Towards Predictive ADME Profiling of Drug Candidates: Lipophilicity and Solubility

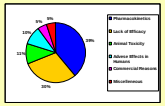
Gennadiy Poda[1], Igor Tetko[2,3] and Douglas C. Rohrer[1], [1]Pfizer Global Research & Development, St. Louis Laboratories, Pfizer Inc, 700 Chesterfield Parkway West, St. Louis, MO 63017, USA, [2]MIPS, Institute for Bioinformatics, Munich, Germany, and [3]Institute of Bioorganic and Petroleum Chemistry, Ukrainian Academy of Sciences, Kiev, Ukraine

Abstract

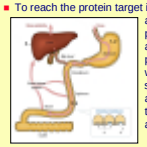
In order to reach the target enzyme or receptor in the human body, drugs have to pass numerous membrane barriers by passive diffusion or carrier-mediated uptake. To achieve that drugs have to be soluble both in water and in lipids. This requirement makes lipophilicity and solubility the two major properties responsible for absorption and bioavailability of drugs. The 1-octanol-water partition coefficient, logP, is well known as one of the major parameters to estimate lipophilicity (or solubility in lipids) of chemical compounds and, to a large degree, determines their ADME properties. The logP is also one of the standard properties identified by Lipinski in the "Rule of 5" for druglike molecules. Aqueous solubility is usually measured as its logarithm of intrinsic or pH-dependent solubility, logS. Reliable predictions of logD and logS can significantly facilitate selection of drug candidates from virtual libraries during the drug design process. Within the ALOGPS approach, a statistical ensemble of associative neural networks trained on the dataset of publicly available data globally maps input parameters to the target property. The final tuning of the model is done using a self-learning feature of the ALOGPS based on a user-defined set of the data and was shown to remarkably improve the accuracy in logD and solubility predictions for proprietary compounds. Thus, the ALOGPS combines the best properties of both global and local models.

Introduction

- Poor pharmacokinetics (PK) is the major reason of failing of drug candidates in clinical trials (Kennedy T. *Drug Disc. Today* **2**, 436-444 (1997)), thus, computational methods to assess PK properties are of great importance.



- To reach the protein target in the human body a drug has to get absorbed in the gut, travel through the portal vein and reach the blood circulation after the first-pass through the liver while passing numerous cell membranes on its way. To achieve that, the drug has to be soluble both in water and in lipids. Thus, aqueous solubility and lipophilicity are the two major factors responsible for absorption and bioavailability of drugs.



- Lipophilicity is the key physicochemical parameter that to a large degree determines PK properties of drugs. The 1-octanol-water partition coefficient, logP, is well known as one of the major parameters to estimate lipophilicity (or solubility in lipids). If a molecule contains ionizable groups, it becomes ionized and its distribution in octanol-water becomes pH-dependent and is determined by the logD distribution coefficient.
- Aqueous solubility is usually measured as its logarithm of intrinsic or pH-dependent solubility, logS.
- Publicly available databases of logP/D and logS do not cover a wide range of chemical space and largely limited to low molecular weight compounds. Thus, performance of *In Silico* models based on this data is usually poor. At the same time, tremendous amount of data generated by Big Pharma remains publicly unavailable. As a result, local models are being developed to predict PK properties for proprietary compounds.
- A self-learning feature of the ALOGPS (http://www.vcclab.org) combines the best properties of both global and local models. Within the ALOGPS approach, a statistical ensemble of associative neural networks trained on the dataset of publicly available data globally maps input parameters to the target property. The final tuning of the model is done in the so called LIBRARY mode based on a user-defined set of data. In this step the program uses nearest neighbors technique to determine local corrections according to the specific feature of the analyzed chemical series.

Understanding Aqueous Solubility

General Solubility Equation (Yalkowsky, 1980)

$$LogS = 0.5 - LogP - 0.01 (MP - 298)$$

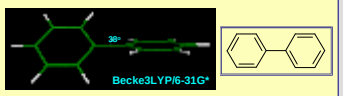
pH Dependence:

$S = S_0 (1 + 10^{(pKa - pH)})$ for bases

$S = S_0 (1 + 10^{(pH - pKa)})$ for acids

where S_0 is the intrinsic solubility (solubility of non-ionized compound)

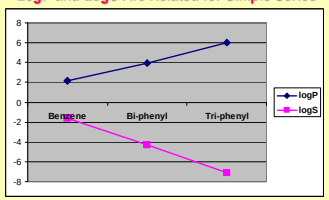
Bi-Phenyl Geometry ?



Torsion angle (degr):	
cff91, e=1	38
cff91, e=4r	33
MMFF, e=1	54
MMFF, e=4r	48
B3LYP/6-31G*	38
~45min	

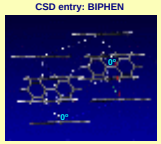
Rotation Barrier (kCal/mol):	0°	min	90°
cff91, e=1	0.8	0.0	1.5
cff91, e=4r	1.1	0.0	1.4
B3LYP/6-31G*	2.1	0.0	2.5
~6h45min			

LogP and LogS Are Related for Simple Series



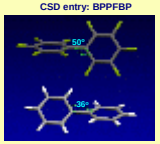
What Crystal Structure Says about Bi-Phenyl Geometry ?

CSD entry: BIPHEN



Crystal Packing: Planar

CSD entry: BPPFBP



Crystal Packing Disrupted: Non-Planar close to DFT geometry

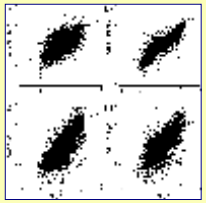
ALOGPS 2.1 Testing on LogD Data Set of Proprietary Compounds

- **Two datasets:**
  - **NlogD** dataset: 669 legacy Pharmacia compounds (nitrogen detector)
  - **ElogD** dataset: 18,889 Pfizer compounds (ElogD method, F. Lombardo et al., *J. Med. Chem.* **44**, 2490-2497, 2001)
- **The two sets are not overlapping**
- **Experimental error** in logD measurements is about 0.3-0.5 log units
- **Average values used** for compounds with multiple measurements and stereoisomers
- **Structural duplicates removed**, data averaged: NlogD dataset 640 compounds; ElogD dataset 17,861 compounds
- **Three prediction protocols:**
  - **"blind"** prediction for both datasets
  - 50% randomly chosen compounds used as a **LIBRARY** (local correction) and prediction made for the rest 50%
  - 75 E-state descriptors used to build an **ASNN model** based on randomly chosen 50% of NlogD dataset and prediction made for the rest 50%

Performance of ALOGPS, ACD LogD and Pallas PrologD

Calculated vs experimental logD values for ElogD dataset:

(A) ALOGPS "blind" prediction
(B) ALOGPS LIBRARY
(C) ACD LogD
(D) Pallas PrologD



Prediction Performance of Programs for ElogD Data Set

Methods Description	N[*]	RMSE	MAE	% cpds within RSME range 0–0.3	0–0.5	0–1.0	0–2.0
ACD Labs LogD, pH 7.4	17,341	1.32	0.97	21	35	63	89
ACD Labs LogP	17,848	1.38	1.08	19	30	55	85
Pallas PrologD, pH 7.4	17,800	1.41	1.06	19	31	58	87
Pallas PrologP	17,860	1.52	1.21	15	25	50	80
ALOGPS "as is"	17,861	1.17	0.92	21	35	62	91
ALOGPS LOO for all cpds used as LIBRARY	17,861	0.64	0.43	50	70	91	98
ALOGPS LOO for 50% cpds used as LIBRARY	8,931	0.69	0.48	45	65	88	98
ALOGPS prediction of 50% remaining cpds	8,930	0.69	0.48	46	66	88	98

[*] different numbers of compounds in this column are due to failure to process some chemical structures by ACD Labs LogD or Pallas PrologD suites

Prediction Performance of Programs for NlogD Data Set

Methods Description	N	RMSE	MAE	% cpds within RMSE range 0–0.3	0–0.5	0–1.0	0–2.0
ACD Labs LogD, pH 7.4	576	0.99	0.69	27	48	79	95
ACD Labs LogP	639	1.14	0.80	27	45	74	92
PALLAS PrologD, pH 7.4	640	1.52	1.29	8	15	41	84
PALLAS PrologP	640	1.46	1.20	10	19	46	86
ALOGPS "as is"	640	1.33	1.09	15	22	50	89
ALOGPS LOO for all cpds used as LIBRARY	640	0.65	0.42	54	70	90	98
ALOGPS LOO for 50% cpds used in "random" LIBRARY	320	0.66	0.44	52	68	88	98
ALOGPS prediction of 50% remaining cpds	320	0.68	0.45	52	73	89	98
ALOGPS blind prediction using ElogD set as LIBRARY	640	1.58	1.29	14	23	43	77
ASNN LOO for 50% cpds used to retrain ASNN	320	0.49	0.37	52	75	95	100
ASNN prediction of 50% test set	320	0.57	0.42	49	73	94	99

Testing of Equilibrium Solubility Models

- **Dataset** compiled from data generated at 5 Pfizer sites, Ann Arbor, Groton, Kalamazoo, La Jolla and St. Louis: 3,142 compounds total (1,969 neutral and 1,218 ionized @ pH 7)
- **Experimental error** in logS measurements is about 0.5 log units
- **Average values used** for compounds with multiple measurements, different salt codes, stereoisomers
- **Structural duplicates removed** by ALOGPS and data averaged
- **Three prediction protocols:**
  - **"blind"** prediction for the whole dataset
  - **LIBRARY** mode for the whole dataset and **neutral** and **ionized** compounds separately
  - **ASNN model** for the whole dataset based on 75 E-state and QikProp descriptors
- Comparison of the ALOGPS (http://vcclab.org/lab/alogps) and QikProp (http://www.schrodinger.com/Products/qikprop.html) solubility models

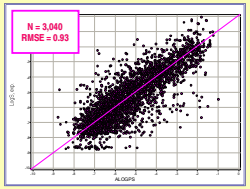
Prediction Performance for Equilibrium Solubility Data Set

Methods Description	N[*]	RMSE	MAE	% cpds within RMSE range 0–0.3	0–0.5	0–1.0	0–2.0
ALOGPS "as is" all	3,141	1.95	1.58	13	22	40	64
ALOGPS "as is" neutral	1,932	1.82	1.40	17	28	49	73
ALOGPS "as is" ionized	1,209	2.15	1.88	8	13	25	50
ALOGPS LIBRARY all	3,040	0.93	0.68	32	49	79	95
ALOGPS LIBRARY neutral	1,880	0.98	0.71	30	47	76	94
ALOGPS LIBRARY ionized	1,160	0.82	0.60	34	54	82	98
QikProp QPlogS	3,131	1.85	1.46	13	22	42	72
ASNN Model 34 QikProp desc.	3,130	0.78	0.59	35	53	83	98
ASNN Model 75 E-state desc.	3,039	0.73	0.54	38	57	86	98

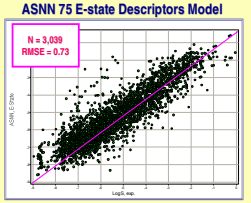
[*] different numbers of compounds in this column are due to failure to process some chemical structures by ALOGPS and QikProp model

Performance of Different Solubility Models

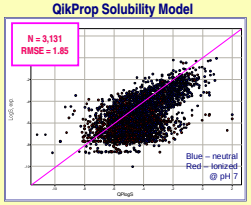
ALOGPS LIBRARY Mode Prediction



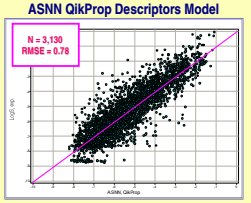
ASNN 75 E-state Descriptors Model



QikProp Solubility Model



ASNN QikProp Descriptors Model



- None of the "blind" predictions was sufficiently good
- "Blind" predictions are in particular poor for ionized at pH 7 compounds
- "Blind" predictions overestimate solubility for compounds with high melting points
- Cluster analysis of the AquaSol dataset indicates its significant limitations in chemical space
- Obtaining a reliable solubility model requires a diverse dataset that covers a large area of chemical space for drug-like molecules
- LIBRARY mode of ALOGPS significantly reduced the RMS error without re-training the model
- Self-learning feature of ALOGPS (LIBRARY mode) provides local correction in chemical space, thus, combines best features of global and local models
- Re-training ASNN resulted in ~20% reduction of RMSE compared to LIBRARY
- 75 E-state descriptors afforded a better ASNN model compared to 34 QikProp descriptors

Conclusions

- In the LogD prediction, ALOGPS 2.1 in the "blind" mode produced superior results compared to widely used ACD LogD and PALLAS PrologD on a large ElogD dataset of >17k compounds
- LIBRARY mode significantly improves prediction abilities of ALOGPS for new compounds for both logD and logS predictions by nearly 50%
- Re-training of the ASNN model to predict logD resulted in ~25% reduction of RMSE compared to the LIBRARY mode
- Prediction of aqueous solubility is a lot more challenging problem compared to LogD as seen from the RMS errors produced by different models
- "Blind" prediction overestimates solubility of compounds with high melting points
- Models based on the AquaSol dataset are not suitable to predict solubility for proprietary compounds
- Reliable prediction of solubility for proprietary compounds requires a model based on a large diverse set of drug-like molecules
- 75 E-state descriptor ASNN model produced superior results compared to the corresponded 34 QikProp descriptor model
- Self-learning feature of ALOGPS (LIBRARY mode) provides local correction in chemical space, thus, combines best features of global and local models
- Methods like ALOGPS that can improve prediction ability by self-learning on user-specific data will find significant applications in the pharmaceutical industry in the near future

Acknowledgement

GP is happy to acknowledge the invaluable work of Franco Lombardo and Marina Shalaeva (Pfizer Global R & D, Groton Laboratories, CT) who collected experimental data for the ElogD dataset. Contribution of the Pfizer Global Solubility Working Group is highly acknowledged, in particular, Howard Ando, Cornel Catana, Marcel de Groot, Hua Gao, Eric Gifford, Jason D. Hughes, Kjell Johnson, Jarek Kostrowicki, Pil H. Lee, Shaughn Robinson and Hongzhou Zhang. The ALOGPS development was partially supported by "Virtual Computational Chemistry Laboratory" INTAS Grant # 00-0363.

229th ACS National Meeting – San Diego, CA, March 13 – 17, 2005