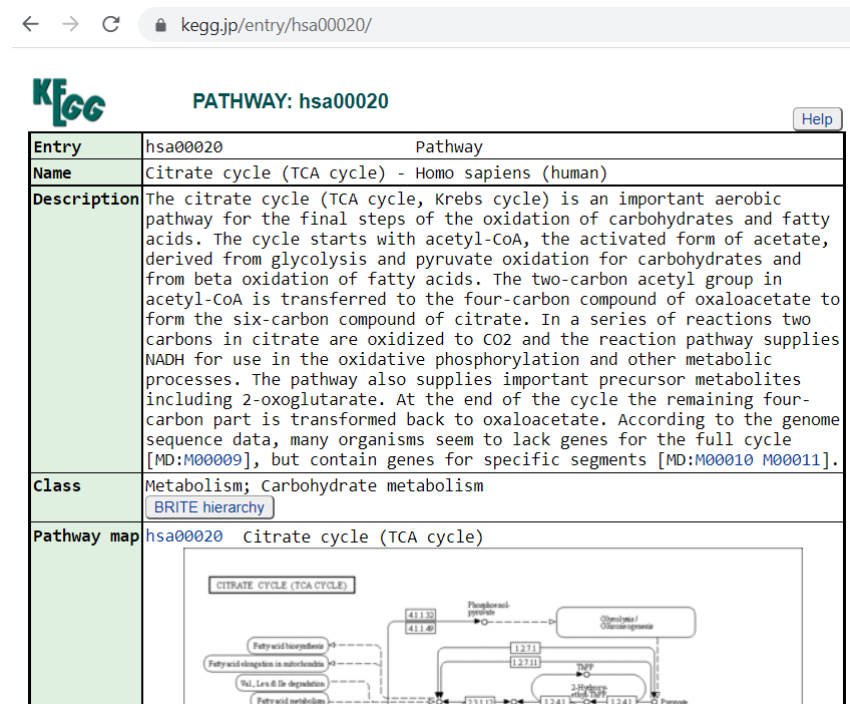
KEGG Pathways Dataset

Συλλέγουμε δεδομένα από τη βάση δεδομένων Pathway του KEGG.

Κάθε entry είναι ένα map, είναι γενικό και μπορεί να αφορά πολλούς οργανισμούς. Το url για την προβολή του είναι όπως παρακάτω:



Αν θέλουμε να ασχοληθούμε συγκεκριμένα με τι ισχύει στον ανθρώπινο οργανισμό, μπορούμε να κρατήσουμε τον κωδικό του map (εδώ 00020) και να αλλάξουμε το url ως:



Η ίδια πληροφορία από το rest api KEGG:

←	→	↺	⚠ Not secure rest.kegg.jp/get/hsa00020
ENTRY	hsa00020	Pathway	
NAME	Citrate cycle (TCA cycle) - Homo sapiens (human)		
DESCRIPTION	The citrate cycle (TCA cycle, Krebs cycle) is an im of acetate, derived from glycolysis and pyruvate oxidation for oxaloacetate to form the six-carbon compound of citrate. In a s other metabolic processes. The pathway also supplies important According to the genome sequence data, many organisms seem to l		
CLASS	Metabolism; Carbohydrate metabolism		
PATHWAY_MAP	hsa00020 Citrate cycle (TCA cycle)		
MODULE	hsa_M00003 Gluconeogenesis, oxaloacetate => fructo hsa_M00009 Citrate cycle (TCA cycle, Krebs cycle) hsa_M00010 Citrate cycle, first carbon oxidation, hsa_M00011 Citrate cycle, second carbon oxidation, hsa_M00307 Pyruvate oxidation, pyruvate => acetyl-		
DRUG	D10691 Bempedoic acid (JAN/USAN/INN) D11090 Ivosidenib (USAN/INN) D11834 Vorasidenib (USAN/INN) D11835 Vorasidenib citrate (USAN)		
DBLINKS	GO: 0006099		
ORGANISM	Homo sapiens (human) [GN:hsa]		
GENE	1431 CS; citrate synthase [K0:K01647] [EC:2.3.3.1] 47 ACLY; ATP citrate lyase [K0:K01648] [EC:2.3.3.8 50 1688 Citrate synthase [K0:K01647] [EC:2.3.3.1]		

Από το api, μας δίνεται πρόσβαση στο αρχείο kgml (αν υπάρχει) του κάθε map. Είναι ένα αρχείο xml το οποίο περιλαμβάνει entries και relations μεταξύ αυτών.

←	→	↺	⚠ Not secure rest.kegg.jp/get/hsa00020/kgml
This XML file does not appear to have any style information associated with it. The document tree is shown below.			
<pre><!-- Creation date: Sep 22, 2021 10:21:10 +0900 (GMT+9) --> ▼<pathway name="path:hsa00020" org="hsa" number="00020" title="Citrate cycle (TCA cycle)" image="https://www.kegg.jp/kegg/ ▼<entry id="33" name="hsa:1738" type="gene" reaction="rn:R07618" link="https://www.kegg.jp/dbget-bin/www_bget?hsa:1738"> <graphics name="DLD, DLDD, DLDH, E3, GCSL, LAD, OGDC-E3, PHE3" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x= </entry> ▼<entry id="34" name="hsa:4967 hsa:55753" type="gene" reaction="rn:R00621" link="https://www.kegg.jp/dbget-bin/www_bget? <graphics name="OGDH, AKGDH, E1k, KGD1, OGDC, OGDH2..." fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x="661" y </entry> ▼<entry id="35" name="hsa:4967 hsa:55753" type="gene" reaction="rn:R03316" link="https://www.kegg.jp/dbget-bin/www_bget? <graphics name="OGDH, AKGDH, E1k, KGD1, OGDC, OGDH2..." fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x="530" y </entry> ▼<entry id="36" name="hsa:1743" type="gene" reaction="rn:R02570" link="https://www.kegg.jp/dbget-bin/www_bget?hsa:1743"> <graphics name="DLST, DLTS, KGD2, PGL7" fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x="403" y="579" width="46 </entry> ▼<entry id="37" name="hsa:8801 hsa:8802 hsa:8803" type="gene" reaction="rn:R00405" link="https://www.kegg.jp/dbget-bin/w <graphics name="SUCLG2, G-SCS, GBETA, GTPSCS..." fgcolor="#000000" bgcolor="#BFFFFB" type="rectangle" x="260" y="579"</pre>			

Με ένα script (prep.py), κατεβάζουμε όλα τα αρχεία kgml που υπάρχουν από την βάση δεδομένων. Βρίσκουμε ότι 207/555 maps δεν αφορούν τον ανθρώπινο οργανισμό, άλλοι 35/555 maps έχουν καταγεγραμμένα entries αλλά δεν έχουν πίνακα relations μεταξύ τους. Έτσι, τα χρήσιμα maps είναι 313/555, δηλαδή έχουν και entries και relations μεταξύ τους. Για κάθε map, σώζουμε σε ξεχωριστό αρχείο τα entries και τα relations που το αφορούν.

Το αρχείο entries έχει τη μορφή:

id	name	type	link	gene_names
18	hsa:226 hsa:229 hsa:230	gene	https://www.kegg.jp/dbget-bin/	ALDOA, ALDA, GSD12, HEL-S-87p
42	hsa:217 hsa:219 hsa:223 h	gene	https://www.kegg.jp/dbget-bin/	ALDH2, ALDH-E2, ALDHI, ALDM..
45	cpd:C00033	compound	https://www.kegg.jp/dbget-bin/	C00033
46	path:hsa00030	map	https://www.kegg.jp/dbget-bin/	Pentose phosphate pathway

Το αρχείο relations έχει τη μορφή:

entry1	entry2	link	value	name	pathway
73	75	ECrel	90	compound	hsa00010
73	74	ECrel	90	compound	hsa00010
67	47	maplink	88	compound	hsa00010
67	46	maplink	94	compound	hsa00010

Μια εξήγηση των παραπάνω υπάρχει στο link: <https://www.kegg.jp/kegg/xml/docs/>

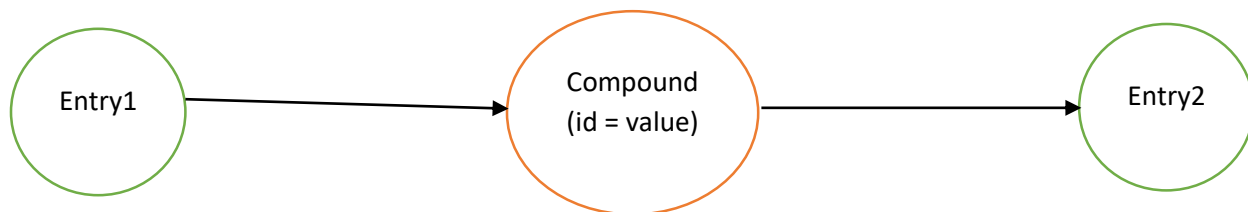
Μερικές απο τις πληροφορίες που μας αφορούν:

attribute name	data type	explanation
entry1	idref.type	the first (from) entry that defines this relation
entry2	idref.type	the second (to) entry that defines this relation
type	relation-type.type	the type of this relation

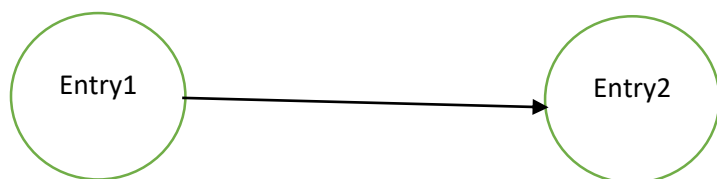
attribute value	explanation
ECrel	enzyme-enzyme relation, indicating two enzymes catalyzing successive reaction steps
PPrel	protein-protein interaction, such as binding and modification
GErel	gene expression interaction, indicating relation of transcription factor and target gene product
PCrel	protein-compound interaction
maplink	link to another map

name	value	ECrel	PPrel	GErel	Explanation
compound	Entry element id attribute value for compound.	*	*		shared with two successive reactions (ECrel) or intermediate of two interacting proteins (PPrel)
hidden compound	Entry element id attribute value for hidden compound.	*			shared with two successive reactions but not displayed in the pathway map
activation	-->		*		positive and negative effects which may be associated with molecular information below
inhibition	--		*		
expression	-->			*	interactions via DNA binding
repression	--			*	
indirect effect	..>		*	*	indirect effect without molecular details
state change	...		*		state transition
binding/association	---		*		association and dissociation
dissociation	+-		*		
missing interaction	-/-		*	*	missing interaction due to mutation, etc.
phosphorylation	+p		*		molecular events
dephosphorylation	-p		*		
glycosylation	+g		*		
ubiquitination	+u		*		
methylation	+m		*		

Από τα παραπάνω και με επαλήθευση με παραδείγματα, συμπαιράναμε ότι όταν relation name == compound τότε ισχύουν οι εξής σχέσεις



Επομένως όταν το ονομα του relation είναι compound, τότε καταγράφουμε τις 2 σχέσεις που φαινονται παραπάνω, αλλιώς καταγράφουμε μία σχέση:



Τα παραπάνω γίνονται με το script `curate.py` και αποθηκεύονται σε ένα αρχείο (`All_relations-Curated.csv`) με μορφή:

	head id	head name	tail id	tail name	pathway	link type	relation name	relation value	entry1	entry2
0	hsa:130589	GALM	cpd:C00267	C00267	hsa00010	ECrel	compound	90.0	73.0	75.0
1	cpd:C00267	C00267	hsa:3098	HK1	hsa00010	ECrel	compound	90.0	73.0	75.0
3	cpd:C00267	C00267	hsa:2645	GCK	hsa00010	ECrel	compound	90.0	73.0	74.0
5	cpd:C00267	C00267	hsa:2538	G6PC1	hsa00010	ECrel	compound	90.0	73.0	76.0

Τέλος με το `GrabNames.py`, βρίσκουμε τα πλήρη ονόματα των ουσιών, βάση των ids και τα συμπληρώνουμε.

Προκύπτει το τελικό αρχείο (`All_relations-Curated-full names.csv`) με μορφή:

head id	head name	tail id	tail name	pathway	link type	relation name	relation value	entry1	entry2	head full name	tail full name
hsa:130589	GALM	cpd:C00267	C00267	hsa00010	ECrel	compound	90.0	73.0	75.0	GALM, galactose mutarotase	alpha-D-Glucose
cpd:C00267	C00267	hsa:3098	HK1	hsa00010	ECrel	compound	90.0	73.0	75.0	alpha-D-Glucose	HK1, hexokinase 1
cpd:C00267	C00267	hsa:2645	GCK	hsa00010	ECrel	compound	90.0	73.0	74.0	alpha-D-Glucose	GCK, glucokinase
cpd:C00267	C00267	hsa:2538	G6PC1	hsa00010	ECrel	compound	90.0	73.0	76.0	alpha-D-Glucose	G6PC1, glucose-6-phosphatase catalytic subunit