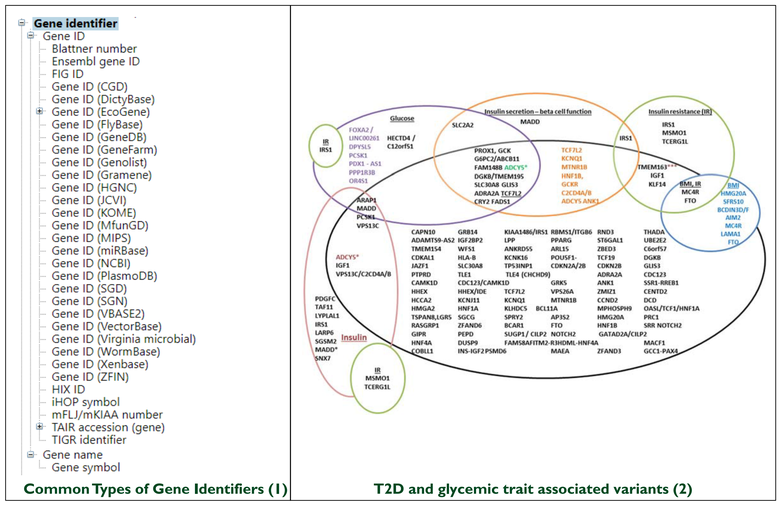
Conversion and Validation of Biomedical Data Using Translation by Identifiers

Brendan T Kearney, Jeremy J Jay
Bioinformatics and Genomics
UNC CHARLOTTE

Background

In biomedical literature, a reader may frequently encounter identifiers for genes, proteins, or other entities. Though some identifiers may be easy to interpret at first glance, attempting to use the resulting datasets for further analysis is often complicated by data integration issues such as ambiguous references and a lack of standardization across identifier types.

Common identifiers types, such as those from the Entrez, Ensembl, GenBank and UniGene databases **(Figure 1A)**, provide uniqueness and stability. However, there are no tools that can directly integrate results from articles or published data without significant manual intervention.



Common Types of Gene Identifiers (1)

T2D and glycemic trait associated variants (2)

Figure 1A **Figure 1B**

Using this translational tool, bio-entities such as these can be categorized into clusters of diseases, mutations, or other groups. Using prediction methods, the tool can use common relationships such as pathways (**Figure 1B**) to compensate for poorly described or missing associations.

Methods

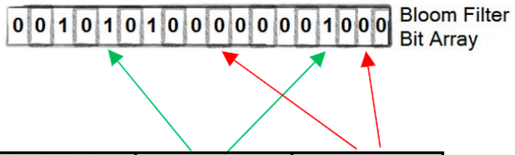
Data Preparation

- Entailed retrieving and converting source data formats to allow our code to efficiently process and perform analysis
  - A sample of 1,396 supplementary data files (Excel, CSV, etc.) were extracted from the PubMed Central Open-Access repository.
- Translation was conducted using multiple multiple data sources:
  - gene2ensembl – NCBI and Ensembl-annotated gene association fetched from the NCBI FTP
  - Homo_sapiens.gene_info – Human Gene annotation from NCBI FTP
  - Ensembl Genes - Human GRCh38.p13 Gene from BioMART

Validation: determining accuracy/sources of ambiguity

- Python for translation and data management
  - Bloom filter package – pypi.org/project/bloom-filter
- R for statistical analysis and visualization

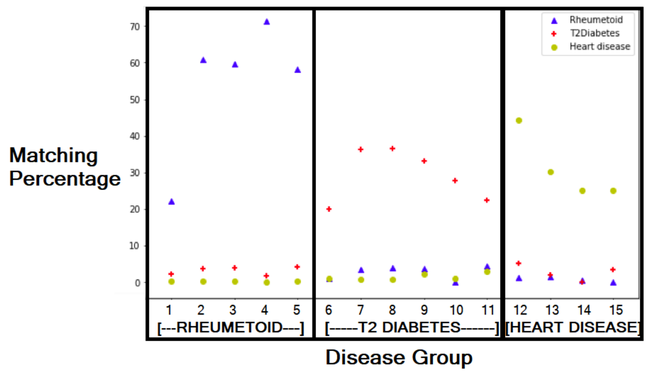
We developed a python tool using Bloom filters to quickly identify and translate common biomedical identifiers.



QUERY:	'KCNJ11'	'retinol'
Hash Function	4,14	9,16
Result	Maybe in set	Definitely not in set

Bloom filters work using a probabilistic bit array test for set membership: Test values are hashed to a set of key indexes and then checked in the bit array to determine if the value cannot possibly be in the set.

Results were tested across "gold standard" disease-specific subsets to test and validate the accuracy and specificity of the methods.



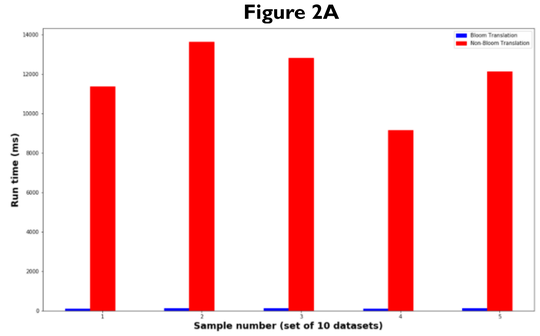
Using disease-specific gene resources, our methods correctly classify the best candidate disease grouping based only on set membership tests (no database identifier lookups necessary). These methods work in real-time on data extracted from PubMed Central.

Based on this work, we plan to build a simple and practical tool to automatically integrate new data into existing analysis workflows.

Results

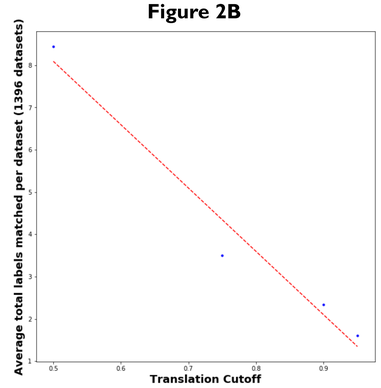
Performance of Bloom filter: In 10 randomly-selected PMC files, the Bloom filter algorithm operated with a 98% average decrease in run time (**Figure 2A**) over five trials when compared to a standard Python matching function.

Figure 2A



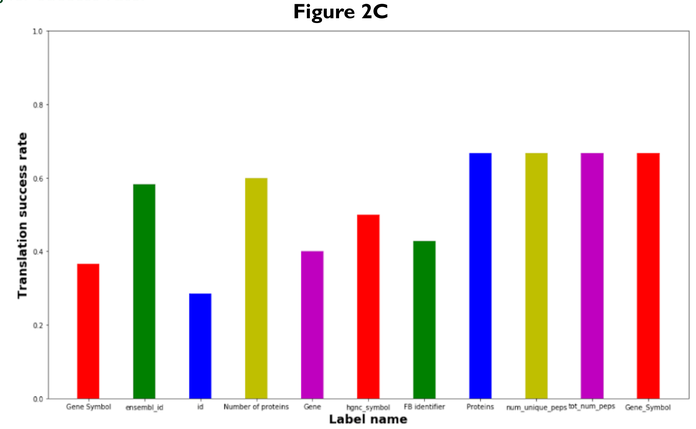
Choosing the optimal cutoff threshold: Increasing the cutoff (percentage of data entries that must be able to matched to a reference dataset via translation algorithm) has a linear effect on the decreasing total labels that pass the cutoff criteria (**Figure 2B**). Lower cutoffs may increase the number of matching columns but introduce more risk of false positives.

Figure 2B



Common Identifiers have varied translational success: At 75% cutoff, the most frequent identifiers and column names in the 1,396 supplementary PMC files do not necessarily translate at a high rate (**Figure 2C**), although they still successfully match much higher than the overall average (8%). Some identifiers may be manually modified before translation to attain a much higher success rate.

Figure 2C



References

1. (2020) EDAM bioinformatics operations, types of data, data formats, identifiers, and topics. *Bioportal*
2. Prasad,R.B. and Groop,L. (2015) Genetics of Type 2 Diabetes-Pitfalls and Possibilities. *Genes*, **6**(1), 87-123
3. Fuchsberger,C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**(7614), 41-47