Short Communication

For reprint orders, please contact: reprints@future-science.com

Future Medicinal Chemistry

Benford's law in medicinal chemistry: Implications for drug design

Alfonso T García-Sosa*,[1]

[1]Institute of Chemistry, University of Tartu, Ravila 14a, 50411 Tartu, Estonia

*Author for correspondence: alfonsog@ut.ee

Aim: The explosion of data based technology has accelerated pattern mining. However, it is clear that quality and bias of data impacts all machine learning and modeling. **Results & methodology:** A technique is presented for using the distribution of first significant digits of medicinal chemistry features: log*P*, log*S,* and pK_a, experimental and predicted, to assess their following of Benford's law as seen in many natural phenomena. **Conclusion:** Quality of data depends on the dataset sizes, diversity, and magnitudes. Profiling based on drugs may be too small or narrow; using larger sets of experimentally determined or predicted values recovers the distribution seen in other natural phenomena. This technique may be used to improve profiling, machine learning, large dataset assessment and other data based methods for better (automated) data generation and designing compounds.

Lay abstract: Machine learning and other technology depends critically on quality of data
Benford's law can indicate data follows natural phenomena easy, fast, statistical
Drug design impacted by FSD of experiment and predicted logP, pKa, solubility distributions
Method suited for large datasets

First draft submitted: 10 January 2019; Accepted for publication: 13 June 2019; Published online: 4 October 2019

Keywords: bias • chemical library • data science • distribution • drug design • drug discovery • drug likeness • filters • machine learning

Benford's (also called Newcomb-Benford, anomalous numbers, or first digit) law is a description of the distribution of first significant digits (FSD) in data and their counter-intuitive proportions [1]. A wide variety of natural and human phenomena have been proven to observe Benford's law as seen by the distribution of FSDs following a frequency of 1 (30.1%) >> 2 (17.6%) > 3 (12.5%) > 4 (9.7%) > 5 (7.9%) > 6 (6.7%) > 7 (5.8%) > 8 (5.1%) > 9 (4.6), and in very specific ratios [1]. The probability P of a number having the FSD d is thus given by [1]:

$$P(d) = \log_{10}(1 + 1/d) \quad \text{(Eq. 1)}$$

Newcomb observed that in tables of physical constants, numbers are more likely to begin with a smaller rather than a larger digit [1]. A possible reasoning is that for a number with an FSD of 1 to increase to an FSD of 2, it requires a 100% increase, whereas from FSD 2 to FSD 3 only a 50% increase is needed, 3–4 a 30% increase, and so on.

Such phenomena seen to follow Benford's law include those in geology and seismology [1], biological pathway kinetic rate constants [2], inhibition constants (IC_{50}) and PhysProp solubility values [3,4], molecular dynamics of fluids and Lorenz chaotic systems [5], among others. The larger the natural variation and amplifications in the data, the better it will comply to Benford's law since several orders of magnitude will better show the distribution of FSDs. Also, the prevalence of small objects versus large ones has been proposed as a cause for Benfords' FSD distribution [6]. In fact, even human activity follows Benford's law and this has been used to detect fraudulent data by comparing the distribution of frequencies and their deviation from expected Benford's law values [7]. There has been a very intense effort in chemistry and pharmacy to profile compounds according to their drug-likeness [8–10], oral biovailability [11], blood–brain barrier penetration [12], as well as antitarget [13–15], ADME [16],

newlands press

and pan-assay interference properties [17,18]. Famous and widely-followed rules include Lipinksi's 'rule of five' for oral bioavailability using molecular mass (MW), octanol/water partition coefficient (log*P*), number of donors, and number of acceptors [11]. Also popular are the 'rule of three' for lead compounds [19], and ligand efficiency metrics [20–22]. However, underlying all of these attempts are the actual compounds used to profile compound sets as well as their diversity. The importance of these profiles is justified by an early identification of compounds that may produce problems further along in optimization [14]. But for such profiling to work, the compounds used for training need to be well-sampled and relevant, as well as covering as much chemical space and relevant structural features as possible.

The present work shows how in medicinal chemistry, important features for describing and characterizing compounds as well as their profiling into popular tools, such as log*P*, acid dissociation equilibrium constant (pK_a), and solubility (*S* and log*S*), are distributed by their FSDs and how these features correspond to other observed natural phenomena distribution. This may help to improve compound (such as drug-like) profiling, as well as to assess the suitability of compound datasets and their (automatic) composition such as for machine learning [23,24], and as an added benefit, to indicate how well computational tools can predict the expected distribution of molecular phenomena.

Methods

Data collection

Experimentallydetermined octanol/water partition coefficient (log*P*) values for approved-drug compounds were downloaded from PubChem [25], and calculated values for drug compounds (ALOGPS and JChem) were downloaded from the DrugBank [26]. The same sources were used for pK_a values (minus logarithm of the acid dissociation equilibrium constants) for drugs (predictions available only by JChem), as well as for log*S* values (logarithm of solubility, only predictions by ALOGPS were available) and also the separately-reported experimental 'solubility' (in mg/ml). Experimentally-determined log*P* values for all of the compounds in the National Cancer Institute (NCI) compound database were also downloaded [27].

Statistics

Logarithmic values (pK_a, log*P*, log*S*) were converted to base 10 numbers and absolute values taken. The separately-reported solubility was left unchanged. The FSD was extracted through bash shell scripting.

χ^2 (Excel 2010) and Wilcoxon rank sum (R v. 3.4.4) statistics and tests were both performed as two-sided unpaired samples.

Results & discussion

The number of values for approved drugs obtained from the PubChem was n = 192 for log*S*, n = 452 for pK_a, and n = 1000 for log*P*. N = 685 for Solubility.

pK_a

Acid dissociation constants are important features of compounds given their relevance to binding, solubility, metabolism, and transport at the different pH conditions in the microenvironment where the drug may be present or act. It also dictates the ratio of ionization states and relative abundance of their (active) forms in the several physiological conditions (gut, blood, intracellular, etc.).

The distribution of FSDs for drug compounds shows that the predicted values follow Benford's law more closely than the experimental compounds (Figure 1), though the latter also follow a monotonically decreasing trend. The reason for this could be the larger size of the training set for the predicted values (JChem) than the experimental set. The statistics for the distributions (Figure 2) also are closer between Benford's law and the predicted JChem value distributions, than between Benford's law and the experimental values.

The statistical χ^2 and Wilcoxon ranks sum tests can be applied to compare the distributions and to discard or not the null hypothesis that the deviation of the distributions are due to random variation [6]. These statistics were not statistically-significant at the 95% level (i.e., $p > 0.05$) for any of the pK_a distributions (Table 1), indicating that all of the FSD distributions: experimental and predicted, do not statistically deviate from Benford's law.

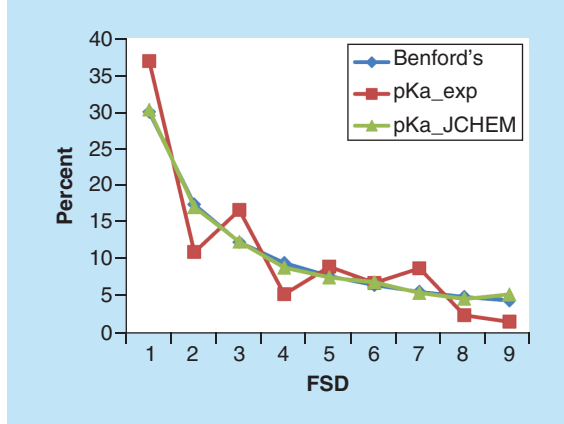

Figure 1. Distribution (percent, %) of first significant digits of pK_a values (transformed to 10^{-pKa}).
Benford's: According to Benford's law; pK_a_exp: Experimentally determined values; pK_a_JCHEM: Predicted by JChem.

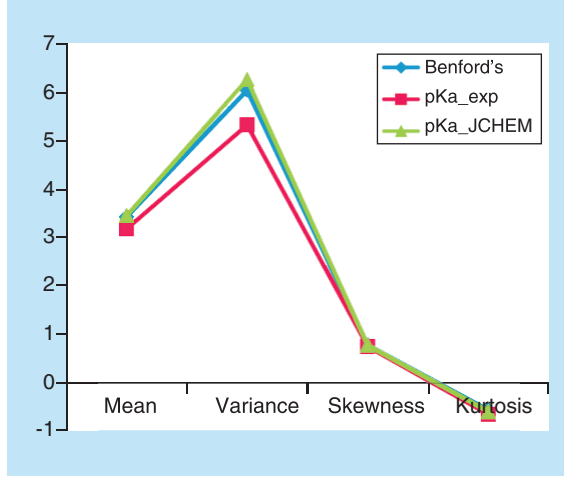

Figure 2. Statistic values for the distributions of pK_a.
Benford's: According to Benford's law; pK_a_exp: Experimentally determined values (transformed to $10^{-pKa}{}_{exp}$); pK_a_JCHEM: Predicted by JChem (transformed to $10^{-pKa}{}_{JCHEM}$).

Table 1. p-values for χ^2 and Wilcoxon rank sum (*W*) test statistics for first significant digit distributions

FSD distribution†	n	χ^2 p-value	W p-value
Experimental pK_a	452	0.1420	0.8633
JChem pK_a	9080	0.9999	1.0000
Experimental log*S*	192	0.9796	1.0000
ALOGPS log*S*	9290	0.9999	1.0000
Experimental solubility	685	0.5571	0.3865
ALOGPS solubility	9080	0.8498	0.4894
Experimental drugs' log*P*	1000	0.1108	0.8251
ALOGPS log*P*	9080	0.9999	1.0000
JChem log*P*	9290	0.9999	1.0000
Experimental NCI log*P*	3576	0.9939	0.9314

† Distributions of FSD were taken from the transformed values, p*K*a to 10^{-pKa}, log*S* to 10^{-logS}, log*P* to 10^{-logP}.

log*S*

Solubility (reported as log*S*) is also a critical feature of drug compounds, which can impact their efficacy as importantly as pharmacodynamic properties. It has a direct effect on the quantity of substance required for producing an effect, as well as on the ease of delivering the substance to its place of action.

Here, only ALOGPS had predicted values, and both these and the experimentally reported values had FSD distributions that followed Benford's law (Figure 3 & Supplementary Figure 1) with statistical test p-values that did not reject the null hypothesis of random variation with respect to Benford's law (Table 1).

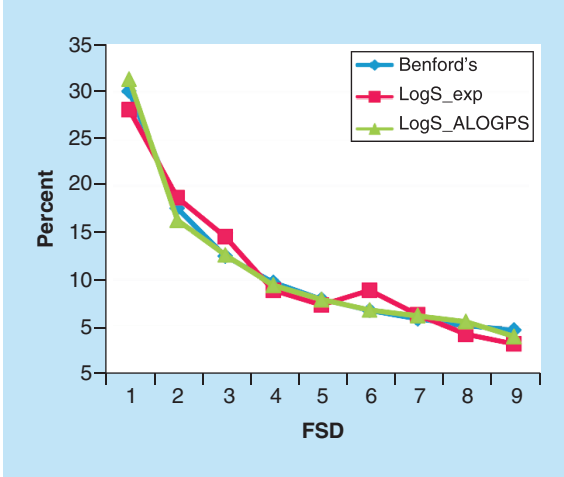

Figure 3. Distribution (percent, %) of first significant digits of log*S* values (transformed to 10^{-logS}).
Benford's: According to Benford's law; LogS_ALOGPS: Predicted by ALOGPS (transformed to $10^{-LogS_{ALOGPS}}$); LogS_exp: Experimentally determined values (transformed to $10^{-LogS_{exp}}$); .

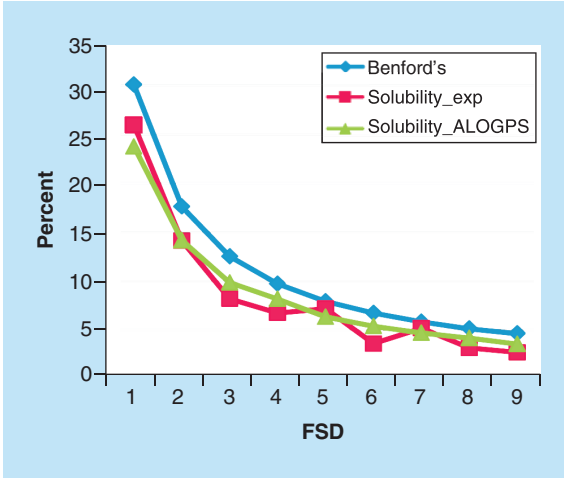

Figure 4. Distribution (percent, %) of first significant digits of solubility values.
Benford's: According to Benford's law; Solubility_ALOGPS: Predicted by ALOGPS; Solubility_exp: Experimentally determined values.

Solubility

The separately-reported solubility (in mg/ml) presented a similar case as for log*S*, with FSD distributions following Benford's law (Figure 4 & Supplementary Figure 2) with statistically undistinguishable (Table 1) ALOGPS and experimental values with respect to Benfords' law, respectively.

Octanol/water partition coefficient (log*P*)

Log*P* is one of the most widely-used properties to profile compounds since log*P* values can correlate to multiple chemical, pharmacological, and toxicological activities [28]. It is also used as a model for membrane permeability since it describes the ratio of distribution in moderately hydrophobic and aqueous environments.

The distribution of FSD for log*P* values are shown in Figure 5, where experimental values for log*P* followed Benford's law, but not as well as the predicted JChem or ALOGPS values. This is also shown in Supplementary Figure 3, where the values for mean, variance, skewness, and kurtosis were closer to Benford's law for JChem and ALOGPS. In any case, these distributions were not statistically different from Benford's law (Table 1).

This effect is most likely due to the number of values sampled, since both the predicted JChem (n = 9290) and ALOGPS (n = 9080) log*P* sets have a larger number of datapoints and magnitudes. Such data motivated the inclusion of a larger experimentally-determined set of log*P* values, which was found in the NCI database [27]. Including the NCI experimental log*P* values (see Figure 5 & Supplementary Figure 3), which is a larger dataset (n = 3576) than the number of drug compounds, shows a much better similarity to Benford's law distribution.

The results in the present work highlight the fact that making assumptions or conclusions on a small-sized sample, such as the number of approved drugs, can give imperfect distributions. The drug compounds do not have the variation of a larger sample of compounds, but this is most likely due to the limited number of drug compounds

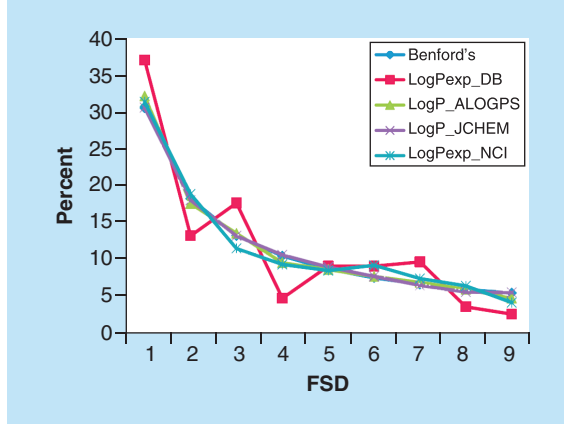

Figure 5. Distribution (percent, %) of first significant digits of log*P* values (transformed to $10^{-\log P}$).
Benford's: According to Benford's law; LogP_ALOGPS: Predicted by ALOGPS; LogPexp_DB: Experimentally determined values for drugs in DrugBank; LogP_JCHEM: Predicted by JChem; LogPexp_NCI: Experimentally determined values for drugs in the National Cancer Institute.

available, rather to any drug-likeness variance. This also prompts the need for better compound profiling, including better drug compound profiling. The ever-growing number of approved drugs that extend beyond well-known parameters (such as the number of non-Lipinski complying 2018 FDA-approved drugs), as well as more diverse datasets, which can be produced automatically, can help in redressing these unbalances.

Reassuring is the fact that predicted values, if generated from large samples of training sets with values of multiple magnitudes, such as those of JChem and ALOGPS, can generate reasonable predictions that follow distributions seen in other natural phenomena. All data used in this work is available at https://hermes.chem.ut.ee/~alfx/index.html in the download section.

Conclusion

The distribution of FSDs of important compound features in medicinal chemistry such as log*P*, pK_a, and solubility (as log*S*, and in mg/ml) show that they all agree with Benford's law, being statistically undistinguishable. Predicted properties complied just as well or even better than experimental values for drug compounds, given that they were trained on large training sets. Including larger datasets of experimentally determined values such as log*P* from the NCI database recovered a better fit to Benford's law distributions than smaller-sized sets. Profiling of compounds, especially of drug compounds, should be revised to include larger datasets with more diversity in magnitudes that can better cover the observed ranges of natural phenomena, as these have statistically-closer conformance to Benford's law of observed distributions.

The results of this work can be used to check if predicted values are reasonable in their imitation of natural phenomena distributions. In addition, they can also help in measuring how well a sample set, such as a set of drug or candidate compounds, agrees with natural phenomena distribution, which can be useful for machine learning, data mining, and drug design. Using sample sets that are too small may introduce artefacts or biases in distributions, which are especially relevant to be tracked if they are widely used as drug-likeness, oral bioavailability filters, or for machine learning, as is the case in practice.

Future perspective

Machine learning and artificial intelligence will continue to grow at a fast pace in medicinal chemistry and in other fields. The concern over the appropriate data to use and to (automatically) generate for machine learning and modeling will be ever stronger. Methods such as the one presented here can assess the distribution and adequateness (lack of bias) of data for use by further tools and may help improve technology based on data. One example is in the continuous testing of the FSD of generated data to observe at what point the distribution of FSD is statistically indistinguishable to natural phenomena. This represents a fast and simple way to measure quality of data generation and compilation, especially useful with very large datasets.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.4155/fmc-2019-0006

Financial & competing interests disclosure

A.T.G.-S. thanks Haridus- ja Teadusministeerium for grant no. IUT34-14. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Summary points

- Machine learning and other technologies are growing in use and relevance in medicinal chemistry and other fields
- These methods depend critically on the quality of the data underlying them
- Quality of data depends on the dataset size, diversity, and magnitudes
- Benford's law of first significant digit distribution and their ratios, as seen in many natural phenomena, can improve data assessment
- Experimental and predicted log*P*, pK_a, and solubility distributions better approximate Benford's law, statistically and qualitatively, the better the quality of their data
- Profiling for drug compound properties must extend present limits

References

Papers of special note have been highlighted as: • of interest

1. Sambridge M, Tkalčić H, Jackson A. Benford's law in the natural sciences. *Geophys. Res. Lett.* 37, L22301 (2010).
2. Grandison S, Morris RJ. Biological pathway kinetic rate constants are scale-invariant. *Bioinformatics* 24(6), 741–743 (2008).
3. Orita M, Moritomo A, Niimi T, Ohno K. Use of Benford's law in drug discovery data. *Drug Discov. Today* 15(9–10), 328–331 (2010).
4. Orita M, Hagiwara Y, Moritomo A, Tsunoyama K, Watanabe T, Ohno K. Agreement of drug discovery data with Benford's law. *Expert Opin. Drug Discov.* 8(1), 1–5 (2013).
5. Tolle CR, Budzien JL, LaViolette RA. Do dynamical systems follow Benford's law? *Chaos* 10(2), 331–336 (2000).
6. Formann AK. The Newcomb-Benford law in its relation to some common distributions. *PLoS ONE* 5(5), e10541 (2010).

• **Suitability of Chi-squared test for comparing distributions of first significant digits with respect to Benford's law**

7. Diekmann A. Not the first digit! Using Benford's law to detect fraudulent scientific data. *J. Appl. Stat.* 34(3), 321–329 (2007).
8. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* 1(1), 55–68 (1999).
9. García-Sosa AT, Oja M, Hetényi C, Maran U. DrugLogit: Logistic discrimination between drugs and non-drugs including disease-specificity by assigning probabilities based on molecular properties. *J. Chem. Inf. Model.* 52(8), 2165–2180 (2012).
10. García-Sosa AT, Maran U, Hetényi C. Molecular property filters describing pharmacokinetics and drug binding. *Curr. Med. Chem.* 19, 1646–1662 (2012).
11. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46(1–3), 3–26 (2001).
12. OpenTox. A project funded by the 7th Framework Programme of the European Commission. Creates computational methods to predict toxicity. http://www.toxcreate.net/predict
13. García-Sosa AT, Maran U. Drugs, non-drugs, and disease category specificity: organ effects by ligand pharmacology. *SQER* 24(4), 319–331 (2013).
14. García-Sosa AT, Maran U. Improving the use of ranking in virtual screening against HIV-1 integrase with triangular numbers and including ligand profiling with anti-targets. *J. Chem. Inf. Model.* 54(11), 3172–3185 (2014).
15. García-Sosa AT. Designing ligands for *Leishmania*, *Plasmodium*, and *Aspergillus* N-myristoyl transferase with specificity and anti-target-safe virtual libraries. *Curr. Comput. Aided Drug Des.* 14(2), 131–141 (2018).
16. vls3d.com: Directory of tools & databases collected over 10 years. Bruno Villoutreix. http://www.vls3d.com/index.php/links/chemoinformatics/admet/admet-and-physchem-predictions-and-related-tools
17. Dahlin JL, Nissink JW, Strasser JM *et al.* PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. *J. Med. Chem.* 58(5), 2091–2113 (2015).
18. Baell JB, Nissink JWM. Seven year itch: Pan-Assay Interference Compounds (PAINS) in 2017-utility and limitations. *ACS Chem. Biol.* 13(1), 36–44 (2017).
19. Congreve M, Carr R, Murray C, Jhoti H. A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* 8(19), 876–877 (2003).
20. Kuntz ID, Chen K, Sharp KA, Kollman PA. The maximal affinity of ligands. *Proc. Natl Acad. Sci. USA* 96(18), 9997–10002 (1999).

- **One of the first concepts relating ligand affinity to size.**

21. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead system. *Drug Discov. Today* 9(10), 430–431 (2004).
22. García-Sosa AT, Sild S, Maran U. Docking and virtual screening using distributed grid technology. *QSAR Comb. Sci.* 28(8), 815–821 (2009).
23. Gómez-Bombarelli R, Wei JN, Duvenaud D *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* 4(2), 268–276 (2018).

- **A good account of machine learning and automated chemical data generation.**

24. Yosipof A, Guedes RC, García-Sosa AT. Data mining and machine learning models for predicting drug likeness and their disease or organ category. *Front. Chem.* 6, 162 (2018).
25. PubChem. National Institutes of Health (NIH). https://pubchem.ncbi.nlm.nih.gov/
26. The DrugBank Database. Canadian Institute of Health Research. https://www.drugbank.ca/
27. NCI Database Download Page. Downloadable Structure Files of NCI Open Database Compounds. Release 4. National Cancer Institute. https://cactus.nci.nih.gov/download/nci/
28. QSARDB Repository. Search Term = logP. http://qsardb.org/repository/discover