Importing data into the BioAtomspace

Getting Started

To get a feel for what the BioAtomspace looks like with minimal code wrangling, you can run the SingularityNet Gene Annotation Service locally in docker containers:
https://github.com/MOZI-AI/annotation-service

Start by installing the Atomspace: https://github.com/opencog/atomspace

When we have a good docker image it will link here.

This repo must be installed for custom atom types (GeneNode and MoleculeNode)
It adds the “bioscience” scheme module (note that agi-bio also exists in the github/opencog project but the github/mozi-ai is most up-to-date
https://github.com/MOZI-AI/agi-bio

Current knowledge base import scripts are here:
https://github.com/MOZI-AI/knowledge-import

This is the scheme backend for the annotation service. It has code to generate subgraphs based on an input gene list:
https://github.com/MOZI-AI/annotation-scheme

This repo collects code for bio-atomspace experiments. It requires current versions of the github/opencog repos for cogutil, atomspace, ure, and pln.
https://github.com/ngeiswei/reasoning-bio-as-xp

Papers predicting COVID19 drugs using network analysis

https://www.biorxiv.org/content/10.1101/2020.03.22.002386v1
The virus-human protein interactions in this paper has biogrid data import linked below

https://www.nature.com/articles/s41421-020-0153-3

Existing Scheme files of Atomese translations

The generated scheme files from the knowledge-import repo and elsewhere can be downloaded from https://mozi.ai/datasets. Most of these tar balls contain multiple scheme files ready for importing into an Atomspace with a bioscience guile module installed.

This spreadsheet details individual files and who is using them for what experiment

current_*date-stamp*.tar.gz This is the current contents of the gene annotation service atomspace

Experimental datasets. These don't have protein and RNA level MoleculeNodes and remove text name and description and reference link information to reduce the atomspace size and simplify the semantics for initial inference rule development.

gene-level-dataset_*date-stamp*.tar.gz includes the 3 GO domain DAGs, SMPDB and Reactome pathway database gene sets, and BioGrid gene interactions

gene-level-dataset_without-location_*date-stamp*.tar.gz this doesn't include Reactome pathway specific subcellular location information contained in the above file set.

go-plus_*date-stamp*.tar.gz this adds part_of/has_part and regulates/is_regulated_by

string_dataset_*date-stamp*.tar.gz protein interaction db with 7 link types. This has a protein level version that replaces **BioGrid** in the **current** and **gene-level** file sets, and a gene level version for the **gene-level** file sets

This links to a spreadsheet with the specific scheme files generated from external databases. It will also contain complete list of versioned atomspace scheme files that are in use among all ongoing experiments:

atomspace scheme file table

TODO:
Work-in-progress schema doc for importing chinese medical treatments, with comments

Work out method for keeping track of which files are used in which experiments, so far there is an "experiment" column.

DBs to IMPORT

Experimental data sets

- MS spectrum database this includes multiple db links for each molecule (ChEBI, PubChem, etc)
- ProteinAtlas (in progress)

Ontologies

The following three have portions imported via GOplus

- ChEBI
- Cell Ontology
- UBERON

Reference catalogs?

- TCM (traditional chinese medicine) treatments & ingredient gene interactions
- Ensembl
- NCBI
- KEGG

Missing/broken semantics

- Multiple species (include viruses, then other model organisms - mouse, fly, worm, etc)
- Evidence codes and weights
- Scientific metadata (types of experiments and measuring devices, types and representations of models, institutions and researchers, etc)
- Change STRING EvaluationLink tv strength from 1 to 0.95
- Change STRING EvaulationLink tv strength to context dependent value

Causal inference structures

- GO-CAMs
- Reactome SBML pathways

Atomese translation schema

1. Gene ontology (GO)

Source: http://snapshot.geneontology.org/ontology/go.obo

Scheme version: GO.scm from https://mozi.ai/datasets/current.tar.gz

Last updated on March 25, 2020

Import script here

Example scheme representation in atomese format for GO term `GO:0000187`

```
;; GO:0000187 name
(EvaluationLink
      (PredicateNode "GO_name")
      (ListLink
      (ConceptNode "GO:0000187")
      (ConceptNode "activation of MAPK activity")
      )
)
;; GO:0000187 namespace
(EvaluationLink
      (PredicateNode "GO_namespace")
      (ListLink
      (ConceptNode "GO:0000187")
      (ConceptNode "biological_process")
      )
)
;; GO:0000187 definition
(EvaluationLink
      (PredicateNode "GO_definition")
      (ListLink
      (ConceptNode "GO:0000187")
      (ConceptNode "The initiation of the activity of the inactive
enzyme MAP kinase (MAPK).")
      )
)
;; GO:0000187 parent go term
```

```
(InheritanceLink
      (ConceptNode "GO:0000187")
      (ConceptNode "GO:0032147")
)
```

2. **Gene members of GO terms**

   Source:

   http://current.geneontology.org/annotations/goa_human.gaf.gz

   Scheme version: GO_annotation.scm from https://mozi.ai/datasets/current.tar.gz

   Last updated on March 25 2020

   Import script here

   Sample scheme representation in atomese format for gene IGF1

```
;; IGF1 is a member of go term GO:0000187
(MemberLink
   (GeneNode "IGF1")
   (ConceptNode "GO:0000187"))

;; IGF1 name
(EvaluationLink
    (PredicateNode "has_name")
      (ListLink
        (GeneNode "IGF1")
        (ConceptNode "Insulin-like growth factor I"))
))
```

3. **Protein (Uniprot) members of GO terms**

   Source:

   http://current.geneontology.org/annotations/goa_human_isoform.gaf.gz

   Scheme version: uniprot2GO.scm from https://mozi.ai/datasets/current.tar.gz

   Downloaded on March 25 2020

   Import script here

   Sample scheme representation in atomese format

```
;; Protein P05019 is a member of GO:0005179
(MemberLink
    (MoleculeNode "Uniprot:P05019")
    (ConceptNode "GO:0005179")
)
```

4. **Gene express Protein**

   Source:

   https://gitlab.com/opencog-bio/ocImport/raw/master/data/entrez2uniprot.csv.xz

   Scheme version: entrez_to_protein.scm from https://mozi.ai/datasets/current.tar.gz

   Last updated on March 25 2020

   Import script here

   Sample scheme representation in atomese format

```
;; Gene IGF1 expresses protein P05019
(EvaluationLink
  (PredicateNode "expresses")
    (ListLink
       (GeneNode "IGF1")
```

```
      (MoleculeNode "Uniprot:P05019")
))
;; Gene IGF1 has entrez id 3479
(EvaluationLink
  (PredicateNode "has_entrez_id")
    (ListLink
      (GeneNode "IGF1")
      (ConceptNode "entrez:3479")
))
```

5. Go-plus: Fully axiomatised version of the GO

GO-plus includes a complete set of relationship types which are not in the GO. includes has_part/part_of, [positively/negatively] regulates/regulated_by relationships between GO terms.

Source: here or here

Scheme version: Go-Plus.scm from https://mozi.ai/datasets/go-plus-2020-04-13.tar.gz

Last updated on April 14 2020

Import script here

Sample scheme representation in atomese format

```
;;
(EvaluationLink
    (PredicateNode "GO_positively_regulates")
    (ListLink
        (ConceptNode "GO:1903896")
        (ConceptNode "GO:0036498")))

(EvaluationLink
    (PredicateNode "GO_negatively_regulates")
    (ListLink
        (ConceptNode "GO:1903895")
        (ConceptNode "GO:0036498")))
(EvaluationLink
    (PredicateNode "GO_has_part")
    (ListLink
        (ConceptNode "GO:0016592")
        (ConceptNode "GO:0070847")))
```

GO-PLUS also includes cross ontology relationships (axioms) and imports additional required ontologies including ChEBI, cell ontology and Uberon.

Atomese format example:

```
(EvaluationLink
    (PredicateNode "has_name")
    (ListLink
        (MoleculeNode "ChEBI:3638")
        (ConceptNode "chloroquine")))
(InheritanceLink
    (MoleculeNode "ChEBI:3638")
    (ConceptNode "GOCHE:35842"))
(InheritanceLink
    (MoleculeNode "ChEBI:3638")
    (MoleculeNode "ChEBI:36683"))
(InheritanceLink
    (MoleculeNode "ChEBI:3638")
    (ConceptNode "GOCHE:88230"))
(EvaluationLink
    (PredicateNode "has_role")
    (ListLink
        (MoleculeNode "ChEBI:3638")
        (MoleculeNode "ChEBI:88230")))
```

```
(EvaluationLink
    (PredicateNode "has_name")
    (ListLink
        (ConceptNode "CL:0002370")
        (ConceptNode "respiratory goblet cell")))
(InheritanceLink
    (ConceptNode "CL:0002370")
    (ConceptNode "CL:0000160"))
```

```
(EvaluationLink
    (PredicateNode "has_name")
    (ListLink
        (ConceptNode "UBERON:0001456")
        (ConceptNode "face")))
(EvaluationLink
    (PredicateNode "UBERON_contributes_to_morphology_of")
```

```
    (ListLink
        (ConceptNode "UBERON:0001456")
        (ConceptNode "UBERON:0000033")))
(EvaluationLink
    (PredicateNode "UBERON_has_part")
    (ListLink
        (ConceptNode "UBERON:0001456")
        (ConceptNode "UBERON:0000165")))
(EvaluationLink
    (PredicateNode "UBERON_contributes_to_morphology_of")
    (ListLink
        (ConceptNode "UBERON:0004089")
        (ConceptNode "UBERON:0001456")))
```

6. Small molecule pathway database (SMPDB)

a. **Genes or/and Proteins database**

Source: http://smpdb.ca/downloads/smpdb_proteins.csv.zip

Scheme version: smpdb_protein.scm from https://mozi.ai/datasets/current.tar.gz

Last updated on March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Gene MAP2K4 is member of small molecule pathway SMP0000358
(MemberLink
  (GeneNode "MAP2K4")
  (ConceptNode "SMP0000358")
)

;; Protein P45985 is member of small molecule pathway SMP0000358
(MemberLink
  (MoleculeNode "Uniprot:P45985")
  (ConceptNode "SMP0000358"))
;; Small molecule pathway SMP0000358 name
(EvaluationLink
  (PredicateNode "has_name")
  (ListLink
    (ConceptNode "SMP0000358")
```

```
    (ConceptNode "Fc Epsilon Receptor I Signaling in Mast
Cells")
))
;; Protein P45985 name
(EvaluationLink
  (PredicateNode "has_name")
  (ListLink
    (MoleculeNode "Uniprot:P45985")
    (ConceptNode "Dual specificity mitogen-activated protein
kinase kinase 4")
))
```

b. **Metabolites (ChEBI) database**

Source:

http://smpdb.ca/downloads/smpdb_metabolites.csv.zip

Scheme version: smpdb_chebi.scm from https://mozi.ai/datasets/current.tar.gz

Last updated on March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; ChEBI 30915 is a member of pathway SMP0000055
(MemberLink
  (MoleculeNode "ChEBI:30915")
  (ConceptNode "SMP0000055"))
;; ChEBI 30915 name
(EvaluationLink
(PredicateNode "has_name")
  (ListLink
    (MoleculeNode "ChEBI:30915")
    (ConceptNode "2-oxopentanedioic acid")
))
```

7. **Reactome pathway**

Sources: here and here

Scheme version: reactome.scm from https://mozi.ai/datasets/current.tar.gz

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Pathways hierarchy
(InheritanceLink
   (ConceptNode "R-HSA-114608")
   (ConceptNode "R-HSA-76005"))
;; Pathway R-HSA-114608 name
(EvaluationLink
(PredicateNode "has_name")
   (ListLink
     (ConceptNode "R-HSA-114608")
     (ConceptNode "Platelet degranulation "))
)
```

8. Physical Entity (PE) Identifier mapping

a) NCBI genes to reactome mapping

Source: https://reactome.org/download/current/NCBI2Reactome_PE_Pathway.txt

Scheme version:NCBI2Reactome_pathway.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Gene location in a context of a pathway
(ContextLink
  ;; IGF1 is a Member of reactome pathway R-HSA-114608
  (MemberLink
    (GeneNode "IGF1")
    (ConceptNode "R-HSA-114608"))
  ;; IGF1 location
```

```
  (EvaluationLink
  (PredicateNode "has_location")
    (ListLink
      (GeneNode "IGF1")
      (ConceptNode "platelet alpha granule lumen"))
  )
)
```

b) Uniprot to reactome mapping

Source: https://reactome.org/download/current/UniProt2Reactome_PE_Pathway.txt

Scheme version:UniProt2Reactome_pathway.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Protein location in a context of a pathway
(ContextLink
  ;; Protein O00194 is a member of reactome pathway R-HSA-114608
  (MemberLink
    (MoleculeNode "Uniprot:O00194")
    (ConceptNode "R-HSA-114608"))
  ;; Protein O00194 location
  (EvaluationLink
    (PredicateNode "has_location")
    (ListLink
        (MoleculeNode "Uniprot:O00194")
        (ConceptNode "platelet dense granule membrane")))
)
```

c) ChEBI to reactome mapping

Source: https://reactome.org/download/current/ChEBI2Reactome_PE_Pathway.txt

Scheme version:ChEBI2Reactome_pathway.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; ChEBI location in a context of a pathway
(ContextLink
  ;; ChEBI 30915 is a member of R-HSA-1614558
  (MemberLink
    (MoleculeNode "ChEBI:30915")
    (ConceptNode "R-HSA-1614558"))
  ;; ChEBI 30915 location
  (EvaluationLink
    (PredicateNode "has_location")
    (ListLink
        (MoleculeNode "ChEBI:30915")
        (ConceptNode "mitochondrial matrix")))
)
```

the human readable locations are from reactome and are specific to the particular protein isoform present in a pathway (map from gene to protein is one to many).

11. The biogrid Protein interaction database

Source: here

Scheme version: biogrid_gene_gene_3.5.182.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Gene interaction and the pubmed references
(EvaluationLink
   (PredicateNode "has_pubmedID")
   (ListLink
       (EvaluationLink
       (PredicateNode "interacts_with")
           (SetLink
               (GeneNode "ARFIP1")
               (GeneNode "ARF3")))
       (ListLink
           (ConceptNode
"https://www.ncbi.nlm.nih.gov/pubmed/?term=10413101")
           (ConceptNode
"https://www.ncbi.nlm.nih.gov/pubmed/?term=9038142")))
)
```

In addition, the interaction between humans and CORONAVIRUS proteins/genes has been imported.

Source: here

Scheme Version: COVID-19-biogrid_3.5.184_2020-04-30.scm

Import script

https://github.com/MOZI-AI/knowledge-import/blob/master/coronavirus_biogrid.py

Sample scheme representation in atomese format

```
(EvaluationLink
    (PredicateNode "has_entrez_id")
    (ListLink
        (GeneNode "SLC44A2")
        (ConceptNode "entrez:57153")))
(EvaluationLink
    (PredicateNode "has_entrez_id")
    (ListLink
        (GeneNode "E")
        (ConceptNode "entrez:43740570")))
(EvaluationLink (stv 1.0 0.95)
    (PredicateNode "interacts_with")
    (SetLink
        (GeneNode "E")
        (GeneNode "SLC44A2")))
(EvaluationLink
    (PredicateNode "expresses")
    (ListLink
        (GeneNode "SLC44A2")
        (MoleculeNode "Uniprot:Q8IWA5")))
(EvaluationLink
    (PredicateNode "expresses")
    (ListLink
        (GeneNode "E")
        (MoleculeNode "Uniprot:P0DTC4")))
(EvaluationLink
    (PredicateNode "from_organism")
    (ListLink
        (GeneNode "E")
        (ConceptNode "TaxonomyID:2697049")))
(EvaluationLink
    (PredicateNode "from_organism")
```

```
    (ListLink
        (MoleculeNode "Uniprot:P0DTC4")
        (ConceptNode "TaxonomyID:2697049")))
```

12. The STRING Protein interaction database

Source: here

Scheme version: https://mozi.ai/datasets/string_dataset.tar.gz

Last updated March 10 2020

Import script https://github.com/MOZI-AI/knowledge-import/blob/master/string_PPI.py

Sample scheme representation in atomese format

```
;; Gene interaction
(EvaluationLink(stv 1.0 0.913)
    (PredicateNode "reaction")
    (SetLink
        (GeneNode "ARF5")
        (GeneNode "YKT6")))
;; Protein interaction
(EvaluationLink(stv 1.0 0.913)
    (PredicateNode "reaction")
    (SetLink
        (MoleculeNode "Uniprot:O15155")
        (MoleculeNode "Uniprot:P84085")))
```

13. RNA transcripts

a) Coding RNA

Source: here

Scheme version: codingRNA.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
(EvaluationLink
```

```
   (PredicateNode "transcribed_to")
   (ListLink
       (GeneNode "OR4F5")
       (MoleculeNode "ENST00000335137")))

(EvaluationLink
   (PredicateNode "translated_to")
   (ListLink
       (MoleculeNode "ENST00000335137")
       (MoleculeNode "Uniprot:Q8NH21")))
```

b) Non-coding RNA

Source: here

Scheme version: non-codingRNA.scm from here

Last updated March 25 2020

Import script here

Sample scheme representation in atomese format

```
;; Gene DDX11L1 transcribes a non coding RNA NR_046018
(EvaluationLink
   (PredicateNode "transcribed_to")
   (ListLink
       (GeneNode "DDX11L1")
       (MoleculeNode "NR_046018")))
```

14. Traditional Chinese Medicine Integrated Database (TCMID)

Source: here

Import script here

Sample scheme representation in atomese format

```
;; The herb “TIAN MEN DONG” has the property of sweet and affects the
lung meridian according to TCM
(EvaluationLink
```

```
  (Predicate “has_property”)
  (ListLink
    (ConceptNode “TIAN MEN DONG”)
    (ConceptNode “TCM:sweet”)))
(EvaluationLink
  (Predicate “meridian_affinity”)
  (ListLink
    (ConceptNode “TIAN MEN DONG”)
    (ConceptNode “TCM:lung”)))
```

15. Other supporting databases

a) Update the Gene symbols with the current approved gene symbol from https://www.genenames.org/download/custom/

- Scheme version: current_symbols.scm from here
- Import script
- Last updated March 25 2020
- Example representation

```
(EvaluationLink
   (PredicateNode "has_current_symbol")
   (ListLink
       (GeneNode "NOV")
       (GeneNode "CCN3")))
```

b) Biogrid gene-to-uniprot mapping

Inorder to infer the Protein-Protein-Interaction from Gene-Gene-Interaction, i.e. biogrid_genes mapped to their coding uniprots though biogrid_id

- Scheme version: biogridgene2uniprot.scm from here
- Last updated on March 25, 2020
- Import script here
- Example representation

```
(EvaluationLink
   (PredicateNode "expresses")
```

```
   (ListLink
       (GeneNode "MAP2K4")
       (MoleculeNode "Uniprot:P45985")))

(EvaluationLink
   (PredicateNode "has_biogridID")
   (ListLink
       (MoleculeNode "Uniprot:P45985")
       (ConceptNode "Bio:112315")))

(EvaluationLink
   (PredicateNode "has_biogridID")
   (ListLink
       (GeneNode "MAP2K4")
       (ConceptNode "Bio:112315")))
```

16) PharmaGKB pathways

https://www.pharmgkb.org/

17) covid-2019

https://thebiogrid.org/220839/publication/a-sars-cov-2-human-protein-protein-interaction-map-reveals-drug-targets-and-potential-drug-repurposing.html

Converted to scm COVID-19-biogrid_3

Representation of pathways

Pathways consist of interactions between chemicals, proteins, genes and effects of interactions.

All chemicals represented by (MoleculeNode “db:id”) format.
For chemicals use these databases:
`ChEBI, PubChem, DrugBank`
`ChEBI is preferred to PubChem and PubChem is preferred to DrugBank.`

A pathway is defined in atomspace as (ConceptNode “id”). (ConceptNode “id”) inherits from (ConceptNode “pathway”).

Predicates used to define interactions in a pathway:
`(PredicateNode "activation_of")`

```
(PredicateNode "catalysys_of")
(PredicateNode "conversion_of")
(PredicateNode "from_organism")
(PredicateNode "has_chebi_id")
(PredicateNode "has_drugbank_id")
(PredicateNode "has_location")
(PredicateNode "has_name")
     Applied to MoleculeNodes to specify it's human-readable name.

(PredicateNode "has_pubchem_id")
(PredicateNode "inhibition_of")
(PredicateNode "reaction")
(PredicateNode "transport_of")

# todo:
Example of receptor activation by a drug from pharmagkb source file
 (EvaluationLink
    (PredicateNode "activates")
    (ListLink
        (MoleculeNode "ChEBI:16865")
        (MoleculeNode "GABAa receptor inactive")   - GABA receptor is
a set of proteins so should be representented as concept
          ))
```

Database ID to URL mappings

Here are reference links to append the node ID and create hyperlink.
- For small molecules (Chebi's) "https://www.ebi.ac.uk/chebi/searchId.do?chebiId=" + ChebiID
- For Proteins "https://www.uniprot.org/uniprot/" + uniprotID
- For SMPDB pathways "http://smpdb.ca/view/"+ smpdbID
- For reactome pathways "http://www.reactome.org/content/detail/" + reactomePtahwayID
- For GO terms "http://amigo.geneontology.org/amigo/term/" + GoID
- For Genes "https://www.ncbi.nlm.nih.gov/gene/?term=" + GeneIDNote, for chebi's and Uniprots, the prefixes will not be included. An example for each of the above will look like:
- ChEBi:17661 will have a reference link https://www.ebi.ac.uk/chebi/searchId.do?chebiId=17661
- Uniprot:Q02080 will become https://www.uniprot.org/uniprot/Q02080Others are forward
- SMPDB pathway SMP0000575 will have a reference link http://smpdb.ca/view/SMP0000575
- Reactome pathways R-HSA-164843 http://www.reactome.org/content/detail/R-HSA-164843
- GO terms GO:0031435 http://amigo.geneontology.org/amigo/term/GO:0031435
- Gene IGF1 https://www.ncbi.nlm.nih.gov/gene/?term=IGF1