Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Empirical prediction of peptide octanol-water partition coefficients

Channa K. Hattotuwagama and Darren R. Flower *

The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20 7NN, UK; Darren R. Flower - Email: darren.flower@jenner.ac.uk; Phone: +44 1635 577954; Fax: +44 1635 577908; * Corresponding author

received November 11,2006; accepted November 22, 2006; published online November 24, 2006

Abstract:

Peptides are of great therapeutic potential as vaccines and drugs. Knowledge of physicochemical descriptors, including the partition coefficient P (commonly expressed in logarithm form: logP), is useful for screening out unsuitable molecules and also for the development of predictive Quantitative Structure-Activity Relationships (QSARs). In this paper we develop a new approach to the prediction of LogP values for peptides based on an empirical relationship between global molecular properties and measured physical properties. Our method was successful in terms of peptide prediction (total $r^2 = 0.641$). The final model consisted of 5 physicochemical descriptors (molecular weight, number of single bonds, 2D-VDW volume, 2D-VSA hydrophobic and 2D-VSA polar). The approach is peptide specific and its predictive accuracy was high. Overall, 67% of the peptides were able to be predicted within +/-0.5 log units from the experimental values. Our method thus represents a novel prediction method with proven predictive ability.

Key Words: peptide**;** log P**;** partition coefficient**;** octanol-water**;** regression**;** physicochemical descriptor**;** hydrophobicity

Background:

Peptides have over time received a bad press, at least pharmaceutically speaking, gaining a reputation as very poor drug candidates. However, such criticism cannot eclipse their pre-eminent role in biological systems. Naturally-occurring peptides often have a limited half-life and thus therapeutic peptides are often delivered parenterally, which can be impractical and expensive. However, peptides can be highly specific, reducing unwanted side-effects, and have low toxicity.

QSAR has focussed on experimentally-determined partition coefficients as the main descriptor of lipophilicty or hydrophobicity, and thus of many other ADMET properties. **[1]** Predicted partition coefficients are routinely used to filter or select compounds for screening and to develop QSARs. The partition coefficient, P, is the ratio between the concentration of a drug or other chemical substance in two phases: one aqueous, the other an organic solvent:

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

Traditionally, experimental logP 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. The organic solvent used is typically, but not exclusively, 1-octanol. The partition coefficient can range over 12 orders of magnitude, and is usually quoted as a logarithm: logP.

The experimental determination of logP values is expensive, time consuming, and labour intensive. Accurate methods for the prediction of peptide logP values would thus be most useful. During the past three decades, many methods for predicting logP have been reported. At present, the most widely used method is known as a fragmental, fragment-based, or additive approach: a molecule is dissected into fragments (functional groups or atoms) and its logP value is obtained by summing the contributions of each fragment. 'Correction factors' are also introduced to rectify the calculated logP value when special substructures occur in the molecule.

There have been various studies carried out on peptide logP prediction. The most convincing approach is based on the direct quantification of hydrophobicity for peptides. **[2, 3]** They carefully measured partition coefficients for many peptides, specifically targeting non-ionizable side chains and obtained different linear-regression models for different types of peptides, resulting in good correlations between observed and predicted log*P* values. This and subsequent work by Akamatsu was incorporated into PLOGP, a peptide LogP prediction program. **[4]** Here, a training set of 219 peptides, varying between 2 and 5 amino acids, was used and the method tested using another 10 peptides.

In this paper we look at prediction of logP values for peptides. Our main motivation is to better understand basic physico-chemical properties in the design of peptide vaccines. **[5]** Using a data-set of experimentally-determined peptide logPs, we have developed a new logP prediction method, for both blocked and unblocked peptides, using Partial Least Squares (PLS) **[6]** as implemented in GOLPE (Generating Optimal Linear PLS Estimations - version 4.5.12; Multivariate Infometric Analysis), based on molecular descriptors calculated using PreADME **[7]**, a web-based application able to calculate large numbers of diverse molecular descriptors including constitutional,

ISSN 0973-2063
Bioinformation 1(7): 257-259 (2006)

Bioinformation, an open access forum
© 2006 Biomedical Informatics Publishing Group

topological, physico-chemical and geometrical descriptors for ADME prediction.

Methodology:

Data-set

A set of peptides with known experimental logP values was compiled from the primary literature **[8]**, through exhaustive, semi-manual searching of a variety of databases: PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi), Web of Science (http://wos.mimas.ac.uk/), Medline (http://medline.cos.com/), and ScienceDirect (http://www.sciencedirect.com/). Both keyword and author searches, as well as retrospective searching, and citation matching of key authors, particularly those describing the development of an assay system, were used to identify papers detailing quantitative experimentally-derived values. The availability of measured LogP values for peptides was limited. Data consisted of 340 peptides (2-16 amino acids in length). The set included 141 blocked peptides, 158 unblocked peptides, and 41 cyclic peptides. See URL: http://www.jenner.ac.uk/Bioinformatics/peptide_structures.htm).

Peptide Additive Method

Individual *.sdf* files of amino acids were submitted to PreADME and corresponding descriptor values calculated. This information was imported into GOLPE and PLS calculations undertaken. PLS is a robust multivariate statistical extension of Multiple Linear Regression (MLR). Experimental logP values were used as the dependent variable. A variable selection procedure within GOLPE, known as "D-Optimal Selection", was chosen to evaluate the effects of individual variables on the model's ability to determine which variables are relevant to the problem. Initially, a small number of descriptors were extracted from a large amount of redundant information. Extraction of descriptors continues until a good statistical model is obtained. Model validity was explored using Cross-Validation (CV). Leave-One-Out Cross-Validation (LOO-CV) was used to assess its predictive ability using the following parameters: cross-validated coefficient (q^2) and by calculating the standard deviation of error of prediction (SDEP)[45], which indicates the error distribution between the observed and predicted values in the regression models. The optimal number of components (NC) from LOO-CV is then used in the non-cross validated model which was assessed using standard MLR validation terms, such as r^2.

Results and Discussion:

Fourteen carefully selected molecular descriptors were calculated for each whole peptide using PreADME: ten constitutional descriptors (molecular weight, number of rotatable bonds, rigid bonds, rings, aromatic bonds, single bonds, double bonds, aromatics, hydrogen bond acceptors and hydrogen bond donors) and four geometrical descriptors (2D-van der Waals surface area, 2D-van der Waals volume, 2D-VSA hydrophobic and 2D-VSA polar). LogP values for each amino acid residue, for both blocked and unblocked peptides, were related to a subset of these descriptors.

For Blocked peptides:

LogP = 0.04983 – 0.04222*molecular weight + 0.02717*single bonds + 0.09814*2D-VDW volume – 0.04452*2D-VSA hydrophobic – 0.04673*2D-VSA polar (2)

LOO-CV parameters are q^2 = 0.814, SDEP = 0.485 and NC = 5, while the non-cross validation parameters are r^2 = 0.836.

For unblocked peptides:

Log P = -2.478 – 0.03751*molecular weight + 0.02338* single bonds + 0.08308*2D-VDW volume -0.03108* - VSA hydrophobic – 0.04204*2D-VSA polar (3)

LOO-CV parameters are q^2 = 0.819, SDEP = 0.350 and NC = 5, while the non-cross validation parameters are r^2 = 0.837.

The results from the PLS model are very promising statistically. Both final models contain the same descriptors: molecular weight, number of single bonds, 2D-van der Waals volume, 2D-VSA hydrophobic and 2D-VSA polar. To calculate logP values for other peptides, standard PreADME amino acid descriptor values are concatenated according to the peptide sequence and a correction applied if the peptide is unblocked. Using the two resulting models for the blocked and unblocked peptides, the non-CV method was validated for a total of 236 linear (86 blocked and 150 unblocked) peptides. Prediction accuracy for all peptides (r^2 = 0.666) was good, but not excellent; for blocked peptides performance was relatively poor (r^2 = 0.381); but for unblocked peptides performance was superior (r^2 = 0.787). We then compared previously-described LogP prediction methods **[8]** with our method: 67% of the peptides were predicted within +/- 0.5 log units, and a further 21% between +/-0.5 and 1.0 log units. 88% predicted within 1 log unit represents the best accuracy of all the methods we compared. **[8]**

ISSN 0973-2063
Bioinformation 1(7): 257-259 (2006)

Bioinformation, an open access forum
© 2006 Biomedical Informatics Publishing Group

There are clear failings in the work we report here. Our principal concern is the paucity of quality data for peptide partition coefficients; indeed, the lack of reported experimental studies prevents us from obtaining a data set of sufficient size. Moreover, we would like to obtain LogD rather than LogP values. Likewise, the peptides we examine here are short and have heavily biased sequence compositions. Longer peptides are of most interest, at least in terms of epitope design and discovery, yet they are under-represented here for experimental reasons. The average length of peptide studied was three amino acids. As many biologically important peptides are much longer than three amino acids, the data set is likely to compromise our ability to perform adequate QSAR analysis.

Conclusion:

We have shown that the empirical relationship between the octanol-water partition coefficient of a peptide and its structure can be easily rationalised by properties of the whole peptide, such as volume and surface area, rather than by the more common fragmental approach. However, the data we analysed is both sparse, compared to the potential size of the dataset, and heavily biased due to experimental constraints. There is an obvious case for dedicated experimental work to be undertaken to support the development of accurate *in silico* methods. We need quality data to work with: existing data is seldom of sufficient quality. Computational chemists can no longer exist solely on morsels swept contemptuously from the experimentalists' table. What we require are experiments which specifically address the kind of predictions that need to be made. Such problems would be 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 our predictions, in a way that would be impossible using solely literature data. To ensure we produce useful, quality *in silico* models and methods, and not poor models and methods, we need to value the prediction generated by them and conduct experiments appropriately.

Acknowledgement:

The Jenner Institute (Formally, The Edward Jenner Institute for Vaccine Research) wishes to thank its erstwhile sponsors: GlaxoSmithKline, the Medical Research Council, the Biotechnology and Biological Sciences Research Council, and the UK Department of Health.

References:

[01] D. Eros, *et al., Curr Med Chem.,* 9:1819 (2002) [PMID: 12369880]

[02] M. Akamatsu & T. Fujita, *J Pharm Sci.,* 81:164 (1992) [PMID: 1545357]

[03] M. Akamatsu, *et al., J Pharm Sci.,* 83:1026 (1994) [PMID: 7965659]

[04] T. Peng, *et al., J. Mol. Model.,* 5:189 (1999)

[05] P. Guan, *et al., J Med Chem.,* 48:7418 (2005) [PMID: 16279801]

[06] S. Wold & H. Van de Waterbeemd, Chemometric methods in molecular design. VCH, Weinheim, 195 (1995)

[07] S. K. Lee, *et al.,* 15th European Symposium on Quantitative Structure – Activity relationships & Molecular Modeling 9 (2002)

[08] S. J. Thompson, *et al.,* Bioinformation 1:237 (2006)

Edited by P. Kangueane

Citation: Hattotuwagama & Flower, Bioinformation 1(7): 257-259 (2006)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

ISSN 0973-2063

Bioinformation 1(7): 257-259 (2006)

Bioinformation, an open access forum

© 2006 Biomedical Informatics Publishing Group