NIH RePORTER and PubMed API Methods

OVERVIEW

* The jupyter notebook is used to download publication data from PubMed and grant data from NIH RePORTER.
* The main block of code consists of a robust function which takes a PubMed URL, an NIH RePORTER search parameters JSON, an Output dataframe, and 4 files destination addresses for storing and retrieving raw data.
* The function first uses a list of genes and synonyms to construct and edit the PubMed url and NIH JSON to make robust searches for each gene and its synonyms. The function then retrieves number of publications, number of grants, and overall funding amount for grants and publications which mention a gene and respective keywords in “titles” and “title or abstract”. A copy of these searches are stored in .txt files to be read on subsequent runs, and the values for each gene are recorded in an output dataframe which is stored at the end of each block.
* 4 of the 7 blocks use the exact same initial function with slightly different parameters based on the specific control being searched. The other 3 use an almost identical structure save for the removal NIH searches in the case of figure 2D, and the removal of the default synonym constructor for 3B and 3C

Step By Step Breakdown of Main function, using the default search for ERBB2 as an example

Changes between steps are highlighted

1. Initial steps; loading of packages, connecting directories. The main function is called and the 7 parameters are all passed.
   1. PubMed URL (At this stage): [https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[ti]+AND+(\"+\_gene\_+\"[ti]))&retmax=20](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer%5bti%5d+AND+(\%22+_gene_+\%22%5bti%5d))&retmax=20)
   2. NIH JSON(At this stage): {"criteria": {"advanced\_text\_search": {"operator": "advanced", "search\_field": "projecttitle,abstracttext","search\_text": ("cancer AND ")}}, "include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
2. After initial steps of loading packages and directories and reading in the list of genes and synonyms, the function loops through each gene included in the synonym file. Each synonym for the gene is then added to the search if it is not in the list of excluded synonyms for that gene. The search terms are formatted to the respective database and added to the PubMed URL and NIH JSON
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[ti]+AND+(“ERBB2"[ti]+OR+"HER-2 "[ti]+OR+"HER-2/neu"[ti]+OR+"HER2+"[ti]+OR+"NEU"[ti]+OR+"NGL"[ti]+OR+"ERBB-2"[ti]))&retmax=20
   2. NIH JSON: {"criteria": {"advanced\_text\_search": {"operator": "advanced", "search\_field": "projecttitle,abstracttext","search\_text": ("cancer AND ("ERBB2" OR "CD340" OR "HER-2" OR "HER-2/neu" OR "HER2" OR "MLN 19" OR "NGL" OR "TKR1" OR "NEU“ OR “ERBB-2”)")}}, "include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
3. Once the search parameters for both databases are established, the PubMed search is performed. First, the code searches the number of “title” publications. If there is already a stored pubmed “title” search under that gene, the stored search is read, otherwise, a new search is accessed through the URL and the result is stored in its respective folder. Once the title search is performed, the [ti] search term identifier is changed to [tiab] in the URL, and the same process is repeated for title/abstract searches.
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[tiab]+AND+(“ERBB2"[tiab]+OR+"HER-2 "[tiab]+OR+"HER-2/neu"[tiab]+OR+"HER2+"[tiab]+OR+"NEU"[tiab]+OR+"NGL"[tiab]+OR+"ERBB-2"[tiab]))&retmax=20
   2. NIH JSON: Unchanged
4. After the PubMed searches are performed and stored, the NIH searches are done much in the same way. First we pass the search parameters for title/abstract searches, retrieving the first 500 grant results using the FindAwardAmount function. If there are more than 500 grants, then up to 9500 additional grants will be retrieved. After title/abstract searches are retrieved, the parameters of the Search JSON are changed to a title only search, and the search is then made again. A freeze of the data is stored same as in step 3.
   1. PubMed URL:Unchanged
   2. NIH JSON: {"criteria": {"advanced\_text\_search": {"operator": "advanced", "search\_field": "projecttitle","search\_text": ("cancer AND ("ERBB2" OR "CD340" OR "HER-2" OR "HER-2/neu" OR "HER2" OR "MLN 19" OR "NGL" OR "TKR1" OR "NEU“ OR “ERBB-2”)")}}, "include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
5. After all three values (# of publications, # of grants, total grant $) for both title and title/abstract searches are retrieved for a given gene, totalling in 6 values, the values are stored in a dataframe under the corresponding gene name. Once this whole process has been done for each gene in the list, then the dataframe is saved and output to its appropriate locations.

Variations:

Differences from the default searches are highlighted

1. Search 2 AACR publications and NCI grants: this search differs from the default one in that the PubMed url includes search tags for each journal associated with the American Association for Cancer Research, while the NIH search parameters include a section that specifies grants from the National Cancer Institute. The keyword “cancer” has been removed, as the context is represented by the institutions
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=( (\"blood cancer discovery\"[ta]+OR+\"cancer discovery\"[ta]+OR+\"Cancer Epidemiology Biomarkers and Prevention\"[ta]+OR+\"cancer immunology research\"[ta]+OR+\"cancer prevention research\"[ta]+OR+\"cancer research\"[ta]+OR+\"clinical cancer research\"[ta]+OR+\"molecular cancer research\"[ta]+OR+\"molecular cancer therapeutics\"[ta]+OR+\"cancer research communications\"[ta])+AND+(\"+\_gene\_+\"[ti]))&retmax=20
   2. NIH JSON: {"criteria": {"agencies": ["NCI"],"advanced\_text\_search": {"operator": "advanced", "search\_field": "projecttitle,abstracttext","search\_text": ("")}}, "include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
2. Search 3 searches without “cancer” keyword: Same as default save for the removal of the “cancer” keyword which is present in all other searches; search terms for this block only contain gene names and synonyms.
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(\"+\_gene\_+\"[ti])&retmax=20
   2. NIH JSON: {"criteria": {"advanced\_text\_search": {"operator": "advanced","search\_field": "projecttitle,abstracttext","search\_text": str("")}},"include\_fields": ["ApplId","ProjectTitle", "AwardAmount", "DirectCostAmt","IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
3. Search 4 publications limited to the United States: A search for all PubMed publications that are associated with or were published in the United States. Since this search only retrieves publication info, it only uses a PubMed URL modified to include articles which have denote a publishing location or at least one author with US affiliation.
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[ti]+AND+("United States"[pl] OR "USA"[Affiliation] OR "US"[Affiliation] OR "United States"[Affiliation] OR "United States of America"[Affiliation] OR "U.S.A"[Affiliation] OR "U.S."[Affiliation])+AND+(\"+\_gene\_+\"[ti]))&retmax=20
4. Search 5 Publications and grants with animal keywords: This search retrieves grants and publications which include terms denoting different classes of animals, specifically canine, bovine, equine, feline, and poultry.
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[ti]+AND+((Canine[mh])+OR+(feline[mh])+OR+(Bovine[mh])+OR+(equine[mh])+OR+(poultry[mh]))+AND+(\"+\_gene\_+\"[ti]))&retmax=20
   2. NIH JSON: {"criteria": { "advanced\_text\_search": {"operator": "advanced","search\_field": "projecttitle,abstracttext","search\_text": str("cancer AND (Canine OR Bovine OR feline OR equine OR poultry) AND ")}},"include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }
5. Search 6 Search without Synonyms: While the initial PubMed URL and NIH JSON are exactly the same as the default search, this block does not actually perform the which finds synonyms and formats them to the search, as only the main gene name is included (main gene names which are excluded take the first synonym from their list of synonyms instead. Example using ERBB2 is shown below
   1. PubMed URL: https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=(cancer[ti]+AND+(“ERBB2"[ti]))&retmax=20
   2. NIH JSON: {"criteria": {"advanced\_text\_search": {"operator": "advanced", "search\_field": "projecttitle,abstracttext","search\_text": ("cancer AND ("ERBB2")")}}, "include\_fields": ["ApplId", "ProjectTitle", "AwardAmount", "DirectCostAmt", "IndirectCostAmt","ProjectStartDate", "ProjectEndDate"],"offset": 0, "limit": 500, }