# Notes: Definition of a training set for the biomedical extraction of key entities from full text articles in Europe PMC.

Currently the **core Europe PMC text mining pipeline** annotates the following entities:

1. Genes/Proteins
2. Organisms
3. Diseases
4. Chemicals
5. Gene ontology terms
6. Accession numbers (patterns)
7. Grants (patterns)
8. Resource names

In addition:

1. For the **Open Targets (OTAR) project** we identify human gene-disease relationships. This is currently a basic implementation, asserting that a gene and disease are related if the co-occur in the same sentence.
2. For the **EMERALD project**, we will be seeking to mine biome information from metagenomics publications, gene names and gene function information, secondary metabolite key words.

## Dataset 1: Named entity recognition for priority entity types

We wish to develop a training dataset for machine learning and benchmarking that is manually annotated for following types on full text articles:

1. Gene/Proteins, each annotation resolving to UniProt
2. Organisms, each resolving to an NCBI Taxonomy record
3. Diseases, resolving to UMLS

## Dataset 2: OTAR

4. OTAR Diseases, resolving to EFO Disease and Phenotype branches
5. OTAR Drugs (resolve to Drugbank)

6. EMERALD Biome terminology (constrain to human biomes - may need organisms, anatomy, disease/phenotype/age/gender? Define with MGnify curators)
7. EMERALD methods (subset of above?)
8. EMERALD Gene function (perhaps geneRiFs and GO will suffice)
9. EMERALD secondary metabolites (is this a mix of CHEBI +?)

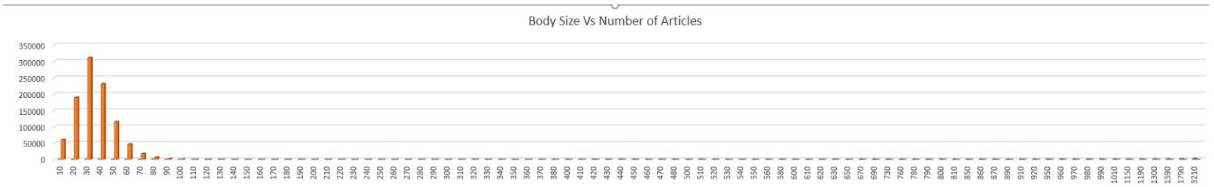
Others concepts that may be of interest

1. Protein-protein interactions (resolve to UniProt)
2. Cell lines (resolve to Cellosaurus)
3. Mutations - for pattern recognition?

# Methods

### Preparation of corpus

1. Articles were selected from the Open Access subset of Europe PMC, as archived on the 31 August 2018 (**v.2018.09**). Only articles from this set with a parsable CC-BY license were used, in order to make the traininset developed usable by anyone, academic or commercial. This gave us a starting pool of 991529 articles.
2. Papers were further filtered by the article body size to remove articles with abnormal length. By binning the article <body/> sections of the CC-BY XML files into 10KB size categories, most articles are in the range 25-50 KB. We therefore further refined the dataset to “typical” articles between 25 and 50 KB size; constraining the article size also makes the work of annotators more consistent. The number of articles in the candidate pool is now 503950 articles.




3. Based on existing Europe PMC annotations, these remaining “standard-sized” articles were categorised into High, Medium and Low frequency occurrence in all permutations of each entity type (genes/proteins, diseases, organisms). The boundaries of High, Medium and Low were determined by establishing the tertiles (33% and 66% percentiles) of frequency of total annotations (not types) in a given article. A total of 27 permutations of articles ranging from High-High-High to Low-Low-Low were established and the articles categorised accordingly.

4. Paper with the Low-Low-Low profile low occurrences of all the entity types are removed as they add little value to the training dataset. We then, in average, randomly sample around 12 articles from each frequency profile to generate an initial training set of 300 articles.
5. We also checked the papers in each of the frequency 26 bins for links to data records and occurrence of Open Targets annotations. A much higher than typical proportion of papers were linked to data and OTAR annotations. We are therefore confident that these papers are a solid basis for the core training set, having been cited in key deposition and added-value databases.

Table 1. Abundance of key entities used to establish Tertile boundaries.

| Entity | Tertile 1 (Low) | | Tertile 1 (Low) | | Tertile 1 (Low) | |
|---|---|---|---|---|---|---|
| | Lower | Upper | Lower | Upper | Lower | Upper |
| Genes/Proteins | 0 | 11 | 12 | 80 | 81 | 2408 |
| Organisms | 0 | 9 | 10 | 57 | 58 | 3108 |
| Diseases | 0 | 4 | 5 | 32 | 33 | 678 |

First we identified from the subset of open access articles archived on August 31 2018 those that have parsable CC-BY licenses (around 1M). Therefore, all the articles selected from this procedure can be reused by both academic and commercial entities. Then we analysed the size of the XML for those articles and established that most of them fall between 25 and 50 KB (503,950 articles) and follow a normal distribution. For these “typically sized” articles, we identified the tertile boundaries of abundance (see Table 1. categorize them into Low, Medium, and High) for the three key entities gene/proteins, diseases and organisms. The articles were then binned into 27 classes representing every combination of Low, Medium, and High gene/proteins, diseases and organisms (3x3x3). We then also overlaid information regarding article links to data and Open Targets (OTAR) gene-disease relationship annotations. Overall, this led to a set of 503,950 standard sized articles [Link to spreadsheet], exhibiting a range of abundances of the the three entities, about 13.4% of which have links to data and 26% of which have OTAR annotations [Link to spreadsheet].

## Draft guidelines for the annotators

1. Definitions of types we interested (i.e. what should be considered as a target entity?)
2. Some examples
3. Select the span of entity based on its syntax e.g. noun phrase/adjective phrase
4. Where (which database/platform) to check/validate whether the terms belong to a specific type and how to use the platform
5. How to communicate if there are unsure terms
6. How to use the annotation platform

## Annotation work begins:

1. Pre-annotate articles using Europe PMC annotation pipeline
2. Present the pre-annotated articles on the annotation platform
3. Let annotators start annotate articles:
    a. Following the guidelines

    b. Annotate entities missed by the pre-annotation
    c. Correct wrong annotations made by pre-annotations
    d. If issues arise, report to lead annotator
4. Lead annotator reviews as many as annotations made by primary annotators
    a. Summarizes issues
    b. Communicates/discusses with primary annotators
    c. Generate specific guidelines that tackle specific issues
    d. Repeat these steps

Depends on available resources and time frame, the annotation steps may vary.

1. If resources and time allows, three primary annotators work on the same set of articles (for agreement score) and then annotations can be checked/corrected by a lead annotator if a lead annotator is available.
2. If (1) is not practical, each primary annotator works on different articles and then annotations are checked or corrected by a lead annotator. As such a lead annotator is crucial for the quality of annotations.

Where an annotation tool is required we will use a combination of Europe PMC annotations and Hypothes.is, which allows annotations of articles from the Europe PMC interface - it's easier for annotators to annotate while reading a well-formatted paper in a browser.

- It requires alteration to the Europe PMC annotations feedback form plus:
    - Hypothes.is user account for each annotator
    - use of a Hypothes.is group "Europe PMC annotations" for marking false negatives.
- We need to ensure the output formats are parsable and able to be integrated with Europe PMC annotations .
- We need to switch off the entity auto-delete function.

As a bonus this approach will also identify bugs in the SciLite viewer as a false negative may indeed have been a true positive but not displayed correctly by SciLite. The approach leads to two tasks: validate pre-tagged entities/relationships and identify missed entities/relationships.

For annotators to validate existing annotations: needs a pop up that allows annotators to report:

- Entity is correct
- Both entities in a relationship are correct AND
    - The relationship is positive and certain
        - Significant transcriptome alterations are detected in the brain of patients with amyotrophic lateral sclerosis (ALS), including carriers of the C9orf72 repeat expansion and C9orf72-negative sporadic cases PMC5886204
    - The relationship is positive and possible
    - The relationship is negative and certain
        - Furthermore, serum levels of OPG did not show differences between patients with osteonecrosis treated intravenously for cancer compared to those treated orally for osteoporosis. PMC5694175
    - The relationship is negative and possible
    - The relationship is indirect

        - e.g. Based on the comparison of serum levels of RANKL and OPG, and the RANKL/OPG ratio among the different stages of osteonecrosis, no significant differences were observed ($p > 0.05$). PMC5694175
        - We determined whether serum levels of Receptor Activator for Nuclear Factor κ B Ligand (RANKL), Osteoprotegerin (OPG), and the RANKL/OPG ratio could be useful biomarkers for the severity of oral lesions in bisphosphonate-related osteonecrosis of the jaw PMC5694175
        - Effect of Degarelix, a Gonadotropin-Releasing Hormone Receptor Antagonist for the Treatment of Prostate Cancer, on Cardiac Repolarisation in a Randomised, Placebo and Active Comparator Controlled Thorough QT/QTc Trial in Healthy Men. PMC5569649

- Report problem with annotation: needs
    - Entity
        - is wrong/false positive
        - is only partially annotated
            - State what the full annotation should be
    - Relationship
        - One/both of the entities (gene/protein -- disease) is a false positive
        - Is not actually a relationship, not even indirect. e.g. It is merely a co-occurrence of terms.

### Identify missed entities/relationships (via Hypothes.is)

Missed entities are reported through Hypothes.is that allows annotators to annotate text and take note in the corresponding comment box. Annotators are required to report the type of the entity/relationship.

- Entities
    - Highlight the full entity phrase and state the type in the Hypothes.is comment box
        - GP:cyclin-dependent kinase 6
        - ORG:Rattus rattus (note: only binomials)
        - DIS:eastern equine encephalitis (clinical presentation)
- Relationships
    - Highlight the text containing the relationship and state
        - Neg/Pos (Relationship is positive or negative)
        - Cert/Like/Remo (Certain --- quite likely --- remotely)
        - Contra (Contradicts (previous work))

# Brief: Construction of Training Set for Key Entities in Europe PMC

Europe PMC is a database of over five million open access full text research articles, most of which are available as XML. We wish to develop machine learning algorithms to extract

useful information for curator workflows. We therefore need to develop a collection of marked-up articles (a training set or Gold standard) for benchmarking and training of algorithms.

The overall goal is to develop an open-access training set of around 300 articles, with the potential to expand this set both in quantity and in entity type in later stages.

We already tag XML articles with gene/proteins, diseases, organisms, and gene-disease relationships using the dictionary and rule-based Europe PMC pipeline. Gene/protein, disease and organisms are annotated by dictionaries using Uniprot names, NCBI Taxonomy names and UMLS terms. We will use these tags as a starting point for manual annotation of a training set of articles.

The gene-disease relationships are tagged using EFO Diseases and the human and mouse gene/protein subset of UniProt names and are based simply on sentence-level co-occurrence of a gene and disease.

## Selection of the articles for the training set

We have identified a subset of around 503K articles that satisfy the following criteria:

- CC-BY license, to make the dataset re-usable in the future
- Articles of “typical” size (XML file size between 25-50KB) to remove articles with abnormal size, so constraining the manual annotation task
- Known to contain a range of abundances of three key entities of interest: genes/proteins, disease/phenotypes and organisms.
- These articles also frequently include known OTAR gene-disease relationship annotations (26%) and/or links to data (13.4%)

These 503,950 articles have been divided into 27 bins based on relative abundance (High, Medium, Low) of the three key entities (gene/proteins, diseases, organisms). After removing the (Low, Low, Low) category, which is sparse and adds little value to the training dataset, 26 bins remain. From these, about 300 articles will be randomly selected, in proportion to the number of articles within each bin (ie 5-15 articles from each in real terms).

About 27.6% of candidate articles in our 26 bins contain OTAR gene-disease relationships. Therefore we would expect about 83papers to contain OTAR annotations which should also be evaluated.

For: Brief for Annotation development see:
https://docs.google.com/document/d/1TpICfvIfK0AFLe90Wqcos2fovzFRzfj1uhrgwDdtNwY/edit
I have separated it out in order to share with Molecular Connections.