Machine Learning in Toxicology: Fundamentals of Application and Interpretation

Sean Ekins, Ph.D., D.Sc. Email <u>collaborationspharma@gmail.com</u> Phone 215-687-1320



Collaborations Pharmaceuticals, Inc.

Laboratories past and present



Lavoisier's lab 18th C



Author's lab $21^{th}\ C$





Edison's lab 20th C

+ Network of global collaborators

What is Cheminformatics?

Cheminformatics combines the scientific working fields of chemistry, computer science and information science

How and why do we use it for drug discovery?

3

- Learn from data to suggest compounds to make or avoid
- Can increase efficiency / cost effectiveness
- Minimize use of animals and costly materials
- Predict failure



Speeding drug discovery with machine learning



- Molecular pattern recognition of biological data
- Descriptors identify these patterns
- Define active and inactive features
- Used to generate predictions for drug activity at a certain target (organism, protein of interest)

What is Machine Learning?

- Find patterns in data to create insights -
- ▶ We use examples of the correct output for a given input
- The algorithm learns from this input data
- The program created by the algorithm recognizes the correct response
- It works on objects similar to what it was trained on as well as new examples



What can we model?

- Data with responses
- Image recognition
- ► Voice recognition
- Text recognition



- e.g molecules tested against a target / disease
- Complex data from genomics, proteomics, metabolomics
- e.g. microarray data for 100s, 1000s of compounds





What algorithms do we use?



Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

MathWorks

The Workflow





How machine learning works





Which Machine Learning Tribe are you?



Bayesians

Likelihood Prior

Assess the

likelihood of

probabilistic

occurrence for

Posterior Margin

What are the five tribes?



Use symbols, rules, and logic to represent knowledge and draw logical inference

Favored algorithm Rules and decision trees inference Favored algorithm

Naive Bayes Neural

Connectionists



Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm networks

Evolutionaries

Generate variations and then assess the fitness of each for a given purpose

Favored algorithm Genetic programs

Analogizers

Optimize a function in light of constraints ("going as high as you can while staying on the road")

Favored algorithm Support vectors

Source: Pedro Domingos, The Master Algorithm, 2015

or Markov

Bayesian Machine Learning

Bayesian classification is a simple probabilistic classification model. It is based on Bayes' theorem

 $p(h|d) = \frac{P(d|h)P(h)}{P(d)}$ *h* is the hypothesis or model *d* is the observed data *p(h)* is the prior belief (probability of hypothesis *h* before observing any data) *p(d)* is the data evidence (marginal probability of the data) *p(d|h)* is the likelihood (probability of data *d* if hypothesis *h* is true) *p(h|d)* is the posterior probability (probability of hypothesis *h* being true given the observed data *d*)

A weight is calculated for each feature using a Laplacian-adjusted probability estimate to account for the different sampling frequencies of different features.

The weights are summed to provide a probability estimate

Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

Ekins, Williams and Xu, Drug Metab Dispos 38: 2302-2308, 2010

Naïve Bayes

How It Works

A naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It classifies new data based on the highest probability of its belonging to a particular class.

Best Used...

- For a small dataset containing many parameters
- When you need a classifier that's easy to interpret
- When the model will encounter scenarios that weren't in the training data, as is the case with many financial and medical applications



Bayesian Models - Examples

PXR

Human sodium taurocholate co-transporting polypeptide (NTCP), Human Multidrug And Toxin Extrusion Proteins, MATE1 and MATE-2K Human apical sodium-dependent bile acid transporter Cytochrome P450 3A4 Time-Dependent Inhibition Human organic cation/carnitine transporter Volume of distribution hERG TB

- IB
- DILI

BBB, nephrotox, malaria, S Aureus, microsomal stability, cytotoxicity etc

Comparison of SVM and Bayesian ADME/Tox classification models generated with the same molecular descriptors.

Model	Reference	SVM 5 fold cross validation ROC	Bayesian 5 fold cross validation ROC	Bayesian Cross validation ROC
DILI (N = 532)	Ekins, Williams et al. (2010)	0.88	0.63	0.74
PXR (N = 312)	Kortagere et al. (2009)	0.81	0.78	0.84
$5HT_{2B}$ (N = 238)	Chekmarev et al. (2008)	0.83	0.82	0.87
hERG (N = 134)	Chekmarev et al. (2008)	0.82	0.71	0.74
hERG (N = 806)	Wang et al. (2012)	0.88	0.84	0.87
hERG (N = 305,616) ^a	Du et al. (2011)	0.83	0.85	0.86
BBB (N = 1968)	Martins et al. (2012)	0.90	0.91	0.92
AMES (N = 6512)	Hansen, Mika et al. (2009)	0.86	0.84	0.84
Nephrotoxicity (N = 104)	Lin and Will (2012)	0.53	0.64	0.65
Clearance iv $(N = 512)^b$	Gombar and Hall (2013)	0.73	0.74	0.76
VDss iv $(N = 556)^c$	Gombar and Hall (2013)	0.84	0.81	0.87

NT not tested. Note DILI model used ECFC_6 fingerprint descriptors instead of FCFP_6.

^a Random forest Tree Out-of-bag training data ROC score: 0.78.

^b SVM (regression)–5 fold cross validation q² = 0.17.

^c SVM (regression)-5 fold cross validation q² = 0.47.

J Pharmacol Toxicol Methods 69: 115-140 (2014)

hMATE1 Bayesian Model Features

+ve

-ve



ROC = 0.88, leave out 50% x 100 ROC = 0.82 Bad features pyrole -low basicity Charge important for increasing interaction with transporter Astorga et al., JPET 341: 743-755 (2012)



Model resources for ADME/Tox

QSAR TOOLBOX









ToxPredict



Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

eTOXIab, an open source modeling framework for implementing predictive models in production environments Pau Carrió, Oriol López, Ferran Sanz and Manuel Pastor*

* Corresponding author: Manuel Pastor manuel.pastor@upf.edu

Author Affiliations

Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del Mar Medical Research Institute), Dr. Aiguader 88, Barcelona, E-08003, Spain

For all author emails, please log on.

Journal of Cheminformatics 2015, 7:8

doi:10.1186/s13321-015-0058-6

The electronic version of this article is the complete one and can be found online at: http://www.jcheminf.com/content/7/1/8

This article is part of the series Jean-Claude Bradley Memorial Series.

Database

Highly accessed Open Access

QSAR DataBank repository: open and linked qualitative and quantitative structure-activity relationship models

V Ruusmann, S Sild and U Maran^{*}

* Corresponding author: U Maran <u>uko.maran@ut.ee</u>

Author Affiliations

Institute of Chemistry, University of Tartu, Ravila 14a, Tartu, 50411, Estonia

For all author emails, please log on.

Journal of Cheminformatics 2015, 7:32 doi:10.1186/s13321-015-0082-6

The electronic version of this article is the complete one and can be found online at: http://www.jcheminf.com/content/7/1/32

What's stopping us?

- Plenty of data available today... incorrectly formatted
- Vague details of experiments
- Minor & major errors in supplied SMILES/structures
- How do we know this structure is correct?
- How do we share results?
- How can the average scientist use this technology?

17

Quality of data

- Free from structure errors
- Free from data errors
- Free from experimental errors
- Structure Validation and Standardization
- Curation
- Annotation
- Structure filters
 - Incorrect valency, atom labels, aromatic bonds, stereochemistry, salts, duplication
- Structure standardization guidelines
 - Provided by the FDA (Substance Registration System Unique Ingredient Identifier (UNII): http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSyste m-UniqueIngredientIdentifierUNII/default.htm)
- Need a record of molecule provenance

Drug Discov Today. 2012 Jul;17(13-14):685-701. doi: 10.1016/j.drudis.2012.02.013. Epub 2012 Mar 8.

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation.

Williams AJ¹, Ekins S, Tkachenko V.

OPEN & ACCESS Freely available online

Enhancing Hit Identification in Mycobacterium tuberculosis Drug Discovery Using Validated Dual-Event **Bayesian Models**

Sean Ekins^{1,2}*, Robert C. Reynolds³", Scott G. Franzblau⁵, Baojie Wan⁵, Joel S. Freundlich^{4,6}, Barry A. Bunin

1 Collaborative Drug Discovery, Burlingene, California, United States of America, 2 Collaborations in Overnistry, Fuguay-Varina, Noth Carolina, United States of America, 3 Southern Research Institute, Birningham, Alabama, United States of America, 4Decastment of Pharmacology & Phylology, UMDNJ - New Jersey Medical School Newark, New Jensey United States of America, 5 institute for Tubercalouis Research, University of Illinois at Orlogo, Orlogo, Illinois, United States of America, 6Decartment of Medicine, Center for Emerging and Reemerging Pathogens, UMDNJ - New Jersey Medical School, Newalk, New Jersey, United States of America

Abstract

Ce

High-throughput screening (HTS) in whole cells is widely pursued to find compounds active against Mycobacterium tuberculasis (Mtb) for further development towards new tuberculosis (TB) drugs. Ht rates from these screens, usually conducted at 10 to 25 µM concentrations, typically range from less than 1% to the low single digits. New approaches to urgently needed to learn from past scree increase the efficiency of hit identification are industry has for many years taken advantage of computational approaches to optitesting, a practice not fully embra ced by academic laboratories in the search for new TB where we manuful a land when the set of the approaches, we analist Mth has

Bayesian Models Leveraging Bioactivity

Lisa K. Woohiser,⁶ Anne, J. Lenaerts,⁶ Barry A. Bunin,¹ Nancy Co.

*Correspondence: ekinssean@yahoo.com (S.E.), fraundjs@umdnj.edu (J.S.F.)

Identification of unique leads represents a significant

challenge in drug discovery. This hurdle is magnified

in neglected diseases such as tuberculosis. We have

leveraged public high-throughput screening (HTS)

data to experimentally validate a virtual screening

approach employing Bayesian models built wi

bioactivity i formation (single-event model) as will

two orders of magnitude. This initial dual-event

Bayesian model identified compounds with antitu-

bercular whole-cell activity and low mammalian cell

cytotoxicity from a published set of antimalarials.

The most potent hit exhibits the in vitro activity an

Modern drug discovery must be more time and cost efficient

In discovering novel therapeutics. These challenges are felt

even more significantly in the search for neglected disease

treatments, where public-private partnerships coordinate drug

culosis (TB), caused by Mycobacterium tuberculosis (Mtb),

and results in 1.7-1.8 million deaths annually (Lienhardt et al.,

2012a) New drugs active against Mtb are urgently needed to

combat a pandemic heavily affected by resistance to available

therapies and coinfection with HIV/ADS (Numberger et al.,

2010). TB drug discovery is challenging, reflected in the lack of

370 Chemistry & Biology 20, 370-378, March 21, 2013 @ 2013 Elsevier Ltd All right's reserved

which infects approximately one-third of the world's population

discovery with very limited resources. A prime example is tuber

in vitro/in v to safety profile of a drug lead. The

Bayesian models offer signitime and cost to dug discover

IN TRODUCTION

as bloactivy and opticative information (de-event mode). We virially powered a compary library and experimentally control active -hit rates exceeding typical HTS results by one

Les K. Wochper, "Anne J. Lenardry," Barry A. Burg Tockboortien (B. Dieswerk) 283 Bar Mon Highs, Sai "Cokboortien is Chempy, 254. Hits) Net Strave F. dd. "Southern Face and high says. South Error & Sauth Err "Oppartment of Madeiru, "Gar are for energin and Ruman "Oppartment of Chemp anong and Provide Ward analy Madeiri Solthor, Grivenith of Madema and Dari Madeiri Solthor, Grivenith of Madema and Dari David analy Madeiri Solthor, Grivenith of Madema and Dari

These authors contributed equally to this work

http://dx.doi.org/10.1016/j.chembiol.2013.01.011

Birmingham, AL 35294-1240, USA

SUMMARY

and Cytotoxicity Information for Drug Discovery

Sean Ekins, 1, 27,* Robert C. Reynolds, 3/8 Hiyun Kim, 4 Mi-Sun Koo, 4 Marilyn Ekonomidis, 4 Meliza Talaue, 4 Steve

*Department of Microbiology, Immundiogy and Pathology, Colorado State University, 200 West Lake Street, Fort Collins, CO 8

*Present address: Department of Chemistry, College of Arts and Sciences, University of Alabama at Birmingham, 1530 Third Av

10

and inactives

a new TB focused therapeutic approved in over 40 years (Gros- (Schneider, 2010). These and other cheminformatics metho

DADED

Introduction

Chemistry &

set et al., 2012; Sacchettini et al., 2008). One response has be

to screen very large compound libraries (Ananthan et al., 20

Maddry et al., 2009; Reynolds et al., 2012), hoping to deli

on the promise of chemical diversity (O' Connor et al., 201

Phenotypic whole-cell high-throughput screens of commerc

Ibrartes have searched for inhibitors of mycobacterial grow

at a cost of millions of dollars, with resultant low single-digit

less) hit rates (Macarron et al., 2011; Magnet et al., 2010; N

et al., 2012; Stanley et al., 2012). The campaig is have result

in numerous hits, but resource constraints have limited folio

and Forument processing and the order of the second second

another resulted in the early-phase candidate SQ109 (L

et al., 2003). Although SQ109 arose directly from a library

congeners of the frontline drug ethembutol, high-through

screening (HTS) typically does not deliver a clinical candida

Exhaustive optimization of a screening bit must occur, initia

following whole-cell activity and then considering pharmace

hits due to resource limitations but the entire data set of activ

inactives, combined with machine learning models, can sign

cantly focus compound selection and improve screening e

clency (Brins and Freundlich, 2011; Brins et al., 2010c, 2010

2011), as practiced in the pharmaceutical industry (Prathip

et al., 2008, to improve the performance of virtual screen

We hypothesize that prior knowledge of Mtb actives a

PLOS ONE

A collaborative database and computational models for tuberculosis

www.rsc.org/molecularbiosystems | Molecular Bio

drug discoverv[†]

Sean Ekins,*abcd Justin Bradford," Krishna Dole," Anna Spektor," Kellan Gregory," David Biondeau," Moses Hohman" and Barry A. Bunin"

Received 28th August 2009, Accepted 21st December 2009 First published as an Advance Article on the web 9th February 2016 DOI: 10.1039/b917766c

The development of anti-infectious agents is widely recognized

as complex and incredibly difficult for a number of reasons.

tic h is cally remant chem to that are only against b) as urgery stated to o

dings are the born of the first set of a manufact of Pelsions, dings are the grow hit can strong high throughput are dury access, as ell as the set to obtain very domain the h is set of the field of the set of

The search for molecules with activity against Mycohacterium tuberculasis (Mtb) is employing many approaches in parallel including high throughput screening and computational meth We have developed a database (CDD TB) to capture public and private Mtb data while enabling data mining and collaborations with other researchers. We have used the public data along with several cheminformatics approaches to produce models that describe active and inactive compounds. We have compared there datasets to those for known FDA approved drugs and between Mtb active and inactive compounds. The distribution of polar surface area and p.K. of active compounds was found to be a statistically significant determinant of activity against Mtb. Hydrophobicity was not always statistically significant. Bayesian classification models for 220 463 molecules were generated and tested with external molecules, and enabled the discrimination of active or inactive substructures from other datasets in the CDD TB.

mine hubber some of the Mth data with a standing unique per able to map to e or shinaadonaf the database, data it andro. Bayesian and narrowophere win gibb into additionation constraint andro. Bayesian and narrowophere nacophores baæd on known Mtb drugs were able to map to and r

> the shortage of compounds in the research and dev pipeline for this disease, as for example 48 antimycol were classified as either candidates, bads, hits or tool con

ing par neter to drug roper ery nyo opho iser in the Lipinski i s oʻdrug-likenesi) v leof 5⁶ w to drug th many. ring do lead-like b multiple S and more ular we ds were classified a < 400 upor ized that past of the dif progressing molecules from hits to clinical candida either target-based or whole cell screening for Mtb attributed to these simple physicochemical prop particular the generally high hydrophobicity of or in many of the libraries texted. Although necessary for molecules into Mtb, hydrophobicity may actually also solubility and potential to proceed through further tes is analogous to the general observations by Lipins that plagued the pharmaceutical industry in the e of combinatorial chemical libraries and high th

Computational and statistical analyses can help understand the physicochemical properties which should be present in molecules tested against Mtb to better lead candidates. For example there have bee ligand-based⁹ or protein-based¹⁰ studies aimed at fil

(ici:10.1007/ci1095-012-0741-5) contains applementary material, which is available to authorized user. M Sarler - C. Takott - P. Madrid - S. Chope 90 International

> Minlo Park, California 94025, USA B.A. Bunin -S. Fiking Collaborative Drug Discovery 1633 Baythore Highway, Suite 342 Burlingume, California 94010, USA

333 Reveneycod Avenue

agents.2 Although these have been many phenotypic screens carried out in various organizations before, until recently there has not been any central depository of the screening results. However, during the last two years there have been changes as large quantities of phenotypic screening data were published and a common database for Mtb drug discovery has been developed.3 In this paper we greatly extend our initial analysis of the Molecular libraries Small Molecule Repository (MLSMR) dataset⁴ and analyze the latest data published by the National Institute of Allergy and Infectious Disease (NIAID) which were generated by the Southern Research Institute (SRI) on over 100 000 compounds purchased from commercial sources (Tuberculosis Antimicrobial Acquisition and Coordinating Facility, TAACF-NIAID-CB25). Our initial analysis of an earlier TAACF set, NIAID, GVKbio, MLSMR and Ballel datasets (Table 1) suggested that the mean value for various mole cular descriptors is statistically different to that of FDA approved drugs,3 The mean value of polar surface area descriptors was frequently higher in active compounds, compared to the inactive molecules across four datasets. The molecular weight, logP, and Lipinski score were statistically significantly higher in the most active compounds in the MLSMR screening data, while the PSA is slightly lower compared to the inactive compounds. This analysis

his journal is & The Royal Society of Chemis

Pharm Res (2012) 29:2115-2127 DOI 10.1007/r11095-012-0741-5

RESEARCH PAPER.

ABSTRACT

cule modulators.

Combining Cheminformatics Methods and Pathway Analysis to Identify Molecules with Whole-Cell Activity Against Mycobacterium Tuberculosis

Malabita Saker - Carolyn Talzott - Pater Madrid - Sidharth Chopra - Bany A. Bunin - Gyanu Lamichhane - Joal S. Freundlich - Sean Búns

Received: 31 August 2011 / Accepted: 16 March 2012 / Published online: 4 April 2012 C Springer Science+Business Media, ILC 2012

New strategies for developing inhibitors of Mycobac-

generate of tuberculosis (TB) drugs. Our approach leverages interaction of intensive data mining and curation and computational approaches, inducing cheminformatics combined with

bioinformatics, to suggest biological targets and their small mole-

Methods We now discribe an approach that uses the TBCyc

pathway and genome database, the Collaborative Drug Dis-

covery database of molecules with activity against MIb and their

associated targets, a 3D phermacophore approach and Bayes-

ian models of TB activity in order to select pathways and

metabolites and ultimately prioritize molecules that may be

Results In this study we combined the TB cheminformatics and

pathways databases that enabled us to computationally search

>80,000 vendor available molecules and ultimately test 23 com-

counds in vitro that resulted in two compounds (N-(2-furvimethyl)-

N-[(5-nitro-3-thiany);arbony[]thicuma and N-[(5-nitro-3-thiany])

arbony[-N'-(2-this nyimethy(thioures) proposed as mimits of D-

fuctorse 1,6 bisphosphate, (MC of 20 and 40 µg/ml, respectively).

Conclusion This is a simple yet novel approach that has the

potential to identify inhibitors of bacterial growth as illustrated by

compounds identified in this study that have activity against Mtb.

Bectronic supplementary material The online vention of this attide

ading as substrate mimics and exhibit activity against TB.

advecutors (Mb) are required in order to identify the next

KEY WORDS Bioinformatics - cheminformatics -

Colaborativedrug discovery - Mycobacterium tuberculoris ohermacrohore

INTRODUCTION

Mysohasterian tabecalasis (Mtb), the causative agent of tuberculosis (TB), is estimated to maintain latent infection in approximately one-third of the world's population and kill 1.7-1.8 million people each year (1). The survival of Mtb relies on an array of cellular functions carried out by metabolites, enzymes, structural and regulatory proteins and RNAs. These essential functions can be targeted to kill or suppress the proliferation of Mtb. Soon after the genome sequence of the Mth H37Ry strain was published (2), various laboratories focused on identifying genes essential for growth under invite and in nice conditions (3). Classification of essential genes as targets is based on forward genetic approaches that consider a protein as a potential target if an egential gene encodes it (4). A target should be egential for growth and viability of the pathogen at least under the condition of host infection. During infection, Mtb appears to reside predominantly within the host lung

G. Lanithave Department of Medicine, Johns Hopkins School of Medicine 1550 Orleans St. Room 103 Baltimore, Maryland 21287, USA

INTRODUCTION

r than for known TB drugs drugs, lengthy treatment, and

S. Birs.

S. Frandlich

Mycobacterium tuboraclosis (Mtb)

tuberculosis (IB), infects appr

world's population, and 1.7-1.8

this disease annually (1). Agents the

are urgently needed to combat thi

heavily influenced by resistance to

Department of Pharmacoutical Sciences

Department of Pharmacology, Robert

University of Medicine & Duntistry of Pacataway, New Jensey 08854, USA

Department of Biochemistry and Bi

College Station, Texas 77843, USA

Batimore, Maryland, USA

and anaerobic hits) and to

Novatis compounds were

SMARTS alerts to identify

pounds as a test set for the

>4.0-fold enrichment over

obic hits not in the compu-

Id enrichment was observed

s in the FDA drugs database.

d The online version of this article

tains supplementary material

ds failed the Abbott SMARTS

🕙 Syringer

increased probability of inhibition of whole cell Mtb activity. compounds are see organisms such as one of the proven methods to discover new reads for anti-MLD berouk; v (Mtb) the live agent of

osis and occur and 1.7-1.8 million people die from it.1 Almost all of the currently used drugs against Mtb were discovered more than 50 years ago and because of the emergence of resistance and the prolonged treatment duration of the current therapies, new urgently needed. Phenotypic excening where



Department of Pharmaceutical Sciences, University of Maryland, Baltimore, M.D. USA

School, University of Medicine & Dentistry of New Jersey, Pisantaway, New Jerney 08854, USA

10 Connshire Drive, Waterford, Connecticut, 06385, USA Division of Biocomputing, University of New Mexico, Albuquerque,

SureChem, The Macmillan Building, 4 Crinan Street, London

Electronic supplementary information (ESI) available: Supplementary figures and methods. The Bayesian models crusted in Discovery Studio are available from the authors upon written request. See DOI: 10.1039/ dm100104i

2316 Mol. Blo5vst., 2010, 6, 2316-2324

adverse processing with the second second processing processing of the second s PA 19046, USA expected failure of ~85% of clinical candidates (Lectord, 20 and the growth of TB drug resistance necessitate new clini Department of Pharmacology, Robert Wood Johnson Medical submissions, which ultimately require the discovery of no hts and leads. We assert that the TB field should further levers existing HTS data, focusing on not just the few most promisi

Introduction

Approx nately me t

Mycobacti

Global Alliance for TB Drug Development, 40 Wall Street, 24th floor, New York, NY 10005, USA

NM 87131

UK NI 9XW

t of Pharn M.D., USA cal Sciences, University of Maryland Departs Baltime acology, Robert Wood Johnson Mediasl nool. Universely of Medicine & Dentistry of New Jersey. Pirantosay, NJ 08854, USA † Electronic supplementary information (ESI) available Include supplemental tables for Bayesian models, supplemental figure showing important descriptors in Bayasian models, examples of molecules common between two datasets or models. in the CDD database, and pharmacophore arronning moults from datahase searching. See DOI: 10.1039/5917766

840 Mol. BioSvit. 2010. 6.840-851

active against Mtb in whole cells screened under the same in vitro conditions. Variou hit molecules were also examined by various filtering rules used widely in the pharma industry to identify compounds with potentially reactive moieties. We found differen the number of compounds flagged by these rules in Mtb datasets, malaria hits, FDA drugs and antibiotics. Combining these approaches may enable selection of compour

lasse number of commercially available molecules to This journal is a The Royal Society of Chem







ycobacterium tuberculosis

How do we describe Molecules?

- As fingerprints
- As properties
- as experimental measurement/s e.g. logP
- OD, 1D, 2D, 3D, 4D descriptors





Open Extended Connectivity

Fingerprints



Collected, deduplicated, hashed Sparse integers

- Invented for Pipeline Pilot: public method, proprietary details
- Often used with Bayesian models: many published papers
- Built a new implementation: open source, Java, CDK
 - stable: fingerprints don't change with each new toolkit release
 - well defined: easy to document precise steps
 - easy to port: already migrated to iOS (Objective-C) for *TB Mobile* app

Clark et al., J Cheminform 6:38 2014



- Skipped targets with > 100,000 assays and sets with < 100 measurements</p>
- Converted data to -log
- Dealt with duplicates
- 2152 datasets

http://molsync.com/bayesian2

- Cutoff determination
- Balance active/ inactive ratio
- Favor structural diversity and activity distribution

Clark and Ekins, J Chem Inf Model. 2015 Jun 22;55(6):1246-60

EMBL-EBI

What do 2000 Chembl models look like



PolyPharma mobile app (iOS)

- Uses Tox21 data
- Enables prediction
- Visual output from Bayesian models
- Atom contributions coloring
- ▶ free~!





Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

aromatase (Toxicity) HSE-bla (Toxicity) ARE-bla (Toxicity) P53-bla (Toxicity) DT40 WT (Toxicity) DT40 Rev3 (Toxicity) ATAD5 (Toxicity) VDR-bla antagonist (Toxicity) VDR-bla agonist (Toxicity) TR-beta-luc antagonist (Toxicity) PPAR-gamma-bla antagonist (Toxicity) PPAR-gamma-bla antagonist (Toxicity) PPAR-delta-bla antagonist (Toxicity)

mitochondria toxicity (Toxicity) AhR-luc (Toxicity) AR-bla agonist (Toxicity) AR-bla antagonist (Toxicity) AR-MDA-luc agonist (Toxicity) ER-BG1-luc agonist (Toxicity) ER-BG1-luc antagonist (Toxicity) ER-bla antagonist (Toxicity) ER-bla antagonist (Toxicity) FXR-bla antagonist (Toxicity) GR-bla agonist (Toxicity) GR-bla agonist (Toxicity) PPAR-delta-bla agonist (Toxicity)

⊕ **1** * ■





- A tool for building and sharing Bayesian models built with biological data from screens
- Assay Central can be used to generate predictions for new molecules (ADME/ Off targets etc)
- Provides model statistics and information on features contributing to activity

www.assaycentral.org

Support by NIH grant 1R43GM122196



X

Targets

MOOT (ICOU)	- 00
ACC1 (Ki)	- 683
ACC2 (BindingDB)	- 68
ACC2 (Combo)	- 88
ACC2 (IC50)	- 68
ACC2 (Ki)	- 68
Capthepsin K (C50)	8
Capthepsin K (Ki)	88
Capthepsin S (C50)	- 68
Capthepsin S (Ki)	- 88
Chagas Disease (Ekins): binary	8
Chagas Disease (Ekins): calculated	- 68
Chagas Disease: Broad (EC50)	8
Chagas Disease: Broad (EC50/Cytotox)	- 88
Chagas Disease: ChEMBL	8
Chagas Disease: Cruzain	8
Chagas Disease: Cruzipain	8
Chagas Disease: Public Collections	-88
Dengue	8
Dengue: ChEMBL	8
Dengue: PubChem	8
Ebola (Entry)	8
Ebola (Replication)	8
Ebola: HeLa Cells (50µM)	8
Ebola: Vero cells	-8
 HIV: Reverse Transcriptase (Bind/IC50) 	8
HIV: Reverse Transcriptase (Bind/Ki)	8
HIV: Reverse Transcriptase (Funct/Ki)	8
L. donovani: Amastigotes	8
L. donovani: Amastigotes, Axenic (DNDi)	8
L. donovani: Amastigotes, Intracellular (DNDi)	8
L. infantum: Amastigotes	- 68

C8%2C41%2... 🕁

All None

ER data

		<u>Title</u> ↑	<u>Target</u>	<u>Organism</u>	<u>Size</u>	ROC	<u>F1</u>	<u>Kappa</u>	MCC	<u>Domain</u>	<u>Invalid</u>
Data	Model	ER (Funct/Ki)	Estrogen Receptor	Human	327	0.8774	0.8118	0.6082	0.6089	0.1649	
Data	Model	ER: CERAPP Agonists	Estrogen Receptor	Human	1677	0.7691	0.4251	0.3042	0.3319	0.2830	
Data	Model	ER: CERAPP Antagonists	Estrogen Receptor	Human	1677	0.6299	0.0911	0.0487	0.0956	0.2830	
Data	Model	ER: CERAPP Binding	Estrogen Receptor	Human	1677	0.7745	0.4360	0.3078	0.3330	0.2830	
Data	Model	ER: METI	Estrogen Receptor	Human	254	0.9202	0.6739	0.6068	0.6337	0.1811	
Data	Model	ER: Tox21 Agonists	Estrogen Receptor alpha	Human	7351	0.8366	0.2973	0.2436	0.3083	0.3764	
Data	Model	ER: Tox21 Antagonists	Estrogen Receptor alpha	Human	7351	0.7628	0.1726	0.1143	0.1848	0.3764	
Data	Model	ER: alpha (All Binding)	Estrogen Receptor alpha	Human	2347	0.9428	0.8737	0.7426	0.7426	0.2932	
Data	Model	ER: alpha (Bind/IC50)	Estrogen Receptor alpha	Human	1127	0.9709	0.9395	0.8698	0.8703	0.2627	
Data	Model	ER: alpha (Bind/Ki)	Estrogen Receptor alpha	Human	2347	0.9426	0.8755	0.7512	0.7512	0.2932	
Data	Model	ER: alpha (Funct/Ki)	Estrogen Receptor alpha	Human	488	0.8843	0.8169	0.6273	0.6285	0.1934	
Data	Model	ER: beta (All Binding)	Estrogen Receptor beta	Human	1806	0.8840	0.8427	0.6645	0.6656	0.2615	
Data	Model	ER: beta (Bind/IC50)	Estrogen Receptor beta	Human	969	0.9621	0.9357	0.8635	0.8656	0.2574	
Data	Model	ER: beta (Bind/Ki)	Estrogen Receptor beta	Human	1806	0.8822	0.8478	0.6690	0.6690	0.2615	
Data	Model	ER: beta (Funct/Ki)	Estrogen Receptor beta	Human	337	0.8907	0.8055	0.6559	0.6559	0.1713	

© 2017 Collaborations Pharmaceuticals Inc. Non-Proprietary Slides.

26

An Example of ER model external testing

ER: METI

Origin:	Assay Central
Field	ER/METI_v2
Comments:	Domain Compatibility: 0.1811452
Training Actives:	35 / 254
ROC	0.9202 (five-fold)
Curve:	
Truth Tabla	Dradiated
	Ves No
	Actual No 26 193
Precision :	0.5439
Recall:	0.8857
Specificity:	0.8813
F1 score:	0.6739
Kappa:	0.6068
MCC:	0.6337

subvalidation





- Focused on providing a suite of ADME/Toxicity models
- Tox21 data, hepatotoxicity, cytotoxicity, mutagenicity, cardiotoxicity, drug-drug interactions, microsomal stability, Pregnane X receptor (PXR) and likelihood of causing drug-induced liver injury (DILI)

	<u>Title</u> ↑	<u>Target</u>	<u>Organism</u>	Actives	<u>Size</u>	ROC	<u>F1</u>	<u>Kappa</u>	MCC	<u>Domain</u>	<u>Invalid</u>
Data Model	CytochromeP450: 1A2	CytochromeP450 - 1A2	Human	310	810	0.9064	0.7945	0.6551	0.6580	0.2985	
Data Model	CytochromeP450: 2C19	CytochromeP450 - 2C19	Human	246	789	0.8835	0.7410	0.6203	0.6206	0.2972	
Data Model	CytochromeP450: 2C9	CytochromeP450 - 2C9	Human	458	1203	0.8594	0.7394	0.5589	0.5634	0.3056	
Data Model	CytochromeP450: 2D6	CytochromeP450 - 2D6	Human	732	1668	0.8965	0.8049	0.6379	0.6408	0.3024	
Data Model	CytochromeP450: 3A4	CytochromeP450 - 3A4	Human	1106	2135	0.8919	0.8113	0.6129	0.6131	0.3224	

Deep Learning uses

- facial recognition algorithms
 - Facebook tagging photos
- self-driving cars
- robot assistants
- Speech recognition
- Stock markets
- Fraud detection



http://tinyurl.com/hak4lcv





http://tinyurl.com/y8vjv8lp

Deep Learning in Pharmaceutical Research

- Bioinformatics
 - Protein disorder
 - Refine docking complexes
 - Model CLIP-seq data
 - High content image analysis data
 - Biomarkers
 - Protein contacts
 - Cancer diagnosis

Pharm Res (2016) 33:2594-2603 DOI 10.1007/s11095-016-2029-7

PERSPECTIVE

- Pharmaceutical
 - Solubility
 - Gene expression data
 - Formulation

datasets

QSAR – Merck DL out performed random forests in 11 /15 and 13/15

Output

layer

Hidden layers

Input

layer

► Tox21

The Next Era: Deep Learning in Pharmaceutical Research

Sean Ekins 12 🕞

molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

¹ Comparison of Deep Learning With Multiple Machine Learning ² Methods and Metrics Using Diverse Drug Discovery Data sets

3 Alexandru Korotcov,[†] Valery Tkachenko,^{*,†} Daniel P. Russo,^{‡,§} and Sean Ekins^{*,‡}

4 [†]Science Data Software, LLC, 14914 Bradwill Court, Rockville, Maryland 20850, United States

s [‡]Collaborations Pharmaceuticals, Inc., 840 Main Campus Drive, Lab 3510, Raleigh, North Carolina 27606, United States

6 [§]The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States

model	data sets used and references	cutoff for active	number of molecules and ratio
solubility	119	Log solubility = -5	1144 active, 155 inactive, ratio 7.38
probe-like	120	described in ref 120	253 active, 69 inactive, ratio 3.67
hERG	121	described in ref 121	373 active, 433 inactive, ratio 0.86
KCNQ1	PubChem BioAssay: AID 2642 ¹²²	using actives assigned in PubChem	301,737 active, 3878 inactive, ratio 77.81
bubonic plague (Yersina pestis)	PubChem single-point screen BioAssay: AID 898	active when inhibition \geq 50%	223 active, 139710 inactive, ratio 0.0016
Chagas disease (Typanosoma cruzi)	Pubchem BioAssay: AID 2044	with EC_{50} < 1 μ M, > 10-fold difference in cytotoxicity as active as described in 88	1692 active, 2363 inactive, ratio 0.72
TB (Mycobacterium tuberculosis)	in vitro bioactivity and cytotoxicity data from MLSMR, CB2, kinase, and ARRA data sets ²²	<i>Mtb</i> activity and acceptable Vero cell cytotoxicity selectivity index = (MIC or IC_{90})/ $CC_{50} \ge 10$	1434 active, 5789 inactive, ratio 0.25
malaria (Plasmodium fakiparum)	CDD Public data sets (MMV, St. Jude, Novartis, and TCAMS) ¹²³⁻¹²⁵	3D7 EC ₅₀ < 10 nM	175 active, 19604 inactive, ratio 0.0089

Datasets preparation:

- Datasets were split into training (80%) and test (20%) datasets (default settings)
- Spit datasets maintain equal proportions of active to inactive class ratios (stratified splitting)
- 4-fold cross validation (default settings) on training data for better model generalization
 Korotcov et al., Molecular Pharmaceutics 2017

AUC for all tested datasets (FCFP6, 1024 bits)

AUC values	BNB	LLR	ABDT	RF	SVM	DNN-2	DNN-3	DNN-4	DNN-5	Clark et al.
solubility train	0.959	0.991	0.996	0.934	0.983	1.000	1.000	1.000	1.000	0.866
solubility test	0.862	0.938	0.932	0.874	0.927	0.935	0.934	0.934	0.933	
probe-like train	0.989	0.932	1.000	0.984	0.995	1.000	1.000	1.000	1.000	0.757
probe-like test	0.636	0.662	0.658	0.571	0.665	0.559	0.563	0.565	0.563	
hERG train	0.930	0.916	0.992	0.922	0.960	1.000	1.000	1.000	1.000	0.849
hERG test	0.842	0.853	0.844	0.834	0.864	0.840	0.841	0.841	0.840	
KCNQ train	0.795	0.864	0.809	0.764	0.864	1.000	1.000	1.000	1.000	0.842
KCNQ test	0.786	0.826	0.801	0.732	0.832	0.861	0.856	0.852	0.848	
Bubonic plague train	0.956	0.946	0.985	0.895	0.992	1.000	1.000	1.000	1.000	0.810
Bubonic plague test	0.681	0.767	0.643	0.706	0.758	0.754	0.752	0.753	0.753	
Chagas disease train	0.812	0.847	0.865	0.815	0.926	1.000	1.000	1.000	1.000	0.800
Chagas disease test	0.731	0.763	0.768	0.732	0.789	0.790	0.791	0.790	0.789	
Tuberculosis train	0.721	0.737	0.760	0.735	0.800	1.000	1.000	1.000	1.000	0.727
Tuberculosis test	0.671	0.681	0.676	0.679	0.695	0.687	0.684	0.688	0.685	
Malaria train	0.994	0.993	0.999	0.979	0.998	1.000	1.000	1.000	1.000	0.977
Malaria test	0.984	0.982	0.966	0.953	0.975	0.975	0.975	0.974	0.974	

Clark et al. J Chem Inf Model 2015

Deep learning wins using rank normalized score by metric or by dataset

Korotcov et al., Molecular Pharmaceutics 2017

Summary

- Machine learning can be used over a broad array of:
 - projects, diseases, targets, whole cell assays, Tox endpoints etc..
- Bayesian algorithm demonstrates wide utility
- Plenty of scope to pursue other machine learning methods with toxicology data
- Spillover of machine learning to new areas

Where to learn more..









Collaborations Pharmaceuticals, Inc.

Kim Zorn Dr. Maggie Hupcey Dr. Tom Lane [soon to be Dr.] Dan Russo <u>Consultants</u> Dr. Alex Clark (Assay Central) Valery Tkachenko (Deep Learning) Dr. Alex Korotcov (Deep Learning)

> Joel Freundlich Alex Perryman Steve Wright And many colleagues

NIH NIAID R41AI108003-01, NIH NIGMS R43GM122196, NIH NCATS R21TR001718, NIH NINDS, 1R01NS102164-01,³⁵ NIH NCATS 1UH2TR002084-01, One NC Small Business program grant

collaborationspharma@gmail.com





