Introduction to Data Mining in Biomedicine

Piyush More
02.12.2025

Understanding biological information

What's the *first thing* you do when you start any research project?

Literature Review

- There are ~37M articles indexed in PubMed[1]
- In case of NGS, there are ~7M gene expression profiles deposited in GEO[2]

Need for a systematic method to access and process the information

Data Mining

[1] Publications Output: U.S. Trends and International Comparisons (https://ncses.nsf.gov/pubs/nsb202333/publication-output-by-region-country-or-economy-and-by-scientific-field)
[2] GEO overview (https://www.ncbi.nlm.nih.gov/geo/)

What is Data Mining?

Mining: Extracting ***key information*** from large datasets
Followed by finding interesting relationships → meaningful associations

Frequently used techniques

- Correlation and regression
- Clustering
- Classification
- Anomaly detection
- Natural Language Processing
    - AI and Large Language Models

* Image obtained from https://static.vecteezy.com/system/resources/previews/009/403/372/original/businessman-digging-database-and-data-mining-concept-cartoon-illustration-vector.jpg

Types of biomedical data

- Genomic Data
  - DNA sequences (mutations and variants), gene expression, chromatin accessibility
- Proteomic and Metabolomic Data
  - Protein levels, metabolic pathways
- Imaging Data
  - Medical imaging (MRI, CT scans, etc.) processed using machine learning techniques
- Clinical Data
  - Patient demographics, medical history, diagnoses, treatment plans, outcomes

Databases and Tools useful for Data Mining

- General
  - NCBI
    - PubMed (https://pubmed.ncbi.nlm.nih.gov/)
    - PubChem (https://pubchem.ncbi.nlm.nih.gov/)
    - MeSH (https://meshb-prev.nlm.nih.gov/)
    - more…
- Genomics
  - GEO (https://www.ncbi.nlm.nih.gov/geo/)
  - CCLE (https://sites.broadinstitute.org/ccle/)
- Pharmacology
  - DrugComb (https://drugcomb.org/)
  - Genomics of Drug Sensitivity in Cancer (https://www.cancerrxgene.org/)
  - Therapeutic Target Database (https://idrblab.net/ttd/)
- R, Python, SQL, and other programming environment

Medical Subject Heading (MeSH)

- Method for categorizing PubMed articles by assigning medical terms
- Uses standardized keywords to describe main topic and subtopics of the article

Hierarchical structure

- Anatomy [A]
- Organisms [B]
- Diseases [C]
  - Infections [C01]
  - Neoplasms [C04]
  - Musculoskeletal Diseases [C05]
  - Digestive System Diseases [C06]
  - Stomatognathic Diseases [C07]
  - Respiratory Tract Diseases [C08]
  - Otorhinolaryngologic Diseases [C09]
  - Nervous System Diseases [C10]
  - Eye Diseases [C11]
  - Urogenital Diseases [C12]
  - Cardiovascular Diseases [C14]
  - Hemic and Lymphatic Diseases [C15]
  - Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16]
  - Skin and Connective Tissue Diseases [C17]
  - Nutritional and Metabolic Diseases [C18]
  - Endocrine System Diseases [C19]
  - Immune System Diseases [C20]
  - Disorders of Environmental Origin [C21]
  - Animal Diseases [C22]
  - Pathological Conditions, Signs and Symptoms [C23]
  - Occupational Diseases [C24]
  - Chemically-Induced Disorders [C25]
  - Wounds and Injuries [C26]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Humanities [K]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographicals [Z]

Drug-regulated CD33-targeted CAR T cells control AML using clinically optimized rapamycin dosing

Jacob Appelbaum [1 2 3 4], April E Price [5], Kaori Oda [1], Joy Zhang [5], Wai-Hang Leung [5], Giacomo Tampella [1], Dong Xia [5], Pauline Pl So [5], Sarah K Hilton [5], Claudya Evandy [1], Semanti Sarkar [1], Unja Martin [5], Anne-Rachel Krostag [5], Marissa Leonardi [1], Daniel E Zak [5], Rachael Logan [1], Paula Lewis [5], Secil Franke-Welch [5], Njabulo Ngwenyama [5], Michael Fitzgerald [1], Niklas Tulberg [1], Stephanie Rawlings-Rhea [1], Rebecca A Gardner [1], Kyle Jones [6], Angelica Sanabria [6], William Crago [6], John Timmer [6], Andrew Hollands [6], Brendan Eckelman [6], Sanela Bilic [7], Jim Woodworth [7], Adam Lamble [1 4], Philip D Gregory [5], Jordan Jarjour [5], Mark Pogson [5], Joshua A Gustafson [1], Alexander Astrakhan [5], Michael C Jensen [1]

Affiliations + expand

PMID: 38502193 PMCID: PMC11060733 DOI: 10.1172/JCI162593

Abstract

Chimeric antigen receptor (CAR) designs that incorporate pharmacologic control are desirable; however, designs suitable for clinical translation are needed. We designed a fully human, rapamycin-regulated drug product for targeting CD33+ tumors called dimerizing agent-regulated immunoreceptor complex (DARIC33). T cell products demonstrated target-specific and rapamycin-dependent cytokine release, transcriptional responses, cytotoxicity, and in vivo antileukemic activity in the presence of as little as 1 nM rapamycin. Rapamycin withdrawal paused DARIC33-stimulated T cell effector functions, which were restored following reexposure to rapamycin, demonstrating reversible effector function control. While rapamycin-regulated DARIC33 T cells were highly sensitive to target antigen, CD34+ stem cell colony-forming capacity was not impacted. We benchmarked DARIC33 potency relative to CD19 CAR T cells to estimate a T cell dose for clinical testing. In addition, we integrated in vitro and preclinical in vivo drug concentration thresholds for off-on state transitions, as well as murine and human rapamycin pharmacokinetics, to estimate a clinically applicable rapamycin dosing schedule. A phase I DARIC33 trial has been initiated (PLAT-08, NCT05105152), with initial evidence of rapamycin-regulated T cell activation and antitumor impact. Our findings provide evidence that the DARIC platform exhibits sensitive regulation and potency needed for clinical application to other important immunotherapy targets.

Keywords: Cancer immunotherapy; Hematology; Leukemias; T cells; Therapeutics.

MeSH terms

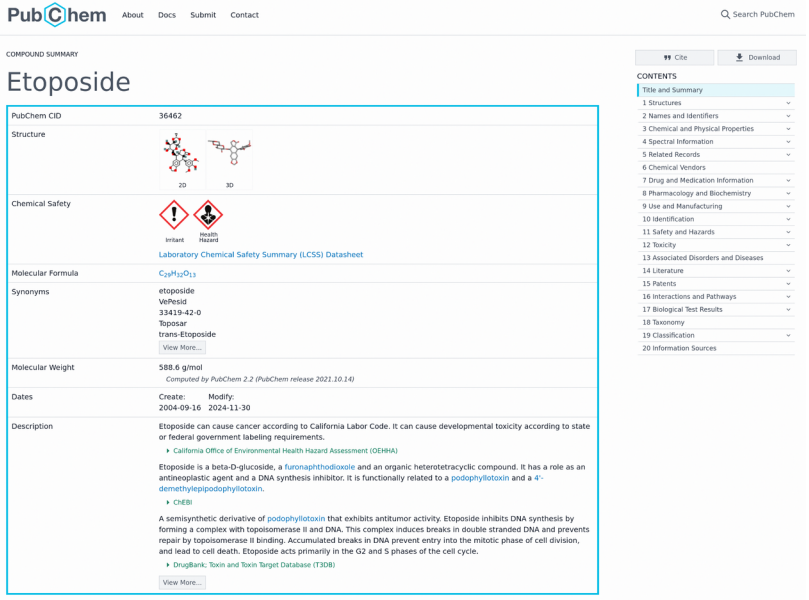
- Animals
- Female
- Humans
- Immunotherapy, Adoptive
- Leukemia, Myeloid, Acute* / drug therapy
- Leukemia, Myeloid, Acute* / immunology
- Leukemia, Myeloid, Acute* / pathology
- Leukemia, Myeloid, Acute* / therapy
- Male
- Mice
- Receptors, Chimeric Antigen / immunology
- Sialic Acid Binding Ig-like Lectin 3* / immunology
- Sialic Acid Binding Ig-like Lectin 3* / metabolism
- Sirolimus* / administration & dosage
- Sirolimus* / pharmacology
- T-Lymphocytes* / drug effects
- T-Lymphocytes* / immunology
- Xenograft Model Antitumor Assays

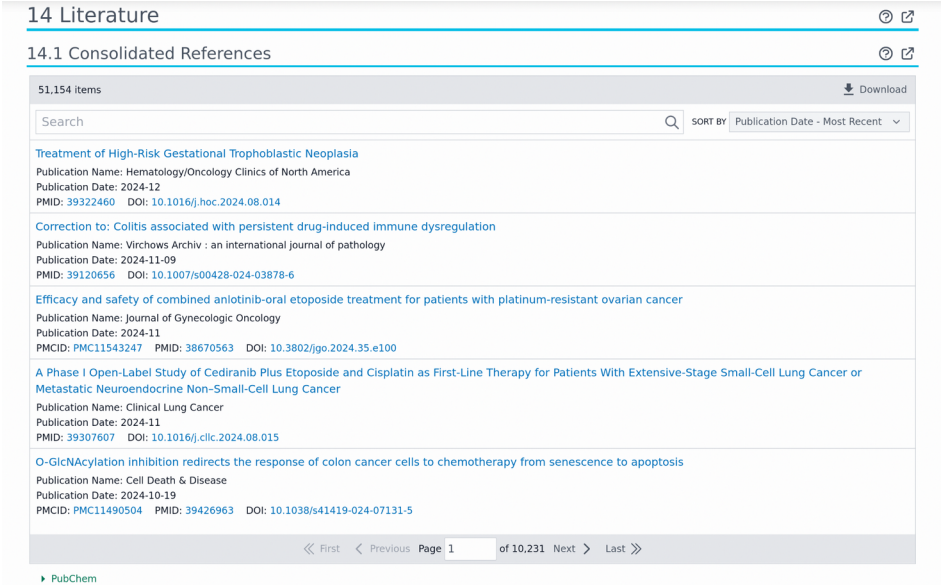
Substances

- CD33 protein, human
- Receptors, Chimeric Antigen
- Sialic Acid Binding Ig-like Lectin 3
- Sirolimus

PubChem

A curated database regarding chemical information including physical and chemical properties, biological activities, and literature evidence

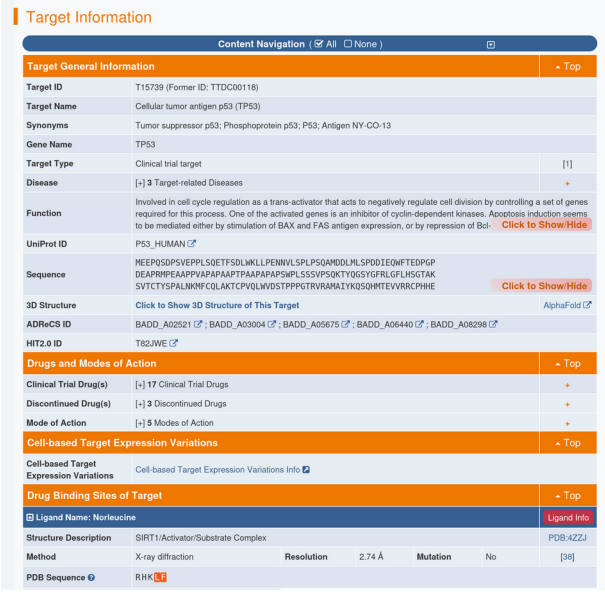




Access at https://pubchem.ncbi.nlm.nih.gov/

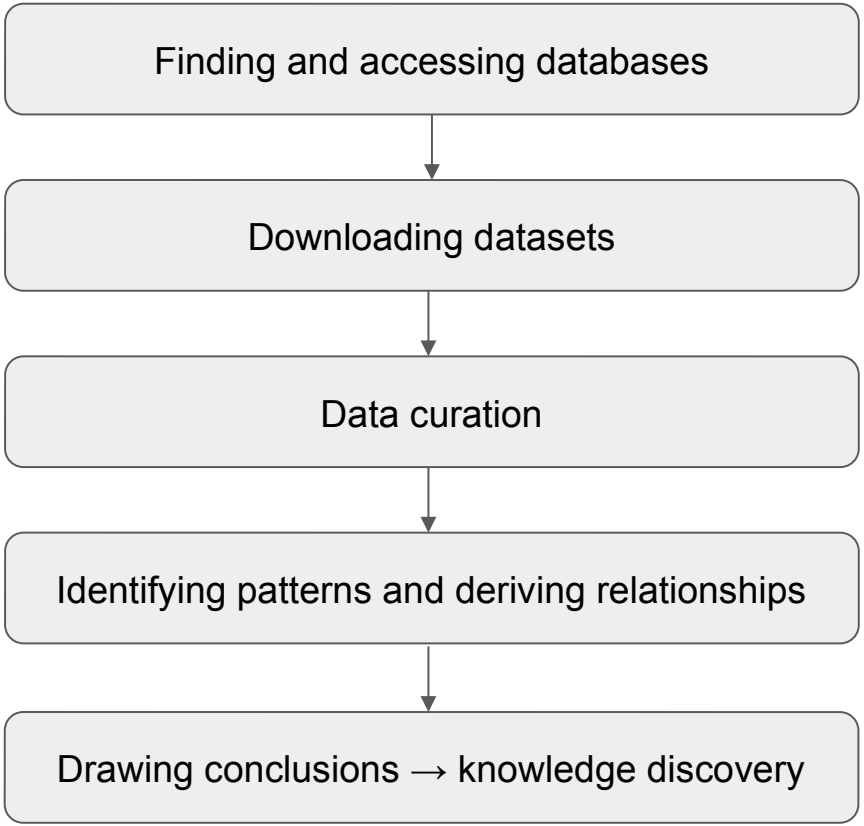
Therapeutic Target Database (TTB)

A curated database concerning therapeutic protein, pathways, diseases, and corresponding drugs



Access at https://idrblab.net/ttd/

Typical data mining process



Typical challenges

- Data Quality
  - Noisy or incomplete data
- Data Integration
  - Combining heterogeneous data types (e.g., genomic, clinical, and imaging data)
- Interpretability
  - Understanding the results of complex models and ensuring they are actionable for biologists/clinicians
- Privacy and Ethics
  - Protecting patient data and ensuring ethical use of medical data.

Application 1: Drug Discovery

Goal: Identifying **new drug candidates** with therapeutic effects or **repurposing existing** drug candidates for novel therapeutic effect

Target identification

High-throughput drug screening

Drug-target interaction

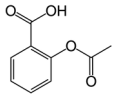
ML models on large datasets to predict new interaction

Biomarker discovery

Mining genetic datasets to predict response

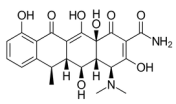
Examples

Aspirin



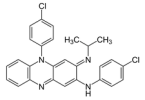
Analgesic → Antiplatelet aggregation

Doxycycline



Antibiotic → Anti-malaria

Clofazimine



Leprosy → Tuberculosis

* Drug structure images obtained from https://en.wikipedia.org/wiki/

Application 2: Personalized Medicine

Goal: **Tailor treatment to individual patients** based on genetic, physiological, and other characteristics

Genetic sequencing

NGS and patient-specific biomarkers

Data integration

Electronic health records to modify treatment

Clinical trials

Identifying patients to be most benefited

Examples

HER2+ breast cancer treatment

Standard treatment
60-80% 5 year OS

Trastuzumab treatment
85-90% 5 year OS[3]

Drug dosing optimization

Statin dosing based on *SLCO1B1* genetic variant[4]

Wearable technologies

Glucose monitoring → real-time insulin adjustments

[3] Piccart-Gebhart et al., 2005, Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. *The New England Journal of Medicine*, 2005;353:1659-1672, DOI: 10.1056/NEJMoa052306

[4] The SEARCH Collaborative Group, 2008, *SLCO1B1* Variants and Statin-Induced Myopathy — A Genomewide Study. *The New England Journal of Medicine,* 2008;359:789-799, DOI: 10.1056/NEJMoa0801936

To remember…

Before starting any experiment, mine the literature and databases to understand the pre-existing data

After obtaining the results, try to integrate with different types of publicly available data to get interesting insights

Exercise

The RNA-Seq data from Hs.505T (Chronic lymphocytic leukemia) cell line shows *BCL2* (Apoptosis regulator Bcl-2) overexpression. Identify drug approved for Chronic lymphocytic leukemia targeting this protein and identify which other genes are co-associated with BCL2 and the identified drug in the literature.

Steps

- Search Target BCL2 in the TTD database (https://idrblab.net/ttd/)
- Identify approved cancer drug and obtain the PubChem ID
- Search PubChem (https://pubchem.ncbi.nlm.nih.gov/) to get PubMed PMIDs of the articles
- Use provided 02_gene2pubmed-subset.csv file and use it with the 04_R-script.R script to identify co-associated genes (https://tinyurl.com/datamining2025)