

Postup analýzy Malaria Dat

Na stránkách <https://www.ebi.ac.uk/chembl/malaria/> jsem vyhledala enzym DHFR. Následně jsem vybrala ta data, která pochází ze Zimničky tropické.

Sponsored by: MMV Medicines for Malaria Venture
Powered by: ChEMBL

Malaria Data

Dihydrofolate Reductase

Compounds Targets Assays Documents

ChEMBL Target Search Results: 30 Please select....

Show / hide columns 10 records per page

ChEMBL ID	Preferred Name	UniProt Accession	Target Type	Organism	Compounds	Bioactivities	
CHEMBL1939	Dihydrofolate reductase	P13922	SINGLE PROTEIN	Plasmodium falciparum K1	157	1753	<input checked="" type="checkbox"/>
CHEMBL2363	Dihydrofolate reductase	Q920D2	SINGLE PROTEIN	Rattus norvegicus	100	303	<input checked="" type="checkbox"/>
CHEMBL202	Dihydrofolate reductase	P00374	SINGLE PROTEIN	Homo sapiens	93	282	<input checked="" type="checkbox"/>
CHEMBL6441	Dihydrofolate reductase	B0BL08	SINGLE PROTEIN	Escherichia coli	82	273	<input checked="" type="checkbox"/>
CHEMBL2426	Dihydrofolate reductase	Q07422	SINGLE PROTEIN	Toxoplasma gondii	18	172	<input checked="" type="checkbox"/>
CHEMBL1926	Dihydrofolate reductase	P16184	SINGLE PROTEIN	Pneumocystis carinii	21	149	<input checked="" type="checkbox"/>
CHEMBL4664	Dihydrofolate reductase	P00376	SINGLE PROTEIN	Mus musculus	36	124	<input checked="" type="checkbox"/>
CHEMBL1075051	Dihydrofolate reductase	P00376	SINGLE PROTEIN	Bos taurus	24	116	<input checked="" type="checkbox"/>
CHEMBL2902	Dihydrofolate reductase	P00381	SINGLE PROTEIN	Lactobacillus casei	35	103	<input checked="" type="checkbox"/>
CHEMBL2675	Dihydrofolate reductase	P00376	SINGLE PROTEIN	Gallus gallus	32	63	<input checked="" type="checkbox"/>

Malaria Data Statistics

- Last Update: CHEMBL_17
- Targets: 5,980
- Compound records: 371,255
- Distinct compounds: 282,295
- Activities: 4,057,545
- Publications: 25,726

Následně jsem vybrala ta data, která měla udnou hodnotu IC50.

Target Components

Component Description	Relationship	Accession
Bifunctional dihydrofolate reductase-thymidylate synthase	SINGLE PROTEIN	P13922

Approved Drugs

ChEMBL ID	Name	Mechanism of Action	References
CHEMBL1201069	CHLOROGUANIDE HYDROCHLORIDE	Dihydrofolate reductase inhibitor	DailyMed
CHEMBL588912	CYCLOGUANIL PAMOATE	Dihydrofolate reductase inhibitor	PubMed
CHEMBL36	PYRIMETHAMINE	Dihydrofolate reductase inhibitor	DailyMed

Target Associated Bioactivities

ChEMBL Activity Types for Target CHEMBL1939

Total: 1753

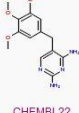
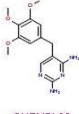
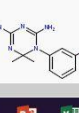
Target Associated Assays

ChEMBL Assays for Target CHEMBL1939

Nakonec jsem stáhla takto získaný set jako bioactivities-16_15_40_42.tab.

ChEMBL Bioactivity Search Results: 340

Show / hide columns 10 records per page

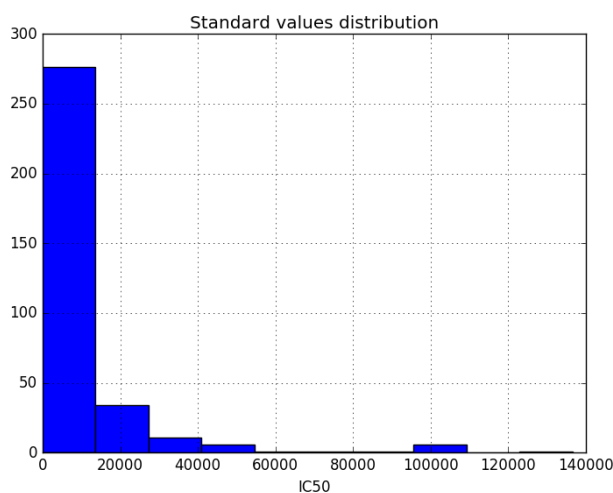
Ingredient	Molweight	Standard Type	Relation	Standard Value	Standard Units	Assay Type	Description	Assay Src Description	Assay Organism	Target Type	Target Name	Target Organism	Reference
 CHEMBL22	290.32	IC50	=	10	nM	B	Inhibition of Plasmodium falciparum DHFR	Scientific Literature	Plasmodium falciparum	SINGLE PROTEIN	Dihydrofolate reductase	Plasmodium falciparum K1	Bioorg. Med. Chem. Lett. (2006) 16:16-43
 CHEMBL22	290.32	IC50	=	136500	nM	F	Antiplasmodial activity IC50 against Plasmodium falciparum K1CB1 DHFR double-mutant (C59R/S10)	Scientific Literature	Plasmodium falciparum	SINGLE PROTEIN	Dihydrofolate reductase	Plasmodium falciparum K1	J. Med. Chem. (2002) 45:6:124
 CHEMBL22	251.72	IC50	=	50	nM	F	In vitro anti-plasmodial activity against Plasmodium falciparum	Scientific Literature	Plasmodium falciparum	SINGLE PROTEIN	Dihydrofolate reductase	Plasmodium falciparum K1	J. Med. Chem. (2004) 47:3:673

Takto získaná data jsem pomocí pythonu načetla.

Nejdříve jsem odstranila data, která neměla vyplněnou hodnotu IC50 a ověřila, zda jsou hodnoty ve stejných jednotkách.

Pomocí rdKitu jsme si spočítala všechny dostupné deskriptory pro dané molekuly a odstranila ty, které neměly pro daná data žádnou vypovídající hodnotu.

Také jsem se podívala jak vypadá rozložení středních hodnot ve všech datech



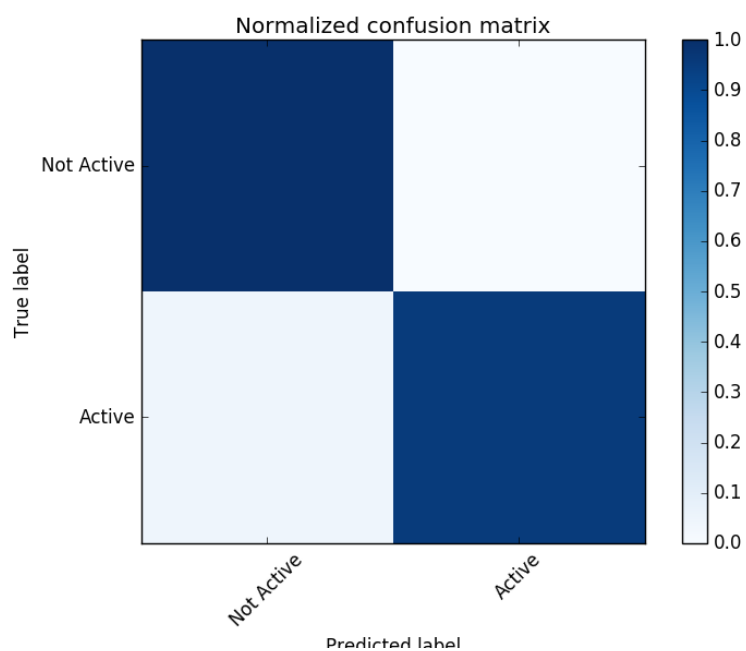
Podle zvoleného thresholdu jsem data rozdělila na aktivní/neaktivní a následně je rozdělila na trénovací/testovací v poměru 70x30. Takto rozdělená data jsem si uložila do souboru train.csv a test.csv.

Spočítala jsem si průměrnou podobnost (pomocí fingerprints a Tanimotovi vzdálenosti) mezi jednotlivými molekulami trénovací sady, která mi vyšla 0,5.

Dále jsem si vypočetla PCA a poté natrénovala SVM model. Tento model jsem trénovala s různými parametry a vybrala nejlepší z nich.

Pro testovací data jsem ověřila, že jsou si podobné více než 0,125.

Na takto přefiltrovaných datech jsem natrénovaný model vyzkoušela a vypsala výsledky spolu s grafem konfuzní matice.



Funkci `qsarIC50`, která počítá celý model lze upravit vstupní parametry a to tak, že lze změnit velikost testovací množiny a treshold pro rozdělení dat na aktivní a neaktivní.