

KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머신러닝 워크샵(온라인)

Chemoinformatics

이민호

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

Chemoinformatics

본 강의에서는 생물학이나 생물정보학 전공자들이 약물이나 소분자 정보의 활용 과정에 필요한 기초 이론, 지식 및 관련 데이터베이스 정보를 전달하는 것을 목표로 한다. 화합물 데이터베이스 종류 및 DB 내에서 활용할 수 있는 정보의 가공 방법 등을 간단히 소개하며, 추후 기계학습 등에 활용할 수 있도록 분자 구조를 다차원 수치 벡터로 변환하는 기법 등을 다룬다.

강의는 다음의 내용을 포함한다:

- Representation of chemical compounds
- File formats in chemoinformatics
- Chemical databases
- Bioassay databases
- Numerical representation
- Molecular fingerprints
- Hadoop/Spark Programming

* 강의: 이민호 교수 (동국대학교 생명과학과)

Curriculum Vitae

Speaker Name: Minho Lee, Ph.D.



► Personal Info

Name Minho Lee
Title Assistant professor
Affiliation Dongguk University

► Contact Information

Department of Life Science, Dongguk University-Seoul, Ilsandong-gu,
Goyang-si, Gyeonggi-do 10326, Republic of Korea
Email MinhoLee@dgu.edu

Research interest : Precision medicine

Educational Experience

2005 B.S. Dept. of BioSystems, KAIST
2013 Ph.D. Dept. of Bio and Brain Engineering, KAIST

Professional Experience

2013-2014 Post Doc, Information & Electronics Research Institute, KAIST
2014-2016 Assistant professor, Dept. of Biological Sciences, Sangji University
2016-2020 Research assistant professor, Catholic Precision Medicine Research Center, College of Medicine, Catholic University of Korea
2020- Assistant Professor, Dept. of Life Science, Dongguk University

Selected Publications (5 maximum)

1. Lee K. et al., Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server, BMC Bioinformatics, 2017
2. Lee M. et al., Genomic structures of dysplastic nodule and concurrent hepatocellular carcinoma, Human pathology, 2018
3. Lee M. et al., Whole-exome sequencing reveals differences between nail apparatus melanoma and acral melanoma, Journal of the American Academy of Dermatology, 2018
4. Lee M. et al., Circulating microRNA expression levels associated with Internet gaming disorder, Frontiers in psychiatry, 2018
5. Lee M. et al., A novel loci of the HR gene in Marie-Unna hereditary hypotrichosis using whole-exome sequencing, Indian Journal of Dermatology, Venereology, and Leprology, 2020

KSBi-BIML 2021

Chemoinformatics

Minho Lee (MinhoLee@dgu.edu)



본 강의 자료는 한국생명정보학회가 주관하는 KSBi-BIML 2021 워크샵 온라인 수업을 목적으로 제작된것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다. 수업 목적으로 배포 및 전송 받은 경우에도 이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없습니다.

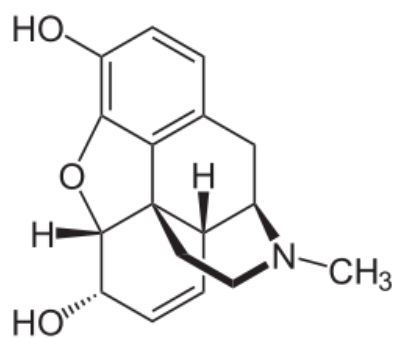
만약 이러한 사항을 위반할 경우 발생하는 모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고합니다.

Topics

- ▶ Representation of chemical compounds
 - ▶ File formats
 - ▶ Chemical databases
 - ▶ Bioassay databases
 - ▶ Numerical representation
 - ▶ Molecular fingerprints
-

How to represent chemical information

- ▶ How can a molecular structure be stored on a computer?
 - ▶ Figure?
 - ▶ Coordinates?



SMILES

- ▶ **Simplified Molecular Input Line Entry System**
 - ▶ Weininger, J Chem Inf Comput Sci, 1988, 28, 31
 - ▶ More recently, a community developed description: <http://opensmiles.org>
 - ▶ Linear format ("line notation") that describes the connection table and stereochemistry of a molecule (i.e. 0D)
 - ▶ Convenient to enter as a query on-line, store in a database
- ▶ **Basic guidelines:**
 - ▶ Hydrogens are implicit
 - ▶ Parentheses indicate branches
 - ▶ Each atom is connected to the preceding atom to its left (excluding branches in-between)
 - ▶ Single bonds are implicit, = for double, # for triple

SMILES examples

SMILES	Name	SMILES	Name
CC	ethane	[OH3+]	hydronium ion
O=C=O	carbon dioxide	[2H]O[2H]	deuterium oxide
C#N	hydrogen cyanide	[235U]	uranium-235
CCN(CC)CC	triethylamine	F/C=C/F	E-difluoroethene
CC(=O)O	acetic acid	F/C=C\F	Z-difluoroethene
C1CCCCC1	cyclohexane	N[C@@H](C)C(=O)O	L-alanine
c1ccccc1	benzene	N[C@H](C)C(=O)O	D-alanine

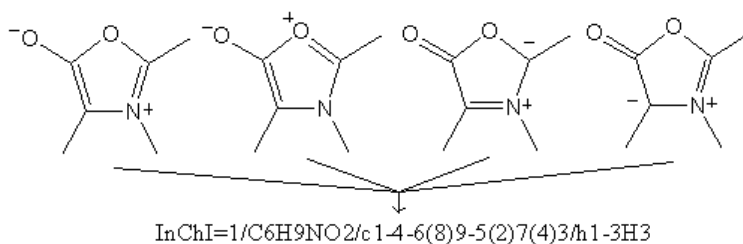
Reaction SMILES	Name
[I-].[Na+].C=CCBr>>[Na+].[Br-].C=CCI	displacement reaction
(C(=O)O).(OCC)>>(C(=O)OCC).(O)	intermolecular esterification

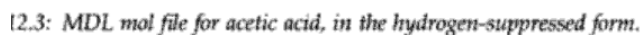
Canonical SMILES

- ▶ In general, many different SMILES strings can be written for the same molecule
 - ▶ Not a unique identifier (one-to-many)
- ▶ Algorithms for producing “canonical SMILES” have been developed
 - ▶ The same unique SMILES string is always created for a particular molecule
 - ▶ One-to-one relationship between structure and representation
 - ▶ Note however, that different software implement different canonicalization algorithms

InChI

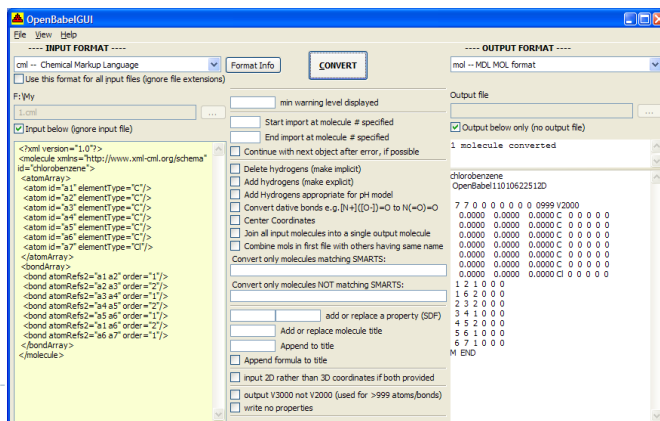
- ▶ International Chemical Identifier
 - ▶ Line notation developed by NIST and IUPAC
 - ▶ Goal: An index for **uniquely** identifying a molecule
 - ▶ Example





Open babel

- ▶ <http://openbabel.org>
- ▶ a chemical expert system mainly used for converting chemical file formats
- ▶ Offers CUI & GUI



Chemical Databases

Database	Content	Size (no. of compounds)	URL
Bioactivity data			
ChEMBL	Bioactivity data from the medicinal chemistry literature	1 360 000	https://www.ebi.ac.uk/chembl/db
PubChem	Biological screening results on small molecules	49 000 000	https://pubchem.ncbi.nlm.nih.gov/
Patents			
IBM	Chemicals from full text patents	2 500 000	http://www-935.ibm.com/services/us/gbs/bao/siip/
SureChEMBL	Chemicals from full text patents	12 400 000	https://www.surechembl.org
Drugs			
DRUGBANK	Drug data and drug target information	7700	http://www.drugbank.ca
FDA/USP SRS	Substances present in FDA regulated products	34 000	http://fdasis.nlm.nih.gov/srs/srs.jsp
Availability			
ZINC	Commercially available compounds	22 700 000	http://zinc.docking.org
emolecules	Commercially available compounds	5 900 000	http://www.emolecules.com
Other			
ChEBI	Database and ontology of Chemical Entities of Biological Interest	27 000	https://www.ebi.ac.uk/chebi/
PDB	Data on biological macromolecular structures	16 000	https://www.ebi.ac.uk/pdbe/

Note: All numbers from Apr 2014.

<http://dx.doi.org/10.1016/j.ddtec.2015.01.005>

PubChem

► <https://pubchem.ncbi.nlm.nih.gov/>

Databases > Upload Services > Help more > Today's Statistics >

PubChem

BioAssay

Compound

Substance

Go

Limits
Advanced

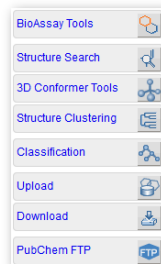
Try the PubChem Search Beta

News The PubChem Data Sources page is now updated. It helps you to find who provided what information in PubChem. [Read more...](#)

News PubChem Widgets 2.0f is released. It substantially updates all table-based widgets and classification widgets, while adding new capabilities and features. [Read more...](#)

[more ...](#)

[Write to Helpdesk](#) | [Disclaimer](#) | [Privacy Statement](#) | [Accessibility](#) | [Data Citation Guidelines](#)
National Center for Biotechnology Information
NLM | NIH | HHS



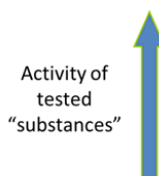
PubChem is organized in 3 separate databases



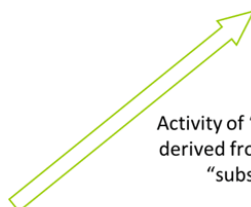
- Depositor-provided
- Unique Identifier: **SID**



- Unique chemical structures
- Unique Identifier: **CID**



Activity of
tested
"substances"



Activity of "compounds"
derived from associated
"substances"

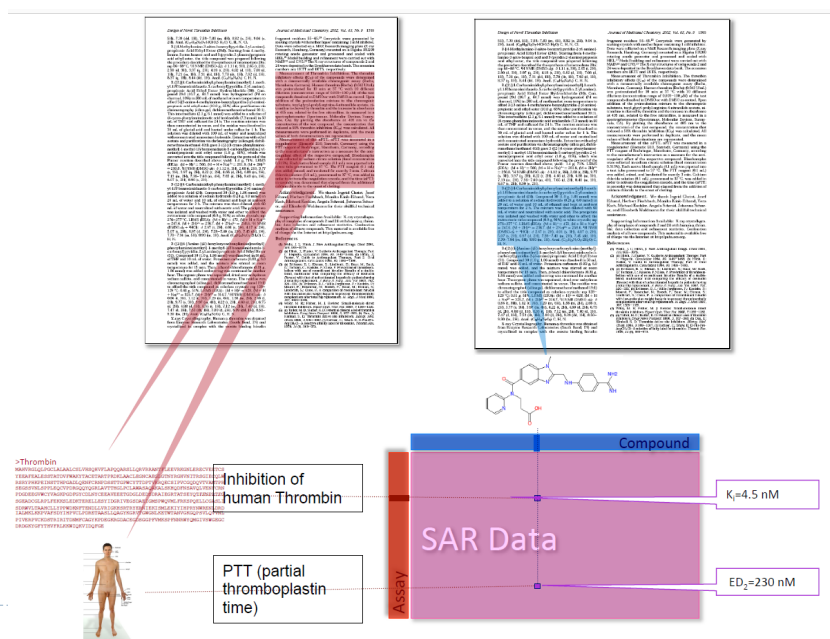


- Biological activity test results
- Depositor-provided
- Unique Identifier: **AID**

ChEMBL

- ▶ Open access database for drug discovery
- ▶ Freely available (searchable and downloadable)
- ▶ Content:
 - ▶ 2D structures & calculated properties (logP, MW, Lipinski, etc.)
 - ▶ Associated bioactivity data extracted from the primary medicinal chemistry journals such as J. Med. Chem.
 - ▶ Deposited data from neglected disease screening (e.g. malaria)
 - ▶ Subset of data from PubChem
- ▶ Covers ~30 years of compound synthesis and testing
- ▶ Annotated FDA-approved drugs

ChEMBL Data generation



ChEMBL Assays – Binding, Functional, ADMET

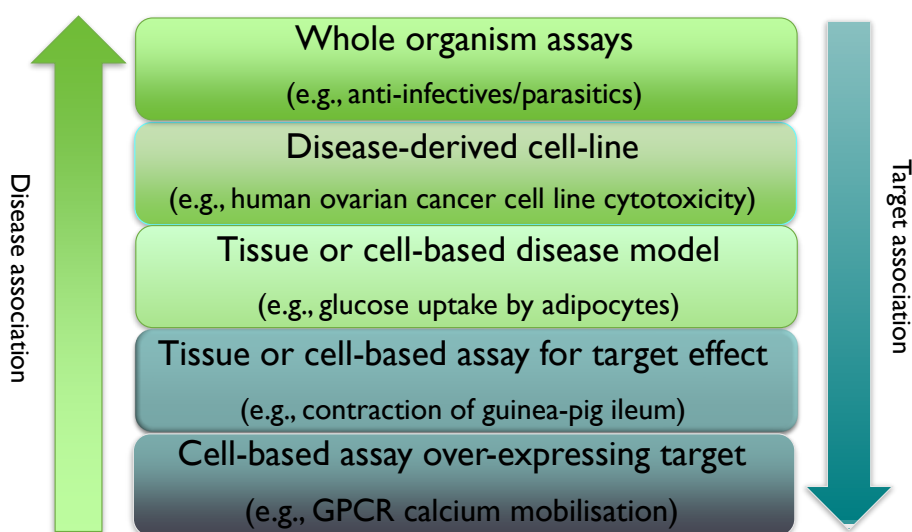
▶ Binding Assays

- ▶ Assays which directly measure the binding of a compound to a particular target
 - E.g., competition binding assays with a radioligand

▶ Various endpoints measured, but most commonly reported are:

- ▶ IC50 (half maximal inhibitory concentration)
- ▶ Ki (binding affinity)
- ▶ MIC (minimum inhibitory concentration)
- ▶ % Inhibition (of activity)

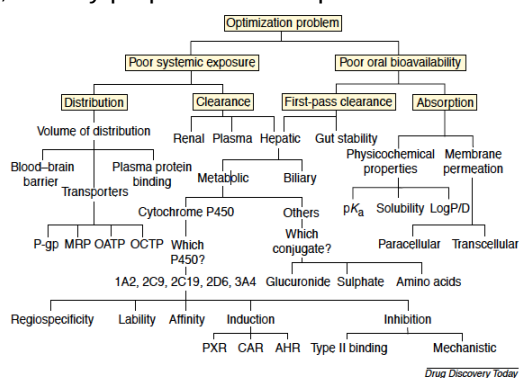
Functional Assays



ADMET Assays

Assays measuring:

Absorption, Distribution, Metabolism, Excretion, Toxicity properties of compounds



Examples include:

- Half-life of compound in rats
- Tissue distribution of compound
- Levels of metabolites

ChEMBL Targets

Protein



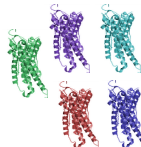
e.g., PDE5

Protein complex



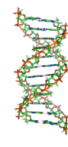
e.g., Nicotinic acetylcholine receptor

Protein family



e.g., Muscarinic receptors

Nucleic Acid



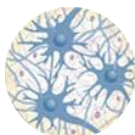
e.g., DNA

Cell Line



e.g., HEK293 cells

Tissue



e.g., Nervous

Sub-cellular Fraction



e.g., Mitochondria

Organism



e.g., Drosophila

Protein Targets

- ▶ Each protein target linked to a sequence in UniProt
- ▶ Information from UniProt used in ChEMBL to allow searching:
 - ▶ Protein name/description
 - ▶ Synonyms and gene names
 - ▶ Organism (and NCBI Tax ID)
- ▶ Proteins in ChEMBL also classified according to family (e.g., Receptor, Kinase, Protease, Transporter etc).
 - ▶ Used for searching by target tree (Browse Targets)

DrugBank example: Acetylsalicylic acid

Aspirin

[Targets \(18\)](#) [Enzymes \(4\)](#) [Transporters \(3\)](#)

IDENTIFICATION			
Name	Acetylsalicylic acid Commonly known or available as Aspirin	Accession Number	DB00945
Description	Also known as <i>Aspirin</i>, acetylsalicylic acid (ASA) is a commonly used drug for the treatment of pain and fever due to various causes. Acetylsalicylic acid has both anti-inflammatory and antipyretic effects. This drug also inhibits platelet aggregation and is used in the prevention of blood clots stroke, and myocardial infarction (MI) ^{Label}. Interestingly, the results of various studies have demonstrated that long-term use of acetylsalicylic acid may decrease the risk of various cancers, including colorectal, esophageal, breast, lung, prostate, liver and skin cancer ¹⁵. Aspirin is classified as a <i>non-selective cyclooxygenase (COX) inhibitor</i> ^{11,14} and is available in many doses and forms, including chewable tablets, effervescent tablets, extended-release formulations, and sublingual tablets ¹⁹.		

DrugBank: ASA targets

TARGETS

1. Prostaglandin G/H synthase 1

... Binding Properties

Details

Kind	Protein	General Function	Prostaglandin-endoperoxide synthase activity
Organism	Humans	Specific Function	Converts arachidonate to prostaglandin H2 (PGH2), a committed step in prostanoid synthesis. Involved in the constitutive production of prostanoids in particular in the stomach and platelets. In gas...
Pharmacological action	Yes		
Actions	Inhibitor		
		Gene Name	PTGS1
		Uniprot ID	P23219
		Uniprot Name	Prostaglandin G/H synthase 1
		Molecular Weight	68685.82 Da

References

1. Flipo RM: [Are the NSAIDs able to compromising the cardio-preventive efficacy of aspirin?]. Presse Med. 2006 Sep;35(9 Spec No 1):1553-60. [PubMed:17078596]
2. Schwartz KA: Aspirin resistance: a review of diagnostic methodology, mechanisms, and clinical utility. Adv Clin Chem. 2006;42:81-110. [PubMed:17131625]
3. Birnbaum Y, Ye Y, Lin Y, Freeberg SY, Huang MH, Perez-Polo JR, Uretsky BF: Aspirin augments 15-epi-lipoxin A4 production by lipopolysaccharide, but blocks the pioglitazone and atorvastatin induction of 15-epi-lipoxin A4 in the rat heart. Prostaglandins Other Lipid Mediat. 2007 Feb;83(1-2):89-98. Epub 2006 Nov 7. [PubMed:17259075]
4. Guthikonda S, Lev EI, Patel R, DeLao T, Bergeron AL, Dong JF, Kleiman NS: Reticulated platelets and uninhibited COX-1 and COX-2 decrease the antiplatelet effects of aspirin. J

ChEBI

- ▶ <http://www.ebi.ac.uk/chebi>
- ▶ Chemical Entities of Biological Interest
- ▶ A freely available, manually curated chemistry database
- ▶ High quality, manually annotated
- ▶ Provides chemical **ontology**


ChEBI

[Home](#)
[Advanced Search](#)
[Browse](#)
[Documentation](#)
[Download](#)
[Tools](#)
[About ChEBI](#)
[Contact us](#)
[Preferences](#)
[Submit](#)

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds.

Search for ☒ only ☐ All in ChEBI

Example: [methylphenidate](#), [H₂O](#), [water](#)

[Advanced Search](#) | [About ChEBI](#)

 **Entity of the month**

1st March 2016
WLL-vs



Funding for the prevention and treatment of [malaria](#) has increased more than 10-fold since the year 2000, with dramatic results. Increased prevention and control measures, including the provision and use of bed nets impregnated with long-lasting (2-3 years) [insecticides](#) to protect people from mosquito bites at night, together with indoor spraying with insecticides, has resulted in a 37% reduction in new malaria cases and a 60% reduction in malaria mortality rates [1].

[Read more ...](#)

 **Documentation**

Tutorial An introduction to the ChEBI database and ontology, showing users how to search and browse the web/programmatic interface.

Statistics Graphs showing the growth of ChEBI, the numbers of curated, submitted and unchecked entries, the numbers of links to other resources and the sources of the information present in ChEBI.

User Manual Learn more about the data fields in the ChEBI and data sources for ChEBI.

[Read more...](#)

 **Downloads**

SDF files ChEBI provides its chemical structures and additional data in structure-data file (SDF) format.

Ontology files ChEBI ontology is provided in the W3C standard [Web Ontology Language \(OWL\)](#) and [OBO](#) formats.

Database files ChEBI is stored in a relational database and we currently provide the ChEBI tables in [flat-file](#), [tab delimited format](#), as an [Oracle binary dumps](#) and a [generic SQL dumps](#) for MySQL and PostgreSQL database.

[Read more...](#)

 **News**

Tweets by @chebi



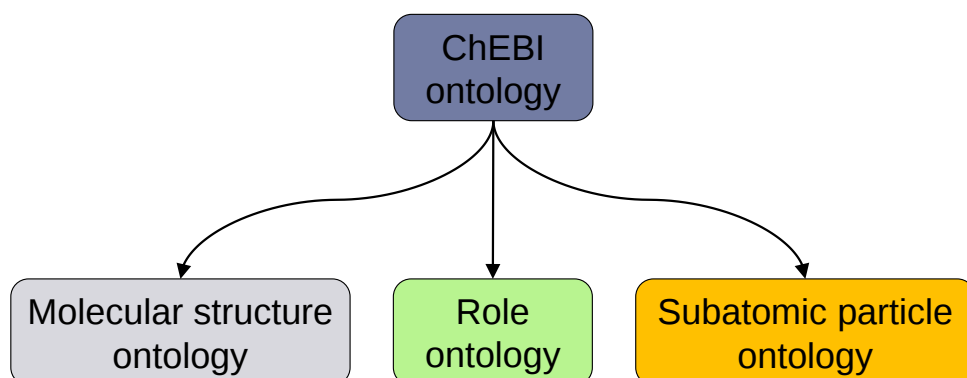
Release 137 is live with 48020 fully annotated entities. See our entity of the month: WLL-vs. [bit.ly/1kLbDP](#)



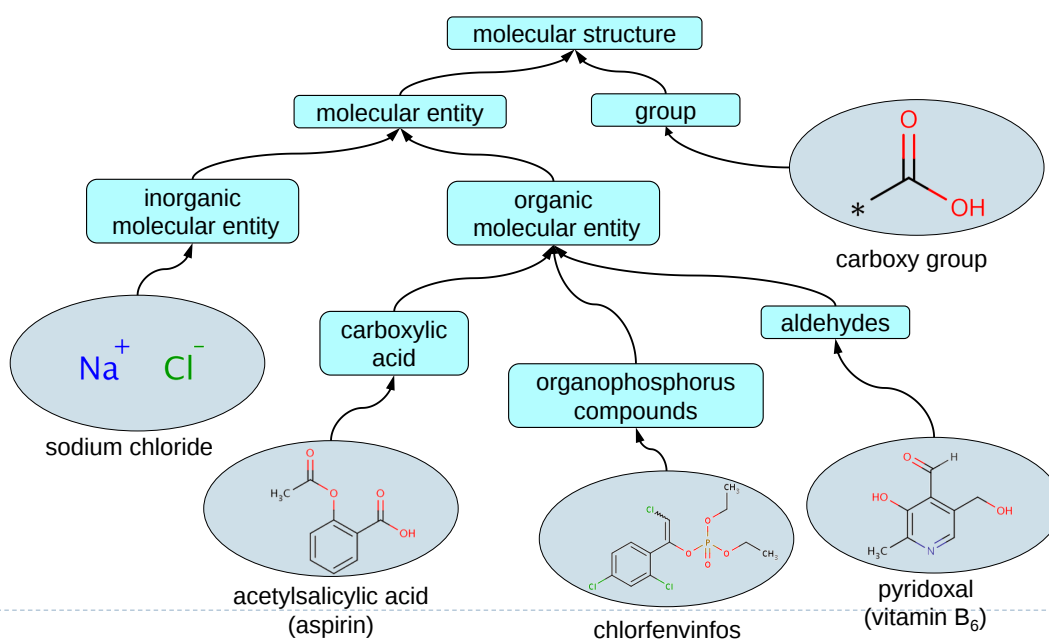
Christoph Steinbeck @steinbeck
European Bioinformatics Institute looking for a [selected author](#) to review the manuscript.

[Archived News...](#)

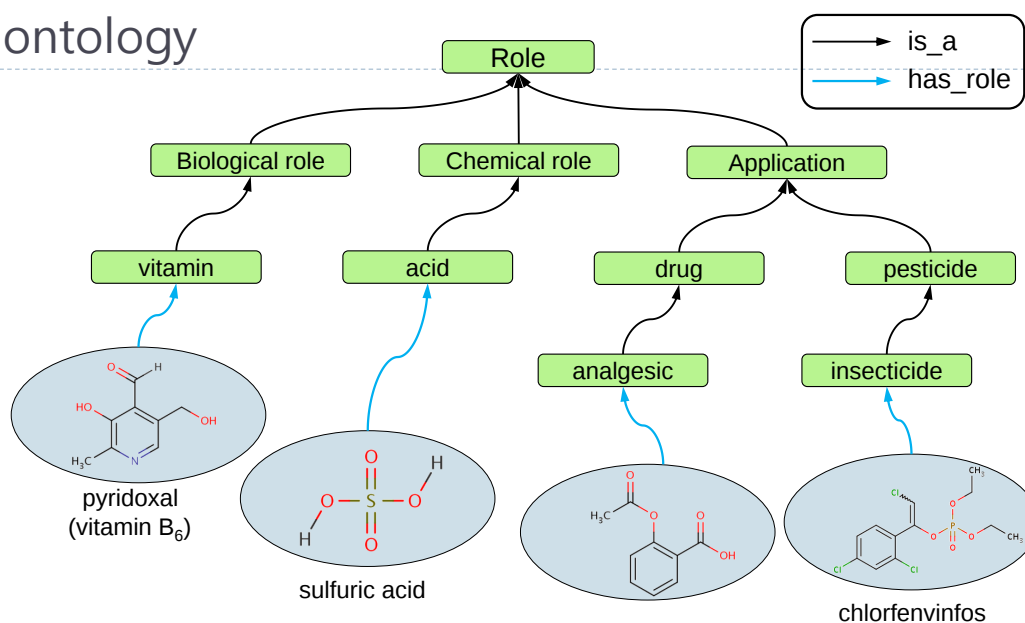
The ChEBI ontology



Molecular structure ontology



Role ontology



ChEBI ontology: aspirin

endoperoxides, precursors of prostaglandins.

[non-narcotic analgesic](#)

A drug that has principally analgesic, antipyretic and anti-inflammatory actions. Non-narcotic analgesics do not bind to opioid receptors.

[View more via ChEBI Ontology](#)

ChEBI Ontology

acetylsalicylic acid ([CHEBI:15365](#)) **has functional parent** salicylic acid ([CHEBI:16914](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** anticoagulant ([CHEBI:50249](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** antipyretic ([CHEBI:35493](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** cyclooxygenase 1 inhibitor ([CHEBI:50630](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** cyclooxygenase 2 inhibitor ([CHEBI:50629](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** drug allergen ([CHEBI:88188](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** EC 1.1.1.188 (prostaglandin-F synthase) inhibitor ([CHEBI:77425](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** non-narcotic analgesic ([CHEBI:35481](#))
[Outgoing](#)
acetylsalicylic acid ([CHEBI:15365](#)) **has role** non-steroidal anti-inflammatory drug ([CHEBI:35475](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** plant activator ([CHEBI:73182](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** platelet aggregation inhibitor ([CHEBI:50427](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** prostaglandin antagonist ([CHEBI:49023](#))
acetylsalicylic acid ([CHEBI:15365](#)) **has role** teratogenic agent ([CHEBI:50905](#))
acetylsalicylic acid ([CHEBI:15365](#)) **is a** acetate ester ([CHEBI:47622](#))
acetylsalicylic acid ([CHEBI:15365](#)) **is a** benzoic acids ([CHEBI:22723](#))
acetylsalicylic acid ([CHEBI:15365](#)) **is a** salicylates ([CHEBI:26596](#))
acetylsalicylic acid ([CHEBI:15365](#)) **is conjugate acid of** acetylsalicylate ([CHEBI:13719](#))

[Incoming](#)
acetylsalicylate ([CHEBI:13719](#)) **is conjugate base of** acetylsalicylic acid ([CHEBI:15365](#))

IUPAC Name

2-(acetyloxy)benzoic acid

INNs

acide acétysalicylique 

ácido acetilsalicílico 

Sources

ChemIDplus

NIST Chemistry WebBook

ZINC

- ▶ <http://zinc.docking.org/>
- ▶ ZINC was originally designed for target based virtual screening (docking)
- ▶ also useful for many other things
 - ▶ finding a compound to purchase
 - ▶ downloading a library in SMILES format for ligand based virtual screening
 - ▶ find compounds by similarity to a starting
 - ▶ find compound ANNOTATED for a particular target (via ChEMBL)
 - ▶ find compounds PREDICTED for a particular target (via ChEMBL or docking)

ZINC subsets

	Lead-Like	Fragment-Like	Drug-Like	All	Shards
Standard Size Updated	Lead-Like 6,053,287 2014-09-29	Fragment-Like 847,909 2015-02-04	Drug-Like 17,999,742 2014-11-24	All Purchasable 22,724,825 2014-11-28	Shards 625,159 2014-05-16
Clean Size Updated	Clean Leads 4,591,276 2014-09-25	Clean Fragments 1,611,889 2014-09-24	Clean Drug-Like 13,195,609 2013-11-05	All Clean 16,403,865 2013-12-18	Clean Shards 325,950 2014-11-24
In Stock Size Updated	Leads Now 3,687,621 2014-06-25	Frgs Now 704,041 2015-02-04	Drugs Now 10,639,555 2014-11-24	All Now 12,782,590 2014-05-01	Shards Now 424,775 2014-09-24
Boutique Size Updated	Boutique Leads 5,114,169 2012-12-24	Boutique Frags 2,755,535 2013-11-08	Boutique Drugs 10,292,210 2012-11-27	All Boutique 12,217,845 2012-11-27	Boutique Shards 80,698 2013-11-08
Comments/Citation	Teague, Davis, Leeson, Oprea, Angew Chem Int Ed Engl. 1999 Dec 16;38(24):3743-3748.		Carr RA, Congreve M, Murray CW, Rees DC, Drug Discov Today. 2005 Jul;10(14):987.		Purchasable chemical space
Filtering Criteria	p.mwt <= 350 and p.mwt >= 250 and p.xlogp <= 3.5 and p.rb <= 7		p.xlogp <= 3.5 and p.mwt <= 250 and p.rb <= 5		p.mwt < 190
			p.mwt <= 500 and p.mwt >= 150 and p.xlogp <= 5 and p.rb <= 7 and p.psa < 150 and p.n_h_donors <= 5 and p.n_h_acceptors <= 10		

Protein Data Bank(PDB): structure database

RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login


An Information Portal to
123622 Biological
Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands
Go

Advanced Search | Browse by Annotations








Welcome

Deposit
Search
Visualize
Analyze
Download
Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

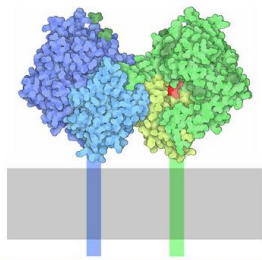
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Events and Activities




October Molecule of the Month



Dipeptidyl Peptidase 4

Latest Entries
As of Tuesday Oct 18
Features & Highlights
News
Publications
Contact Us

PDB file (text format)

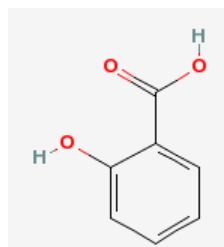
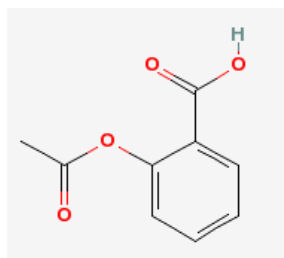
```
HEADER      TRANSFERASE                      17-JUN-02   1M17
TITLE       EPIDERMAL GROWTH FACTOR RECEPTOR TYROSINE KINASE DOMAIN
TITLE       2 WITH 4-ANILINOQUINAZOLINE INHIBITOR ERLOTINIB
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: EPIDERMAL GROWTH FACTOR RECEPTOR;
COMPND      3 CHAIN: A;
COMPND      4 FRAGMENT: TYROSINE KINASE DOMAIN (RESIDUES 671-998);
COMPND      5 SYNONYM: RECEPTOR PROTEIN-TYROSINE KINASE ERBB-1;
COMPND      6 EC: 2.7.1.112;
COMPND      7 ENGINEERED: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE      3 ORGANISM_COMMON: HUMAN;
SOURCE      4 GENE: EGFR;
SOURCE      5 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE      6 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
SOURCE      7 EXPRESSION_SYSTEM_STRAIN: AUTOGRAPHICA
SOURCE      8 CALIFORNICA/T.NICOPLUSIA;
SOURCE      9 EXPRESSION_SYSTEM_CELL_LINE: SF9;
SOURCE      10 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE      11 EXPRESSION_SYSTEM_PLASMID: PVL1392
KEYWDS      TRANSFERASE, TYROSINE KINASE DOMAIN
EXPDTA      X-RAY DIFFRACTION
AUTHOR      J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
REVDAT      2 25-FEB-03 1M17 1 JRNL
REVDAT      1 04-SEP-02 1M17 0
JRNL        AUTH J.STAMOS,M.X.SLIWKOWSKI,C.EIGENBROT
JRNL        TITL STRUCTURE OF THE EPIDERMAL GROWTH FACTOR RECEPTOR
JRNL        TITL 2 KINASE DOMAIN ALONE AND IN COMPLEX WITH A
JRNL        TITL 3 4-ANILINOQUINAZOLINE INHIBITOR.
JRNL        REF J.BIOL.CHEM. V. 277 46265 2002
JRNL        REFN ASTM JBCHA3 US ISSN 0021-9258
REMARK      1
REMARK      2
REMARK      2 RESOLUTION. 2.60 ANGSTROMS.
```

Molecular similarity

- ▶ Structurally similar molecules tend to have similar properties
- ▶ If we can measure similarity somehow
 - ▶ Can construct a distance matrix
 - ▶ Such matrices can be used to cluster compounds
 - ▶ Can use to find molecules in a database similar to a particular query
 - ▶ Can find unknown molecules with a similar property
 - ▶ Can use to see whether a particular property is correlated with molecular similarity

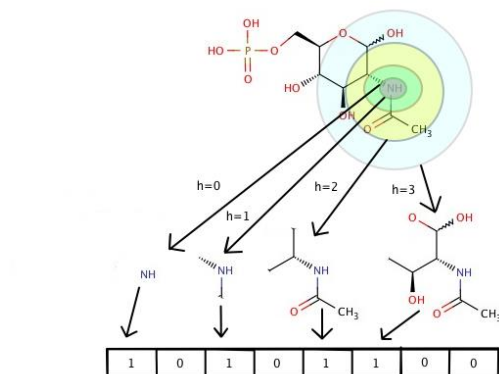
...But how to measure similarity?

- ▶ How similar are aspirin (A) and salicylic acid (B)?

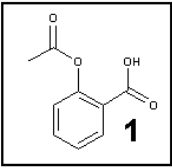
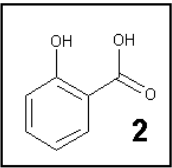


Chemical fingerprint

- ▶ A molecular fingerprint is an encoding of the molecular structure onto a (long) binary string



Tanimoto coefficient

	 1				 2			
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

A = Number of bits set in both = 3

B = Number of bits set in (1), but not in (2) = 2

C = Number of bits set in (2), but not in (1) = 0

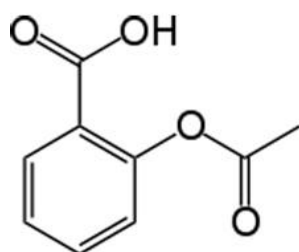
$$\text{TANIMOTO COEFFICIENT} = A / (A + B + C)$$

$$= 3 / (3 + 2 + 0) = 0.6 \text{ or } 60\%$$

Types of fingerprint

- ▶ PubChem
- ▶ Daylight
- ▶ Extended Connectivity Fingerprint (ECFP)
- ▶ ...

PubChem fingerprint



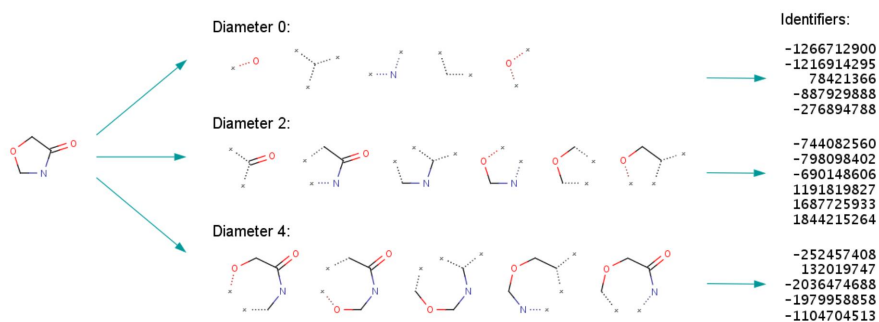
Aspirin Structure

Bit Position	Bit Substructure
0	≥ 4 H
1	≥ 8 H
2	≥ 16 H
3	≥ 32 H
4	≥ 1 Li
5	≥ 2 Li
6	≥ 1 B
7	≥ 2 B
8	≥ 4 B
...	...

Pubchem Molecular Fingerprint

ECFP

► Extended Connectivity Fingerprint



► ECFP_2, ECFP_4, ECFP_6, ... : depending on diameter

<https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP>

Identifier list representation:

- ▶ Bit string length
 - ▶ Larger length decreases the likelihood of bit collision

- ▶ Chemistry Development Kit (CDK)

- ▶ JAVA
 - ▶ <https://cdk.github.io/>
- ▶ RDKit
 - ▶ C++, Python
 - ▶ <https://www.rdkit.org/>
- ▶ R packages
 - ▶ rcdk in CRAN
 - ▶ Rcpin in bioconductor