# DataSets

Provide gene-mutation-drug relations for the advancement of personalized medicine

* Genomics of Drug Sensitivity in Cancer (GDSC)
* Cancer Cell Line Encyclopedia (CCLE)
* Cancer Therapeutics Response Portal (CTRP)

Databases contain gene-mutation-drug relations extracted from manually curated literature on clinical studies:

* ClinVar
* My Cancer Genome
* MD Anderson Personalized Cancer Therapy Knowledgebase

Computational methods that automatically extract gene-mutation-drug relations from the literature are urgently needed to assist in the curation process.

# NER tools to identify different entities in text

* identify mutations
  + tmVar
  + EMU
  + MutationFinder
* identify genes
  + BANNER
  + GNormPlus
* identify drugs
  + ChemSpot
  + tmChem
* identifies gene, disease, drug and cell line names
  + BEST Biomedical Entity Extractor: BEST returns, as a result, a list of 10 different types of biomedical entities including genes, diseases, drugs, targets, transcription factors, miRNAs, and mutations that are relevant to a user’s query

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070740/>

<http://best.korea.ac.kr/>

BEST uses a dictionary-based approach to extract biomedical entities from texts, and indexes the entities along with the source texts. The BEST dictionary consists of 12 different databases each covering different subsets of entity types (we will describe it in the Methods section).

* + BioBert : <https://github.com/dmis-lab/biobert>. This repository provides fine-tuning codes of BioBERT, a language representation model for biomedical domain, especially designed for biomedical text mining tasks such as biomedical named entity recognition, relation extraction, question answering, etc. Please refer to our paper [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](http://arxiv.org/abs/1901.08746) for more details.

# Relations between entities

* Relations between entities using not only sentence-level co-occurrence but also information from external databases
  + PharmGKB
  + DrugBank
  + Doughty extracted gene/protein and mutation names from texts and mapped them using a protein sequence filter in addition to co-occurrence information. Their gene-filtering tool checks amino acid sequences from NCBI RefSeq and compares them with wild type amino-acid information containing mutation names
* Pre-defined rules with trigger words to find relations between entities
  + SNPshot: sentence-level co-occurrence and pre-defined keywords to identify relations between entities
  + Mahmood et al. used a series of natural language processing (NLP) modules with part-of-speech tagging to find syntactic structures and specific pre-defined keywords in sentences containing mutations
* Machine Learning
  + DeepDive to extract gene-gene interactions from sentences and achieved reasonable precision on a large-scale literature test set
  + Singhal used a machine learning approach to identify mutation-gene-disease relations in the literature. They extracted simple general features such as the distance between a mutation and a disease, frequency of disease occurrence, and frequency of co-occurrence of mutation-disease pairs. They also used the sentiment scores between a mutation and a disease when they appeared in the same sentence. Using these features, they trained a decision tree classifier, and achieved better performance than state-of-the-art approaches used for finding gene-disease associations. Moreover, since this approach is independent of specific sentence structures, it can be used to identify other associations such as mutation-drug associations.

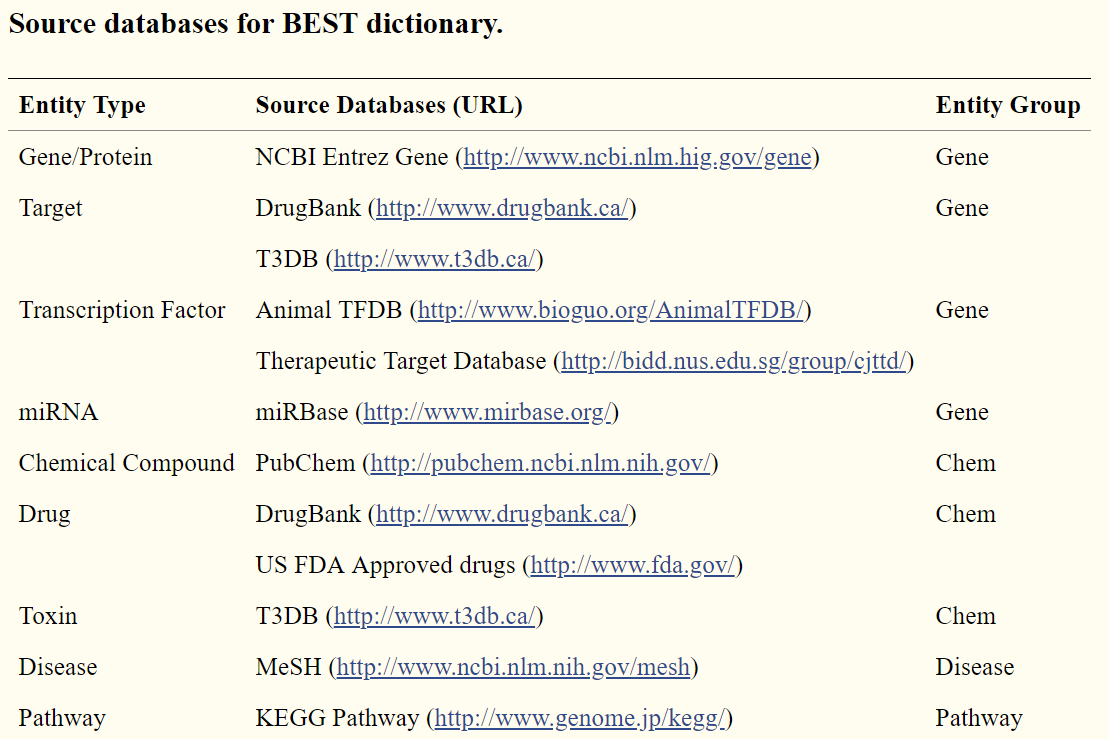
**Using all the PubMed articles as our background domain knowledge**

**Variant-entity relation extraction task**: many of the relations have different forms and some of them are described in a complicated way in documents => Deep convolutional neural network (CNN) which is a deep learning technique that uses multiple layers of neurons and convolutional layers for classification

# Tratamiento a nivel de documento

BRONCO: which is a manually curated mutation-gene-disease-drug relation corpus. BRONCO contains >400 variants and their relations with genes, diseases, drugs and cell lines in the context of cancer, all of which were extracted from 108 full-text articles

BEST EE: BEST own dictionary-based named-entity extraction module which is available at [http://infos.korea.ac.kr/bioentityextractor](http://infos.korea.ac.kr/bioentityextractor/)



BEST Search Tool: Apache Solr v.4.9, a Lucene-based search platform, where the Solr’s indexing structure and ranking system were logically redefined to score entities. BEST indexing subsystem performs entity extraction, entity indexing, and meta-information indexing. Biomedical entities into the following four groups: gene\_group, chem\_group, disease\_group, and pathway\_group. In the gene\_group, genes/proteins, targets, transcription factors, and miRNAs are included. Chemical compounds, drugs, and toxins are included in the chem\_group. Diseases and pathways are included in the disease\_group and the pathway\_group, respectively

For example, when we enter a query to find the relation between a mutation and a drug, we check all the biomedical entities such as gene names, disease names, and cell line names that appear in the same sentence.

After entering this query, we obtained the result list of entities with their scores. This score is called BSSM.

The second method uses not only the mutation itself but also the other biomedical entities that appear near the mutation to generate the query.

PROCESO 1: (NER) Obtener tuplas (gen, mutación) candidatos de un documento.

PROCESO 2: (CLASIFICACIÓN) La IA decidirá si esta tupla es correcta o incorrecta, considerando la frecuencia y el contexto.

PROCESO 3: (CLASIFICACIÓN DE FRASES) Fácil de decidir si la frase relaciona la tupla. PARRAFOS¿?

PROCESO: BEST es interrogado sobre una consulta lanza el **extractor BEST EE** y obtiene las entidades relacionadas y los artículos de PubMed que se corresponden a la búsqueda. Ver el score que produce: <http://best.korea.ac.kr/help/BEST_Guide.pdf>

Desde BRONCO se localiza el nombre normalizado de la Entidad Biomédica “mutación” de la consulta (p.e. de la mutación Val600Glu, BRONCO nos devuelve la mutación normalizada V600E) y se obtienen las entidades relacionadas (BRAF y melanoma)

1. Partimos de un texto: ***“****The oral* ***BRAF*** *inhibitors* ***vemurafenib*** *(formerly PLX4032) and* ***dabrafenib*** *(formerly GSK2118436) induce a high frequency of tumor regressions in patients with***BRAF** **V600E***mutant metastatic* ***melanoma*** *and* ***vemurafenib*** *improves overall survival compared with chemotherapy****”.*** Utilizando **BEST EE** obtenemos
   1. La mutaciónV600E.
   2. Las posibles relaciones con (V600E, vemurafenib) y (V600E, dabrafenib).
   3. Entidad Biomédicas cercanas: (gen BRAF, enfermedad melanoma)

BSSM: Buscar en BEST con el nombre normalizado (V600E)

1. Buscar por todas las Entidades Biomédicas de la frase, excluyendo las del mismo tipo de la puesta en la consulta.
   1. BSSA: Combinar todas con AND (V600E and BRAF and melanoma)
   2. BSSO: Combinar todas con OR (V600E or BRAF or melanoma)
   3. BSSAO: Combinar la entidad de la consulta con cada una de las entidades relacionadas (V600E and BRAF or V600E and melanoma)
2. Lanzar las búsquedas en BEST Search y guardar los scores