ELSEVIER

feature

# Use of Benford's law in drug discovery data

**Masaya Orita,** masaya.orita@jp.astellas.com, **Ayako Moritomo, Tatsuya Niimi and Kazuki Ohno**

**Benford's law states that the distribution of the first digit of many data sets is not uniform. The first digit of any random number will be 1 almost 30% of the time, and larger digits occur as the first digit with lower and lower frequency, to the point where 9 occurs as a first digit only 5% of the time. Here, we demonstrate that several data sets in the field of drug discovery follow Benford's distribution, whereas 'doctored' data do not. Our findings indicate the applicability of Benford's law in assessing data quality in the field of drug discovery. We also propose a useful index of evaluating data quality based on Benford's law.**



In 1881, the American astronomer Simon Newcomb discovered the statistical principle now known as Benford's law [1]. Newcomb calculated that the probability that the first digit of any value is $d1$, given by

$$P(d1) = \log[1 + (1/d1)] \quad (1)$$

According to this formula, 1 occurs as the leading digit with a probability of 30%, a value that is considerably greater than the expected value of approximately 11.1% (1 in 9). Although Newcomb's efforts and findings were ignored at the time, physicist Frank Benford published a paper in 1938 regarding this same phenomenon [2], with 20 examples using 20,229 different sets of data including areas of rivers, baseball statistics and numbers in magazine articles. A reasonable mathematical explanation for this phenomenon remained elusive until Hill constructed a rigorous mathematical proof in 1996 [3].

Interestingly enough, manipulated, unrelated or artificially created numbers usually do not follow Benford's law, owing – partly – to the layperson's misunderstanding of randomness and real data distributions [4]. Building on this idea, Varian suggested that Benford's law could be used, therefore, to assess the honesty of purportedly random scientific data [5]. In the late 1980s, several statisticians using Benford's law discovered data manipulation in the accounting departments of several companies [6,7]. Since this incident, the phenomenon has been used more and more frequently to detect fraud across a number of fields [8–11]. In the field of natural science, Benford's law has also been used as a quality control tool. For example, Brown et al. applied Benford's law or Zipf's law (which is often considered to be a generalized form of Benford's law) in the screening of analytical data on pollutant concentrations in ambient air [12,13].

## Why should we care about data quality?

A superfluity or excessive amount of data can lead to a reduction in data quality owing to data handling errors and computer bugs. For example, analysis of results from high-throughput screening requires examination of tens of millions of data points, a task to which computational chemists often apply novel internal analysis tools. However, occasional program bugs in these internal tools can seriously impair the accuracy of the analysis.

Data fabrication is another increasingly serious issue in the scientific community. In the field of drug discovery and development, several papers have been published proposing methods of identifying possible instances of data fabrication [14,15]. Al-Marzouki *et al.* [16] proposed the use of statistical methods for detecting data fabrication by comparing the baseline data from one study (a diet trial) with those from another. In the diet trial, they observed a combination of features in the baseline data (variances, means and digit preference) that were strongly suggestive of data fabrication. In this manner, the data quality assessment framework is extremely important to maintaining data quality.



1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2010.03.003

**TABLE 1**

**Data set in the field of drug discovery and agreement with Benford's law.**

| Data set | Probability of first digit | | | | | | | | | $\chi^2$ value | Number |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | |
| Benford's law | 30.10% | 17.60% | 12.50% | 9.70% | 7.90% | 6.70% | 5.80% | 5.10% | 4.60% | | |
| McMaster HTS [16] DHPR inhibitors % inhibition | 26.20% | 14.90% | 12.70% | 11.00% | 9.50% | 8.10% | 7.00% | 5.60% | 5.10% | 0.02 | 49990 |
| Fontaine [17] Factor Xa inhibitors IC50 | 26.50% | 18.40% | 12.10% | 8.30% | 10.40% | 6.80% | 7.00% | 5.60% | 4.90% | 0.018 | 412 |
| Sutherland [18] DHPR inhibitors IC50 | 30.20% | 17.00% | 11.20% | 11.50% | 10.90% | 5.40% | 3.30% | 4.80% | 6.00% | 0.034 | 672 |
| Sutherland [18] COX2 inhibitors IC50 | 28.30% | 16.90% | 13.50% | 9.20% | 11.40% | 6.80% | 6.00% | 4.10% | 3.90% | 0.021 | 414 |
| Sutherland [18] benzo-diazepone receptor inhibitors IC50 | 27.90% | 17.10% | 15.00% | 10.50% | 8.40% | 6.60% | 6.30% | 3.30% | 4.80% | 0.015 | 333 |
| PysProp solubility | 32.70% | 16.50% | 11.80% | 9.50% | 8.60% | 6.50% | 6.20% | 4.70% | 3.50% | 0.0071 | 5574 |

## Is Benford's law valid in the field of drug discovery research?

Hoyle et al. introduced the chi-squared ($\chi^2$) statistic as an indicator to show whether data follow Benford's law. Using this statistic, these authors clearly demonstrated that the distribution of mRNA transcription data from a large number of organisms, measured using a range of experimental platforms, closely followed Benford's law [17].

The $\chi^2$ statistic is a nonparametric statistical technique used to determine whether an observed frequency distribution differs from the theoretically expected one. The $\chi^2$ statistics uses nominal or ordinal level data, employing frequency analysis rather than means and variances. The value of the statistic is given by

$$\chi^2 = \Sigma[(O-E)^2/E] \quad (2)$$

where $\chi^2$ is the chi-squared statistic, $O$ is the observed frequency and $E$ is the expected frequency. Hoyle et al. used the $\chi^2$ value as an indicator of whether the observed distribution follows Benford's law; namely, when $\chi^2 = 0$, the observed distribution obeys Benford's law completely.

To investigate the applicability of Benford's law in the field of drug discovery, several data sets were collected through McMaster University's high-throughput screening laboratory homepage (http://hts.mcmaster.ca/) and Chemoinfomatics.org (http://www.cheminformatics.org/) [18–20]. First digits probabilities and the $\chi^2$ value for these data sets are summarized in Table 1. Given the small value for the observed $\chi^2$ statistics, these data sets can be said to follow Benford's law closely. Particularly noteworthy is the fact that the percent inhibition of a large data set (approximately 50,000 entries), solubility of a large data set (approximately 44,000 entries) and the half maximal inhibitory concentration of a small data set (approximately 300–700 entries) obey Benford's law (see other data sets in Supplementary Data). Thus, in this sense, Benford's law is robust. If a data set does not follow Benford's law, the possibility exists that the data were fabricated or manipulated or that some error occurred during data processing. We suggest, therefore, that the $\chi^2$ statistic, which can be easily calculated, be employed in data quality assessment.

## Can Benford's law improve the quality check of *in silico* prediction?

### *Test set selection for a quantitative structure–activity relationship/quantitative structure–property relationship (QSPR/QSAR) study, based on Benford's law*

An exponential increase has been seen in recent years with regard to development of computational tools and methods for predicting biological activity and pharmacokinetic properties of compounds during new drug discovery endeavours, such as quantitative structure–property relationship/quantitative structure–activity relationship (QSPR/QSAR) studies. These *in silico* techniques not only shorten the research-to-market cycle but also drastically reduce efforts wasted during pharmaceutical research and development with regard to optimizing lead compounds, thereby reducing the overall cost of drug development. Both commercial and proprietary systems can be successfully applied to the pharmaceutical industry [21].

Modern *in silico* prediction techniques are characterized by their use of multiple descriptors of chemical structure combined with the application of both linear (multiple linear regression [22] with variable selection, partial least squares [23], among others) and nonlinear (*k*-nearest neighbours [24,25], artificial neural networks [26], among others) optimization approaches, with a strong emphasis on rigorous model validation to ensure that the models have acceptable predictive power. Selection of training sets from an available data set is a key step in obtaining predictive models, and if an initial training set is not appropriate for model building, any statistical methods of data analysis and correlation will lead to meaningless results, although such selection is generally determined by the researchers' experiences and interests. However, we argue that too intense selection of training sets to obtain apparent predictive accuracy of the model can disrupt the applicability of Bedford's rule to the training set data.

Figure 1a shows a plot of experimental aqueous solubility (log *S*; data of compounds whose experimental values were measured at between 15 °C and 37 °C) versus calculated topological polar surface area (TPSA) [27], drafted using commercial software (Molecular Operating Environment version 2008. 10.; Chemical Computing Group Inc., Quebec, Canada). Data from

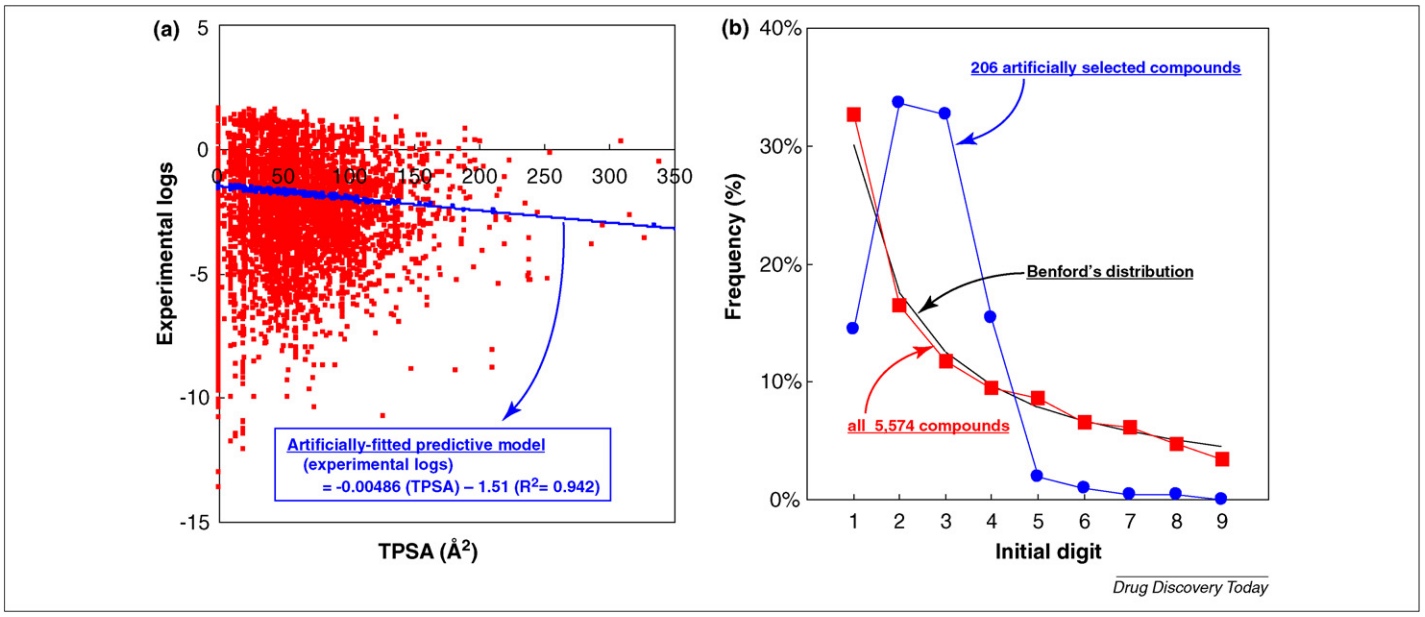


**FIGURE 1**

Distribution of the first digit for 'doctored' data (a) Plot of experimental log *S* vs. calculated topological polar surface area (TPSA). Red squares are 5574 molecules derived from the PhysProp database (commercial data set of experimental data for physical properties) and blue circles are 206 compounds artificially selected for our case study. Using the data from the 206 artificially selected compounds, we fit a predictive model (blue line, $R^2 = 0.942$; fitted equation: [experimental log *S*] = 0.00486*[TPSA] − 1.51). (**b**) Plot of first digit frequencies of all 5574 molecules in the PhysProp database (red squares) and the 206 artificially selected compounds (blue circles), as well as Benford's distribution (black line).

5574 molecules from the PhysProp database (http://www.syrres.com/), a commercial data set of experimental data for physical properties, were used to develop this plot. The aqueous solubility is not only a fundamental molecular property but also important with regard to pharmacology, toxicology and medicinal chemistry. In general, aqueous solubility improves with the reduction in molecular size or the size of the lipophilic part of the compound, or with a reduction in the size of the polar part for which TPSA is used as one index. However, many factors must be considered for the accurate prediction of water solubility—a task that is by no means easy. As shown in Figure 1a, almost no correlations can be observed between these two parameters ($R^2 = 0.003$). Here, we assume that the blue points in Figure 1a (208 compounds) are artificially selected as a training set, and an intentional predictive simple linear model is built using the data. Because these data were specifically chosen to be good data, we are able to achieve a fit, and a good coefficient of correlation between the predicted and experimental log *S* for the training set is obtained ($R^2 = 0.942$; fitted equation: [experimental log *S*] = −0.00486*[TPSA] − 1.51). However, because aqueous solubility is known to be unpredictable based on a single simple descriptor, and because log *S* and TPSA are not expected to be negatively correlated, this fitted expression is not acceptable.

With regard to our present case study, we encountered a serious problem in the representative points of the training set not being distributed with the whole area occupied by the entire data set. However, a third party unfamiliar with the entire data set will probably find it difficult to identify artificial aspects of training set selection. Figure 1b shows the plot of first digit frequencies of all 5574 molecules in the PhysProp database and the 206 compounds plotted as blue points in Figure 1a and artificially selected as a training set in our case study. Also plotted here is Benford's distribution. On closer examination of this plot, we can see that all 5574 molecules in the Phys Prop database are distributed according to Benford's law ($\chi^2 = 0.007$), whereas the 206 artificially selected compounds are not ($\chi^2 = 0.819$). These findings therefore suggest that Benford's law can be used to test for random selection in a training set. Furthermore, the law also holds true in the chemoinformatic field of drug discovery research. This technique might greatly assist third-party investigators who cannot confirm the methods of selecting a training set by an author who proposes a good predictive model.

## Conclusion

In recent years, researchers have had to deal with more and more data, thereby increasing concern regarding a decrease in data quality owing to data-handling errors and computer bugs. Equally important are efforts to stamp out data manipulation and outright fabrication. We proposed a simple data quality check protocol based on Benford's law by demonstrating that several data sets in the field of drug discovery follow Benford's distribution and that a manipulated data set does not follow the distribution. These results indicate the usefulness of Benford's law in assessing data quality. Despite the need for further research into methodology for data quality checks in the field of drug discovery, few reports have reviewed this topic. We believe that our proposal is simple to execute, easy to understand and therefore extremely useful for data quality assessment.

## Acknowledgements

We thank Naoko Katayama, Makoto Oku, Kenichi Mori, Hideyoshi Fuji and Yuzo Matsumoto for carefully reviewing the manuscript.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2010.03.003.

### References

1 Newcomb, S. (1881) Note on the frequency of use of dierent digits in natural numbers. *Am. J. Math.* 4, 39–40
2 Benford, F. (1938) The law of anomalous numbers. *Proc. Am. Philos. Soc.* 78, 551–572
3 Hill, T. (1995) A statistical derivation of the significant-digit law. *Stat. Sci.* 10, 354–363

4 Hill, T.P. (1999) The difficulty of faking data. *Chance* 26, 8–13
5 Varian, H.R. (1972) Benford's law. *Am. Stat.* 26, 65–66
6 Carslaw, C. (1988) Anomalies in income numbers: evidence of goal oriented behavior. *Account. Rev.* 63, 321–327
7 Thomas, J.K. (1989) Unusual patterns in reported earnings. *Account. Rev.* 54, 773–787
8 Nigrini, M.J. and Mittermaier, L.J. (1997) The use of Benford's law as an aid in analytical procedures. *Audit. J. Pract. Theory* 16, 52–67
9 Nigrini, M.J. (1999) I've got your number. *J. Account.* 187, 79–83
10 Hassan, B. (2002) Assessing data authenticity with Benford's law. *Inform. Syst. Control* 6
11 Hassan, B. (2003) Examining data accuracy and authenticity with leading digit frequency analysis. *Ind. Manage. Data Syst.* 103, 121–125
12 Brown, R.J.C. (2007) The use of Zipf's law in the screening of analytical data: a step beyond Benford. *Analyst* 132, 344–349
13 Brown, R.J.C. (2005) Benford's law and the screening of analytical data: the case of pollutant concentrations in ambient air. *Analyst* 130, 1280–1285
14 Buyse, M. *et al.* (1999) The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat. Med.* 18, 3435–3451
15 Taylor, R.N. *et al.* (2002) Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. *Drug Inform. J.* 36, 115–125
16 Al-Marzouki, S. *et al.* (2005) Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 331, 267–270
17 Hoyle, D.C. *et al.* (2002) Making sense of microarray data distributions. *Bioinformatics* 18, 576–584
18 Elowe, N.H. *et al.* (2005) Experimental screening of dihydrofolate reductase yields a "Test Set" of 50 000 small molecules for a computational data-mining and docking competition. *J. Biomol. Screen.* 10, 653–657
19 Fontaine, F. *et al.* (2005) Anchor-GRIND: filling the gap between standard 3D QSAR and the grid-independent descriptors. *J. Med. Chem.* 48, 2687–2694
20 Jeffrey, J. *et al.* (2003) Spline-fitting with a genetic algorithm: a method for developing classification structure–activity relationships. *J. Chem. Inf. Comput. Sci.* 43, 1906–1915
21 van de Waterbeemd, H. and Gifford, E. (2003) ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204
22 Clementi, S. and Wold, S. (1995) *Chemometrics Methods in Molecular Design* (van de Waterbeemd, H., ed.), pp. 319–338, VCH
23 Wold, S. (1995) *Chemometrics Methods in Molecular Design* (van de Waterbeemd, H., ed.), pp. 195–218, VCH
24 Zheng, W. and Tropsha, A. (2000) Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* 40, 185–194
25 Hoffman, B. *et al.* (1999) Quantitative structure–activity relationship modeling of dopamine $D_1$ antagonists using comparative molecular field analysis. Genetic algorithms-partial least-squares, and k nearest neighbor methods. *J. Med. Chem.* 42, 3217–3226
26 Ajay, (1993) A unified framework for using neural networks to build QSARs. *J. Med. Chem.* 36, 3565–3571
27 Ertl, P. *et al.* (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* 43, 3714–3717



**Masaya Orita*,**
**Ayako Moritomo,**
**Tatsuya Niimi,**
**Kazuki Ohno**
*Chemistry Research Labs, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan*
**Corresponding author:*