Hindawi Publishing Corporation
Dataset Papers in Biology
Volume 2013, Article ID 976758, 4 pages
http://dx.doi.org/10.7167/2013/976758

Dataset Paper

A Benchmark Dataset Comprising Partition and Distribution Coefficients of Linear Peptides

Matthew N. Davies[1] and Darren R. Flower[2]

[1] *Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Hospital Campus, 4th Floor South Wing Block D, Westminster Bridge Road, London SE1 7EH, UK*
[2] *School of Life and Health Sciences, University of Aston, Aston Triangle, Birmingham B4 7ET, UK*

Correspondence should be addressed to Darren R. Flower; d.r.flower@aston.ac.uk

Received 28 March 2013; Accepted 17 April 2013

Academic Editors: L. McGuffin and L. A. Ponce Soto

Peptides have a dominant role in biology; yet the study of their physical properties is at best sporadic. Peptide quantitative structure-activity relationship (QSAR) lags far behind the QSAR analysis of drug-like organic small molecules. Traditionally, QSAR has focussed on experimentally determined partition coefficients as the main descriptor of hydrophobicity. A partition coefficient (log P) is the ratio between the concentrations of an uncharged chemical substance in two immiscible phases: most typically water and an organic solvent, usually 1-octanol. A distribution coefficient (log D) is the equivalent ratio for charged molecules. We report here a compilation of partition and distribution coefficients for linear peptides compiled from literature reports, suitable for the development and benchmarking of peptide log P and log D prediction algorithms.

1. Introduction

Peptides abound in nature, functioning as hormones, including bradykinins, insulin, gastrins, oxytocins, and various growth factors; as neuropeptides [1], such as encephalins and endorphins; as MHC-bound epitopes, the principal recognition element in cellular immunology [2]; as intermediates in the degradation of proteins [3]; as bacteriocins [4], such as microcins; as antimicrobial host defence peptides [5], such as dermcidins and defensins; and as venom peptides [6], such as α-, μ-, and ω-conotoxins, χ-conopeptides, conantokins, and contulakins; to name but a few of their many important and diverse roles and functions. However, peptides have historically been regarded by the pharmaceutical industry as poor drug candidates [7], not least as they are thought to lack desirable Lipinski-like qualities, such as possessing a low molecular weight [8]. Naturally occurring peptides also often have a limited half-life. If administered orally, they are rapidly broken down by endo- and exopeptidases within the gut, reducing oral bioavailability. For this reason, therapeutic peptides are often delivered parenterally, which can be both impractical and expensive. The use of peptides also has many potential advantages, if these practical problems can be overcome. Peptides can be highly specific, thus reducing unnecessary side effects; whilst naturally occurring peptides are likely to exhibit low toxicity.

Thus, many peptides have been licensed as drugs [7], including captopril, nesiritide, ceruletide, and exenatide. Peptide vaccines are, likewise, an area of active research in both clinical and preclinical environments [9]. So-called cell-penetrating peptides form another avenue being vigorously investigated, this time as drug delivery agents [10]. In all these scenarios, a proper understanding of the physicochemical properties that underlie and determine peptide function would be advantageous. To quote Jacob Bronowski: "We gain our ends only through the laws of nature; we control her only through understanding her laws." In this case, the means to this understanding is provided by QSAR.

QSAR has, traditionally, focussed on experimentally determined partition coefficients (log P) as a principal descriptor of lipophilicity or hydrophobicity and, as a consequence, of many other ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicological) properties, which are heavily dependent upon lipophilicity. Knowing accurate

values for experimental or predicted partition and distribution coefficients is useful for filtering combinatorial libraries and compound collections and also allows Quantitative Structure Activity Relationships (QSARs) to be developed, underpinning our growing predictive understanding of ADMET properties [11]. Thus, a proper understanding of lipophilicity is a key requirement of modern compound design, whether such compounds are synthetic small molecules or peptides. Despite longstanding problems with properly measuring log P values, extant peptide partition and distribution coefficients represent a potentially useful source of descriptors for QSAR studies of peptides.

As is well known, the partition coefficient, P, is the ratio between the concentration of a drug or another chemical substance in two phases: one aqueous and the other an organic solvent. Traditionally, experimental log P measurement involves dissolving a compound within a biphasic system comprised of aqueous and organic layers and then determining the molar concentration of the compound in each layer:

$$P = \frac{[\text{drug}]_{\text{organic}}}{[\text{drug}]_{\text{aqueous}}}. \quad (1)$$

The organic solvent used is typically, but not exclusively, 1-octanol. Although 1-octanol is, in reality, a poor choice for the organic phase as it contains a considerable amount of dissolved water and thus does not even effectively separate hydrophobic from other intermolecular interactions, it remains in common usage. Its general relevance to biological systems is also open to question, and many have suggested that measuring the partition into phospholipids bilayers or micelles would be more appropriate. There are many examples of measurements of peptide partition into other organic phases, such as phospholipids bilayers and micelles [12], and these may, ultimately, prove to be a more rewarding productive source from which a biologically relevant measure of hydrophobicity for short peptides can be derived.

The partition coefficient can range over 12 orders of magnitude and is usually quoted as a logarithm: log P. The partition constant is distinct from the distribution constant, D, which is dependent upon pH. It is generally assumed that the log P of the neutral species is 2–5 log units greater than that of the ionized form and that this is sufficiently large that the partitioning of the charged molecule into the organic phase can be neglected. For singly ionisable species, log P and log D are related by simple mathematical equations, which correct for the relative molar fractions of charged and uncharged molecules.

For monoprotic acids:

$$\log D_{\text{pH}} = \log P - \log\left[1 + 10^{\text{pH}-\text{pK}_a}\right]; \quad (2)$$

For monoprotic bases:

$$\log D_{\text{pH}} = \log P - \log\left[1 + 10^{(\text{pK}_a-\text{pH})}\right]. \quad (3)$$

However, where molecules possess two or more ionisable centres, the equivalent relationships become ever more complex as the number of charged groups increases. For example, ampholytes, or amphoteric compounds, have both acidic and basic functions; ampholytes fall into two main groups: ordinary and zwitterionic, which are distinguished by the relative acidity of the two centres. In ordinary ampholytes, both groups cannot simultaneously ionize, since the acidic pK_a is greater than the basic pK_a. However, for zwitterions, the acidic pK_a is less than the basic pK_a and so both can be ionized at once, thus forming an electrically neutral internal salt. Consideration of ion-pairing leads to even more complex relations [13]. The necessary correction due to ionization required for distribution coefficients is thus not trivial in the general case of a multiply protonatable molecule.

Generally speaking, peptides are, potentially, multiply charged polyprotic ampholytes, with both N- and C-terminal charges and charges from residue side chains. While it is possible to measure multiple pK_a values using modern spectrophotometric and potentiometric methods, this has not been undertaken systematically for peptides. pK_a values of ionizable groups in both proteins and peptides vary considerably. Previously, we have developed a database of protein—as opposed to peptide—pK_as [14]. The analysis of this data indicates unequivocally that measured protein pK_a values differ significantly from values measured for model compounds. No equivalent database or compilation exists for short, biologically interesting peptides, at least in the open literature; though a brief anecdotal examination of what data is readily available suggests that short peptide pK_a values also vary [15].

In this paper, we report a rigorous and scrupulous compilation of partition and distribution coefficients for linear peptides collated from literature reports, suitable for the development, and benchmarking, of peptide log P and log D prediction. In a previous paper [16], we described a more haphazard, heterogeneous dataset and used it to evaluate the accuracy of extant log P prediction algorithms. Our main motivation is to better understand basic physicochemical properties in the design of peptides as pharmaceutical products, such as peptide drugs or peptide vaccines.

2. Methodology

This compilation was derived through exhaustive, semimanual searching of both public literature resources: Web of Science (http://wok.mimas.ac.uk/), MEDLINE (http://medline.cos.com/), and PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), and a variety of private publisher's databases, such as ScienceDirect (http://www.sciencedirect.com/). We explored extensively the available literature using global keyword and author searches, retrospective searching of cited papers, and forward citation matching of key papers. As an example of our approach, the following was used to search for articles in PubMed:

(*LOGP* OR " logP* " OR " log(P*)" OR " log(P)" OR LOGD* OR " logd* " OR " log(D*)" OR " log(D)" OR "partition coefficient" OR "partition coefficients" OR "distribution coefficient" OR "distribution coefficients" OR octanol**) *AND* (*"peptides" OR "peptide" OR "peptidic" OR peptido* OR "amino acid" OR "amino acids" OR oligopeptide* OR dipeptide* OR tripeptide* OR*

tetrapeptide OR pentapeptide* OR hexapeptide* OR heptapeptide* OR octapeptide* OR nonapeptide* OR decapeptide* OR dodecapeptide* OR nonadecapeptide* OR octadecaneuropeptide* OR neuropeptide* OR oligopeptide* OR polypeptide**).

Our goal was the identification of reports detailing quantitative, experimentally derived values for octanol-water partition and distribution coefficients for linear peptides. Where reports presented ambiguous or inconsistent data, such reports were rejected. All cyclic peptides were also excluded.

For each peptide, several pieces of information were archived, including the explicit structure of each peptide rendered in two dimensions. As many peptides in the dataset have minor, nonstandard chemical modifications, such as blocking groups, it was necessary to use a representation that explicitly defined the full structure rather than using an inaccurate amino acid sequence. The stored 2-dimensional structure was generated from an initial, manually composited SMILES string [17] using the cheminformatics toolkit CACTVS [18] and then adjusted manually where necessary. To allow for full and complete flexibility and proper integration with QSAR software, the compiled and collated data is presented in the form of an SD file [19]. The SD file contains the structure, sequence, log P, log D, experimental method, pH, and the original reference.

Our interest in the current problem stems from our desire to find effective measures of hydrophobicity for use in peptide QSAR studies [20]. Partition coefficients, as opposed to distribution coefficients, are widely regarded with scepticism, but nonetheless represent the only data available in sufficient quantity. Secondly, the peptides we examine here are short and have heavily biased sequence compositions, compared to all possible naturally occurring and synthetically possible peptide sequences. Data are thus both sparse and tendentious in terms of length and sequence properties. In particular, longer peptides are of most interest, yet they are significantly underrepresented here for experimental reasons. Overall, the availability of measured log P values for peptides was very limited; thus the use of a mixture of experimental conditions was unavoidable.

The dataset comprises both log P and log D measurements, and it is not practical to arbitrarily adjust log P values and thus generate log Ds, unless we have access to reliable pK_a values for multiply-ionisable peptides, which will typically possess two or more protonatable groups. Peptides with five or more centres are not uncommon. This involves measuring the pK_a values for each ionisable group in each polyprotic peptide and then correcting the measured partition using the kind of equation given in the introduction, although these may become prohibitively complex as the number of protonatable centres rises and other effects, such as charge pairing, are included.

3. Dataset Description

The dataset associated with this Dataset Paper consists of one item which is described as follows.

Dataset Item 1 (Chemical Structure Data). A compilation of peptides from the primary literature with corresponding experimental partition and distribution coefficient values. It comprises 428 linear, unbranched, noncyclic peptides. These varied in length from 2 to 9 amino acids. The set contains 348 log P values and 80 log D values. 103 peptides were unblocked at the C terminus, and 283 were unblocked at the N terminus. Sixty-two of the reported coefficients were recorded at pH values below 7.0; 309 were recorded between pH 7.0 and pH 8.0; 57 were above pH 8.0. This is in contrast to the dataset from [16], which comprised 340 peptides, 2 to 16 amino acids in length, albeit most peptides were less than 10 amino acids. The set from [16] included 41 cyclic peptides, which were excluded here. The 428 peptides in the new dataset thus represent a 40% improvement over our previously reported heterogeneous dataset; the present dataset is both more consistent and more explicit. Data included in the SD file are the following. <Structure> is the full atomic 2D representation of the peptide structure, including any and all chemical modifications. <EXTREG> is the name of the peptide, essentially the amino acid sequence rendered using the IUPAC 1 letter code. <logP> is the experimentally determined partition coefficient corresponding to the peptide above. <logD> is the experimentally determined distribution coefficient corresponding to the peptide above. <pH> is the pH value at which the coefficient was determined. <method> is the experimental method used to determine the partition or distribution coefficients reported above. <reference> is the full reference to the original paper in the scientific literature reporting the experimental measurement above.

4. Concluding Remarks

There is a clear paucity of quality data for partition and distribution coefficients for peptides, due to the unequivocal lack of reliable and relevant experimental measurements. There is at least thirty times as much data for drug-like small molecules. There is thus an obvious case for dedicated experimental work to be undertaken to support the development of accurate *in silico* methods. Computational biology cannot be based solely on heterogeneous and sporadic sources of data generated parenthetically and peripherally by individual research projects. Experiments are needed that address specifically the types of predictions that we wish to make. Such problems would be largely resolved by a properly designed training set. Our potential ability to combine *in vitro* and *in silico* analysis would allow us to improve both the scope and power of predictions, in a way that would be impossible through the sole use of data from the literature: to use the old adage "garbage in, garbage out." To ensure the production of useful, quality *in silico* models and methods, and not poor ones without practical utility, we need to value the prediction for its own sake and conduct dedicated experiments accordingly. In the meantime, making the dataset described here available is unprecedented and should provide an impetus to the development of more sophisticated and more robust prediction methods for peptides, within a profound contingent effect on the design of peptide drugs and vaccines.

Dataset Availability

The dataset associated with this Dataset Paper is dedicated to the public domain using the CC0 waiver and is available at http://dx.doi.org/10.7167/2013/976758/dataset.

Disclosure

The authors declare that they have no competing interests.

Acknowledgment

The authors thank Drs. Sarah J. Thompson, Channa K. Hattotuwagama, John D. Holliday, Antony Williams, and Greg Pearl for discussions.

References

[1] C. J. Grimmelikhuijzen and F. Hauser, "Mini-review: the evolution of neuropeptide signaling," *Regulatory Peptides*, 177, pp. S6–S9, 2012.

[2] P. M. Saunders and P. van Endert, "Running the gauntlet: from peptide generation to antigen presentation by MHC class I," *Tissue Antigens*, vol. 78, no. 3, pp. 161–170, 2011.

[3] A. V. Sorokin, E. R. Kim, and L. P. Ovchinnikov, "Proteasome system of protein degradation and processing," *Biochemistry*, vol. 74, no. 13, pp. 1411–1442, 2009.

[4] M. Nishie, J. Nagao, and K. Sonomoto, "Antibacterial peptides, "bacteriocins": an overview of their diverse characteristics and applications," *Biocontrol Science*, vol. 17, no. 1, pp. 1–16, 2012.

[5] L. Steinstraesser, U. Kraneburg, F. Jacobsen, and S. Al-Benna, "Host defense peptides and their antimicrobial-immunomodulatory duality," *Immunobiology*, vol. 216, no. 3, pp. 322–333, 2011.

[6] R. J. Lewis and M. L. Garcia, "Therapeutic potential of venom peptides," *Nature Reviews Drug Discovery*, vol. 2, no. 10, pp. 790–802, 2003.

[7] D. J. Craik, D. P. Fairlie, S. Liras, and D. Price, "The future of peptide-based drugs," *Chemical Biology & Drug Design*, vol. 81, no. 1, pp. 136–147, 2013.

[8] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Advanced Drug Delivery Reviews*, vol. 46, no. 1–3, pp. 3–26, 2001.

[9] M. S. Bijker, C. J. M. Melief, R. Offringa, and S. H. Van Der Burg, "Design and development of synthetic peptide vaccines: past, present and future," *Expert Review of Vaccines*, vol. 6, no. 4, pp. 591–603, 2007.

[10] F. Milletti, "Cell-penetrating peptides: classes, origin, and current landscape," *Drug Discovery Today*, vol. 17, no. 15-16, pp. 850–860, 2012.

[11] D. G. Sprous, R. K. Palmer, J. T. Swanson, and M. Lawless, "QSAR in the pharmaceutical research setting: QSAR models for broad, large problems," *Current Topics in Medicinal Chemistry*, vol. 10, no. 6, pp. 619–637, 2010.

[12] R. E. Jacobs and S. H. White, "The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices," *Biochemistry*, vol. 28, no. 8, pp. 3421–3437, 1989.

[13] M. Miyanaga, K. Imamura, K. Tanaka, T. Sakiyama, and K. Nakanishi, "Analysis for partition equilibrium of amino acid derivatives in aqueous/organic biphasic systems," *Journal of Bioscience and Bioengineering*, vol. 88, no. 6, pp. 651–658, 1999.

[14] C. P. Toseland, H. McSparron, M. N. Davies, and D. R. Flower, "PPD v1.0—an integrated, web-accessible database of experimentally determined protein pKa values," *Nucleic Acids Research*, vol. 34, pp. D199–D203, 2006.

[15] X. Fang, Q. Fernando, S. O. Ugwu, and J. Blanchard, "An improved method for determination of acid dissociation constants of peptides," *Pharmaceutical Research*, vol. 12, no. 10, pp. 1423–1429, 1995.

[16] S. J. Thompson, C. K. . Hattotuwagama, J. D. Holliday, and D. R. Flower, "On the hydrophobicity of peptides: comparing empirical predictions of peptide log P values," *Bioinformation*, vol. 1, no. 7, pp. 237–241, 2006.

[17] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences®*, vol. 28, no. 1, pp. 31–36, 1988.

[18] W. D. Ihlenfeldt, "The cactvs chemoinformatics toolkit: universal chemical information processing with Tcl scripts," in *Proceedings of the 238th ACS National Meeting*, American Chemical Society, Washington, DC, USA, August2009.

[19] A. Dalby, J. G. Nourse, W. Douglas Hounshell et al., "Description of several chemical structure file formats used by computer programs developed at molecular design limited," *Journal of Chemical Information and Computer Sciences*, vol. 32, no. 3, pp. 244–255, 1992.

[20] P. Guan, I. A. Doytchinova, V. A. Walshe, P. Borrow, and D. R. Flower, "Analysis of peptide-protein binding using amino acid descriptors: prediction and experimental verification for human histocompatibility complex HLA-A∗0201," *Journal of Medicinal Chemistry*, vol. 48, no. 23, pp. 7418–7425, 2005.