Plan for EGO experiment

Use chr 22 as an example.

1. There are 359 sydh datasets and 692 Haib datasets.

For each dataset, peaks reported by ENCODE will be used to denote where the annotation occur.

chr 22 start1 end1

chr 22 start2 end2

chr 22 start3 end3

1. There will be 1051 leaf-level annotation through the entire chromosomes. Each base has the potential to carry one or several of the 1051 labels. Each pf the 1051 leaf node can be represented by an EGO term.
2. Build a hierarchical tree for cell types and a hierarchical tree for TFs and combine them. How many new nodes (excluding the leaf node) will be there? Each can be represented by an EGO term.
3. Build a database (likely a graphical database to store these EGO terms and their relationship).
4. Build a query algorithm that is able to enumerate the number of occurrences for each label.

Use cases:

1. Suppose multiple GWAS studies have identified 150 loci in the genome that show genome-wide significance. For each locus, we can specify a 10kb region centered around the GWAS SNP. Then the 1.5MB region can be used to query against EGO to see which EGO terms are enriched in these regions.

In the user case folder, I have a file named diabetes. Columns E and F are chr and position (I assume this is GH38). These can be the center of the 10kb regions. In the “genomic regions” sheet. The last two columns are query region starts and ends.

|  |  |
| --- | --- |
| Chr | Position |
| 1 | 46933362 |
| 1 | 46950769 |
| 1 | 51339059 |
| 1 | 63729372 |
| 1 | 63790347 |
| 1 | 63791317 |
| 1 | 65526457 |

1. The promoters of a group of genes (a couple of hundreds) that show differential expression in a RNA-seq case-control experiments.
2. Enhancer regions identified by chromatin state in a particular cell type. The number of regions can be in the hundreds of thousands.

Working pipeline:

1. Two datasets: 359 SYDH ENCODE ChIP-seq data (from Snyder lab) and 692 HAIB ENCODE ChIP-seq data (from Hudson Alpha).
2. Metadata to ontology mapping:
   1. Metadata includes cell line, transcription factor, and antibody used for TF detection in Chip-seq assay
   2. Get unique TF and Cell line terms and unique combination of TF and cell line terms
   3. Map TF terms to Human transcription factor terminology and Cell line terms to Cell Line ontology (CLO)
   4. For unmapped terms:
      1. TF: try to organize TFs under human transcription factor terminology, mainly based on TF secondary structure
      2. Cell line: try to add into CLO, consider following information, derived from what cell type (CL) or tissue (UBERON), and whether derived from organism that has disease (e.g. cancer).
   5. Current metadata information can be found in the file: HaibSydhTfbs\_terms.xlsx.
3. EGO development:
   1. Using ontoFox get CLO and human TF terms and import into EGO
   2. For TF terms, we may consider to add TF terms in EGO with EGO IDs
   3. Write script to automate EGO term generation based on unique TF and CL combinations in the metadata file (following need to be discussed)
      1. Label: TF binding in CLO cell

TF: may consider to use PR label, abbreviation of PR can be used as alternative label

* + 1. Definition: A TF name binding to DNA in CLO cell
    2. Logical axioms:
       1. Binding function
       2. ‘realized in’ some biological process only ‘occurs in’ some CLO (consider to replace by a shortcut relation)
       3. ‘capable of binding to’ some TF

Notes: I have script can create class with annotation and simple logical axiom (ie. objectProperty some class).

* 1. Base on the use case, we need to add some middle layer terms, such as TF binding site,  helix TF binding site, TF binding in CLO, etc. Or we may get related information by querying Triple store we are planning to build, such as, give all the instances is a type that binding to TFs and TFs are subClassOf  helix TF.

1. Representation of Chip-seq data by instantiation of EGO:
   1. Information associated with instance data
      1. #Chromosome
      2. Binding start position
      3. Binding end position
      4. Data generated lab
      5. Consider in future, antibody used, assay used
      6. Need to make decision that the information will be generate using annotation or object property. For using object property, we need to generate many associated instance. So, suggest ‘lab’ (antibody, assay) use object property. Chromosome#, start position and end position use annotation property.
   2. Pseudo code for automated instance data generation
      1. Get lab, TF, Cell line information from file name

“wgEncode” + labName + “Tfbs” + CellLineName + TfName + …. (Need to discuss with Dr. Qin)

May use Cell line and TF list to find CellLine and Tf name

* + 1. Get annotation information from the opened file
       1. Column 1, chromosome #
       2. Column 2, start position
       3. Column 3, end position
    2. Base on information i and ii can create the instance
       1. Combination of Cell line and TF name is used to find EGO term, which is the class type of the instance
       2. Need to create lab instance which is instance of OBI:organization, with label of lab name
       3. Information in ii are annotation property associated with the instance
       4. I have script that can generate instances using information 1-3. Need to make some modification to make all the steps automatically.

Notes: need to find a strategy to generate instance IRI and can be find the same instance based on some known information, so we can add additional information to the instances when needed. Following combination may be unique:

* + - * + Filename + chromosome# + start position + end position

Questions: Do we need to represent technical/biological replicates results?

1. Instance level data query (triple store SPARQL query)
   1. TF binding sites in the region of interest
      1. TF binding site instances which are on specific chromosome and end position greater than region start position and start position less than region end position. For these instances, we need to retrieve their EGO type, lab information?, counts. We may need to use middle layer EGO terms to perform further statistical analysis

**--**

Note on 10/8/17: we have 150-200 TFs.

Designed to make: 175 \* 60 cell lines = 10,500 new EGO terms.

Ex: HeLa cell x TF1 🡪 TF1 in HeLa *subClassOf* TF1

🡪what is means:

A specific TF in a specific cell type: can tell

Input: a list of chromosome regions (as shown above).

Each region is a peak detected to be specific for a TF.

* Steve: if

**Use case**: There is a dataset D: peaks reported by ENCODE will be used to denote where the annotation occur.

chr 22 start1 end1

🡪 A region that a user likes to query (note: like a user provides one gene, one of many genes to be used for GO enrichment)

🡪 Overall use case: a user provides a list of such regions. Based on the information, we will need to query which EGO term is enriched.

What is the EGO term? A TF binding (note: peak is a signal of such binding).

🡪 Output: query which TF binding is enriched.

🡪

Basis of GO enrichment: 🡪 one-dimension

GO terms in 3 branches of hierarchies. Each GO term is linked to a list of genes, and each gene is linked to different GO terms 🡪 ***how to generate the base***: manually curated from literature that provides experimental data support.

Basis of EGO enrichment: 🡪 two-dimension

Each EGO term is linked to a list of regions, and each region is linked to different EGO terms (e.g., TF in a cell type. Note: since TF is cell type specific, we need both. The overall base is all cell types, where we want to say which cell type is enriched.). 🡪

***how to generate the EGO base***:

Steve group will generate the data where each base will be linked to TF in a specific cell type.

Example of EGO enrichment analysis: “TF in HeLa” is enriched?

We have 60 cell types 🡪 we can generate 60 different EGO data.

Example with GO enrichment: human, mouse, …

In EGO: we may need to cross different cell types. 🡪 enrichment Results: CDCF binding is enriched in HeLa cell 🡪 Rank #1, in U20 cell enriched 🡪 Rank #2 …

EGO terms: CDCF binding in HeLa, CDCF binding in U20, …

Knowledge: base x (in a region) is bound to CCF in HeLa, or U20, …

“Alpha-helix TF in HeLa”, …

* EGO general CLO + TF
* For each cell type,
* EGO: base-specific, GO: gene-specific. 🡪 One advantage
* EGO: having different cell types in EGO; GO: neutral. 🡪 Another EGO advantage.

One possibility: EGO for HeLa, EGO for U20? But Steve insists to put all together, which is another advantage for EGO …

🡪 TF-related: a TF binding site may be located inside the region.

Modification: only added if there is data.

🡺 Query: enrichment of EGO terms (e.g., a combination of TF and cell type).